Manufacturing the Future

Concepts, Technologies & Visions

Manufacturing the Future

Concepts, Technologies & Visions

Edited by

Vedran Kordic

Aleksandar Lazinica

Munir Merdan

pro literatur Verlag

Published by the plV pro literatur Verlag Robert Mayer-Scholz

plV pro literatur Verlag Robert Mayer-Scholz
Mammendorf
Germany

www.ars-journal.com
Additional copies can be obtained from:
publication@ars-journal.com

First published July 2006

Typeface Palatino Linotype 10/11/12 pt

Printed in Croatia

A catalog record for this book is available from the German Library.

Manufacturing the Future: Concepts, Technologies & Visions/ Edited by Vedran Kordic, Aleksandar Lazinica, Munir Merdan.

p. cm.

ISBN 3-86611-198-3

1. Distributed Manufacturing. 2. Modern assembly systems. 3. Supply Chain I. Kordic, Vedran. II. Lazinica, Aleksandar. III. Merdan, Munir

Contents

Preface

The primarily goal of this book is to cover the state-of-the-art development and future directions in modern manufacturing systems.
This interdisciplinary and comprehensive volume, consisting of 30 chapters, covers a survey of trends in distributed manufacturing, modern manufacturing equipment, product design process, rapid prototyping, quality assurance, from technological and organisational point of view and aspects of supply chain management.
The book's coverage reflects the editor's belief that modern industrial applications stem from the synergy of general understanding of the big picture of technologies with an in-depth knowledge of a targeted application area and its consequent impact on business development.
This book is the result of inspirations and contributions from many researchers worldwide.
We would like to thank all the researchers and especially to the authors of the chapters who entrusted us with their best work. It is their work that enabled us to collect the material for this book. We hope you will enjoy reading the book as much as we have enjoyed bringing it together for you.
Further, the editors would like to acknowledge and express their appreciation to the reviewers.
Their contribution is highly appreciated and it has helped to make this book of significantly greater value to its readers.

Editors
Vedran Kordic
Automation and Control Institute
Vienna University of Technology

Aleksandar Lazinica
Institute for Production Engineering
Vienna University of Technology

Merdan Munir
Automation and Control Institute
Vienna University of Technology

1

Multi-Agent Based Distributed Manufacturing

J. Li, J.Y H. Fuh, Y.F. Zhang and A.Y.C. Nee

1. Introduction

Agent theory is developed from distributed artificial intelligence, which is regarded as a prospective methodology suitable for solving distributed complex problems, and it has been applied in many areas including manufacturing engineering. In this chapter, some basic issues for agent theory are described and an example of one agent-based distributed manufacturing system is presented.

1.1 Agent and multi-agent system

Jennings and Wooldridge (Jennings and Wooldridge 1998) have defined an agent as "a computer system situated in some environment and capable of autonomous action in this environment, in order to meet its design objectives". Some of the main properties of agents are autonomy, socialability, reactivity, and proactiveness (Wooldridge and Jennings 1995):

- Autonomy: Autonomy characterizes the ability of an agent to act on its own behalf. Agents can operate without direct intervention of humans or other agents, and have a some kind of control over their actions and internal states (Castelfranchi 1995).
- Socialability: Agents can interact with other agents via agent communication languages (Gensereth and Ketchpel 1994)
- Reactivity: Agents can perceive the changes of their environment, which may be the physical world, a collection of other agents, the Inter net, and other fields, and respond to make the related decision accordingly in real time.
- Proactiveness: Agents do not only act in response to their environment, but also exhibit goal-directed behavior by taking the initiative

All of these properties are necessary for agents to act as autonomous, loosely coupled and self coordinating entities in an open distributed system; which forms a multi-agent based system (MAS). A MAS consists of a group of agents that play individual roles in an organizational structure (Weiss 1999). The most important characteristic of MAS is the agents' capabilities of communication and cooperation, which make them to interact with other agents to achieve their individual objectives, as well as the common goals of the system (Wooldridge and Jennings 1995). Other important characteristics of the agent-based systems include scalability, modularity and re-configurability.
In an MAS model, every agent is a representative of a functional cell. For instance, in order to agentify a complex system, it will be divided into some subsystems, each of which is further encapsulated into an agent. Each agent conquers its individual problem, and cooperates with other related agents to solve the whole problem. In the distributed system modeling, an agent is the representative of a distributed cell which solves its own problems and can cooperate with other agents to fulfill a task if necessary. A comprehensive book on multi-agent theory can be found in (Weiss 1999).

1.2 The architecture of MAS

The architectures of multi-agent based systems provide the frameworks within which agents are designed and constructed (Shen 2002). Similar with the organization of the distributed manufacturing system, there are three types of architecture for multi-agent based systems, which are hierarchical (A), heterarchical (B) and hybrid structures (C), as shown in the figure 1.

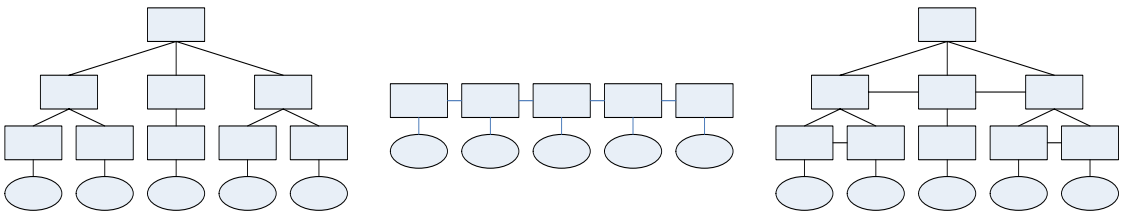


Figure 1. Architecture of MAS

In the hierarchical architecture (A), the agents are connected with layered relationship and all of the control modules are organized into a hierarchical man-

ner. Each agent will have only one direct supervisor at its directly upper layer and several subordinate agents at its directly lower layer. The agent executes the commands and plans only from its supervisor agent and gathers the feedback information from its subordinate agents. The main advantage for this structure is that global optimization can be achieved possibly as the complete information and status of the system can be collected by the agent at the highest layer; while the main disadvantages resides in less adaptability and reliability because the system may be malfunction once the central controller agent breaks down.

Heterarchical architecture (B) is another different style compared with the previous one because there is no central controller in this kind of structure and the relationship of the agents is peer to peer. Each of the agents is autonomous and has its own decision-making mechanism. The cooperation work among agents is to be realized by negotiation: the related agents will negotiate and make tradeoff for a variety of factors. The advantage for this type architecture is its high robustness because breakdown of one agent will not influence others and the rest can still work. The main problem for this architecture lies in the difficulty to achieve a global optimization as no single agent can collect the full information and status from others. Furthermore, another shortcoming is that the execution efficiency is relative low in such framework because the negotiation process may be inefficient and less effectiveness, especially for those tasks need to be completed by several cooperative agents.

The third type (C) is the hybrid architecture, which can be regarded as a compromise of the above two kinds. The hierarchy of the system enhances its efficiency and effectiveness on a global basis while achieving some advantages of the heterarchical architecture to keep the good adaptability and autonomy. In this architecture, the agents at the lower level are also intelligent and have some degree of autonomy, which can be viewed as a heterarchical structure. But the agents also have their upper layered supervisor agent, which can collect the information and distribute tasks to some capable subordinate agents. As the upper level supervisor agent can get a global view for its subordinate agents, some global optimal decision can be achieved. At the same time, as the lower level agents are autonomous, some decisions can be made locally and will not impact other agents, which can improve the robustness and adaptability of the whole system.

1.3 The coordination methodology for MAS

The methodology of negotiation and coordination is one of the bases for effective management and control in a distributed system. Presently, the well-known Contract Net Protocol (CNP) (FIPA 1997; FIPA 2000(1)) is adopted as the coordination and negotiation protocol in most of multi-agent systems. CNP method was proposed by smith (Smith 1980; Davis and Smith 1983; Smith 1988) and recommended by FIPA(The Foundation for Intelligent Physical Agents)(FIPA 2000(1); FIPA 2000(2)), an international organization that is dedicated to promoting the industry of intelligent agents by openly developing specifications supporting interoperability among agents and agent-based applications. A standard process for the CNP involves four basic steps as shown in Figure 2:

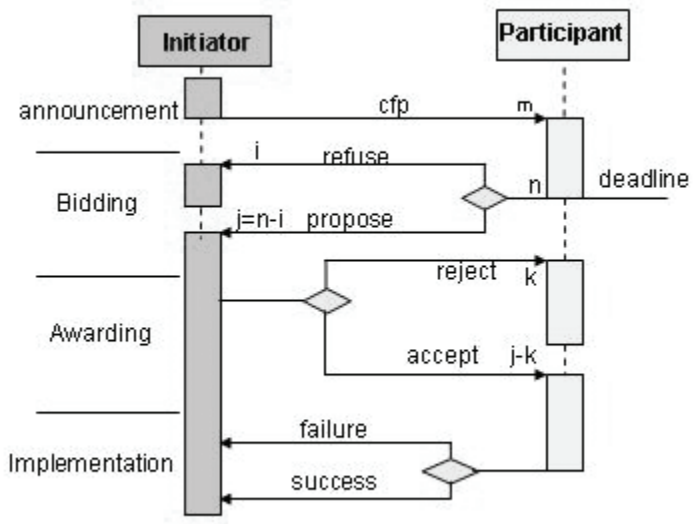


Figure 2. FIPA Contract Net Protocol (FIPA 2000(1))

- Task announcement: The initiator agent broadcasts an announcement to the participant agents to call for proposal (cfp).
- Bidding: Those participants that receive the announcement and have the appropriate capability to make the evaluation on the task, and then reply their bids to the initiator agent.
- Awarding: The initiator agent awards the task to the most appropriate agents according to the proposals they have submitted.

- Implementation: The awarded participant agent performs the task, and receives the benefits predefined.

Currently, the CNP method is widely used for negotiation and coordination among agent systems, and has been proved to be effective to solve distributed problems.

1.4 The development platform for agent-based systems

Most of the intelligent agent and multi-agent systems are working under distributed and heterogeneous environments, and C++ and Java are the two most-adopted programming languages. At the early stage, some works were developed from scratch, which were rather difficult to deal with. Recently, useful platforms and templates have been provided by some institutes, which can provide some basic and necessary modules such as communication, interface design, agent kernel template, etc. The adoption of these platforms facilitates the development and let the designers focus on the functional modules programming, thus to reduce the workload and difficulty of agent applications development. Among these development platforms, JADE (F. Bellifemine, Caire et al. 2006) and Jatlite (JATLite) are two typical and widely applied systems.

JADE (Java Agent DEvelopment Framework) is a software framework developed in Java language by TILAB (JADE 2005). It is composed of two parts, one is the libraries (Java classes) required to develop the agent applications and functions and the other is a run-time environment providing some necessary services for the agents' execution. The platform can be executed in a distributed, multi-party application with peer-to-peer communication, which include both wired and wireless environment. The platform supports execution with cross operation system and the configuration can be controlled via a remote GUI; furthermore, the platform also supports hot exchange, moving agents from one machine to another at run-time.

In JADE, middleware acts as the interface of low layer and applications. Each agent is identified by a unique name and provides a set of services. The agent can search for other agents to provide given services according to the middleware if necessary. With the role of middleware, the agents can dynamically discover other agents and to communicate with them by a peer-to-peer paradigm. The structure of a message complies with the ACL language defined by

FIPA and includes fields, such as variables indicating the context a message refers-to and timeout that can be waited before an answer is received, aiming at supporting complex interactions and multiple parallel conversations (F. Bellifemine, Caire et al. 2006). Furthermore, in order to support the implementation of complex conversations, JADE provides a set of skeletons of typical interaction patterns to perform specific tasks, such as negotiations, auctions and task delegation.

Compared with JADE, JATLite is a lighter and easier to use as a platform for agent-based applications. As it provides only some basic and necessary functions for agent applications, JATLite is more suitable for prototype development in agent-based research work. JATLite is composed of some java packages which help to build agent-based applications with Java language. In the package, four different layers: abstract, base, KQML and router layer, covering from the lowest layer with an operation system to the router function. The package is developed according to TCP/IP protocols, which ensures the system can be running in the Internet. In JATLite, the router acts as the key role in the message communication among the agents.

Although the functions of JATLite may not be as powerful as those in JADE, it is still widely used. The platform is simple and provides some reliable basic services for the agent execution. Furthermore, it sill provides some templates for agent execution; thus, the designers can implement their applications easily.

1.5 Application of MAS in manufacturing system integration

With manufacturing systems become distributed and decentralized in different geographical sites, it is necessary to study the solution of specific problems which arise in a distributed environment. As the MAS system shows the promising capability to solve distributed problems, a great amount of efforts have been made to apply the multi-agent theory to the manufacturing system integration, aiming to study the problems of the distributed manufacturing system. In this part, some typical agent-based manufacturing systems are introduced.

MetaMorph

MetaMorph and MetaMorph II are two consecutive projects developed in the University of Calgary (Shen, Maturana et al. 1998; Shen, Xue et al. 1998; Maturana, Shen et al. 1999; Shen 2002). MetaMorph is an adaptive multi-agent

manufacturing system aimed to provide an agent-based approach for dynamically creating and managing agent communities in distributed manufacturing environments (Maturana, Shen et al. 1999). There are two main types of agents in MetaMorph: *resource agents* and *mediator agents*. Resource agents are used to represent manufacturing devices and operations, and mediator agents are used to coordinate the interactions among agents (resource agents and also mediator agents).

Mediator-centric federation architecture is one of the system characteristics, by which the intelligent agents can link with mediator agents to find other agents in the environment. The activity for mediators is interpreting messages, decomposing tasks, and providing processing times for every new task. Additionally, mediators assume the role of system coordinators by promoting cooperation among the intelligent agents. Both brokering and recruiting communication are adopted to find the related agents for specific tasks. Once appropriate agents have been found, these agents can be directly linked and communicate directly without the aid of mediator.

The object of MetaMorph II project is to integrate the manufacturing enterprise's activities such as design, planning, scheduling, and simulation, execution, with those of its suppliers, customers and partners into a distributed intelligent open environment. In this Infrastructure, the manufacturing system is primarily organized at the highest level through 'subsystem' mediators. Each subsystem is connected (integrated) to the system through a special mediator. Each subsystem itself can be an agent-based system (e.g., agent-based manufacturing scheduling system), or any other type of system like the feature-based design system, knowledge-based material management system, and so on. Agents in a subsystem may also be autonomous agents at the subsystem level. Some of these agents may also be able to communicate directly with other subsystems or the agents in other subsystems. Mediators are also agents, called mediator agents. The main difference between a mediator and a facilitator is that a facilitator provides the message services in general, but a mediator assumes an additional role of system coordinators by promoting cooperation among intelligent agents and learning from the agents' behavior.

CIIMPLEX

CIIMPLEX (Consortium for Intelligent Integrated Manufacturing Planning-Execution) (Peng, Finin et al. 1998) was developed by UMBC and some other institutes, which presents an agent-based framework of enterprise integration

for manufacturing planning and execution. The system is composed of name server, facilitator agent and gateway agent and some executive agents. The different functions of the manufacturing process are encapsulated into individual agents. In the system, a set of agents with specialized expertise can be quickly assembled to gather the relevant information and knowledge, and to cooperate with other agents to arrive at timely decisions to deal with various enterprise scenarios.

Different executive agents are designed to perform special functions such as data collection, analysis of plans and schedules, resolving the conflicts; furthermore, some agents are created to integrate the function of the legacy system. With this architecture, the raw transaction data of the low level, such as shop floors activities, can be collected, aggregated, interpolated and extrapolated by agents and made available for other interested agents. Manufacturing planning and execution can thus be integrated through the collaboration of these agents.

The AARIA project

The AARIA project (Autonomous Agents at Rock Island Arsenal) (Parunak, Baker et al. 1998; Parunak, Savit et al. 1998) is an agent-based prototype system based on the Internet-related technologies. In the system, Internet is used as the platform, and distributed scheduling and controlling techniques are developed to realize the distributed manufacturing. All of the agents are tied by Internet to form a virtual manufacturing environment for tasks. With the agent technology, the resource can be redeployed easily to meet the fast changing environment, which increases the agility of the system. Furthermore, the productive resources can be adjusted according to the products' requirement, which make the system meet the customization requirements.

In the system, besides the functional decomposition, physical factors are also considered during the resource agentificaiton process. The main agents of the system include resource brokers, part brokers, and unit process brokers. Resource broker agents manage the constrained resources of the system (e.g. people, machines, facilities, etc.). Part broker agents manage material handling and inventory. Unit process broker agents utilize their knowledge of how to combine resources and parts to make other parts. These three types of agents negotiate among themselves and with the customer along the axes of possible production including price, quality, delivery time, product features, and speed of answers (Baker, Parunak et al. 1999).

DaimlerChrysler manufacturing line control system

One industrial application of agent-based manufacturing line control system is implemented in DaimlerChrysler (Bussmann and Schild 2001; Bussmann and Sieverding 2001), whose objective is to develop a flexible transportation and control system. In this project, each work piece, machine and shifting table is encapsulated into one specific agent. In the execution, the work piece agent will auction off its coming operations to machine agents. Every machine agent's bid include information about its current state of buffer. Once a work piece agent awards a machine agent, it will be the next goal of the work piece. The routing of the work piece will be negotiated by the work piece agent with the shifting tale agent.

The application of agent-based system shows two key advantages for product manufacturing. One is the distributed responsiveness, as the decision making can be much more localized. If unexpected events occur, agents have the autonomy and proactiveness to try alternatives thus can be more responsive to prevailing circumstances. The other advantage is that dynamical control mechanism, which improves the agility of the system. Because the schedules are built up dynamically through flexible interactions, they can be readily altered in the event of delays or unexpected contingencies. The implementation of the testing system has increased throughput and greater robustness to failure (Jennings and Bussmann 2003), which also shows a good prospect for the agent-based manufacturing system.

2. Multi-Agent Based Distributed Product Design and Manufacturing Planning

In this section, one agent-based distributed manufacturing system developed in the National University of Singapore (NUS) (Sun 1999; Jia 2001; Wang 2001; Jia, Fuh et al. 2002; Li 2002; Jia, Ong et al. 2004; Mahesh, Fuh et al. 2005) is presented, which studies a multi-agent based approach to integrate product design, manufacturability analysis, process planning and scheduling in a distributed manner. Under this framework, geographically dispersed entities are allowed to work cooperatively towards overall manufacturing system goals. The system model considers constraints and requirements from the different product development cycles and manufacturing.

The system adopted a federator structure to model the various manufacturing functional departments in a manufacturing process that includes design, manufacturability evaluation, process planning, scheduling and shop floor control. In the system, the different functional departments dispersed in different geographical sites are encapsulated into agents. Facilitator architecture is selected as the system architecture, which comprises a facilitator agent, a console agent and several service agents. The facilitator is responsible for the decomposition and dispatch of tasks, and resolving conflicts of system execution. The console agent acts as an interacting interface between designers and the system. The service agent models the functional modules of different product development phases, including Designing Agent, Manufacturing Resource Agent, Manufacturability Evaluation Agent, Process Planning Agent, Scheduling Agent, etc. Each functional agent represents a participant involved in a different product development and manufacturing phase. Facilitator plays the central and control roles in the whole environment, and each participant can know the status and requirements of other participants in real-time through it.

2.1 System framework design

In a multi-agent manufacturing environment, the isolated and distributed functional sub-systems can be integrated by encapsulating them as interacting agents. Each agent is specifically in charge of a particular design or manufacturing activity. The agents communicate and exchange information to solve problems in a collaborative manner. The components interact dynamically, addressing the different manufacturing planning issues collaboratively, thereby avoiding costly manual iterations. The federated structure adopted as the architecture ensures the openness of the system, which makes the functional agents can join or leave without having to halt or to reinitialize the other agents' work in progress. The different components can interact dynamically in such platform, addressing the product design and manufacturing planning issues efficiently, and the separate domains of expertise may reside at distributed sites on a network but collaborate with others on a common task, which results in great time saving in terms of data transfer and interpretation. Some legacy software tools can also be wrapped into Java-based agents having the capability of interacting with others.

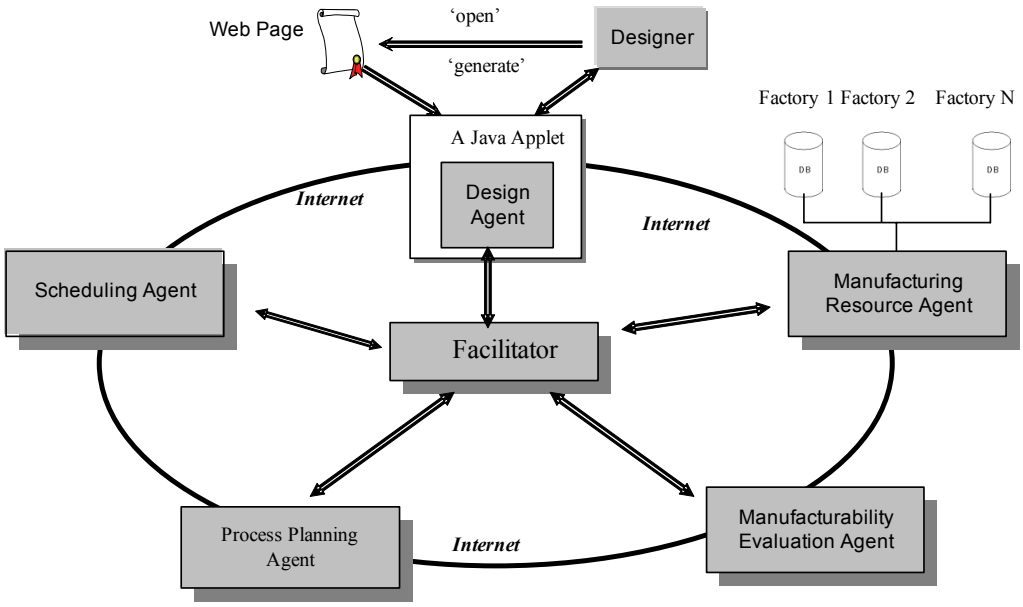


Figure 3. System architecture

The system architecture, as depicted in figure 3, which is composed of six components: Facilitator, Design Agent (D-Agent), Manufacturing Resource Agent (MR-Agent), Manufacturability Evaluation Agent (ME-Agent), Process Planning Agent (PP-Agent) and Scheduling Agent (S-Agent). The last four service agents (encapsulated pre-existing legacy tools) and the D-Agent interact with each other through the Facilitator.

2.2 Agent coordination and individual agents

The function of each agent during this framework is defined as follows:

(1) Facilitator:
It is responsible for the management of interactions and conflict resolution in the agent community. Once any agent joins or leaves the system, it needs to register with status and information to the Facilitator. Thus, the Facilitator "knows" which agent is available, and any function each agent has. Each executive agent receives tasks from the facilitator, and feedbacks the results to it after completing. The Facilitator also routes the requests information received to appropriate agents based on its knowledge of capabilities of each agent, which is known as content-based routing. In performing this task, the Facilita-

tor can go beyond a simple pattern matching by translating messages, decomposing problems into sub-problems, and scheduling the work on those sub-problems.

(2) D-Agent:
It is the interface between the system and the designers, by which the design information of product is submitted to other agents for manufacturability analysis, process plan and scheduling generation. Once the designed parts need further modifications, the information will be also sent back. It also advises the designer to make necessary modifications to the design.

(3) MR-Agent:
This agent manages manufacturing resource models from those different factories of the system, which contain information of available shop-floor resources, including machines and tools, and the capability of these resources. These models are stored in individual databases located at different local sites. The agent is in charge of looking for a suitable capability for manufacturability evaluation.

(4) ME-Agent:
This agent is responsible for the manufacturability evaluation of the product design with the help of acquiring capability information from the MR-Agent. It returns information about conflicts to the Facilitator, as well as suggestions for product redesign or a suitable capability model.

(5) PP-Agent:
This agent is responsible for the generation of an optimal process plan based on the design and selected resources.

(6) S-Agent:
This agent makes the manufacturing scheduling for parts, and feedback to the facilitator.

In order to manage the product and manufacturing information, each agent has a local database, which is used to store and manage messages received from other agents. Furthermore, with the Internet, all of these individual databases are integrated into a distributed database to improve the execution efficiency of the system.

Under such framework, the manufacturing tasks are usually executed by the cooperation of several different related agents. The takes are decomposed firstly into some sub-tasks and dispatched to the destination for process, which needs the cooperation and coordination of the agents in the system. In the project, the agents of the system make negotiations trying to find optimal trade-offs among their local preferences and other agents' preferences and make commitments based on the negotiation results. The task-completing process in the system consists of the following steps:(1)A remote designer submits design information of a product/part to the Facilitator via the D-Agent; (2)The Facilitator decomposes the task into mutually interrelated sub-tasks from a global objective point of view; (3)The Facilitator dispatches the sub-tasks to appropriate executive agents; (4)Executive agents complete their sub-tasks independently; (5)The Facilitator detects conflicts;(6)The Facilitator defines and refines the shared space of interacting agents to remove conflicts; and (7)Conflict resolution.

2.3 Agent definition

2.3.1 Manufacturability evaluation agent (MEA)

This agent is in charge of evaluating the manufacturability of a designed part during the design phrase and sending the modifying information to the design agent if necessary. The agent judges the manufacturability for one part and selects the most preferable machining plan alternatives considering the part's dimensions, tolerances, and surface finishes, along with the availability and capabilities of machine tools and tooling constraints. MEA is, firstly, to check whether the design features are defined correctly; secondly to check if design features can be machined or not based on the current available manufacturing resources; and thirdly to find out all available manufacturing resources that can fabricate the product.

Manufacturability Evaluation

After receiving the feature information from the Facilitator, the ME-Agent carries out a manufacturability evaluation process for the design. It starts with a local manufacturability evaluation on the model in terms of design flaws. Any local conflict detected in the process is notified to the D-Agent by the Facilitator for design modification. Upon the completion of local manufacturability

evaluations, the ME-Agent makes a global manufacturability evaluation on the model by acquiring a factory model from the RC-Agent.

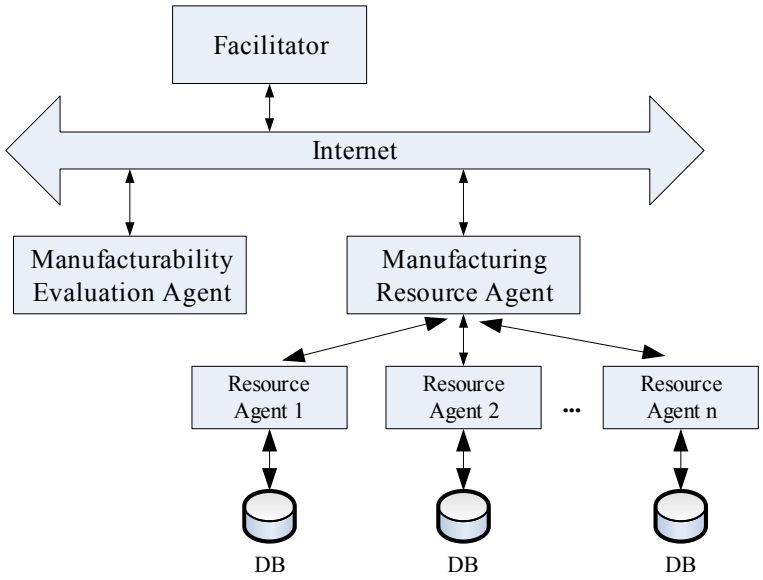


Figure 4. Manufacturability evaluation modules

The RC-Agent checks the availability of various resources required to create the part. Any global conflict is notified to the MR-Agent for it to search for a suitable factory model as a substitute to the former. The two analysis processes are repeatedly executed until no conflict is found on the part model and the agent analyses four manufacturability issues as follows:

(1)Design flaws:
The design flaws refer to those features which are difficult or impossible to machine. The ME-Agent identifies all possible flaws to avoid much higher rectification cost at an advanced stage.

(2)Tool accessibility:
The ME-Agent checks the tool accessibility of each feature. A feature may be inaccessible due to its location and orientation. For those flaws, the cutting tool may not work correctly and need to be modified. (3)Availability of cutters: The ME-Agent checks whether all the required cutting tools to machine the part are available in the factory under consideration. If some machined features ex-

ceed the manufacturing capability of the cutting tools available, then, they will need further revision on the design.

(4)Tolerance and surface finish requirements:
The ME-Agent also need to check the capability of machines contained in the factory against the tolerance and surface finish requirements in the part model.

2.3.2 Resource coordination agent (RCA)

The RCA collects the manufacturing resource information from the work shops and factories with the help of RA. As the system is open and heterogeneous, it is needed that the agent can support flexibility and extensibility in dynamic manufacturing environments, which means that resource coordination, including interaction with users via the other agents, should be sensitive to both the query context and the currently available information. The RCA has the following functionality:

- For a manufacturing resource request, multiple instantiations of the search node may be created;
- Task execution is data-driven;
- Parse the query, and decompose it if appropriate parsing involves getting an ontological model;
- Construct KIF queries based on the SQL queries' contents, and query the Resource Agent using the KIF queries to find relevant resources.

2.3.3 Resource agent

The Resource Agent (RA) manages the data contained in manufacturing information source (e.g., distributed systems database) available to retrieve and update. It acts as an interface between the local data source and other agents, hiding details of the local data organization and representation. To accomplish this task, an RA can announce and update its presence, location and the description of its contents to the RCA. There are two types of information that is of potential interest to other agents:

1. value (ranges) of chosen data objects,
2. the set of operations allowed on the data. The operations range from a single read/update to more complicated data analysis operations. The advertisement information can be sent to the RCA.

RA also needs to answer queries from other agents. It has to translate queries expressing in a common query language (such as KQML) into a language understood by the underlying system. This translation is facilitated by a mapping between the local data concepts and terms, as well as between the common query language syntax, semantics and operators, and those of the native language. Once the queries are translated, the RA sends them to the manufacturing resource database for execution, and translates the answers back into the format understood by the RCA. Additionally, RA and the underlying data source may group certain operations requested by other agents into a local transaction. In addition, RA provides limited transaction capabilities for global resource transaction.

2.3.4 Process planning agent

Process planning agent is developed to generate the optimal or near-optimal process plans for designed part based on the criterion chosen. Under the distributed environment, factories possessing various machines and tools are dispersed at different geographical locations, and usually different manufacturing capabilities are selected to achieve the highest production efficiency. When jobs requiring several operations are received, feasible process plans are produced by available factories according to the precedence relationships of the operations. The final optimal or near-optimal process plan will emerge after comparison of all the feasible process plans. In order to realize and optimize the process plan for the distributed manufacturing systems, the Genetic Algorithm (GA) methodology is adopted as an optimizing method. The GA method is composed of four operations as following: encoding, population initialization, reproduction, and chromosome evaluation and selection.

Encoding

When dealing with a distributed manufacturing system, a chromosome not only needs to represent the sequence of the operations but also indicate which factory this process plan comes from. Therefore, the identity number of the factory will be placed as the first gene of each chromosome no matter how the other genes are randomly arranged. Each other gene comprises the operation ID and corresponding machine, tool and tool access direction (TAD), which will be used to accomplish the operation. As a result, a process plan including

factory and operation information will be represented by a random combination of genes.

Population Initialization

The generation of the initial population in GA is usually done randomly; however, the initial population must consist of strings of valid sequences, satisfying all precedence relations. Once the number of initialized chromosomes is prescribed, the procedures of initialization are given as follows:

(1) Randomly select one factory ID number from the available factory list.
(2) Randomly select one operation among those, which have no predecessors.
(3) Among the remaining operations, randomly select one which has no predecessor or which either predecessor all have already been selected.
(4) Repeat step (3) until each operation has been selected for only once.
(5) Revisit the first selected operation.
(6) Randomly select machines and tools from the selected factory that can be used for performing the operation.
(7) Randomly select one amongst all possible TADs for the operation.
(8) Repeat steps (6) and (7), until each operation has been assigned a machine, tool and TAD.
(9) Repeat steps (1) to (8) until the number of prescribed chromosome is reached.

Reproduction

A genetic search starts with a randomly generated initial population; further generations are created by applying GA operators. This eventually leads to a generation of high performing individuals. There are usually three operators in a typical genetic algorithm, namely crossover operator, mutation operator and inversion operator. In the proposed GA, mutation and crossover operators are used for gene recombination, which is also called offspring generation.

Crossover

In this step, a crossover operator is adopted to ensure the local precedence of operations is met and a feasible offspring is generated. The procedure of the crossover operation is described as follows:

(1) Randomly choose two chromosomes as parent chromosomes.

(2) Based on the chromosome length, two crossover points are randomly gen erated to select a segment in one parent. Each string is then divided into three parts, the left side, the middle segment and the right side according to the cutting points.
(3) Copy the left side and right side of parent 1 to form the left side and right side of child 1. According to the order of operations in parent 2, the opera tor constructs the middle segment of child 1 with operations of parent 2, whose IDs are the same as operations of the middle segment in parent 1.
(4) The role of these parents will then be exchanged in order to generate an other offspring child 2.
(5) Re-assign machines and tools to the operations in the middle segment to legalize the offspring chromosomes according to the factory id.

Mutation

Mutation operator is used to investigate some of the unvisited points in the search space and also to avoid pre-mature convergence of the entire feasible space caused by some super chromosomes. A typical GA mutation makes changes by simply exchanging the positions of some randomly selected genes. However, for the distributed manufacturing system, mutation once is not enough to explore all the feasible operation sequences, as well as compare the different selected factory combination. In the proposed GA process, mutation happens to the chromosomes twice, one is for selected factory (mutation 1) and the other is for the operations (mutation 2).

The procedure of mutation 1 is described as follows:

(1) Randomly select a factory ID from the factory ID list, which is different from the current one.
(2) In order to legalize the chromosome, machines and tools will be re- as signed for all the operations according to the new factory-id.

The procedure of mutation 2 is depicted as follows:
(1) Randomly choose a chromosome.
(2) Choose several pairs of genes stochastically and permute their positions.

Chromosome Evaluation

When all the individuals (process plans) in the population have been determined to be feasible, i.e. an operation precedence is guaranteed, they can be

evaluated based on the objective functions. The objective of the CAPP problem is to obtain an optimal operation sequence that results in optimizing resources and minimizing production costs as well as processing time. In this research, two optimization criteria, i.e. minimum processing times and minimum production cost, are employed to calculate the fitness of each process plan and measure the efficiency of a manufacturing system. After the completion of the manufacturability evaluation, the PP-Agent generates an optimal process plan for the factory supplied by the ME-Agent. The agent first constructs the solution space by identifying all the possible operation-methods (OpM's) for machining each feature and then uses a GA to search for the best plan according to a specific criterion. The criterion can be constructed by using the following cost factors:

Machine cost (*MC*)

$$MC = \sum_{i=1}^{n} MCI_i \qquad (1)$$

where n is the total number of OpM's and MCI_i is the machine cost index for using machine-i, a constant for a particular machine.

Tool cost (*TC*)

$$TC = \sum_{i=1}^{n} TCI_i \qquad (2)$$

where TCI_i is the tool cost index for using tool-i, a constant for a particular tool.

Machine change cost (*MCC*): a machine change is needed when two adjacent operations are performed on different machines.

$$MCC = MCCI \times \sum_{i}^{n-1} \Omega(M_{i+1} - M_i) \qquad (3)$$

where *MCCI* is the machine change cost index, a constant and M_i is the ID of the machine used for operation i.

Setup change cost (*SCC*): a setup change is needed when two adjacent OpM's performed on the same machine have different Tool Approaching Directions (TADs).

$$SCC = SCCI \times \sum_{i=1}^{n-1}((1-\Omega(M_{i+1}-M_i))\times\Omega(TAD_{i+1}-TAD_i)) \qquad (4)$$

where *SCCI* is the setup change cost index, a constant.

Tool change cost (*TCC*): a tool change is needed when two adjacent OpM's performed on the same machine use different tools.

$$TCC = TCCI \times \sum_{i=1}^{n-1}((1-\Omega(M_{i+1}-M_i))\times\Omega(T_{i+1}-T_i)) \qquad (5)$$

where *TCCI* is the tool change cost index, a constant.

2.3.5 Scheduling agent

In a distributed manufacturing environment, every factory has its particular niche areas and can outperform other factories in those specific aspects; therefore, one batch of products are to be finished by the most suitable factory combination with the considerations of low cost and short make span. To meet such requirement, each available candidate factory will submit a feasible process plan for a product in the batch it is capable of processing. The agent then compares all the candidate factories, selects the final factory combination for the products, and meanwhile arranges the manufacturing operations in an optimal sequence. In brief, to generate an optimal schedule in a distributed manufacturing environment, there are two determining factors: the selected factory (or process plan) for every product and operations' sequencing of the machines in the factories. Here, GA is also used as the optimization method to achieve better scheduling results.

The scheduling agent is mainly composed of four major components: the scheduling kernel including the GA engine, the stand-alone scheduling module, scheduling agent module, and the e-scheduling module. Among these four parts, the scheduling kernel can be categorized as the basic component, while the stand-alone module, scheduling agent module and e-scheduling module can be categorized as the application components. The basic component could be combined with any of the three application components to form a scheduling entity, which can carry out the scheduling tasks solely based on a specified scheduling objective.

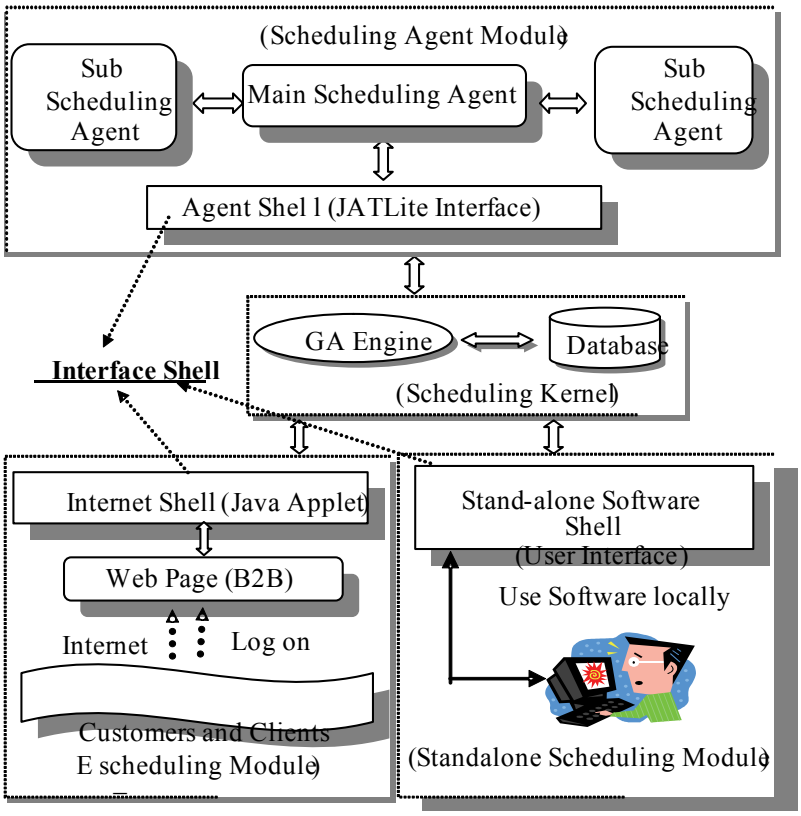


Figure 5. Scheduling agent

Integrating with the scheduling kernel, each of the three application modules has a particular working mode.

The scheduling agent module includes one main scheduling agent (MSA) and many distributed sub-scheduling agents (SSAs). The scheduling agent structure consists of not only the parallel sub-scheduling agents (SSAs) but also a main scheduling agent (MSA). Such a structure can effectively facilitate the information communication among the different production participants within the distributed scheduling system. The MSA, used by the production management department, is responsible for regulating the job production operations, collecting the factories' information from the SSAs and then, making scheduling through the scheduling kernel. After the scheduling, the details about which job is manufactured in which factory in what time and what machine is responsible for what operation can be obtained and sent to each SSA. To finish the job production cooperatively, the SSAs are distributed in the

manufacturing factories, each representing a factory, collecting the factory's working status, detecting its dispatched jobs, and checking the manufacturing progress. In addition, if any contingency happens in any factory, the SSA will send the accident information to the MSA, wait and execute the re-arranged schedule made by MSA.

Genetic Algorithm for distributed Scheduling
Chromosome representation
To handle the distributed scheduling problems, the genes in the GA must comprise the two dominant factors in the distributed manufacturing environment, i.e., the selected factory for every job with the corresponding process plan and the operation processing sequence. Here, a four-element string is used as a gene to represent an operation of a job. The first element is the identification number of the selected factory that is used to process the job, and the next three elements represent the ID of a job. Thus, a schedule can be represented by a combination of genes, which is called "chromosome" in the GA terminology, as long as the combination comprises all the operations of the jobs. Every operation processing sequence can be interpreted according to its occurrence order in the chromosome. As such, for any distributed scheduling problem, a random feasible schedule, including which job goes to which factory and what is the operation processing sequence, can be encoded using a combination of genes.

Chromosome population initialization
To begin the search for the best chromosome and correspondingly, the factory combination and the optimal schedule the chromosome represents, a number of chromosomes are initialized as follows:

(1) Create the lists of ID number of the feasible factories for every job. If job 'j03' can be processed in factories '1', '3' and '5', then the list for job 'j03' will be 1, 3, and 5.
(2) For every job, randomly select a factory ID number from the job's feasible factory list.
(3) According to the job's process plan in the selected factory, produce the job's operation genes. For example: the genes, 1j03-1j03-1j03, mean the first, second and third (last) operation of job 'j03'. All the operations will be manufactured in factory '1'.
(4) Repeat step (2) and step (3), until there is no job left.

(5) Combine and mix all the produced genes together in a stochastic way to form a chromosome.
(6) Repeat step (5), until a prescribed number of chromosome populations is formed.

Genetic operators
The power and potential of GA method come from the gene recombination, including crossover, mutation and inversion, which explore all the possible search space. In this proposed GA, two genetic operators, *crossover* and *mutation*, are employed for the gene recombination, which is called offspring generation. The procedures for the crossover operation are described as follows:

(1) Choose two chromosomes as parents and exchange a random partial string (genes) in the two parents to generate two offspring chromosomes.
(2) Regulate (delete or compensate) genes in each offspring chromosome so that it comprises the operations of all the jobs and inherits the genetic traits of their parents.

Because the sequence of the genes in the chromosome expresses the jobs' operation processing sequence, after the crossover, the processing sequence for every operation (gene) in the schedule (chromosome) is changed. It is important to note that the precedence of a job's operations will not be affected by the crossover because every gene (operation) of a job in the chromosome is not fixed to a specific operation of the job and it is interpreted according to the order of occurrence in the sequence for a given chromosome.

The crossover is carried out under the assumption that a factory has been selected for every job. Yet, in the distributed manufacturing environment, the randomly selected factory is, by and large, not necessarily the most suitable one for the specific job. Therefore, another genetic operator, gene mutation, is employed twice for modifying the operation processing sequence again as well as changing the selected candidate factory for the jobs. The mutation procedures are as follows:

(1) In a chromosome, choose several pairs of genes stochastically and permute their positions (Mutation1).
(2) Select one gene (job) from the chromosome in a random manner. In the meantime, randomly select a factory ID number from the job's feasible fac

tory list, which has been created in step (1) of chromosome population ini tialization.

(3) In the selected chromosome, replace the first element (represents the ID of the selected factories) of all the genes (the gene selected in step 2) with the newly selected factory ID number (Mutation2). The aim of the step is to change the factory selection used to manufacture the job.

(4) To keep the consistency of the chromosomes, apply the same change of fac tory selection for the gene (job) to all the other chromosomes in their gen eration.

In step (1), the first mutation changes the jobs' operation sequences, while from step (2) to step (4), the second mutation changes the factory ID, which is used for a randomly selected job.

After gene crossover and mutation, the parent chromosomes can produce a generation of offspring. In the same way, the offspring could reproduce the next generation of offspring. Thus, through this iteration, numerous chromosomes are produced and can be compared. Correspondingly during this process, many possible factory combinations and job operation processing sequences are formed and analyzed. The application of such gene crossover and mutation in this GA ensures each product (job) to be manufactured in its most suitable factory. In addition, the production schedule of each of the factories in the distributed environment can be generated concurrently.

3. Prototype Implementation

In order to verify the effectiveness of the system, a prototype of the proposed system has been developed for the integration of design, manufacturability analysis, and process planning. The developed system includes a unique facilitator, and several functional agents, which are organized according to the framework of Figure 3. JATLite is selected as the template for the agents' development.

Each agent is composed of the following components: network interface, local knowledge model and domain knowledge model. All the agents in the system use the common communication protocol, KQML, for concurrent negotiations. KQML is conceived as both a message format and a message-handling protocol to support run-time knowledge sharing among agents. It is essentially a wrapper to encode context sensitive information. The KQML language is di-

vided into three layers: the content layer, the message layer, and the communication layer, as shown in figure 6.

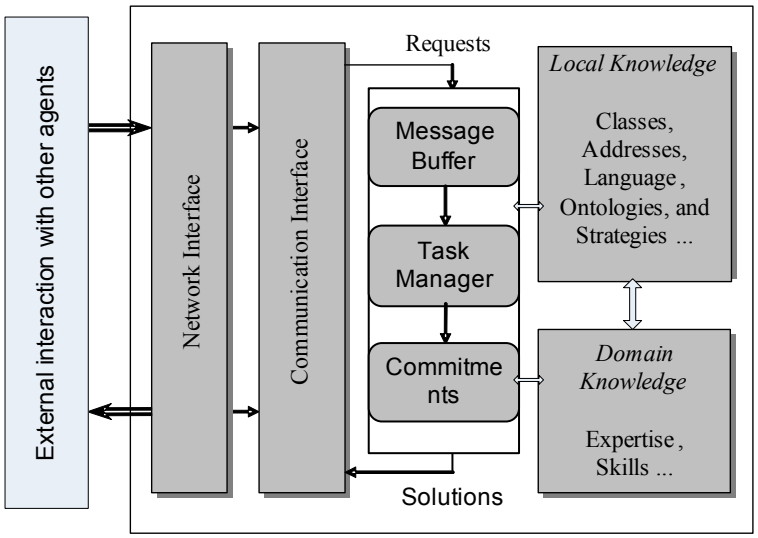


Figure 6. Internal structure of an agent

In the system, agents are autonomous cognitive entities, with deductive, storage and communication capabilities. Autonomy in this case means that an agent can function independently from any other agent. There are three kinds of functional agent in the system. Each has different internal structure, and can be decomposed into the following components:

(1) A network interface: It couples the agent to the network.
(2) A communication interface: It is composed of several methods or functions for communicating with other agents.
(3) A local function module: Resources in this model include Java classes to perform the desired functions, other agent names, messaging types in KQML syntax. The functional module also provides the facility of inference and collaboration facilities. The collaboration facilities are the general plans for coordination behaviour that balances the agent's local interests with global (community) interests.
(4) Agent Knowledge base: The model comprises expertise that is required for an agent to perform functional tasks, and skills that may be methods for ac tivating actions corresponding to the received requests.

A snap shot on the prototype system (PPA and MSA) is shown in figure 7 below.

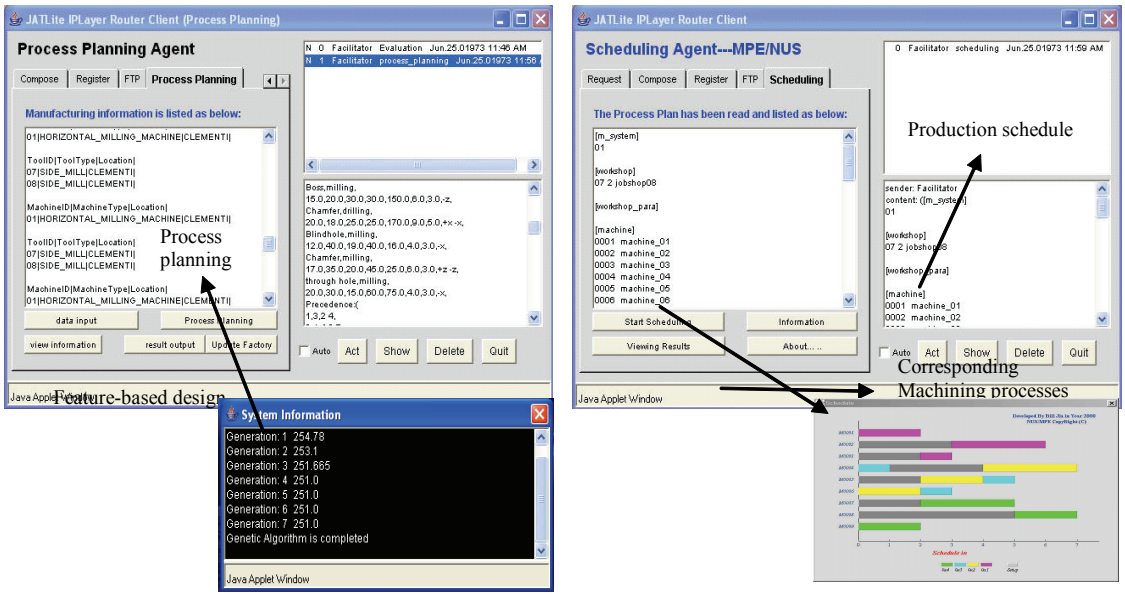


Figure 7. Optimal process plans and production schedules generated from the PPA and MSA respectively

4. Conclusion and Future Work

Agent theory was developed from the distributed artificial intelligence around 20 years before. As its nature and characteristic are suitable for distribute problems solving, agent theory has been viewed as a promising theory and methodology applied to a distributed environment. It has now achieved some promising results in industrial applications. The rapid development of Internet has also provided a good tool and suitable platform for agent's application. With the use of Internet and advancement of communication methods, agents can be dispersed in different geographical places, which make it easy for collaboration of different partners in a manufacturing supply chain.

Except the description to some basic theories and key issues in agent-based systems and the introduction to some typical manufacturing systems, this chapter also introduces one agent-based distributed manufacturing system developed in NUS. The objective of this research is to develop a distributed

collaborative design environment for supporting cooperation among existing engineering tools organized as independent agents on different platforms. A facilitator multi-agent system architecture is discussed for distributed collaborative product design and manufacturing planning, and one prototype for collaborative design and planning on machining processes, which has been developed as a proof-of-concept system, demonstrates the capability of such a multi-agent approach to integrate design, manufacturability evaluation, process planning and scheduling efficiently. This approach could be extended to include other product life cycle considerations, collaboratively, at the design stage. Some of the advantage of the system over the traditional mode can be achieved in several facets as follows:

1) Distributed function execution improves the efficiency and the reliability of the system, thus to increase the responsiveness of the enterprise to the market requirements. In the system, the major functional tasks are distrib uted, and each one agent needs to focus on one task execution, which can improve the efficiency of the process. Furthermore, as the functions are executed dispersedly, once some functional agent malfunction, the rest can still work, which also improves the reliability of the whole system.

2) Open architecture to make the system good adaptability and easy exten sion. As the system adopts a facilitator structure, the newly added agent can execute the system tasks if only registering with the facilitator; at the same time the functional agent can leave the system only need to inform the facilitator and not influence other's progress. Moreover, the newly functional agents can be also integrated into the system without interfering other agents' function. All of these can make the system good adaptability and easy extension, thus to improve the agility of the system.

3) The agent-based system provides a platform to realize the concurrent func tion in the product development. The different departments from design through manufacture to customer service can join together for one product design, thus to improve the quality and efficiency of the final product de velopment.

Although some promising results have been achieved in the prototype and previous research work on the agent theory, there are still difficulties to be

overcome for its wider application in industry. Some of the challenges faced include:

1) Effective coordination and negotiation methods for MAS. Coordination and negation methods are the basis and key issues for intelligent agent sys tems. It has been under study for a long time and a variety of methods have been proposed, but one effective and efficient method for the agent-based system is still needed.

2) Methods to incorporate and agentify the legacy manufacturing systems and tools. Now, there are various computer-aided software systems ap plied in manufacturing system and industrial scenario, but there is still no successful methodology to agentify these legacy modules in the agent sys tems. This is one bottleneck that impedes a wider development of agent-based methods for industrial applications.

3). Agent theory provides a decentralized solution for complex systems, de composing and conquering make the agents easy to deal with sub-tasks. But one problem emerges is that the local optimization can not result in a global optimal result for the whole system. How to achieve a global opti mal solution in the agent-based system still needs a further study.

5. References

Baker, A. D., H. V. D. Parunak, et al. (1999). *Internet-based Manufacturing:* A Perspective from the AARIA Project, Enterprise Action Group.

Bussmann, S. and K. Schild (2001). An agent-based approach to the control of flexible production systems. ETFA 2001. *The 8th International Conference on Emerging Technologies and Factory Automation. Proceedings*, 15-18 Oct. 2001, Antibes-Juan les Pins, France, IEEE.

Bussmann, S. and J. Sieverding (2001). Holonic control of an engine assembly plant: an industrial evaluation. *Proceedings of IEEE International Conference on Systems, Man & Cybernetics*, 7-10 Oct. 2001, Tucson, AZ, USA, IEEE.

Castelfranchi, C. (1995). Guarantees for autonomy in cognitive agent architecture. *Proceedings of the workshop on agent theories, architectures, and languages on Intelligent agents*, Amsterdam, the Netherlands, Springer-Verlag New York, Inc.

Davis, R. and R. G. Smith (1983). "Negotiation as a metaphor for distributed problem solving." *Artificial Intelligence* 20(1): 63-109.

F. Bellifemine, G. Caire, et al. (2006). JADE: A White Paper.

FIPA (1997). FIPA 97 Part 2 Version 2.0: *Agent Communication Language Specification*, FIPA.

FIPA (2000(1)). *FIPA Interaction Protocol Library Specification*, FIPA.

FIPA (2000(2)). *FIPA ACL Message Structure Specification*, FIPA.

Gensereth, M. R. and S. P. Ketchpel (1994). "Software Agents." *Communications of the ACM* Vol. 37(No. 7): 48-53.

JADE (2005). *JADE:Java Agent DEvelopment Framework*, http://jade.tilab. com/.

JATLite http://java.stanford.edu/ .

Jennings, N. R. and S. Bussmann (2003). "Agent-based control systems: Why are they suited to engineering complex systems?" *IEEE Control Systems Magazine* 23(3): 61-73.

Jennings, N. R. and M. Wooldridge (1998). *Applications of intelligent agents*, Springer-Verlag New York, Inc.

Jia, H. Z.,2001, *Internet-based multi-functional scheduling for distributed manufacturing systems*, *M. Eng Thesis* ,National University of Singapore,Singapore.

Jia, H. Z., J. Y. H. Fuh, et al. (2002). "Web-based Multi-functional Scheduling System for a Distributed Manufacturing Environment." *Concurrent Engineering* 10(1): 27-39.

Jia, H. Z., S. K. Ong, et al. (2004). "An adaptive and upgradable agent-based system for coordinated product development and manufacture." *Robotics and Computer-Integrated Manufacturing* 20(2): 79-90.

Li, L.,2002, Agent-based computer-aided process planning for distributed manufacturing systems., *M. Eng Thesis*, National University of Singapore,Singapore.

Mahesh, M., J.Y.H.Fuh, et al. (2005). "Towards A Generic Distributed and Collaborative Digital Manufacturing", *Proceedings of the International Manufacturing Leaders Forum on Global Competitive Manufacturing*, Adelaide, Australia.

Maturana, F., W. Shen, et al. (1999). "MetaMorph: an adaptive agent-based architecture for intelligent manufacturing." *International Journal of Production Research* 37(10): 2159-73.

Parunak, H. V. D., A. D. Baker, et al. (1998). "The AARIA Agent Architecture: from Manufacturing Requirements to Agent-Based System Design".

WorkshopProc. on Agent-Based Manufacturing, ICAA'98, Minneapolis, MN.

Parunak, H. V. D., R. Savit, et al. (1998). Agent-Based Modeling vs. Equation-Based Modeling: A Case Study and Users' Guide. *Proceedings of the First International Workshop on Multi-Agent Systems and Agent-Based Simulation*, Springer-Verlag, London, UK.

Peng, Y., T. Finin, et al. (1998). "A Multi-Agent System for Enterprise Integration." *International Journal of Agile Manufacturing,* vol. 1(No. 2): 201-212.

Shen, W. (2002). "Distributed manufacturing scheduling using intelligent agents." *Intelligent Systems, IEEE [see also IEEE Intelligent Systems and Their Applications]* 17(1): 88-94.

Shen, W., F. Maturana, et al. (1998). Learning in Agent-Based Manufacturing Systems. *Proceedings of AI & Manufacturing Research Planning Workshop*, IAlbuquerque, NM, The AAAI Press,.

Shen, W., D. Xue, et al. (1998). An Agent-Based Manufacturing Enterprise Infrastructure for Distributed Integrated Intelligent Manufacturing Systems. *Proceedings of the 3rd International Conference on the Practical Applications of Agents and Multi-Agent Systems*, London, UK.

Smith, R. G. (1980). "The contract net protocol: high level communication and control in a distributed problem solver." *IEEE Transactions on Computers C*-29(12): 1104-1113.

Smith, R. G. (1988). The contract net protocol: high-level communication and control in a distributed problem solver *Distributed Artificial Intelligence* Morgan Kaufmann Publishers Inc.**:** 357-366

Sun, J.,1999, Agent-based product design and planning for distributed concurrent engineering, *M. Eng Thesis*, National University of Singapore,Singapore.

Wang, G.,2001, Agent-based manufactuirng service system, *M. Eng. Thesis*, National University of Singapore.

Weiss, G. (1999). *Muliagent Systems: A Modern Approach to Distributed Artificial Intelligence*. Cambridge, Massachusetts, The MIT Press.

Wooldridge, M. and N. R. Jennings (1995). "Intelligent Agents: Theories and Practices." *Knowledge Engineering Review*: 115-152.

2

The Cobasa Architecture as an Answer to Shop Floor Agility

Jose Barata

1. Introduction

Shop floor agility is a central problem in current manufacturing companies. Internal and external constraints, such as growing number of product variants and volatile markets, are changing the way these companies operate by requiring continuous adaptations or reconfigurations of their shop floors. This need for continuous shop floor changes is so important that finding a solution to this problem would offer a competitive advantage to contemporary manufacturing companies.

The central issue is, therefore, which techniques, methods, and tools are appropriate to address shop floors whose life cycles are no more static but show high level of dynamics. In other words, how to make the process of changing and adapting the shop floor fast, cost effective, and easy. The long history of industrial systems automation shows that the problem of developing and maintaining agile shop floors cannot be solved without an integrated view, which accommodate the different perspectives and actors involved in the various phases of the life cycle of these systems. Moreover, supporting methods and tools should be designed and developed to accommodate the continuous evolution of the manufacturing systems along their life cycle phases – a problem of shop floor reengineering. The design and development of a methodology to address shop floor reengineering is thus an important research issue aiming to improve shop floor agility, and, therefore, increasing the global competitiveness of contemporary manufacturing companies.

Agility is a fundamental requirement for modern manufacturing companies in order to face challenges provoked by the globalisation, changes on environment and working conditions regulations, improved standards for quality, fast technological mutation, and changes of the production paradigms. The turbulent and continuous market changes have impacts at different levels, from company management to shop floor. Only companies that exhibit highly

adaptable structures and processes can cope with such harsh environments. Furthermore, the capability to rapidly change the shop floor infrastructure is a fundamental condition to allow participation of manufacturing enterprises in dynamic cooperative networks. Networked enterprise associations, such as virtual enterprises, advanced supply chains, etc. are examples of cooperative structures created to cope with the mentioned aspects. Manufacturing companies wishing to join these networked structures need to be highly adaptable in order to cope with the requirements imposed by very dynamic and unpredictable changes. In such scenarios, agility means more than being flexible or lean. Flexibility in this context means that a company can easily adapt itself to produce a range of products (mostly predetermined), while lean essentially means producing without waste. On the other hand, agility corresponds to operating efficiently but in a competitive environment dominated by change and uncertainty (Goldman et al. 1995), which means adaptation to conditions that are not determined or foreseen a-priori. The participation in dynamic (and temporary) organisations requires agile adaptation of the enterprise to each new business scenario, namely in terms of its manufacturing capabilities, processes, capacities, etc.

It is worth noting that the need of methods and tools to manage the process of change was first felt at the company's higher management levels. This is not surprising because the external business conditions are initially felt at managerial levels. Therefore, in past research the processes of change (reengineering/adaptation) have been addressed mostly at the level of business process reengineering and information technology infrastructures. Little attention, however, has been devoted to the changes needed at the manufacturing system level and, yet, the shop floor suffers a continuous evolution along its life cycle and it is subject to ever increasing demands on its flexibility. In fact, despite the efforts put in the creation of agile organisational structures, little attention has been devoted to the agility of the shop floor, even if many research works have been focused on flexible assembly and flexible manufacturing systems (Gullander 1999; Onori 1996; Vos 2001; Zwegers 1998). There are some research works (Huff and Edwards 1999; Koren et al. 1999; Mehrabi et al. 2000), in which shop floor agility is achieved by focusing on the reconfigurability of the individual equipment rather than considering a global agility approach. Nevertheless the situation is that a non-agile shop floor seriously limits the global agility of a manufacturing company even if its higher levels are agile. A good indication of how great the demand for agile shops-floors is within manufacturing companies is the increasing number of shop floor altera-

tion projects (Barata and Camarinha-Matos 2000). As long as people in the shop floor are faced with the need to often change (adapt) their production systems, the need to have methods and tools to cope with such challenge increases significantly.

A particularly critical element in a shop floor reengineering process is the control system. Current control/supervision systems are not agile because any shop floor change requires programming modifications, which imply the need for qualified programmers, usually not available in manufacturing SMEs. To worsen the situation, the changes (even small changes) might affect the global system architecture, which inevitably increases the programming effort and the potential for side-effect errors. It is therefore vital to develop approaches, and new methods and tools that eliminate or reduce these problems, making the process of change (re-engineering) faster and easier, focusing on *configuration* instead of *codification*. Hence this chapter is focused on the reengineering aspects required by the control/supervision architecture, which covers an important part of any global life cycle support methodology.

The proposed architecture to improve shop floor reengineering (CoBASA) aims at accommodating the following requirements:

- **Modularity**. Manufacturing systems should be created as compositions of modularised manufacturing components, which become basic building blocks. The building blocks should be developed on the basis of the processes they are to cater for.
- **Configuration rather than programming**. The addition or removal of any manufacturing component (basic building block) should be done smoothly, without or with minimal programming effort. The system composition and its behaviour are established by configuring the relationships among modules, using contractual mechanisms.
- **High reusability**. The building blocks should be reused for as long as possible, and easily updated for further reuse.
- **Legacy systems migration**. Legacy and heterogeneous controllers should be considered in the global architectures and a process should be found out to integrate them in the new agile architecture.

Reducing the programming effort that is usually required whenever any changes or adaptations take place in the shop floor becomes one of the most important requirements for the proposed architecture. The main question being addressed in this chapter and which the CoBASA architecture intends to answer is highlighted below:

Question
Which methods and tools should be developed to make currentmanufacturing control/supervision systems reusable and swiftly modifiable?

The hypothesis formulated as a basis for CoBASA to address the previous question is defined below:

Hypothesis
Shop floor control/supervision reengineering agility can be achieved if manufacturing systems are abstracted as compositions of modularised manufacturing components (modular approach) that can be reused whenever necessary, and, whose interactions are specified using configuration rather than reprogramming.

The approach followed to tackle the problem raised in the question was the following:

Approach
- The life cycle of shop floor manufacturing systems should explicitly include a new phase: the reengineering phase that captures the time frame in which the systems are being changed or adapted (reengineered).
- Multiagent based systems are a good modelling and implementation paradigm because of their adequacy to create cooperative environments of heterogeneous entities.
- Manufacturing components are agentified (transformed from physical manufacturing components into agents) to become modules that can be used and reused to compose complex systems.
- The different types of manufacturing systems are represented by coalitions or consortia of agentified manufacturing components, which are essentially societies of self-interested and heterogeneous agents whose behaviour is governed by contracts.
- Contract negotiation is the configuration basis required whenever a control/supervision system needs to be changed or adapted.

The proposed architecture Coalition Based Approach for Shopfloor Agility – CoBASA to answer the question raised above is a multiagent based architecture that supports the reengineering process of shop floor control/supervision architectures. In an innovative way, CoBASA uses contracts to govern the relationships between coalition members (manufacturing agents) and postulates a

new methodological approach in which the reengineering process is included within the life cycle. Since the CoBASA approach is based on the concept of manufacturing modules that might be reused, it requires the manufacturing community to structure and classify the process involved, thus leading to a more systematic or structured methodological approach.
Therefore the CoBASA concept considers modularity and plugability as one of its most important foundations principles. The control system architecture being proposed considers that each basic components are modules of manufacturing components that can be reused and plugged or unplugged with reduced programming effort, supporting in this way the plug & produce metaphor.

CoBASA assumes that there is a similarity between the proposed reengineering process and the formation of consortia regulated by contracts in networked enterprise organisations. The problems a company faces in order to join a consortium are analogous to the shop floor adaptation problem. In other words, the formation of a coalition of enterprises to respond to a business opportunity is analogous to the organisation of a set of manufacturing resources in order to perform a given job. The proposed approach is therefore to use the mechanisms and principles developed to support the enterprise integration into dynamic enterprise networks as inspiration for an agile shop floor reengineering process.

2. CoBASA Basic foundations

Human organisations are a good source of inspiration for complex problem solving because they are intrinsically complex and humans are used to creating highly dynamic complex structures to cope with complex problems. The approach followed in the design of CoBASA assumes that there are similarities between the reengineering process and the formation of consortia regulated by contracts in networked organisations. The challenges a company faces to be agile are similar to the shop floor adaptation problem. Furthermore, the problems a company faces in order to join a consortium have some similarity to the adaptation of a manufacturing component (resource) on a shop floor.
Individual companies have a basic set of core competencies or skills. To be able to create/produce complex services or products, when working alone, companies must have a wide range of skills. It is assumed that a service/product is created/produced by the application of a set of skills. However, due to the in-

creasing level of worldwide competition, companies need to focus only on those skills they are best at. The drawback of this decision lies on a lesser capability to create/produce complex services/products by themselves. The solution to survival is cooperating with other companies. Consequently, one or several cooperating partners are called upon to bring the missing skills and resources required to create/produce a complex service/product. At the same time, making cooperation work is not an easy task especially when played by partners that do not have previous knowledge of each other. Some kind of trust is almost mandatory for a successful cooperation.

Accordingly, cooperation can be promoted by a structure called **cluster** or a VE breading environment, already identified in chapter 3. This long-term aggregation of companies with similar interests or affinities, willing to cooperate, increases the trust level and can better accommodate business disturbances. The potential of skills resulting from the whole cluster is bigger than the sum of the skills that were brought in by each individual company because new skills can be composed of the basic ones. This is an interesting characteristic that renders clusters even more attractive, because the whole community being cooperative, enables much more potential to create/produce things. Although the cluster might have a potentially large set of skills, nothing is created/produced by the cluster, which simply possesses a potential for doing things. The cooperating structure that companies use to create/produce things is the **consortium**. A cooperative consortium or Virtual Enterprise is a group of companies that cooperate to reach a common objective. The formation of a consortium is generally triggered by a business opportunity. Different consortia can be formed with subsets of the cluster members. The capabilities of a consortium depend not on the global skills (potential) of each member but on the specific skills they agree to bring into the consortium. This means that the consortium global capabilities might be either larger (because of skill composition in which new skills can be formed from the basic ones) or smaller than the sum of the individual capabilities of its members.

Contracts are the mechanism that regulates the behavioural relationships among consortium members or between consortium members and the "external" client that generated the business opportunity. The same entity constrained by different contracts can have different behaviours. If, for some reason, a company participating in a consortium reduces or increases its core competencies, this change might have an impact on higher-level consortia, which can see their capabilities (skills and capacities) maintained, reduced or

increased. This situation obviously implies a renegotiation of the established contracts.

Similarly, in the manufacturing shop floor the manufacturing components, which are controlled by a diversity of controllers and correspond to companies in the Virtual Enterprise world, are the basic set from which everything is built up. A shop floor can be seen as a micro-society, made up of manufacturing components. The components have basic core capabilities or core competencies (skills) and, through cooperation, can build new capabilities. A robot, for instance, is capable of moving its tool centre point (TCP) and setting different values for speed and acceleration. Its core competencies are represented in Figure 1. A gripper tool, on the other hand, has as basic skills the capability to close (grasp) or open (ungrasp) its jaws. These two components when acting alone can only perform their core skills.

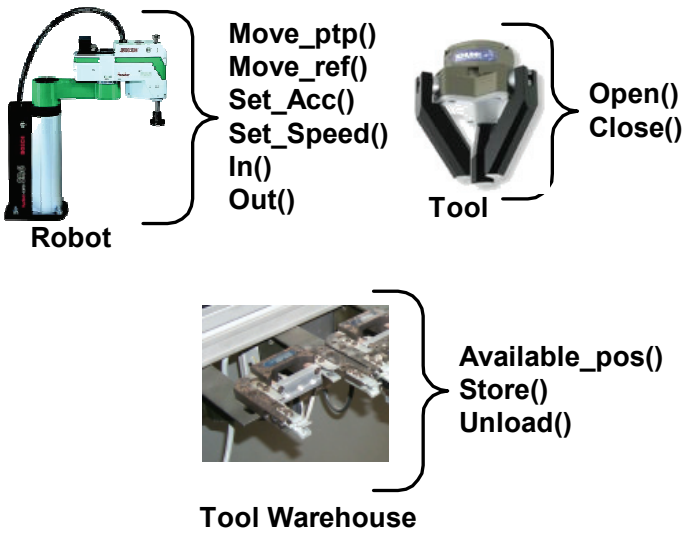


Figure 1. Example of basic manufacturing components and core competencies

However, when they cooperate, it is possible to have a pick-and-place operation that is a composition of the move with the open and close skills. The greater the diversity and complexity of individual capabilities, the greater are the chances of building more complex capabilities. In the architecture being proposed every manufacturing component e.g. robots, tools, fixing devices, is associated to an agent that represents its behaviour (agentified manufacturing component). When these agents interact or cooperate they can generate aggre-

gated functionalities that are compositions of their individual capabilities. This is what happens when, for instance, several manufacturing components are working together in a manufacturing cell.

Definition 1 - Manufacturing component or module
A manufacturing component is a physical piece of equipment that can perform a set of specific functions or basic production actions on the shop floor such as moving, transforming, fixing or grabbing.

Definition 2 – Agentified manufacturing component
An agentified manufacturing component is composed of a manufacturing component and the agent that represents it. The agent's skills are those offered by the manufacturing component, which is connected to the agent through middleware.

Definition 3 – Coalition/Consortium
A coalition/consortium is an aggregated group of agentified manufacturing components, whose cooperation is regulated by a coalition contract, interacting in order to generate aggregated functionalities that, in some cases, are more complex than the simple addition of their individual capabilities.

A coalition is usually regarded in the multiagent community as an organisational structure that gathers groups of agents cooperating to satisfy a common goal. On the other hand, the term consortium is more usual in the business area where it is defined as an association of companies for some definite purpose. The definitions are quite similar because in both situations there is the notion of a group of entities cooperating towards a common goal. This common definition is adapted to the context of the architecture being proposed here. From now on the terms consortium and coalition are used with the same meaning. Nevertheless, to emphasise that the architecture being introduced here is composed of manufacturing components and not of companies the term coalition will be favoured.
The coalition is the basic organisational form of cooperation in the architecture being proposed. A coalition is able to execute complex operations that are composed of simpler operations offered by coalition members. A new coalition

can be established with either individual members or other existing coalitions (Figure 2).

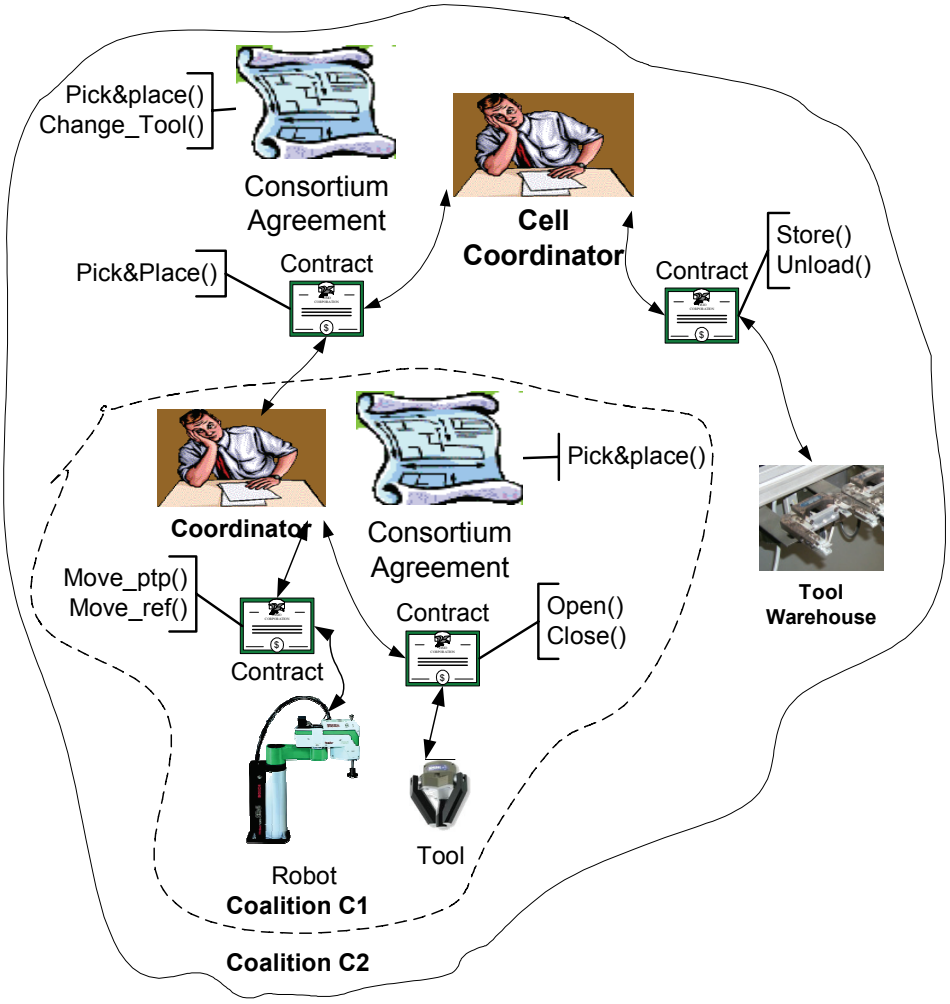


Figure 2. Consortia example

A robot cooperating with a gripper chosen from a tools' warehouse illustrates a simple example of a coalition. The better the way coalitions can be changed, the better the agility of the manufacturing systems they represent will be. If agility is seen as the capability to easily change the shop floor as a reaction to unforeseen changes in the environment, then an easy way to create and change coalitions is an important supporting feature for the manufacturing system's agility.

When forming a group of collaborative agents there are no limitations on the type of agents that can be involved in it but there is an important restriction

which limits their cooperation capability – their spatial relationship. Manufacturing agents that are not spatially related cannot cooperate, as it is in the case of, for instance, a robot and a tool. If the tool is not within the reachability space of the robot it will be impossible to create a cooperative relationship. Another example of constraint is the technological capability. In order to be usable by the robot, the tool has to be technologically compatible with the robot wrist. Therefore, when creating a coalition it is mandatory to know what the available and "willing" to participate agents are that should present some compatibility among them (for instance spatial or technological compatibility). The manufacturing agents that can establish coalitions should be grouped together because of these aspects of compatibility. This is analogous to the long-term collaborative alliances of enterprises. The objective of these clusters is to facilitate the creation of temporary consortia to respond to business opportunities. Similarly, in the case of the architecture being described there is a need for a structure (cluster) that groups the agentified manufacturing components willing/able to cooperate.

Definition 4 - Shop floor cluster
A shop floor cluster is a group of agentified manufacturing components which can participate in coalitions and share some relationships, like belonging to the same manufacturing structure and possessing some form of technological compatibility.

A community of agents belonging to the same physical structure – a manufacturing cell, thus forms a cluster, and when a business opportunity (i.e. a task to be executed by the shop-floor) arises, those agents with the required capabilities (skills and capacities) and compatibility are chosen to participate in a coalition. The limitation for an agentified manufacturing component to be accepted in a shop floor cluster is that it must be compatible with the others physically installed in the cell. For instance, an agentified robot installed far from a cell is not a good candidate to join the cluster that represents that cell, because it can never participate in any coalition. Since all the manufacturing components installed in a cell answer the requirements for compatibility a shop floor cluster is associated with a physical cell. Figure **3** shows how manufacturing agents, cluster, and coalition interrelate. Agentified components in the same "geographical" area of the shop-floor join the same cluster.

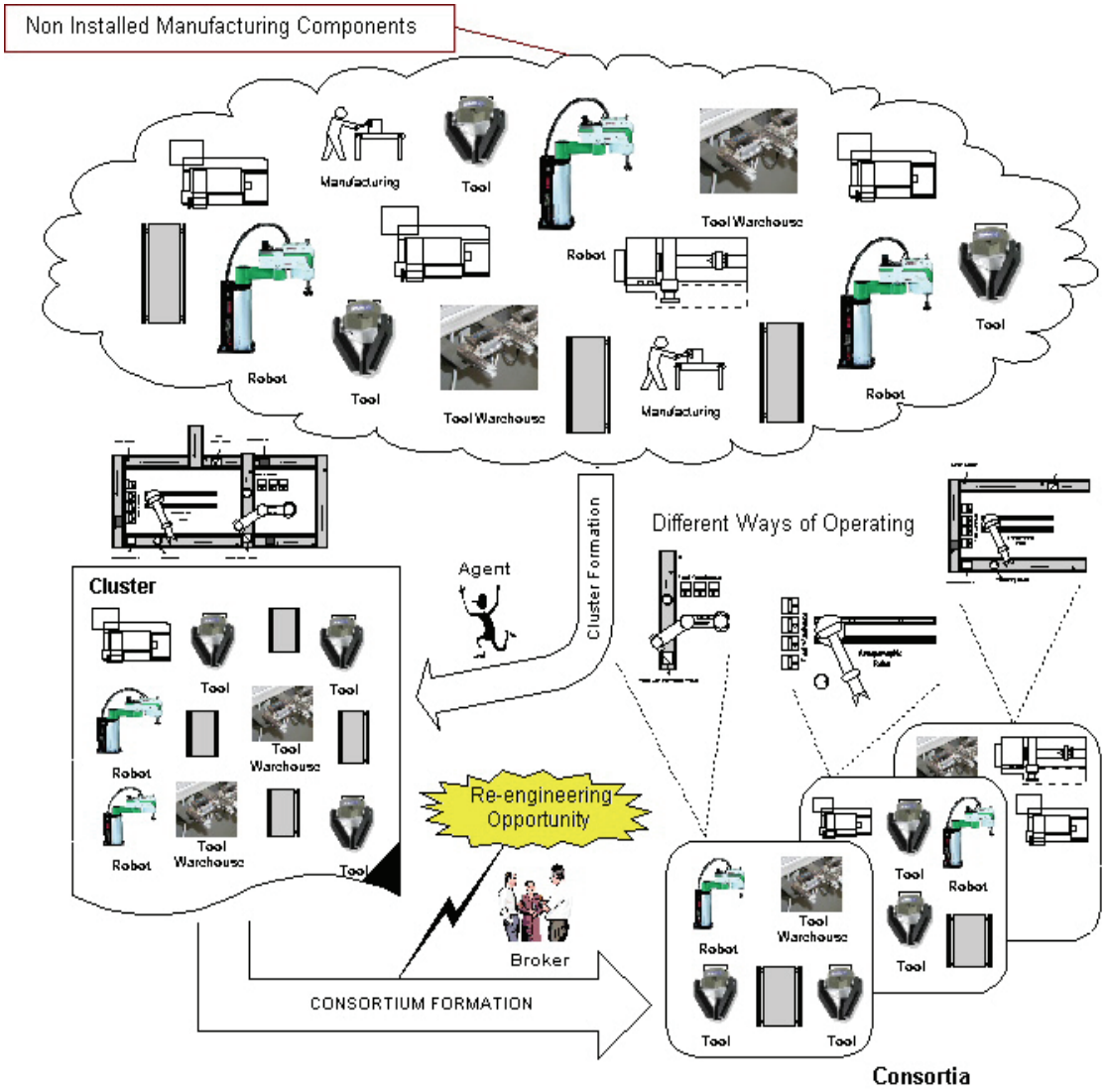


Figure 3. Consortia formation

The different coalitions that can be created out of a cluster represent the different ways of exploiting/operating a manufacturing system. Adding or removing a component from the physical manufacturing system also implies that the corresponding agent must be removed from the cluster, which can also have an impact on the established coalitions. A broker is used to help the formation of coalitions to reduce the complexity of the individual agents in terms of coalition formation. By delegating this responsibility to the broker, the individual

agents can be simpler because all they have to do is negotiate the terms of their participation with the broker rather than carrying out all complex details of coalition formation such as deciding which members are better indicated to answer the requirements of a coalition being formed.
The interactions between the cluster and its members are regulated by a contract. This contract establishes the terms under which the cooperation is established. It includes terms such as the ontologies that must be used by the candidate, the duration, the consideration (a law term that describes what the candidate should give in exchange for joining the cluster, usually the skills that the candidate is bringing to the cluster). The behaviour of a coalition is regulated by another contract that is "signed" by all its members. The important terms of this type of contract, other than the usual ones like duration, names of the members, penalties, etc., are the consideration and the individual skills that each member brings to the coalition. The importance of contracts as a mechanism to create/change flexible and agile control structures (consortia) lays in the fact that the generic behaviours presented by generic agents are constrained by the contracts that each agent has signed. This calls forth the idea that different coalition behaviours can be achieved by just changing the terms of the coalition contract, namely the skills brought to the coalition.
The expectation at this point is that coalitions of agentified manufacturing components, if regulated by contracts, that are declarative and configurable information structures, may lead to significantly more agile manufacturing systems. It is expected that the different ways of exploiting a system depend only on how coalitions are organised and managed. This approach solves the problem of how to create dynamic (agile) structures, but not the problem of how to integrate heterogeneous manufacturing components' local controllers. In order to overcome this difficulty, the process used to transform a manufacturing component into an agent (agentification) follows a methodology to allow their integration (Camarinha-Matos et al. 1997; Camarinha-Matos et al. 1996).

3. CoBASA architecture

The basis for the agility is provided by the way coalitions can be created, changed, and terminated. CoBASA is a contract based multi-agent architecture designed to support an agile shop floor evolution. It is a multiagent system because its components are agents, as defined in the Distributed Artificial Inteligence (DAI) / Multiagent community (Ferber 1999; Franklin and Graesser 1997;

Weiss 1999; Wooldridge and Jennings 1995; Wooldridge 2000; Wooldridge 2002). In addition, it is contract based because the behaviour of coalitions is determined by contractual arrangements. The coordination and cooperation of the coalitions and individual agents is inspired by the works of *social order* in multiagent systems (Conte and Dellarocas 2001). In the specific case of CoBASA its norms are the contracts that regulate the cooperation and behaviour of the involved agents.
Since a CoBASA system is a community of interacting agents some sort of knowledge sharing is needed to guarantee effective communication and coordination. The various concepts needed by CoBASA (contracts, skills, credits, among others) are supported by ontologies, which can be seen as global knowledge engraved in CoBASA agents.
Finally, CoBASA, can be considered a complex adaptive system that displays emergent behaviour (Johnson 2001) mainly because this is essentially a bottom up system, in which complex structures (coalitions) are composed out of simpler manufacturing components. This "movement" from lower level structures to higher-level complexity is called emergence.

3.1 The components

The basic components of the CoBASA architecture are:

- Manufacturing Resource Agents,
- Coordinating Agent, Broker Agent,
- Cluster Manager Agent,
- and Contract.

Definition 5 – Manufacturing Resource Agent (MRA)
The MRA is an agentified manufacturing component extended with agent like skills such as negotiation, contracting, and servicing, which makes it able to participate in coalitions.

An agent called Manufacturing Resource Agent (MRA) models manufacturing components. This agent represents the behaviour of a manufacturing component. In addition it has a social ability (interaction and cooperation with the other agents) to allow its participation in the agent community.
Several types of MRAs, one type for each manufacturing component type, can be conceived. Therefore it is expectable to find robot MRAs, gripper MRAs, tool warehouse MRAs, etc. From a control perspective, each MRA is individu-

alised by its basic skills, which represent the functionality offered by the represented manufacturing component.

Each MRA possesses the following basic abilities:

- Adhere to/ withdraw from a cluster
- Participate in coalitions
- Perform the manufacturing operations associated with its skills.

Each MRA that belongs to a given manufacturing cell can participate in the cluster that represents that cell. Therefore, every agent, independently of its skills, can join a cluster as long as it is compatible with the other cluster's elements. Nevertheless, this adhesion is not always guaranteed because the cluster, before accepting a candidate, evaluates its "values". The candidate's value is given by a concept called *credits*, which represents a kind of curriculum vitae. If the curriculum does not reach a certain level the agent is not accepted. Further details about the credit system are given in the clustering section. A negotiation is held between the MRA and the cluster whenever the agent wants to join the cluster. A MRA can join or leave different clusters when the manufacturing component it represents is installed or removed from different manufacturing cells.
All negotiations related to the creation, changing, and termination of coalitions are performed by the MRA. The agent does not automatically choose the skills the MRA brings in to a coalition, which are instead chosen by a user. The MRA participation in a coalition may terminate either because the coalition successfully reached its end or because of an abnormal condition. Performing the manufacturing operations associated with the represented skills is the kernel activity of the MRA. While the other two activities are more related to its social activity, this one represents real manufacturing work. Whenever a robot MRA, for instance, receives a request to execute a *move* command it reacts by sending the appropriate command to the real robot controller that in turn causes the movement of the physical robot.

Definition 6 – Coordinating Agent (CA)
A CA is a pure software agent (not directly connected to any manufacturing component) specialised in coordinating the activities of a coalition, i.e. that represents the coalition.

Although a coalition is not an agent, it is one of the main concepts that stand in the background of the architecture being presented. A basic coalition, besides being composed of MRAs, includes an agent that leads the coalition – Coordinating Agent (CA). In addition it can include as members other coalitions. The coordinator of a coalition is able to execute complex operations that are composed of simpler operations offered by coalition members.
The CA is, in many aspects, very similar to the MRA. Because it must also be able to join a cluster as well as participating in coalitions, its basic social activity is quite the same. However, there are two differences. First, a CA does not directly support manufacturing operations (skills) but is instead able to create complex skills based on some rules of composition of skills brought in by the members (e.g. MRAs) of the coalition it coordinates. Second, a CA does not offer manufacturing skills to a coalition except when leading a coalition participating in other coalitions.

The CA has two different statuses:

1) free to coordinate, and 2) coalition leader.

When free to coordinate it is just waiting to be a coalition leader. When the CA is eventually chosen to coordinate a coalition its status is changed as well as its situation in the cluster. A CA with a coalition leader status represents a coalition in the cluster.
As members of coalitions, MRAs can only play the member role whilst CAs can play both the coordinator and member roles. A simple manufacturing coalition is composed of some MRAs and one CA. However, a coalition can be composed of other coalitions, creating, in this way, a hierarchy of coalitions. Therefore, a CA can simultaneously coordinate MRAs and others CAs (Figure 4). In this figure CA2 is simultaneously a member of *coalition 1*, and the coordinator of *coalition 2*, composed of MRA B and MRA C. Please note that *coalition 1* is composed of MRA A and CA2. CA1 does not have direct access to the members of *coalition 2*.
A coalition needs a CA, instead of only MRAs to reduce the complexity of a MRA. If the coalition was only composed of MRAs, the complex task of coordinating a coalition would be added to the usual tasks such as controlling the manufacturing component, negotiating cluster adhesion and participating in coalitions, etc. Among other things, a coalition coordinator needs to generate new skills, and should be simultaneously member and coordinator. Please

note that skill generation is not the only problem since the way skills are composed and represented in order to be executed properly is not a trivial task. Separating the functionality related to coordination from the one related to executing commands simplifies the architecture of the individual agents. MRAs become less complex at the expense of introducing another agent type, the CA.

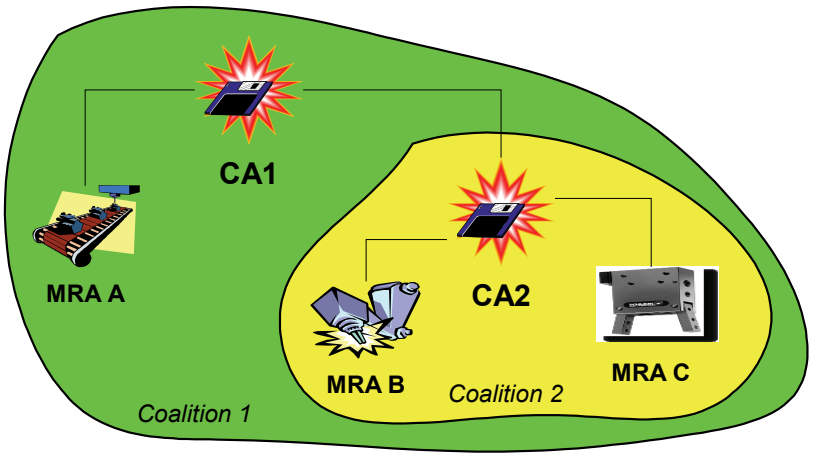


Figure 4. Hierarchy of coalitions/consortia

Definition 7 – Cluster Manager Agent (CMgA)
A cluster manager agent is an agent that supports the activities required by the cluster it represents. This agent stores information about all the MRAs that compose its cluster.

A cluster by itself is not an agent but rather an organisation of agents. However, an agent might model the activities that support cluster management, such as joining the cluster, leaving the cluster, changing skills, etc. An agent called Cluster Manager (CMgA) models the management activities of the cluster.

The CMgA must support the following basic activities:

- Attend requests for cluster adhesion
- Update cluster-related information
- Provide information to the broker.

Whenever the CMgA receives a request from a MRA or CA to join the cluster it starts the negotiation process that ends either with a refusal or acceptance. Based on the credits of the requester the CMgA decides if the requester is accepted or not. A registry of all agents that constitute the cluster is maintained by the CMgA and, whenever necessary, this information is updated by cluster members. The CMgA also provides all the information needed by the broker agent when creating coalitions.

Definition 8 – Broker Agent (BA)
A broker is an agent that is responsible for the creation of coalitions. It gathers information from the cluster and, based on user preferences, supervises/assists the process of creating the coalition.

An agent called broker agent (BA) supports the brokering activity, which is relevant in order to create coalitions. The notion of brokers, also known as middle agents, match makers, facilitators, and mediators is a subject of intense research in the multiagents field (Giampapa et al. 2000; Klusch and Sycara 2001; Payne et al. 2002; Sycara et al. 1997; Wiederhold 1992; Wong and Sycara 2000).
The broker therefore interacts with the human, the cluster, and the candidate members to the consortium. Coalitions/consortia can be created either automatically or manually. At the current stage only the manual option is considered. The main interactions between the concepts that have been referred to are shown in Figure 5. Contracts are the next important CoBASA mechanism, which is used to regulate the MRAs and CAs interaction with a CMgA as well as the behaviour within the coalition.

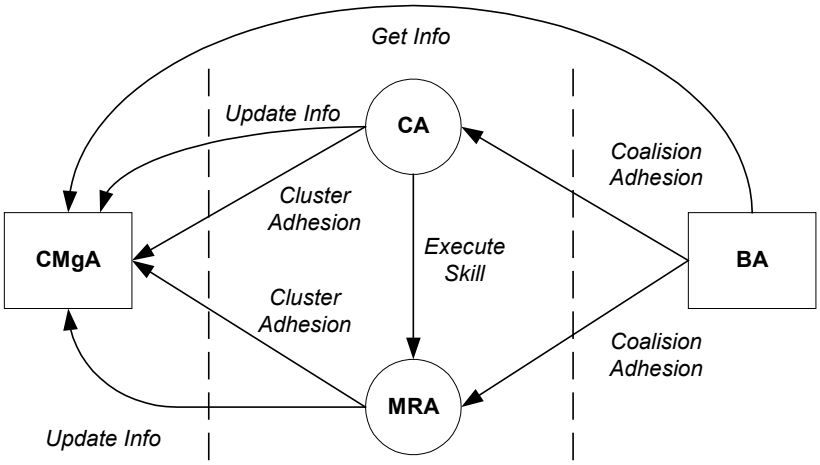


Figure 5. Interactions among the main components

In the CoBASA architecture two type of contracts are considered: **cluster adhesion contract (CAC),** and **multilateral consortium contract (MCC).**

Definition 9 – Cluster Adhesion Contract (CAC)
This contract regulates the behaviour of the MRA when interacting with a cluster. Since the terms imposed by the cluster cannot be negotiable by the MRA the contract type is "adhesion". The CMgA offers cluster services in exchange for services (abilities or skills) from the MRA.

The CAC includes terms such as the ontologies that must be used by the candidate, the duration of the membership, the consideration (a law term that describes what the candidate should give in turn of joining the cluster, usually the skills that the candidate is bringing to the cluster).

Definition 10 – Multilateral Coalition/consortium Contract (MCC)
This contract regulates the behaviour of the coalition by imposing rights and duties to the coalition members. The contract identifies all members and must be signed by them to be effective. The coalition leader (CA) is identified as well as its members. The members are entitled to a kind of award (credit) in exchange for their skills.

The important terms of this type of contract other the usual ones like duration, names of the members, penalties, etc., are the consideration and the individual skills that each member brings to the contract. Note that the skills involved in a specific consortium contract may be a subset of the skills offered by the involved agent when it joins the cluster. The importance of contracts as a mechanism to create/change flexible and agile control structures (consortia) lays on the fact that the generic behaviours exhibited by generic agents are constrained by the contract that each agent has signed. This calls forth that different consortium behaviours can be achieved by just changing the terms of the consortium contract, namely the skills brought to the consortium.
MCCs represent simultaneously a coordination mechanism and a mean to facilitate coalitions/consortia dynamics. Since a coalition/consortium is created, changed, and terminated mainly through contract operations, the task of grouping manufacturing components able to perform certain tasks (coalition) is facilitated. In addition, the introduction of new components to this group involves only contract configurations. Agility is thus achieved since moving components from one organisational form to another involves only configuration instead of programming effort.

3.2 Coalition dynamics

Since CAs are able to generate new skills from the set of skills brought in by its members, coalitions enable the creation of completely different control structures. This could not ever be achieved using a traditional control architecture because of its rigidity. Traditional approaches need to know in advance the logical organisation of the components as well as the complete set of skills that need to be controlled.
Considering this agility at the coalition level and considering also that coalitions can be composed of other coalitions, the next question is what impact a change on a coalition has on the whole structure. This impact might happen because after a change on a coalition (addition or removal of members) the skills its CA is able to perform are likely to change. They can be either increased, reduced, or in some situations they are kept. The last situation occurs when a component that brings no value to the coalition is introduced or removed. If a coalition participating in another coalition looses skills, then it is necessary to verify if any of the missed skills were offered to any other higher-level coalition. If this happens a renegotiation process must be started with the higher-level one, which should then verify the impact and if necessary renego-

tiate with its own higher-level coalition(s). This process is expanded through the whole levels until reaching the upper one. As a conclusion it can be claimed that the removal (or addition) of a manufacturing component (MRA) (its skills) provokes the automatic updating of the higher-level skills that could be directly or indirectly dependent on the ones that were removed (added).
It is important to retain that the skills offered to the coalitions at a higher-level can be a subset of the skills possessed by the CA member agent.
The skills brought to a coalition j led by CAi are the union of the skills brought by all MRAs that belong to the coalition j plus all the skills offered by the various coalitions that might be participating in coalition j. This means that a complex skill can be dependent on another complex one. To understand the next steps of CoBASA operation the following definitions are necessary:

S_{CA_i}	The set of skills of CAi in coalition/consortium i
$S_{MRA_{i,j}}$	The set of skills of MRA i in coalition/consortium j
$S_{CAmembers_i}$	The set of skills brought to the coalition/consortium i, led by CAi, by its members
$S_{CAgenerated_i}$	The set of skills generated by CA i in coalition/consortium i
$S_{CAoffered_{i,j}}$	The set of skills the coalition/consortium i, led by CA i, offers to the coalition/consortium j

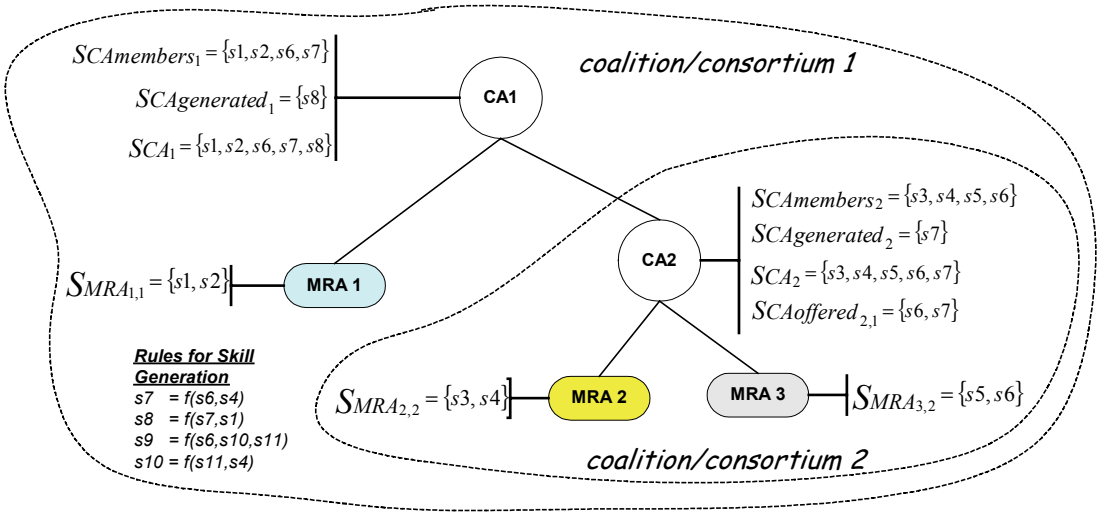


Figure 6. Coalition in its initial situation

Figure 7 shows that the skills offered by the coalition 2 are a subset of the skills the coalition possesses, which is perfectly valid. The skills to be offered are chosen during the coalition creation by the broker. The generation of skills is based on a set of rules that belong to the CoBASA knowledge base. For instance in coalition/consortium 1, according to the rules illustrated in Figure 3 only the rule "s8 = f(s7,s1)" can be fired and thus s8 is the only generated high level skill. All the other rules require input skills that are not present.

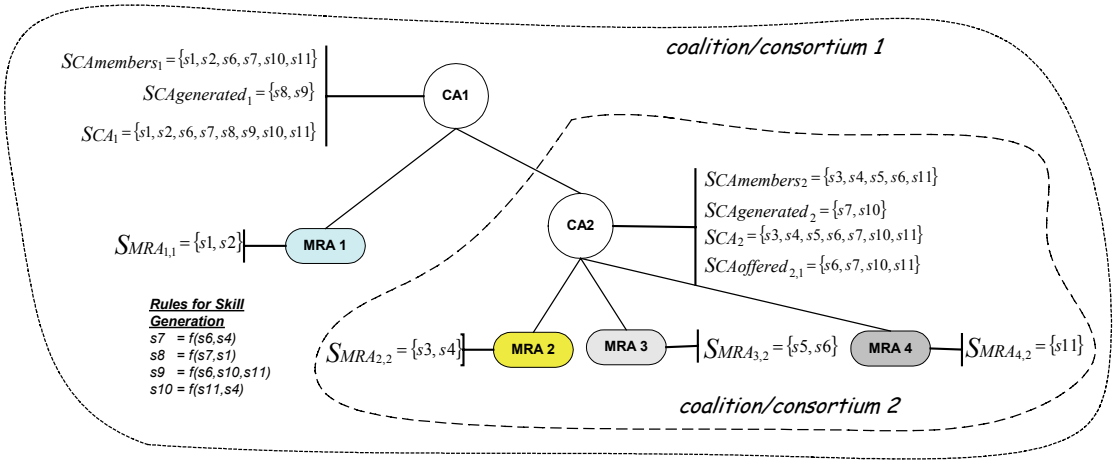


Figure 7. Hierarchy of coalitions after introducing a new element MRA 4

The effect of coalitions dynamics in CoBASA, can be verified by analysing what happens when a new component is added, for instance to coalition 2 (Figure 7). The introduction of MRA 4, which brings in new skill s11 causes an alteration on the set of skills CA2 can handle. It can be seen that the set of skills for the coalition 1 were increased. The update is almost automatic because it has only to do with the generation of complex skills and renegotiation between coalition leaders.

Considering now the removal of a component (MRA 3, for instance), it causes a reduction of skills both in coalition 1 and coalition 2 (

Figure 8).

From this discussion it is now possible to better understand why the CoBASA architecture can be considered a complex adaptive system. In effect coalitions are just an expression of the interaction that occur among coalition/consortium members. The skills owned by the coalition/consortium leader represent the

behaviour that results from its members' interactions. It can be identified a "movement" of low level skills to higher level ones, which allow us to claim that this architecture displays a kind of emergent behaviour (Johnson 2001).
A coalition member must execute all the operations promised by it in the consortium contract, when requested by the coalition coordinator. On the other hand, the coordinator (CA) can create complex operations (services) by aggregation of the individual operations of the members.
Let us now have a first look at the contracts that regulate the behaviour of coalitions and their members.

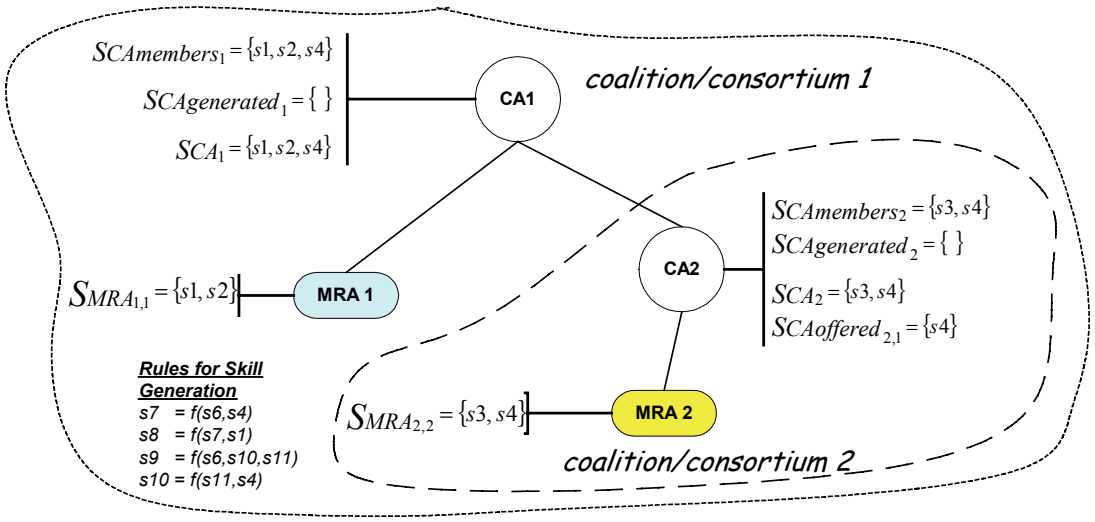


Figure 8. Hierarchy of coalitions after removing MRA 3

Figure 9 shows a hierarchy of two coalitions/consortia in which CA2 is simultaneously the coordinator of *coalition 2* and a member of *coalition 1* led by CA1. As it could be expected there are two multilateral consortium contracts, one for each consortium/coalition. However, each member of a consortium/coalition must have a copy of the contract that regulates the coalition's operation, since the members' behaviour is regulated by that contract. This means that in the case of figure 6 CA2 behaviour is conditioned, in fact, by two contracts instead of one: 1) the contract of *coalition 1*, where CA2 is a member, and 2) the contract of *coalition 2*, where CA2 is the coordinator. To distinguish between these two types of roles, the MCC contracts each CA might be bound to are divided into **membership contracts** and **coordination**

contracts. All contracts in which the agent plays the member role are membership contracts while those in which it plays the coordinator role are coordination ones. Despite this division, the structure of the contracts is the same, since both types are multilateral consortium contract - MCC.
Skills descriptions help the creation of manufacturing coalitions. However this is not their only role, since they are also very important when the coalition is being operated (operational phase). This is so because skills represent also the commands to be used among coalitions/MRAs (services). The important question here is how the CA reacts when it receives a request to perform a certain task according to the skills it offered.

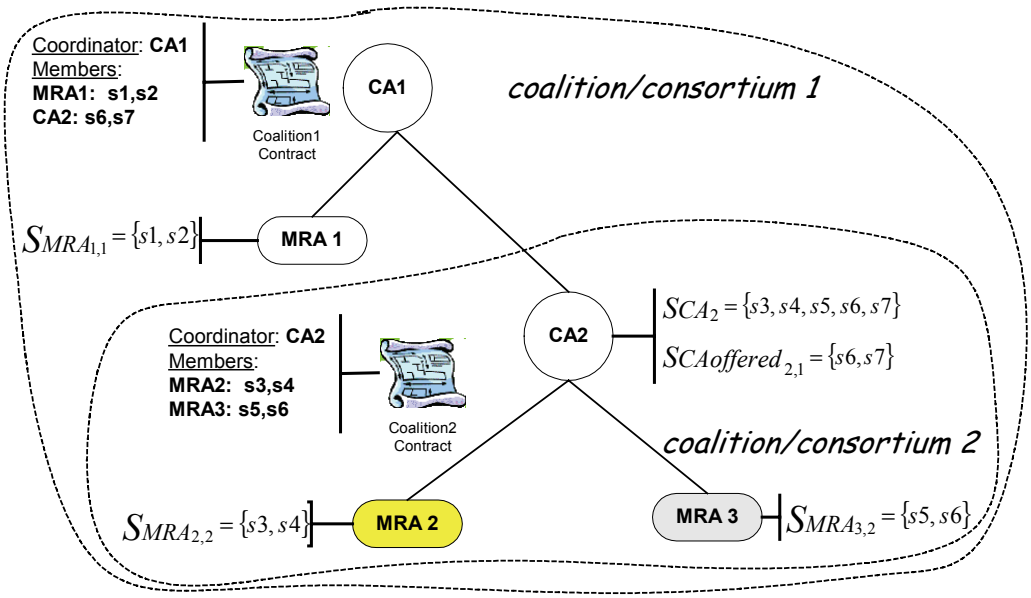


Figure 9. Coalitions contracts

When the CA is requested to perform some task associated to one of its skills, it behaves differently according to the skill type. If the skill was not generated by this CA (simple skill) the action consists simply in redirecting the request to the member of the coalition that has brought it. On the other hand, if the skill is generated by this CA then the procedure is more complex. This is so because the skill is now a composition of the skills brought to the coalition by its members, and this composition can be complex. This means that a model is needed to describe this composition and it should allow the modelling of complex command structures, which are needed to represent those skills that have complex structures. The CA must then execute the model by sending lower

level commands (skills) according to the model structure of the complex skill being executed. This is to conclude that a model is required to represent the structure of the composed skill and then an execution machine is needed as part of the CA to execute the model properly.
If each CA embeds a generic execution machine, like a Petri Net (Zurawski and Zhou 1994) executor, or even a workflow engine (WFMC 2002), able to execute Petri Nets or Workflow models than the CA is transformed into a kind of generic machine that can work with different types of skills.

3.3 Contracts

According to the law of contracts (Almeida 2000; McKendrick 2000), a contract is made up of a promise of one entity to do a certain thing in exchange for a promise from another entity to do another thing. Some law researchers (Almeida 2000) claim that the contractual statements (promises) are performing acts in the sense that they have effects. This means that the existence of a contract between two or more entities imposes constrains on their behaviour and can produce outcomes that were not possible without a contract, mainly due to the performing nature of the statements or promises.
There are several types of contracts, but in this work only two are considered as introduced in previous section: generic **multilateral contracts** and **adhesion contracts**. The main difference between them is the process of formation, which in the case of the adhesion contracts is via standardised forms. The contract offered by the cluster manager agent to the candidate member agents is a typical contract of adhesion, in the sense that the cluster imposes its terms. The only thing an agent can do is accepting or refusing it. Part of the terms of this adhesion contract, namely the "consideration" of the candidate agent, is left open to be filled in by the candidate, when accepting the offer. In terms of the human law systems **consideration** was defined by an 1875 English decision as "some right, interest, profit or benefit accruing to the one party, or some forbearance, detriment, loss or responsibility given, suffered or undertaken by the other". In most of the law systems in order to **create a contract** at least two sequential statements are required: an offer followed by an acceptance. An offer can be followed by a counter-offer, which in turn can also be followed by another counter-offer and so on. The process terminates when one of the partners sends an acceptance. The offer and the acceptance might not be the first and second action but they will be surely the last but one, and the last. Offers may set certain conditions on acceptance and to these, the acceptor is bound. The

acceptance validates and gives life to the contract. The contract starts at the moment the acceptance reaches the offeror.
The cluster manager, and the candidate agents when negotiating the cluster contract will use the offeror-acceptance protocol of real life contracts with some adaptations.
An offer, once made, can be revoked before acceptance. An offer can also expire if a deadline for acceptance passes. If there is no specified deadline, then the offer expires in a "reasonable time", depending on the subject matter of the contract (Almeida 2000). In the approach being followed an offer is made without specifying a deadline. This indicates that it must be answered in a "reasonable time", which is the normal time-out imposed to the global architecture for communication among the agents. An offer that was rejected cannot be subsequently accepted.
An alternative to reach an agreement other than the offer-acceptance protocol is using joint contractual terms, which express the agreements of the parts in only one text. This modality is specially used for creating contracts that involve more than two partners (multi-lateral contracts). In this case the parts reach agreement on the final terms of the contract using different kind of communicative acts in a preliminary phase. Afterwards, the final contract is put on a written form (final agreement) and finally all the partners must subscribe the contract. The contract turns effective when the last partner subscribes the document.
The formation of the coalition contract used in the proposed architecture uses this modality with some adaptations. The human user interacting with the broker will prepare the agreement on the terms of the contract (preliminary phase). It is this user that chooses the skills that each agent will bring to the contract (this user is just configuring the system). The broker agent then sends the final *text* to all partners to be subscribed. When the last agent finally subscribes it, the contract is considered as valid.

3.3.1 Cluster Adhesion Contract - CAC

The cluster adhesion contract is defined externally to the cluster and modelled using a knowledge representation system – Protégé 2000 (Protégé-2000 2000). The cluster manager agent can interact with this system to have access to the contract representation. Whenever it needs to offer an adhesion contract to an agent it just uses the form, waiting afterwards for its acceptance or refusal.

The formation of the contract starts when the cluster manager sends a message to the candidate agent containing an instance of an adhesion contract. The "accept" message from the candidate contains the complete adhesion contract, now filled in with the terms of the candidate (its skills), and when received by the cluster manager the contract turns to be valid. The cluster manager only agrees to negotiate with the candidate agent if it is not on the black list of the cluster. The cluster manager agent then checks for the credits of the candidate, which represents a kind of curriculum vitae. A credit is, for instance, the number of hours working properly, or a number that qualifies the global performance of the agent when working on consortia. Those agents with lower level qualification can sometimes not be accepted as members of the cluster. This is to guarantee that consortia created out of a cluster have a certain level of qualification (Barata and Camarinha-Matos 2002). When the candidate (MRA/CA) does not have sufficient credits, the cluster manager replies with a FAILURE command message (left part of Figure 13). If the credits are accepted, the cluster manager fills in all the cluster adhesion contract (CAC) terms except the skills that will be brought in by the candidate, which should be filled in by the candidate. Then the cluster manager sends a REQUEST message to the candidate asking it to accept the contract. This corresponds to an offer in contract law terms. The MRA/CA evaluates the contract offer and decides if it can accomplish all its terms. If not, the candidate sends a FAILURE message to the CMgA stating that it does not accept the offer. Then a FAILURE message is sent to the candidate stating that the cluster manager did not accept its REQUEST to join the cluster. If, on the other hand the MRA/CA, after evaluating the offer decides for its acceptance, sends an INFORM message stating its acceptance. The cluster manager sends then a final INFORM message to the candidate stating that its initial REQUEST has been accepted (right part of Figure 13).

The commands exchanged between the candidate and the cluster manager follows the FIPA protocols (FIPA 2002).

There is a tight connection between the CAC and credits (agent's curriculum). If credits are regarded as a kind of performance measure it is quite natural that at the end of a contract credits must be updated corresponding to a sort of curriculum updating. This happens independently of the termination type, either normal or abnormal. A contract terminated by performance might be regarded as a successful one because it means the contractee agent (MRA/CA) has accomplished all its promises. Therefore it is natural that this agent could add some good points to its curriculum. On the other hand, if an abnormal termi-

nation is considered, it is normal that a kind of curriculum penalisation takes place. This rewarding/penalisation step at the end of every contract guarantees that the agent's curriculum is a mirror of its performance. When the members of the cluster adhere to a cluster by accepting the CAC they "know" exactly what are the penalisations or rewards they get when the contract is terminated.

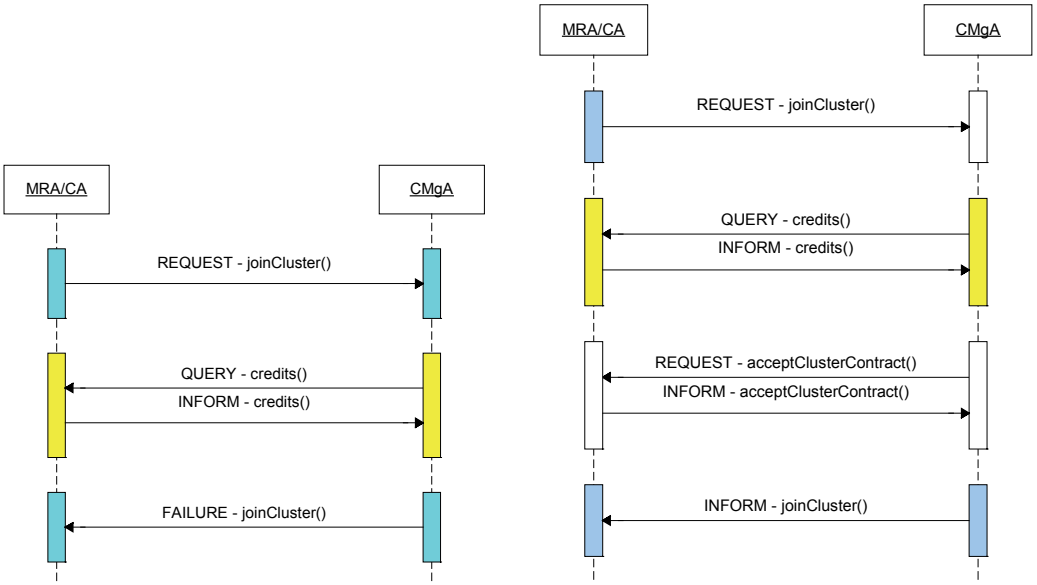


Figure 10. Unsuccessful and successful cluster joining

3.3.2 Coalition Contract - MCC

The broker agent, with the help of a human expert, creates the coalition contract (MCC). The model of this type of contract has many similarities with the previous one but has also some slight differences because it is a multilateral contract instead of a bilateral contract. To support various members and one contractor the contract has one common part dedicated to the contractor (the agent playing the co-ordination role), and another part dedicated to each of the other members. The *members* part of the contract is composed of several *individualConsortia* elements that in turn describe the individual contractual terms of each member of the coalition. The **promise** (declaration or manifestation of an intention in a contract) brought to the contract by each member is a set of manufacturing skills.

The broker creates the contract when a coalition is created. The user configures the different parts of the contract based on the requirements needed by the coalition. For each member the individual part is fulfilled namely by choosing which skills the members bring to the coalition.
The performance of the MCC includes the execution of the contract promises (skills). This is done while the contract is still valid and the coalition is operating. Only promised skills can be asked.
At the end of the contract the CA awards each coalition member with a number that represents the quality of the handed out service. This award or penalisation, if added to the agent credits, can be used to improve (or even reduce) its qualification, and is important for the future participation of the agent on consortia. This mechanism is similar to the one mentioned when CACs have been discussed. Similarly there are three different ways of terminating a MCC: by performance, by frustration, and by breach.
The "good" way of terminating a contract is by performance. In this situation the CA (coordinator) verifies if the participation of any member is within the valid date. If not, the CA asks that member to terminate its participation. Based on the value stored in the individual exception part of the MCC, the award for the participation in the coalition is collected.
Terminating the MCC by a frustration reason is an abnormal way, and consequently the breaking agent may incur in some penalisations. The request to break the contract by frustration is always initialised by the coalition member that detected the frustration. When this happens the member collects the penalisation stored in the contract. Three reasons can lead a coalition member to request to terminate a contract for frustration reasons:

1. The user requests the agent (MRA/CA) to leave (physical move, for instance)
2. A CA participating in another coalition detects their members are not responding
3. A CA/MRA of a lower level could not renegotiate a contract change with its higher level CA.

Terminating by breach is the worst case of termination of a contract from the penalisations point of view. The request to breach the MCC can be started either by the coordinator or by one of the members. A breach of the contract started by the coordinator implies that one of the members misbehaved.

On the other hand a breach started by one of the members means coordinator misbehaviour. A member starting a breach does not incur in penalisations. However when it is "guilty", i.e., the coordinator detected some misbehaviour, it gets penalised. A member shows bad behaviour whenever it does not answers a request from its coordinator to execute one of the promised skills. Likewise if the member, in spite of replying to the request, is not able to perform it properly, i.e., the excuse for the failure is not included in the MCC. A coordinator, on the other hand, shows bad behaviour whenever it does not answer a request from the member, which can be, for instance, a call to renegotiate the contract terms.

4. CoBASA main interactions

The most important functionalities related to CoBASA coalitions are:

1. Creating new coalitions
2. Changing coalitions
3. Coalition dissolution
4. Service execution

4.1 Creating new coalitions

The main actor in creating coalitions is the broker agent (BA). A human user chooses the coalitions based on the logical structure he/she wants to create. The other important actor is the cluster manager agent (CMgA) that provides information about available members. In addition to these two agents others are needed to create a coalition:

1. A CA not currently engaged in any consortium (available to lead a coalition).
2. MRAs, if the coalition will include manufacturing components.
3. CAs leading coalitions that might be included as members of the coalition being created.

Fifure 11 shows the interactions that happen between the different actors involved in creating a coalition.

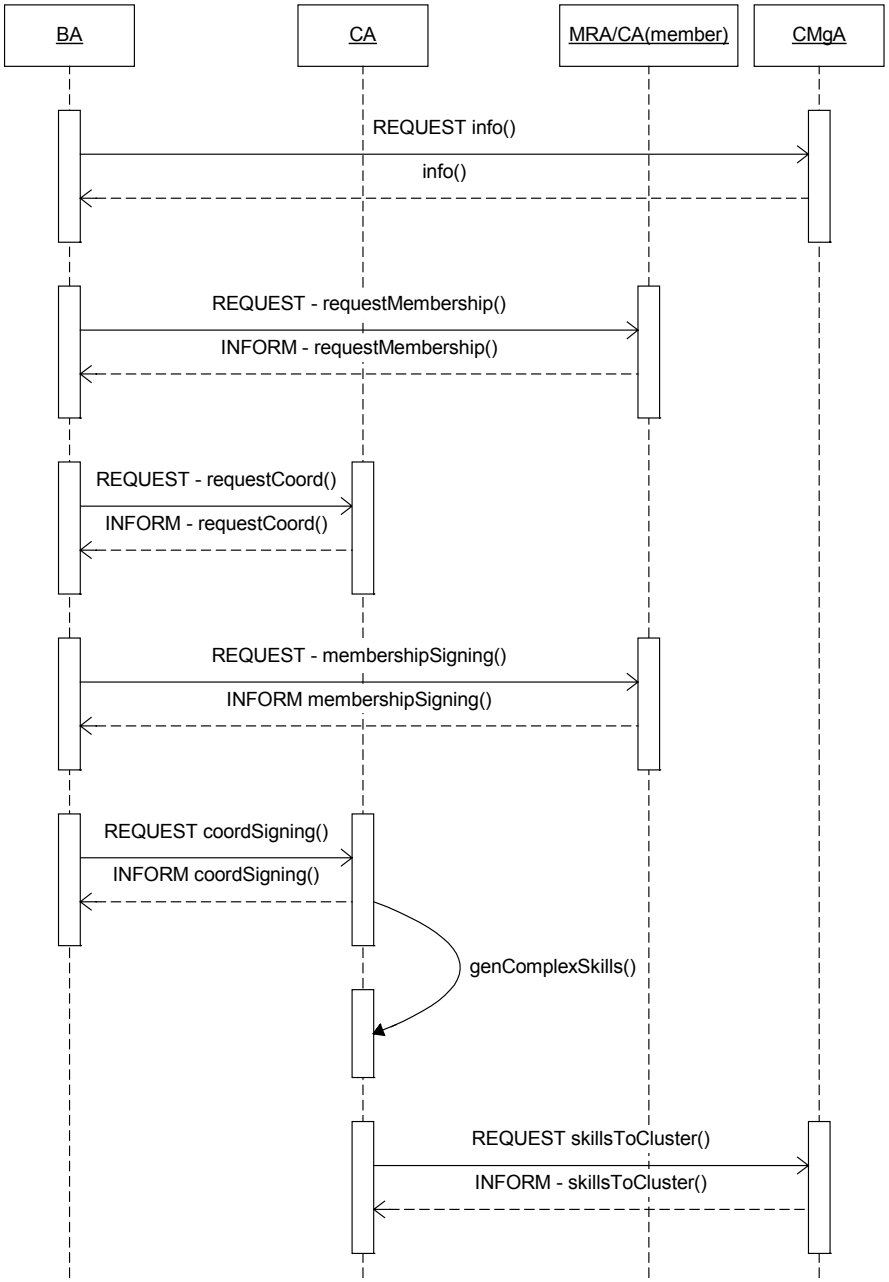


Figure 11. Interactions when creating a coalition

The figure shows the BA agent, the CA agent that has been chosen to be the coordinator, an agent to represent the members of the coalition (*MRA/CA(member)*), and the cluster manager agent (CMgA). Independently of

the type of MRAs or CAs that form the coalition, the behaviour is the one indicated in the figure. All information exchanged between the various actors is shown using the FIPA REQUEST protocol (FIPA 2001).
The broker asks for information about candidate members in the cluster by sending a REQUEST command. After getting the information from the cluster manager (CMgA), the broker shows the available members to the user as well as their individual information and lets him/her compose the coalition and create the contract that regulates it. The broker then asks each member to verify if they accept the contract, what is done by sending a REQUEST *to be member* command. This step is done in order to make sure each individual agent evaluates the contract before accepting it. This corresponds to asking the agent if it is interested in participating in the coalition under those conditions.
After all candidate members, including the coordinator, have expressed their interest in participating in the coalition, the broker starts the process of signing the contract by sending a REQUEST *to sign* command. Signing does not involve a complex formalism because the objective is to indicate to coalition members that the contract is now effective. After the broker requests that the coordinator signs the contract, the coalition is now operating from its point of view. After signing the contract the CA must try to generate its complex skills (*genComplexSkills*) as it has just received a new set of skills from its members. This step is crucial for the agility of the system, because the coalition is now generating automatically its skills based on the skills brought in by the members components are organised, i.e. changing the system's logical control structure, making this phase directly connected to the reengineering phase of the production system. This phase is divided into two different parts: the first one discusses the addition of one member to an existing coalition, and the other discusses the removal of one element. Although the description is made for one element to simplify the diagrams, the addition/removal of several elements is straightforward.
The interactions involved when a new member is added to an existing coalition are shown in Figure 15. As in the previous case, the broker and the cluster manager are important players because it is through the broker that the coalition is altered while the CMgA provides the necessary information. Furthermore, the coalition coordinator (CA) and its members (consMemb), the member to be added (newMember), and the coordinators of the coalitions (CA+1, CA+2), where hypothetically the coalition being changed is participating in, are the other actors.

The process starts with the BA asking the CMgA to provide information about its members that compose it. When the skills are generated the new coalition leader can then ask the CMgA to update its skills and to change its status from free to coordinate to coalition leader. The coalition is now registered in the cluster manager through its leader.

4.2 Changing coalitions

Changing a coalition corresponds to changing the way the manufacturing (Figure 15). Hence, the user, via the broker, selects the coalition to be changed which provokes the BA to ask the coordinator of that coalition to send it its MCC (REQUEST getContract).
This contract is needed because the user needs to configure its individual part with data from the new member as well as possibly changing other parts. After changing the contract, the new member is asked to accept the contract and to sign it. These operations are similar to the ones introduced in the creation phase. The broker now needs to renegotiate the new terms of the contract with the other coalition members to let these members discuss it (REQUEST membershipReneg).
Under normal circumstances these agents accept the changed contract. What happens if one or more members refuses to participate is not shown to keep the figure simpler. In any case, when in this situation, the user through the broker or through the member's GUI has the authority to overcome this situation. The broker then proceeds to the renegotiation phase with the coalition leader (CA). The goal of this phase is to get the new contract version accepted by the CA. This is why this process is called a renegotiation (REQUEST coordReneg). When the broker receives the INFORM stating that the contract was accepted the process is finished from the broker point of view. However, the CA has some other tasks to do before the whole process is concluded. First, it needs to check if the addition of the new element has generated new skills, which is done by activating genComplexSkills.
Next, the CA checks if it is currently engaged in any other coalition as well as if it has got new skills. If yes in both cases, it renegotiates with the leader (CA+1) of that coalition to change the skills it is bringing in (REQUEST *coordReneg*). Finally, after the successful renegotiation, the CA updates the skills of the coalition in the cluster manager (REQUEST *updateSkills*).

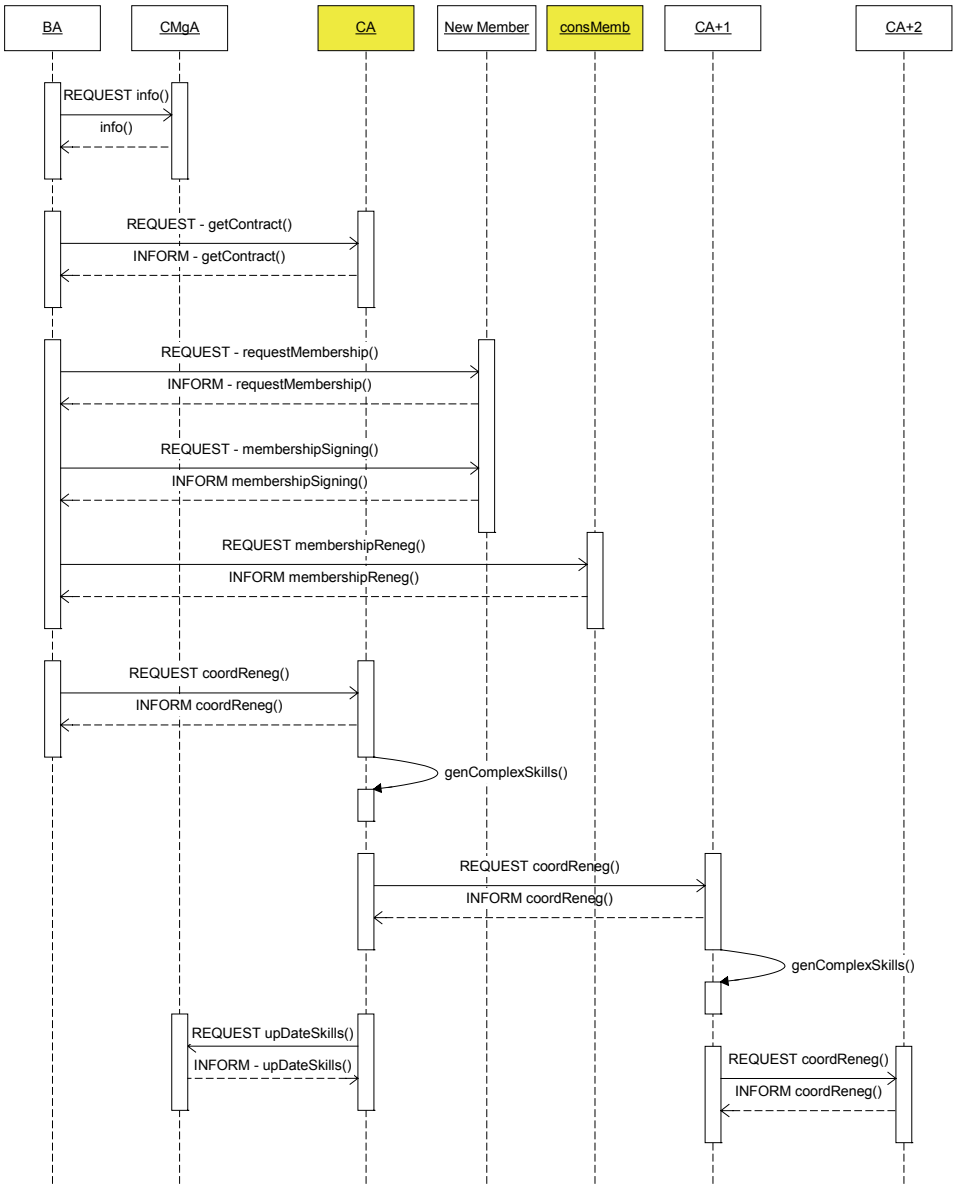


Figure 12. Adding an element to an existing coalition

Figure 12 also shows that if the renegotiation between the CA and CA+1 has impact on CA+1's skills, and if CA+1 is also participating in another coalition led by CA+2, then it will request CA+2 to renegotiate the terms of its participation in that coalition contract. The process is repeated until it reaches the highest-level coordinator in the hierarchy of coalitions. This is a very important

mechanism because whenever a coalition is changed, the impact of this change is automatically propagated to all the coalitions that are directly and indirectly related to it (transitivity).
The removal of one element is not shown because it follows a similar negotiation pattern.

4.3 Coalition dissolution

A coalition can be dissolved either when the system is being dismantled or when it is being reengineered. In the first case, all coalitions need to be terminated and then all cluster contracts must also be terminated. In the second case, the system is suffering such a radical change that it is not worth keeping any of the existing coalitions. Therefore all coalitions are dissolved in order to create completely new ones. Dissolving a coalition is different from changing it (removal of elements) in the way that the coalition coordinator also terminates its activity and changes its status in the cluster from coalition leader to free to coordinate.
Figure 17 illustrates the whole process for a coalition composed of one coordinator and one member.
Since this is a convenient way of terminating, the BA discharges the MCC by performance. It first discharges the CA and then all coalition members (REQUEST *dischargeByPerf*).
After accepting the discharge, the CA updates its credits in the cluster, which have just been increased by the reward it has received, as well as its status, since the CA is now free to coordinate.
Note that now the CA does not generate complex skills because it does not have any member to give it any skill. After discharging the MCC, coalition members collect their rewards and add them to their credits, and then update their credits in the CMgA (REQUEST *upDateCredits*).

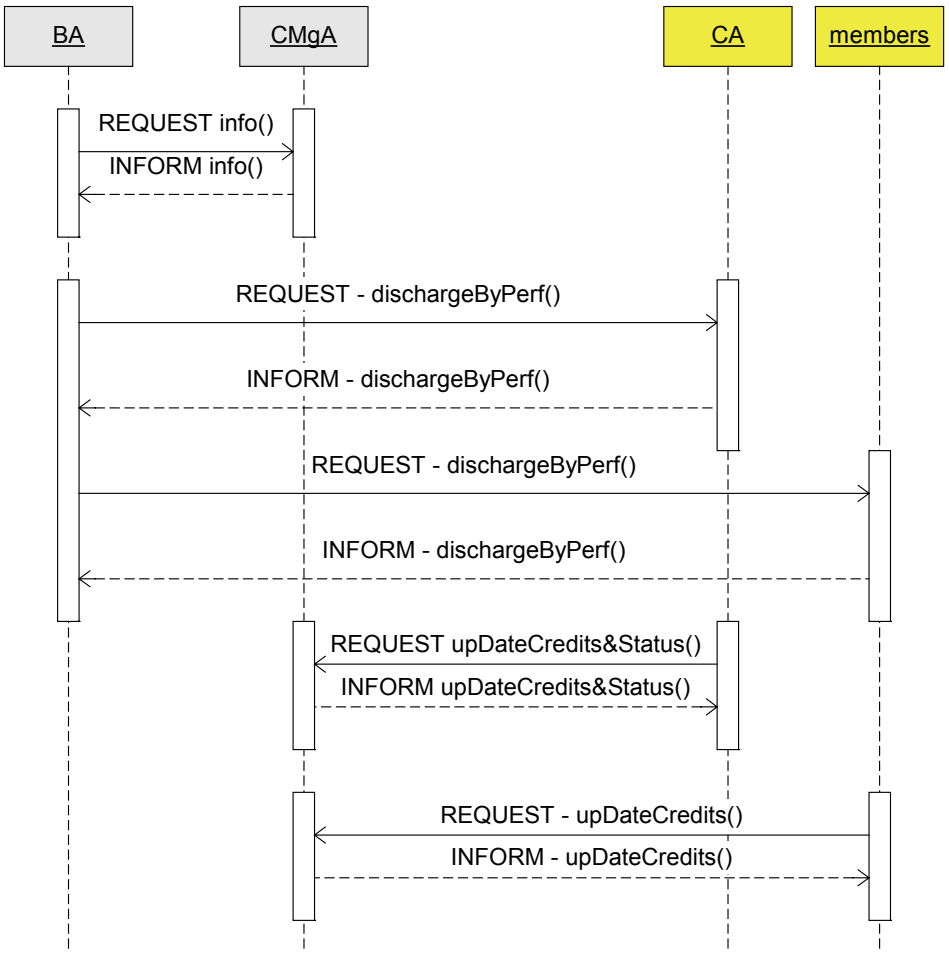


Figure 13. Coalition dissolution

4.4 Service execution

This phase corresponds to the production phase of the production system life cycle, since operating a coalition is asking its members to execute skills (or commands) they have promised in the MCC that regulates that coalition. In addition, asking to perform a skill involves, ultimately, executing some commands in the manufacturing physical component connected to one of the MRAs that belongs to the hierarchy of coalitions. It must be recalled that MRAs are always the lower level participants of any hierarchy of coalitions.

Figure 14shows the execution of skills in the hierarchy of coalitions shown on the left part of the figure. It is considered that CA1 requests the complex skill *s7*, which is offered to *coalition 1* by *coalition 2*. Furthermore, *s7* is composed of *s4* and *s6*, offered to *coalition 2* by *MRA 2* and *MRA 3* respectively. When CA1 needs to execute its skill *s7*, due to, for instance, a higher-level request, the agent finds out in the coalition's MCC that CA2 offered that skill. Then CA1 sends a REQUEST *service* command asking CA2 to execute skill *s7*, since it is offered by *coalition 2*. When CA2 receives the request it validates its origin by looking in the various contracts stored in **membership contracts** to whose coalition or coalitions this skill had been offered to. Next, the leaders in the set of membership contracts in which the skill is offered are checked to validate the request. After this validation, the CA1 decomposes the requested skill into its basic components (*s4* and *s6*), and, then, after verifying which agents offered them, starts sending the requests according to the complex skill structure. When MRA 2 finishes the execution of *s4* it replies to CA2 with an INFORM *service* command or a FAILURE *service* command (not shown in the figure), depending on, respectively, if the request was successfully accomplished, or not. After receiving the INFORM message for the first request (*s4*) CA2 sends the REQUEST service command to MRA 3 asking for *s6* in a way similar to *s4*. After CA2 receives the s6 INFORM message from MRA 3, it sends an INFORM service command to CA1 informing that its request for *s7* has been successfully achieved.

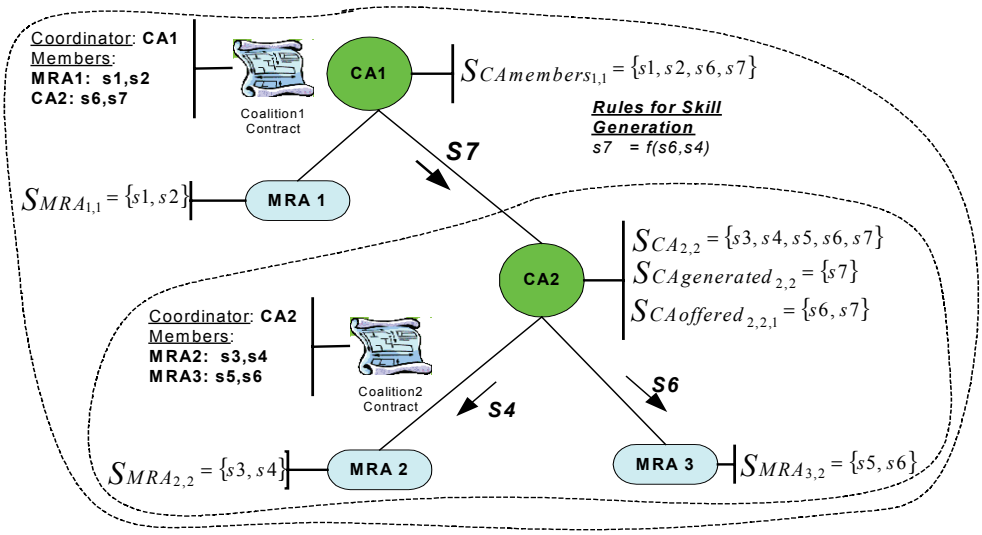


Figure 14. Skills requests in a hierarchy of coalitions

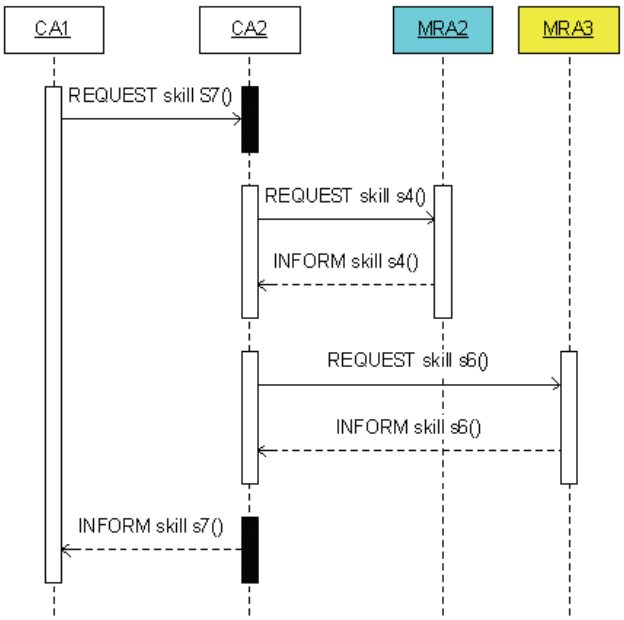


Figure 14. (Continued.) Skills requests in a hierarchy of coalitions

Although abnormal execution situations are not shown, it is important to know when they happen:

1. An agent does not answer a valid request addressed to it from its coalition leader.
2. An agent refuses to execute a valid request from its coalition leader.
3. A request command was not successfully accomplished.

In the first and second situations the agent is immediately expelled from the coalition. The coordinator does this by asking the faulty agent to breach its coalition contract. Although this extreme situation rarely happens, it is considered to be showing agents that the act of refusing something promised on a contract has serious consequences. Eventually the faulty agent asks for user attention after such a situation happens. The third abnormal situation is when the agent who was asked to execute an offered skill replies with a FAILURE message, which denotes that for some reason the agent could not successfully execute the command. The reason is indicated in the message content. Whenever the coalition leader (CA) receives such a message, it first verifies the reason and then decides accordingly. If the reason is acceptable, the CA tries to

find an alternative solution using an error recovery strategy. If the reason is not acceptable the error is so serious that it needs the attention of a user.

5. Practical Implementation

The CoBASA architecture was validated in the NovaFlex pilot assembly cell (Figure 15), which is composed of two robot cells, one automatic warehouse and various conveyors connecting the two robot cells.

Figure 15. Célula NovaFlex

5.1 Development platform and CoBASA prototype

The JADE – Java Agent Development framework (Bellifemine et al. 2001; JADE 2001) was chosen for the experimental work mainly because it is an open source FIPA compliant platform, provides good documentation and support, and it is also recommended by the experience of other research groups with whom the authors have close relationship. Its use of *Behaviours*, and the easy connection to JESS rule processing engine (Jess 2000) helps in reducing the programming effort. Moreover JADE, implements the FIPA-ACL agent communication language. Another interesting feature of JADE is the functionalities

provided to manage the community of agents. It includes a *Remote Monitoring Agent* (RMA) tool, which is used to control the life cycle of the agent platform, and an agent for *white pages* and life cycle services (Agent Management Service - AMS).

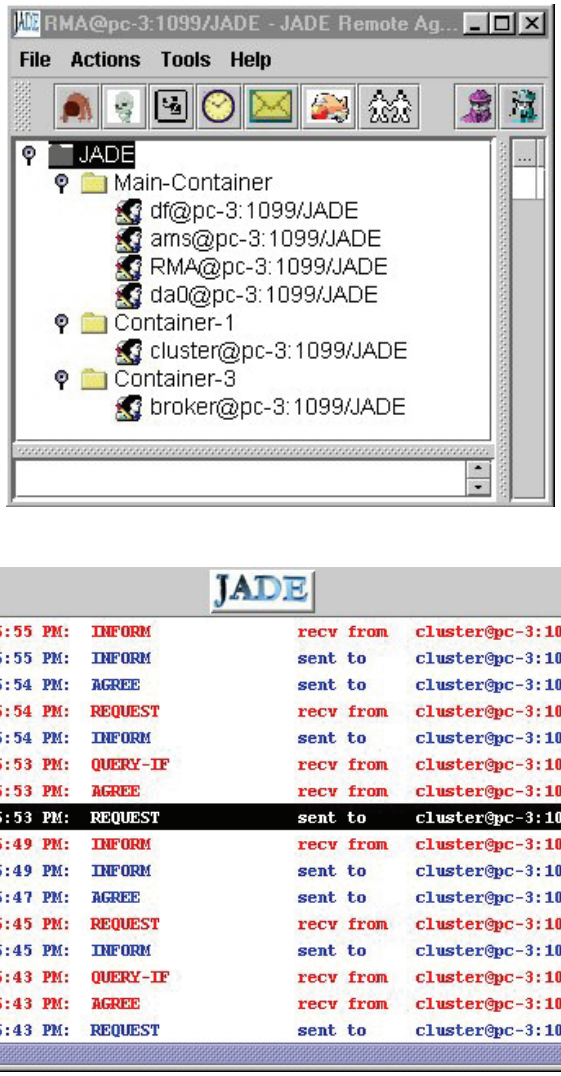


Figure 16. JADE Monitoring tool and messages between the Cluster and the Generic Agent

In Figure 16 (left hand side) the JADE monitoring tool shows the three example agents of the architecture. The agent address is da0@pc-3:1099/JADE. Although all agents were running in the same platform pc-3, this is not at all mandatory. The right hand side of Figure 16 shows the sequence of messages

between the cluster manager (CMgA) and a CA/MRA. This specific case shows the registering sequence in the cluster of two MRAs.
Figure 17 shows the main user interface of the agent (CA/MRA) (left part). The right part shows the window that is opened when the user clicks the *cluster* button. In this window the user verifies the cluster adhesion contract (Figure 18), asks the cluster manager to update the agent's credits and skills, and can terminate the agent's participation in the cluster (*dischargeByFrustration* button).
The agent's interface lets the user access other windows related to its participation in coalitions as well as its execution phase.

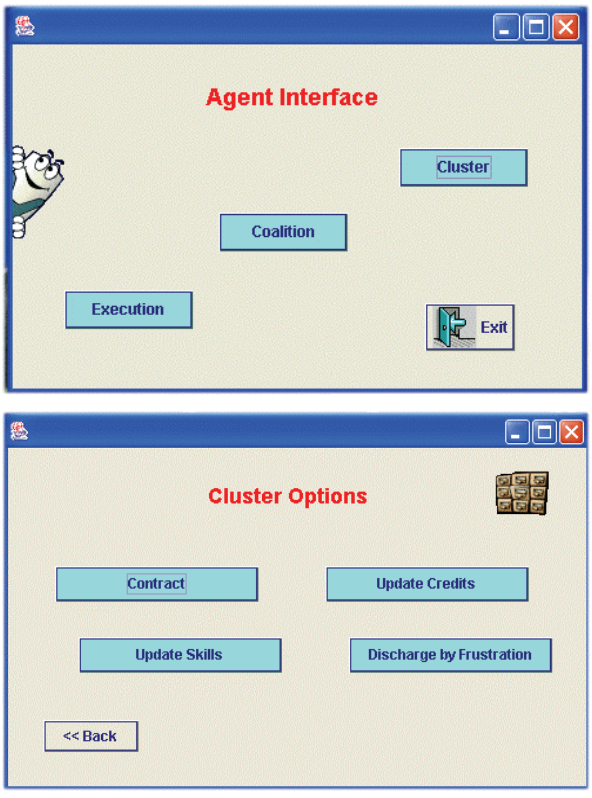


Figure 17. Agent interface and cluster options window

Figure 19 is the basic GUI of the broker. When the user chooses a candidate by selecting it (left column of available members), the broker asks the cluster manager for information about the selected agent. The figure shows that the cluster has five types of manufacturing components: robots, grippers, feeders, fixers, and coordinators (the tabs). When the user clicks on the "tabs" (options)

the members of that type existing in the cluster appear, and when the name is clicked the skills appear in the small window. The right part of the window shows the agents that have been chosen. In this case agents of type robot, feeder, gripper, and a CA, were chosen. When the user clicks on one type, the specific agent names appear in the middle column. In addition if the names in the middle column are selected the skills that were chosen to be brought in to the coalition are shown.

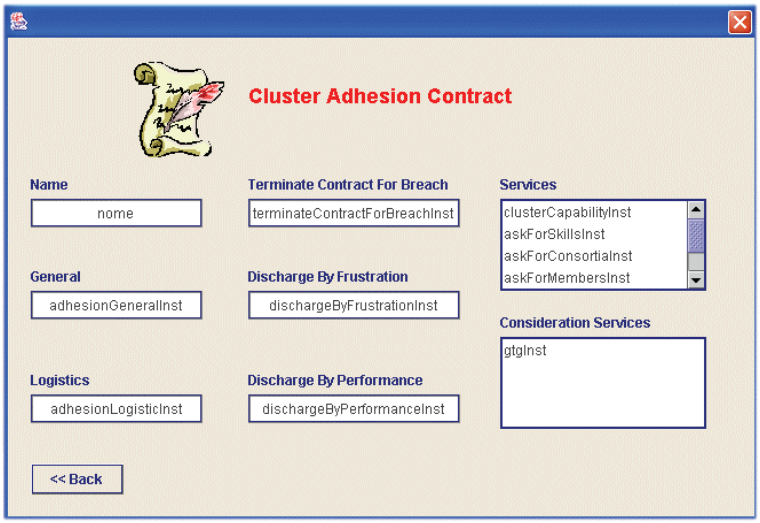


Figure 18. Cluster adhesion contract window

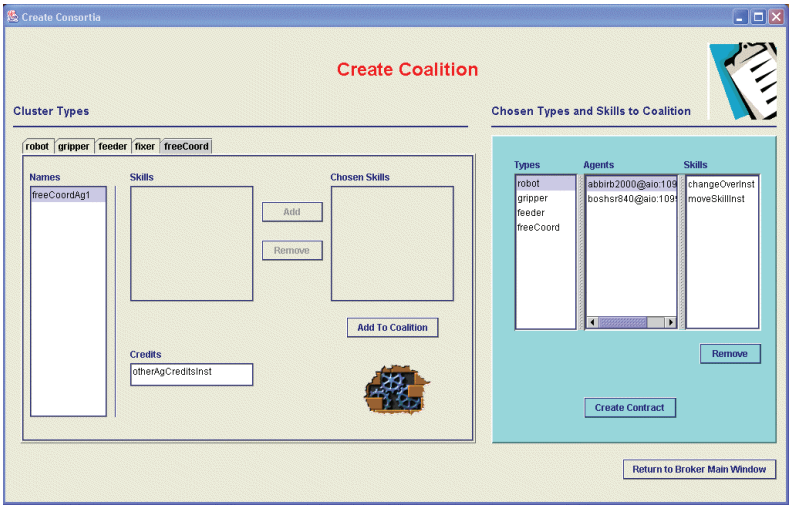


Figure 19. Create coalition/consortium in the broker

5.2 Agentification

Connecting the physical controller to the AMI could be an easy task if every physical component was controlled directly by its own agent. However, outdated legacy controllers with closed architectures control most of existing physical components. To integrate these legacy components in the agents' framework it is necessary to develop a software wrapper to hide the details of each component. The wrapper acts as an abstract machine to the agent supplying primitives that represent the functionality of the physical component and its local controller. The agent machine interface (AMI) accesses the wrapper using a local software interface (proxy), where all services of the wrapper are defined.

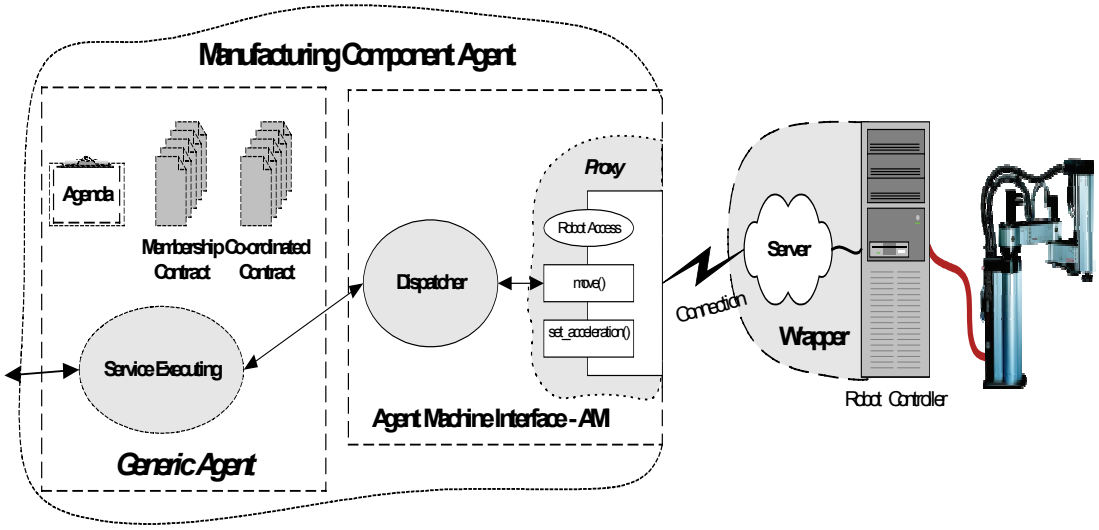


Figure 20. Physical component integration

Figure 20 shows a high level representation of an operative agent indicating how the wrapper integrates a manufacturing component (robot).
In previous works, the wrapper used to integrate physical components during the agentification process has been successfully implemented using two-tier client-server architecture (Barata et al. 1996; Camarinha-Matos et al. 1997; Camarinha-Matos et al. 1996). Recently, the wrappers for our NovaFlex manufacturing system, which is described in (Barata and Camarinha-Matos

1994), were developed using DCOM.
The Agent Machine Interface implementation is generic, i.e. an AMI can be connected to different distributed components (proxy) just by configuring what are the services of that proxy and the name/address of the component.
The generic agent tied to the AMI behaves in a slightly different way from other agents, at the initialisation phase. In this situation the GA reads from a contract representation file an instance of a consortium contract between itself and the AMI, and establishes a coalition. The member promise part (AMI) of the contract contains all the services supplied by the AMI. The agents not connected to an AMI, on the other hand, are configured not to read any contract representation file at initialisation time. This approach is very flexible because it permits to create (generate) any type of manufacturing agent just by configuring an AMI and the consortium contract between the agent and the AMI. The only part of the system that is dependent of the physical component is of course the wrapper.

6. Conclusions

The CoBASA system offers an approach to introduce agility at the shop floor control level. The prototype proved the feasibility of the proposed approach, which seems able to provide a solution to rapid reengineering of shop-floor systems. Based on the concept of generic agent and its various behaviours regulated by contracts, it is possible to change the behaviour of a complex shop floor through the definition of new contracts (configuration) without the need to reprogram the control system. Current developments are devoted to assess the level of agility of the solution and to partially automate the brokerage activities.

7. References

Almeida, C. F. (2000). *Contratos I - Conceitos; Fontes; Formação*, Almedina, Coimbra.

Barata, J., and Camarinha-Matos, L. M. (1994). "Development of a FMS/FAS System." *Studies in Informatics and Control*, 3(2-3), 231-239.

Barata, J., and Camarinha-Matos, L. M. (2000). "Shopfloor Reengineering To Support Agility in Virtual Enterprise Environments." E-Business and Virtual Enterprises, L. M. Camarinha-Matos, H. Afsarmanesh, and R. Rabelo, eds., Kluwer Academic Publishers, London, 287-291.

Barata, J., and Camarinha-Matos, L. M. (2002). "Contract Management in Agile Manufacturing Systems." Collaborative Business Ecosystems and Virtual Enterprises, L. M. Camarinha-Matos, ed., Kluwer Academic Publishers, New York, 109-122.

Barata, J., Vieira, W., and Camarinha-Matos, L. M. "Integration and MultiAgent Supervision of Flexible Manufacturing Systems." *Mechatronics'96 - The 5th UK Mechatronics Forum International Conference*, Guimarães - Portugal, 185-190.

Bellifemine, F., Poggi, A., and Rimassa, G. (2001). "Developing Multi-Agent Systems with a FIPA-Compliant Agent Framework." *Software-Practice & Experience*, 31(2), 103-128.

Camarinha-Matos, L. M., Barata, J., and Flores, L. (1997). "Shopfloor Integration and MultiAgent Supervision." I. Rudas, ed., 457-462.

Camarinha-Matos, L. M., Seabra Lopes, L., and Barata, J. (1996). "Integration and Learning in Supervision of Flexible Assembly Systems." *IEEE Transactions on Robotics and Automation (Special Issue on Assembly and Task Planning)*, 12(2), 202-219.

Conte, R., and Dellarocas, C. (2001). "Social Order in Multiagent Systems." Multiagent systems, artificial societies, and simulated organizations, Kluwer Academic Publishers, Boston, ix, 239.

Ferber, J. (1999). *Multi-Agent Systems : an Introduction to Distributed Artificial Intelligence*, Addison-Wesley, Harlow.

FIPA. (2001). "FIPA Request Interaction Protocol Specification." *XC00026F*, FIPA - Foundation for Intelligent Physical Agents, Geneve.

FIPA. (2002). "The Foundation for Intelligent Physical Agents."

Franklin, S., and Graesser, A. (1997). "Is it an Agent or Just a Program? A Taxonomy for Autonomous Agents." Intelligent Agents III - Agent Theories, Architectures, and Languages, J. P. Muller, M. Wooldridge, and N. R. Jennings, eds., Springer-Verlag, Berlin, 21-35.

Giampapa, J. A., Paolucci, M., and Sycara, K. "Agent Interoperation Across Multagent System Boundaries." *Fourth International Conference on Autonomous Agents (Agents 2000)*, Barcelona - Spain.

Goldman, S. L., Nagel, R. N., and Preiss, K. (1995). *Agile competitors and virtual organizations: strategies for enriching the customer*, Van Nostrand Reinhold, New York.

Gullander, P. (1999). "On Reference Architectures for Development of Flexible Cell Control Systems," PhD Thesis, Gotenborg University, Sweden.

Huff, B. L., and Edwards, C. R. (1999). "Layered Supervisory Control Architecture for Reconfigurable Automation." *Production Planning & Control*, 10(7), 659-670.

JADE. (2001). "http://sharon.cselt.it/projects/jade/."

Jess. (2000). "http://herzberg.ca.sandia.gov/jess/."

Johnson, S. (2001). *Emergence*, Penguin group, London.

Klusch, M., and Sycara, K. (2001). "Brokering and Matchmaking for Coordination of Agent Societies: A Survey." Coordination of Internet Agents: Models, Technologies, and Applications, A. Omicini, F. Zambonelli, M. Klusch, and R. Tolksdorf, eds., Springer-Verlag, Berlin, xxvii, 523.

Koren, Y., Heisel, U., Jovane, F., Moriwaki, T., Pritchow, G., Ulsoy, A. G., and Van Brussel, H. (1999). "Reconfigurable Manufacturing Systems." *CIRP Annals*, 48(2).

McKendrick, E. (2000). *Contract Law*, PALGRAVE, New York.

Mehrabi, M. G., Ulsoy, A. G., and Koren, Y. (2000). "Reconfigurable Manufacturing Systems: Key to Future Manufacturing." *Journal of Intelligent Manufacturing*, 11, 403-419.

Onori, M. (1996). "The Robot Motion Module: A Task-Oriented Robot Programming System for FAA Cells," PhD thesis, The Royal Institute of Technology, Stockholm.

Payne, T., Singh, R., and Sycara, K. "Facilitating Message Exchange through Middle Agents." *The First International Joint Conference on Autonomous Agents and Multi-Agent Systems*.

Protégé-2000. (2000). "http://protege.stanford.edu."

Sycara, K., Decker, K., and Williamson, M. "Middle-Agents for the Internet." *IJCAI-97 International Conference on Artificial Intelligence*, Nagoya - Japan.

Vos, J. A. W. M. (2001). "Module and System Design in Flexible Automated Assembly," PhD Thesis, Delft University Press, Delft.

Weiss, G. (1999). "Multiagent Systems : a modern approach to distributed artificial intelligence." MIT Press, Cambridge, Masschusetts, xxiii, 619.

WFMC. (2002). "Workflow Management Coalition."

Wiederhold, G. (1992). "Mediators in the Architecture of Future Information Systems." *IEEE Computer Systems*, 25(3), 38-49.

Wong, H. C., and Sycara, K. "A Taxonomy of Middle-Agents for the Internet." *Fourth International Conference on MultiAgent Systems*, 465-466.

Wooldridge, M., and Jennings, N. R. (1995). "Intelligent Agents - Theory and Practice." *Knowledge Engineering Review*, 10(2), 115-152.

Wooldridge, M. J. (2000). *Reasoning about Rational Agents*, MIT Press, Cambridge, Massachusetts; London.

Wooldridge, M. J. (2002). *An Introduction to Multiagent Systems*, J. Wiley, New York.

Zurawski, R., and Zhou, M. C. (1994). "Petri Nets and Industrial Applications - a Tutorial." *IEEE Transactions on Industrial Electronics*, 41(6), 567-583.

Zwegers, A. (1998). "On Systems Architecting - a study in shop floor control to determine architecting concepts and principles," PhD Thesis, Eindhoven Technical University, Eindhoven - The Netherlands.

3

Development of Holonic Manufacturing Execution Systems

Fan-Tien Cheng, Chih-Feng Chang and Shang-Lun Wu

1. Introduction

Today, most semiconductor manufacturing companies utilize Manufacturing Execution Systems (MES) (MacDonald, 1993; Samanish, 1993; Nguyen, 1996; Scatt, 1996; MESA, 1997) to deliver information to optimize production activities from order booking through design, production, and marketing to realize the agile manufacturing enterprise. The MES market is composed of several vendors providing an integrated suite of application products (called an integrated MES), and 200, or so, vendors offering individual point solutions (Scott, 1996). An integrated MES may have many advantages, such as a single-logic database, rich functionality, well-integrated applications, and a single model of factories, products, and manufacturing processes. However, integrated MES's are sometimes regarded as monolithic, insufficiently configurable, and difficult to modify. Point solutions can offer best-in-class capabilities for a particular function (such as cell controller, work-in-process (WIP) tracking, statistical process control, scheduling, etc.); the end result is multiple databases, multiple models, and integration nightmares plus maintenance costs (McGehee, et al. 1994; Kadar et al., 1998).

In order to solve the problem of the dichotomy between the integrated MES and point solutions, the concept of the integratable MES has been proposed (Scott, 1996). With the integratable MES, each application can be both a self-sufficient point solution, and can be integrated into a larger suite of products. Therefore, the integratable MES offers an open, modularized, configurable, distributed, and collaborative environment such that rapid implementation, complexity reducing, agility, cost-effective integration, easiness of use, and ownership cost reducing may be achieved (McGehee et al., 1994; Kadar et al., 1998).

McGehee et al. (1994) presented the Texas Instruments Microelectronics Manufacturing Science and Technology (MMST) CIM System Framework, which was based on open-distributed system and object technologies. This re-

engineering effort used the OMT methodology models (Rumbaugh et al., 1991) to express the MMST Framework. Following the MMST CIM System Framework, SEMATECH developed the CIM Framework Specification version 2.0 (SEMATECH, 1998), which is an abstract model for typical semiconductor manufacturing systems.

Several approaches to distributed manufacturing architectures were surveyed by Kadar et al. (1998), and their fundamental features were highlighted. Moreover, an object-oriented simulation framework for development and evaluation of multi-agent manufacturing architectures was introduced by Kadar et al. (1998). Further, Cheng, et al. (1999) applied the distributed object-oriented technologies to develop the MES Framework. This framework has the characteristics of openness, modularization, distribution, reconfigurability, interoperability, and easy maintenance.
Common automatic manufacturing systems have fragility and security problems that also need to be seriously taken into consideration, however these two issues are not considered in the MES frameworks mentioned above. This paper applies the concepts of holon and holarchy to redesign a Holonic Manufacturing Execution System (HMES) Holarchy that not only possesses the characteristics of the MES Framework (Cheng et al., 1999) but also has the properties of failure recovery and security certification.

The concepts of holon and holarchy are originated from mechanisms of social organizations and biological organisms (Valckenaers et al., 1994; Tonshoff et al., 1994; HMS; Van Leeuwen & Norrie, 1997). They have the characteristics of intelligence, autonomy, coordination, reconfigurability and extensibility. Based on these characteristics, the major weakness in the automatic manufacturing systems, fragility, is removed so that the failure recovery feature is attained. Security certification also can be considered.
A typical deployment diagram for HMES in the semiconductor packaging plant is displayed in Fig. 1. HMES includes Shop-Floor Holon, Scheduling Holon, WIP Holon, Data Warehouse, Material Handling, Equipment Holon, Equipment, AGV, AS/RS and so on. The HMES Holarchy will be developed by a systematic approach in this paper. For demonstration purpose, one of the functional holons - WIP Holon - will be designed and implemented. Most of the studies concerning holonic manufacturing systems (Markus et al., 1996; Ramos, 1996; Hino & Moriwaki, 1999) focus on factory architecture and/or how to assign a production task to each manufacturing holon. The purpose of this paper is to propose a systematic approach for developing a workable

Holonic Manufacturing Execution System (HMES) by applying up-to-date software and information technologies. The systematic approach is started with system analysis by collecting domain requirements and analyzing domain knowledge.

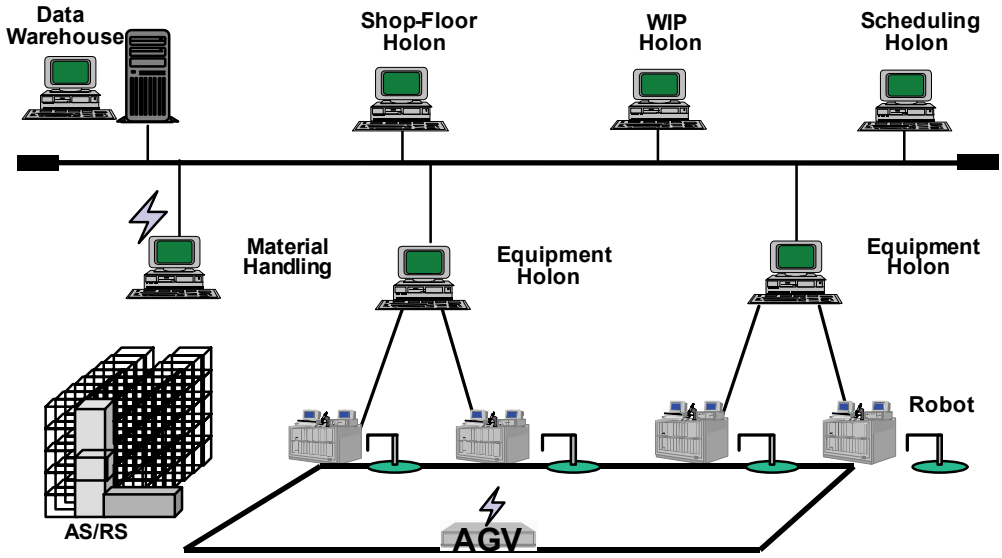


Figure 1. Deployment Diagram for Holonic Manufacturing Execution Systems

The HMES Holarchy is designed by the procedure of constructing an abstract object model based on domain knowledge, partitioning the application domain into components, identifying generic parts among components to form the Generic Holon, developing the Generic Holon, defining holarchy messages and the holarchy framework of HMES, and finally designing functional holons based on the Generic Holon. The technologies (Chen & Chen, 1994; Gamma et al., 1995; Mowbray, 1995; Orfali et al., 1996; Sparks et al., 1996) of distributed object-oriented approach, design pattern, framework, N-tier client/server architecture, and component software are applied to develop the entire HMES and its functional holons.

This paper is organized as follows: Section 2 introduces the characteristics of holon and holarchy. Section 3 describes the development procedure of HMES. This development procedure includes four stages: system analysis, holarchy design, application construction, and system integration and testing. Among those stages, holarchy design needs most elaboration and it is explained in de-

tail in Section 4. Section 5 demonstrates WIP holon design. Section 6 describes application construction and system integration. Section 7 makes comparisons among Legacy MES, Framework MES, and Holonic MES. Finally, this paper ends with summary and conclusions.

2. Characteristics of Holon and Holarchy

Twenty-six years ago, the Hungarian author and philosopher Arthur oestler proposed the word holon to describe a basic unit of organization in biological and social systems. A holon, as Koestler devised the term, is an identifiable part of a system that has a unique identity, yet is made up of sub-ordinate parts and in turn is a part of a larger whole.

The strength of holonic organization, or holarchy, is that it enables the construction of very complex systems that are nonetheless efficient in the use of resources, highly resilient to disturbances (both internal and external), and adaptable to changes in the environment in which they exist. All these characteristics can be observed in biological and social systems.

The stability of holons and holarchies stems from holons being self-reliant units, which have a degree of independence and handle circumstances and problems on their particular level of existence without asking higher level holons for assistance. Holons can also receive instruction from and, to a certain extent, be controlled by higher-level holons. The self-reliant characteristic ensures that holons are stable and able to survive disturbances. The subordination to higher-level holons ensures the effective operation of the larger whole.

The task of the Holonic Manufacturing System (HMS) consortium is to translate the concepts that Koestler developed for social organizations and living organisms into a set of appropriate concepts for manufacturing industries. The goal of this work is to attain in manufacturing the benefits that holonic organization provides to living organisms and societies, e.g., stability in the face of disturbances, adaptability, and flexibility in the face of change, and efficient use of available resources.

As an initial step, the HMS consortium developed the following list of definitions (among others) to help understand and guide the translation of holonic concepts into a manufacturing setting (Van Leeuwen & Norrie, 1997; Ulieru, 1997):

a) **Holon:** An autonomous and cooperative building block of a manufacturing system for transforming, transporting, storing and/or validating in-

formation and physical objects. The holon consists of an information processing part and often a physical processing part. A holon can be part of another holon.

b) **Autonomy:** The capability of an entity to create and control the execution of its own plans and/or strategies.

c) **Cooperation:** A process whereby a set of entities develops mutually acceptable plans and executes these plans.

d) **Holarchy:** A system of holons that can cooperate to achieve a goal or objective. The holarchy defines the basic rules for cooperation of the holons and thereby limits their autonomy.

e) **Holonic Manufacturing System (HMS):** A holarchy that integrates the entire range of manufacturing activities from order booking through design, production, and marketing to realize the agile manufacturing enterprise.

f) **Holonic Attributes:** The attributes of an entity that make it a holon. The minimum set is autonomy and cooperatives.

Based on the above definitions, it is clear that holonic manufacturing systems can be regarded as a unified way to approach the hierarchical control of any manufacturing unit from the production process to the whole enterprise level.
In this work, the concepts of holon and holarchy are adopted to develop the HMES Holarchy so that the functional holons of the HMES can possess the properties of intelligence, autonomy, cooperation, reconfigurability, and extensibility. In addition, the functional holons of the HMES Holarchy can have the capabilities of failure recovery and security certification.

3. Development Procedure of Holonic Manufacturing Execution Systems

As depicted in Fig. 2, the development procedure of HMES includes four stages: (a) system analysis, (b) holarchy design, (c) application construction and (d) system integration and testing. Note that the final step of holarchy design stage is functional holon design and implementation.
The first stage, system analysis, concentrates on collecting domain requirements and analyzing domain knowledge. The second stage, the most important stage, is holarchy design, which is further divided into seven steps as shown in Fig. 2.

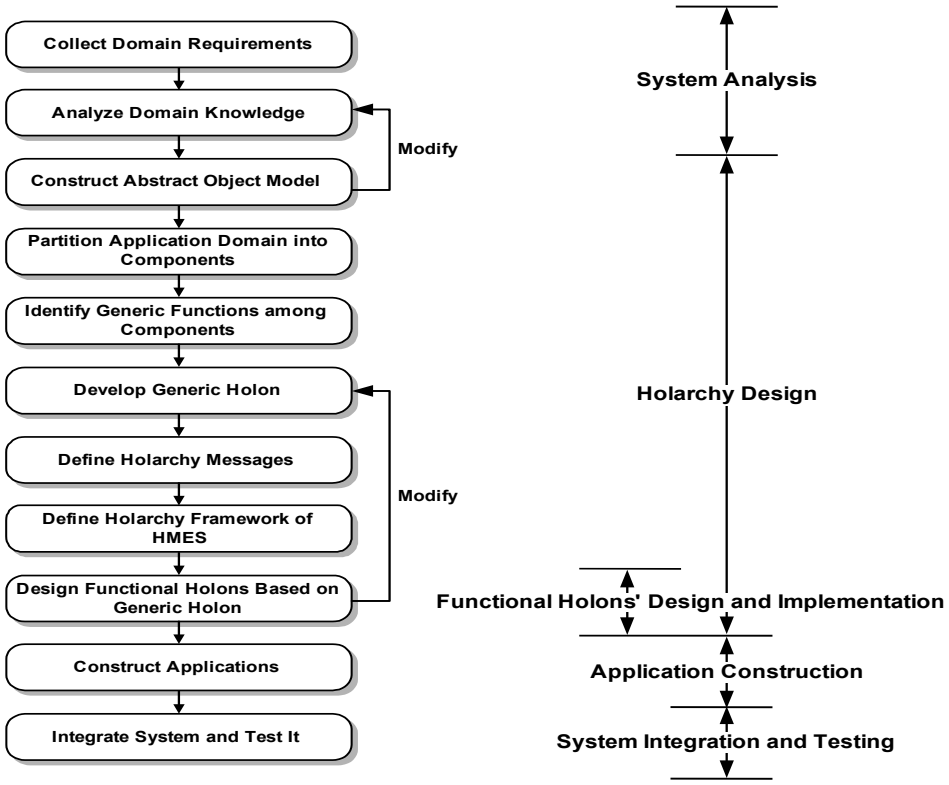


Figure 2. Development Procedure of Holonic Manufacturing Execution Systems

The system's object model is constructed according to the domain knowledge and requirements. The application domain is partitioned into components that will eventually become various functional holons. Within these components, their generic functions are further identified and extracted. Based on these generic functions, the so-called Generic Holon is developed. Holarchy messages among functional holons are defined and holarchy framework of HMES (also denoted HMES Holarchy) is developed. Finally, various functional holons can be designed by inheriting the Generic Holon and implementing the holarchy messages. The third stage of HMES development is application construction. Finally, the development procedure ends with system integration and testing.

4. Holarchy Design

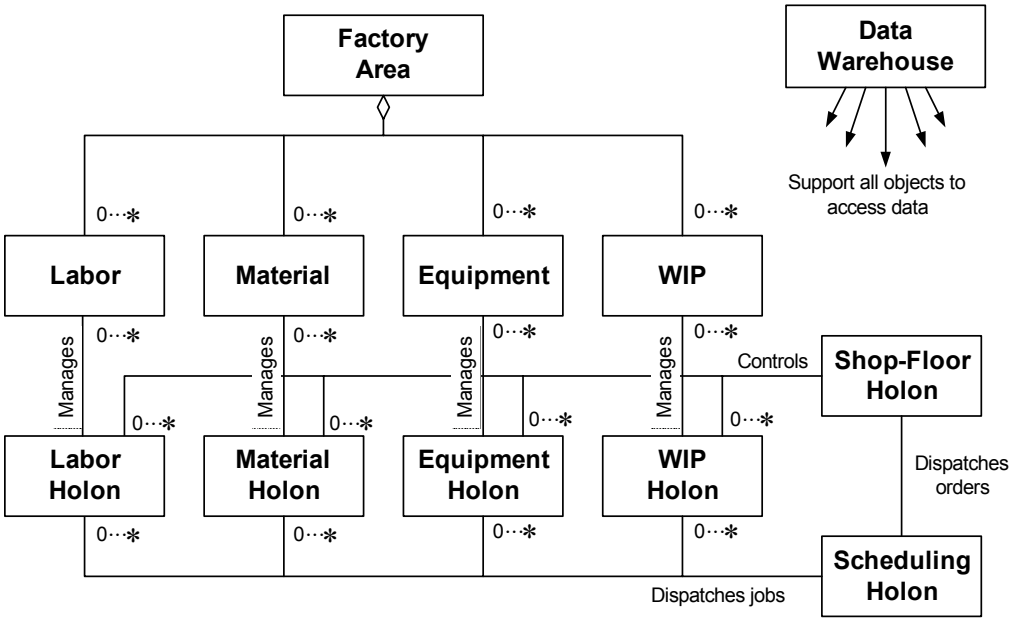
Seven steps are included in the holarchy design stage. They are explained below.

4.1 Constructing an Abstract Object Model

A typical deployment diagram for HMES is shown in Fig. 1. It is well known that MES is composed of several functional modules that handle specifics, e.g. material, equipment, labor, and planning (MacDonald, 1993). The abstract object model is constructed as in Fig. 3(a) (Cheng et al., 1999).
The four key elements of a factory are labor, material, equipment, and work-in-process (WIP). Each element is managed by its specific managing holon. All four of these managing holons are controlled by the Shop-Floor Holon. The Shop-Floor Holon also dispatches orders to the Scheduling Holon. The Scheduling Holon dispatches jobs to the Labor Holon, Material Holon, Equipment Holon, and WIP Holon.

4.2 Partitioning Application Domain into Components

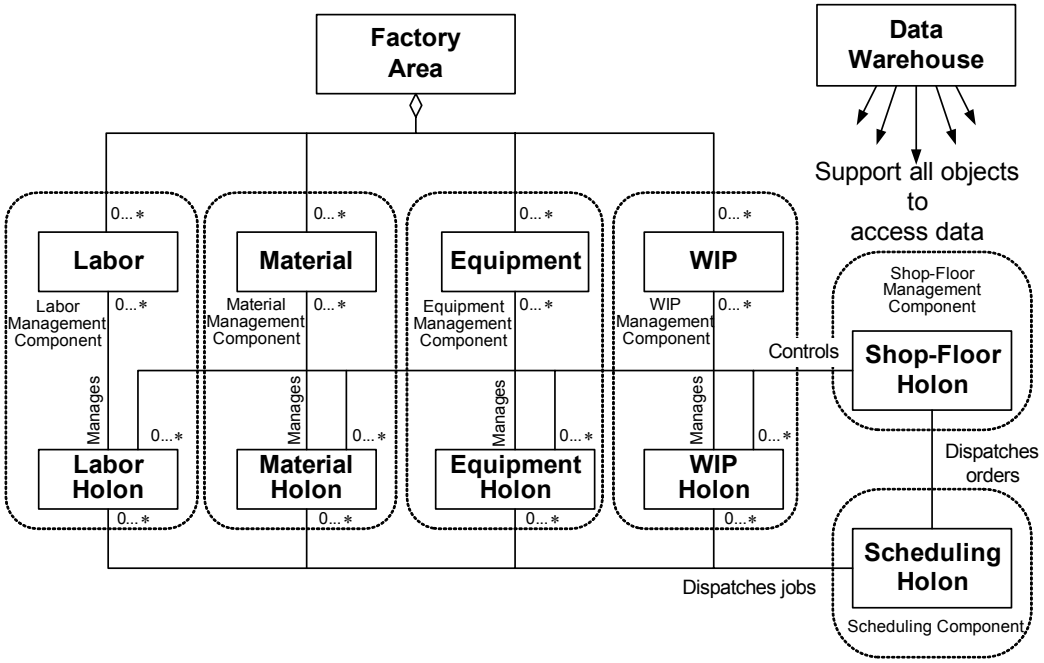
To design a distributed and integratable MES, its application domain is partitioned systematically as depicted in Fig. 3(b). In addition to the data warehouse, the system is divided into six components.



(a) Abstract Object Model

They are labor management, material management, equipment management, WIP management, scheduling, and shop-floor management components. The

labor, material, equipment, and WIP management components handle labor, movements of materials, process equipment, and WIP tracking, respectively. The scheduling component takes care of scheduling and dispatching tasks of the system. The shop-floor management component is in charge of system-level services and management, i.e., order management, life-cycle services, collection services, and query services. Each management component has a specific functional holon, which serves as the manager of that specific management component



(b) Partitioning Application Domain into Components

Figure 3. Object Model of an HMES

As mentioned previously, each management component needs a specific functional holon to serve as the manager of that component.

4.3 Identifying Generic Functions among Components

The purpose of this paper is to apply the concepts of holon and holarchy to design the HMES Holarchy and functional holons that not only possesses the properties of the MES Framework (Cheng et al., 1999) but also has the properties of failure recovery and security certification. Therefore, based on the prin-

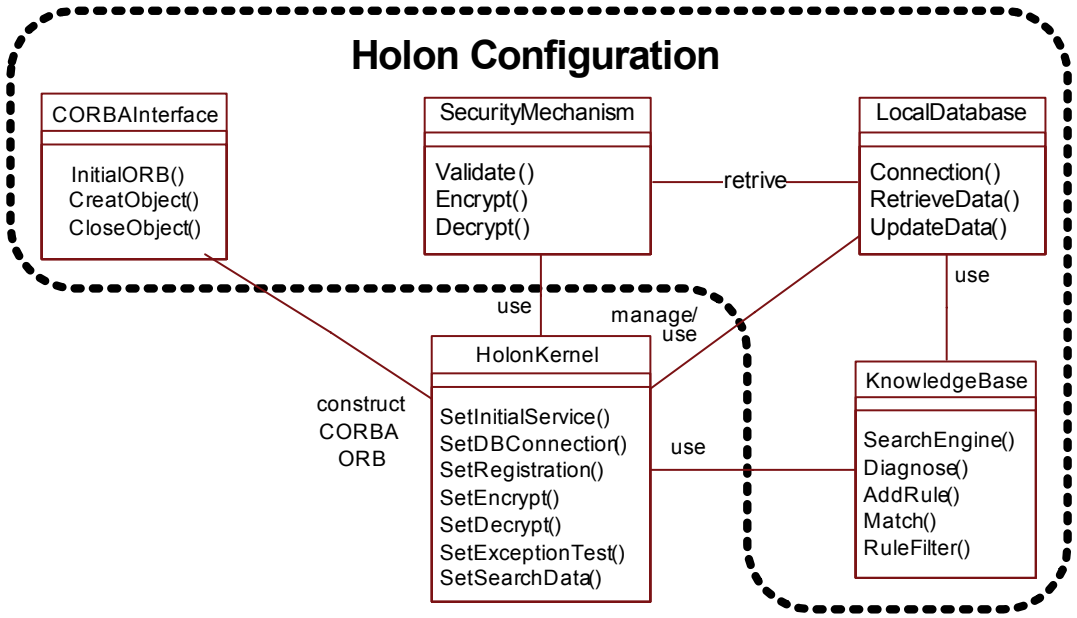
ciple of software reuse (Chen and Chen, 1994; Cheng et al., 1999), the Generic Holon which handles the generic functions of functional holons shall first be devised. After judicious consideration, the authors conclude that in addition to the communication infrastructure, the Generic Holon shall possess security mechanisms, search mechanisms, and intelligence mechanisms to deal with the generic functions that emphasize failure recovery and security certification.

4.4 Developing Generic Holon

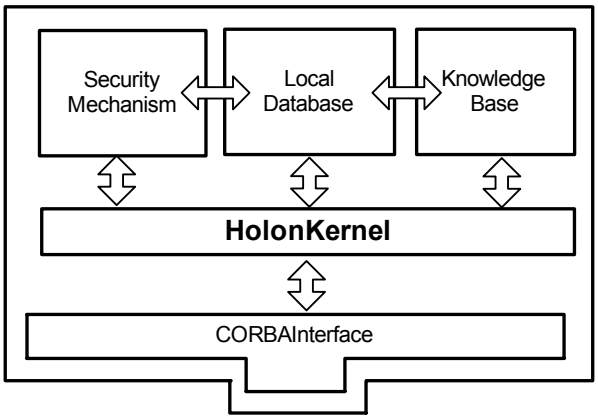
The requirements for developing the Generic Holon are:

a) It can construct the communication infrastructure for communication, collaboration, and extensibility purposes.
b) It provides the intelligence mechanism for exception diagnosis.
c) It provides the search mechanism for collaboration and reconfigurability.
d) It provides the security mechanism for security check and encryption / decryption.
e) It provides the ability to establish database services for information storage / retrieval.

According to these requirements and following the development procedure for object-oriented systems (Eriksson and Penker, 1998; Huang et al., 1999; Cheng et al., 2002), the Generic Holon's class diagram and internal architecture is obtained as shown in Fig. 4. For further illustration, please refer to (Lin, 2000; Chang, 2000) for the detailed designs of the Generic Holon.
Observing Fig. 4(a), the basic structure of the class diagram is HolonKernel manages/uses HolonConfiguration that consists of CORBAInterface, SecurityMechanism, LocalDatabase, and KnowledgeBase. By inheriting HolonKernel, a functional holon can possess all the characteristics of the Generic Holon. CORBAInterface is designed for constructing a communication infrastructure and achieves the collaboration platform. In order to establish secure communication, the SecurityMechanism is created for handling all the operations of security. KnowledgeBase constructs a search engine for searching desired services and a reasoning mechanism for exception diagnosis. The LocalDatabase sets the connection of database for SecurityMechanism and KnowledgeBase to access the database. On the other hand, the internal architecture of the Generic Holon is depicted in Fig. 4(b).



(a) Class Diagram



(b) Internal Architecture

Figure 4. Class Diagram and Internal Architecture of Generic Holon

Observing Fig. 4(b), the Generic Holon owns HolonKernel to communicate with other holons by CORBAInterface. Using LocalDatabase, the Generic Holon can maintain autonomous properties and necessary information. SecurityMechanism can retrieve the related information through LocalDatabase and then check user's authorization for security certification. The intelligence

mechanism for exception diagnosis purposes of the Generic Holon is mainly considered in knowledgeBase that also needs the support of LocalDatabase. After completing the design of the Generic Holon, any functional holon can be designed by inheriting Generic Holon to obtain generic properties of holon and then adding the specific functions of that functional holon.

4.5 Defining Holarchy Messages

After partitioning the application domain into components, we need to define holarchy messages among all the functional holons so that interoperability and collaboration among all the functional holons are enabled. According to Fig. 1 and Fig. 3(b), the holarchy messages of HMES are defined as in Fig. 5.

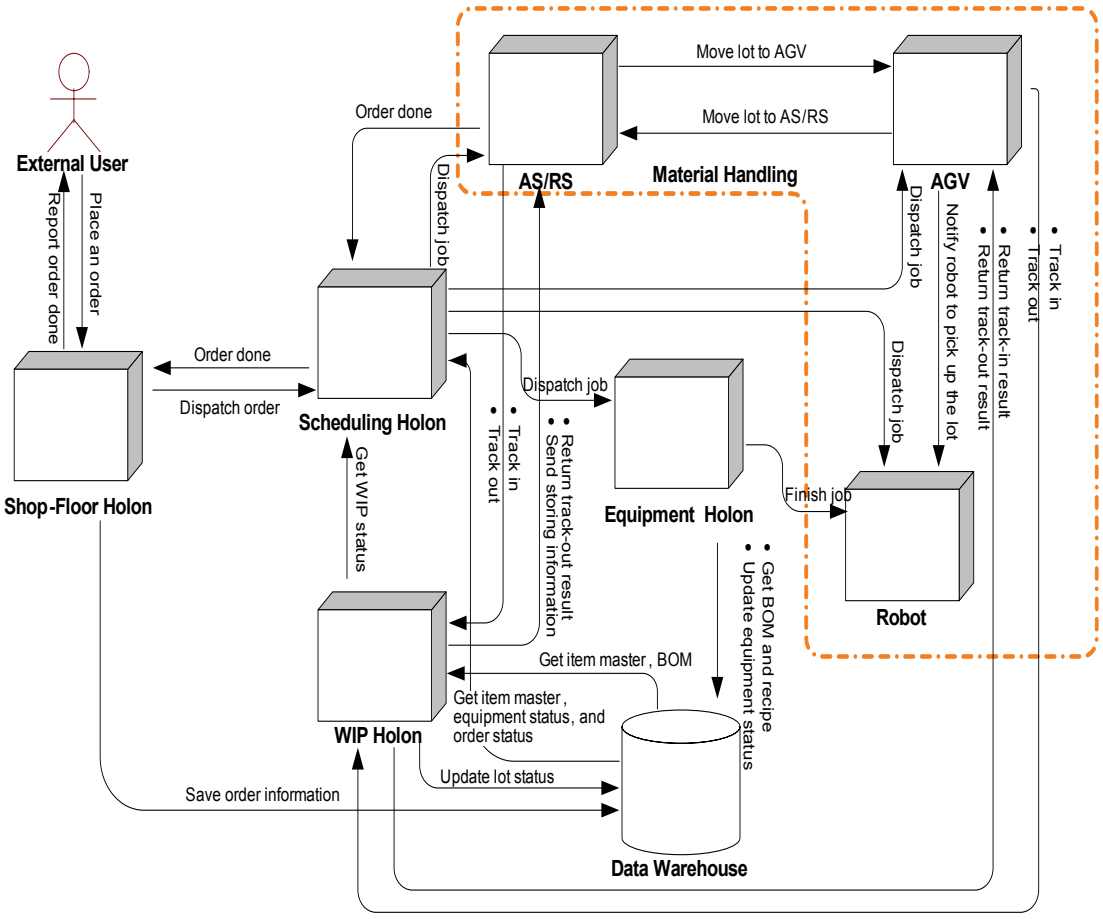


Figure 5. Defining Holarchy Messages

The Shop-Floor Holon receives a place an order message from an external user and the Shop-Floor Holon will reply report order done when the order is done. Based on the received order, the Shop-Floor Holon will send dispatch order to the Scheduling Holon and the Scheduling Holon will reply order done if the order is finished. The Shop-Floor Holon sends save order information to the Data Warehouse to save all the order information. Similarly, the interfacing holarchy messages of Scheduling Holon, WIP Holon, Equipment Holon, Data Warehouse, and Material Handling (which includes AS/RS, AGV, and robot) can be defined as shown in Fig. 5.

4.6 Defining Holarchy Framework of Holonic Manufacturing Execution Systems

After the development of the Generic Holon and holarchy messages, we are ready to define the holarchy framework of HMES (or HMES Holarchy in short).

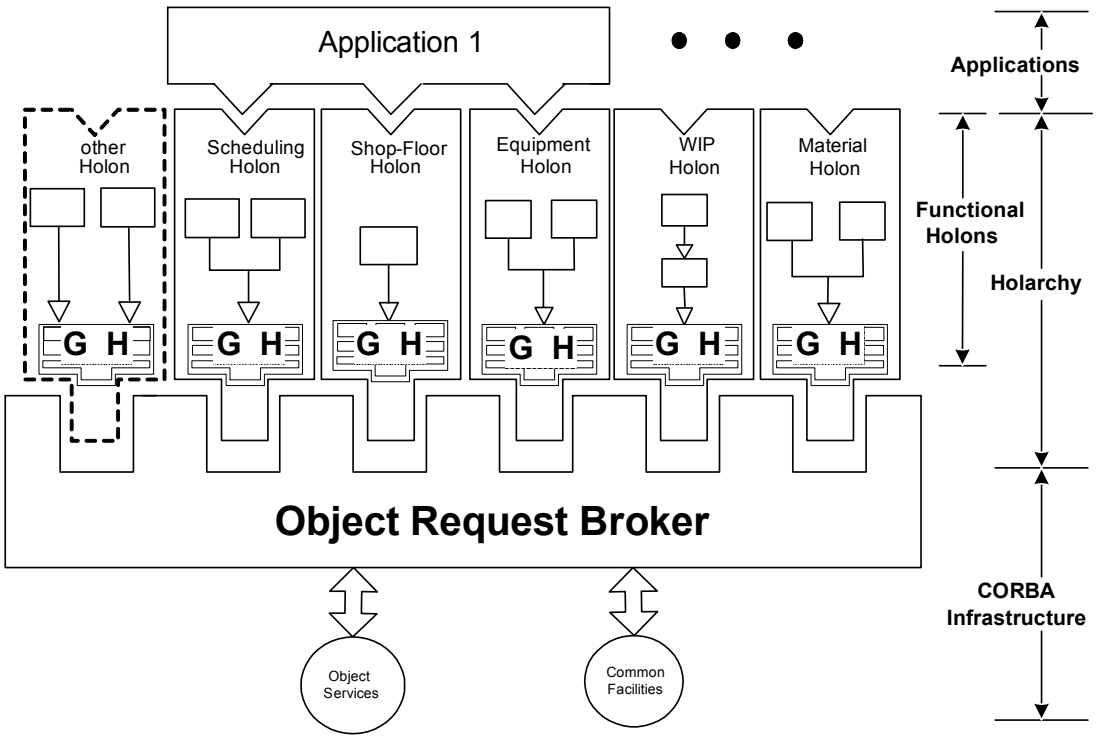


Figure 6. Holarchy Framework of Holonic Manufacturing Execution Systems

The HMES Holarchy is illustrated in Fig. 6 which utilizes CORBA infrastructure (Orfali et al., 1996; OMG, 1998) as the system's communication backbone. Every functional holon shall inherit the Generic Holon so as to possess the properties of a holon as well as the capabilities of failure recovery and security certification. Then, specific functions of each functional holon can be added individually to become a specific functional holon. The holarchy messages of each functional holon can be specified by CORBA IDL (Interface Definition Language) (Orfali et al., 1996; OMG, 1998). Therefore, each functional holon can be integrated into the HMES Holarchy in a plug-and-play fashion.

This HMES Holarchy is expandable. As illustrated on the left side of Fig. 6, other functional holon may also be integrated into the HMES Holarchy if this functional holon inherits the Generic Holon and defines the functional holon's CORBA IDL by the expanded holarchy messages. Finally, applications of the HMES can be easily constructed by invoking the related functional holons as depicted on top of Fig. 6.

4.7 Designing Functional Holons

The final step of holarchy design is to design various functional holons based on the Generic Holon. As mentioned in the previous sub-section, with the HMES Holarchy architecture, it becomes straightforward to design a functional holon by simply inheriting the Generic Holon, adding the functional holon's specific function, and defining its IDL based on the system's holarchy messages. In the following section, the WIP holon is selected as the example to elaborate the design procedure of a functional holon.

5. WIP Holon Design

The functional requirements for WIP holons are:

a) It manages the life cycle of WIP objects.
b) It performs track-in and track-out operations and updates the corresponding WIP information in real-time.
c) It provides WIP information to users and other holons.
d) Its interfaces are in compliance with the HMES Holarchy.
e) It possesses the capabilities of exception recovery and security certification.

Requirements (a) to (c) are the specific functions of WIP holons while Requirements (d) and (e) are the common requirements for the components of HMES Holarchy. It is natural to develop the WIP Holon by inheriting the Generic Holon first to take care of Requirements (d) and (e) and then considering the specific requirements (a) to (c). Based on the above design principle and following the development procedure for object-oriented systems (Eriksson and Penker, 1998; Huang et al., 1999), the class diagram of the WIP Holon is designed and shown in Fig. 7.

The upper portion of Fig. 7 is the Generic Holon that has been designed and illustrated in Fig. 4(a). WIPManager, which is the primary role of the entire WIP Holon, inherits the Generic Holon to accomplish Requirements (d) and (e).
WIPManager uses RecoveryManager to perform specific recovery operations. WIPManager also manages the life cycle of WIP objects and is in charge of track-in and track-out operations of all the WIP. A new WIP object is created when a new lot arrives. The WIP object contains its own specific attributes such as LotID, BOM, and ItemMaster, etc. A WIP object also performs its own Trackin() Trackout() operations and invokes NewVariables() methods of BOM and ItemMaster to obtain the associated production information. UserInterface provides the necessary operations for external users to interface with the WIP Holon.

Observing Fig. 7, the + sign before an operation means the operation is public, and the – sign stands for private. In the WIPManager, public operations stand for the IDL of the system; while in the UserInterface, public operations indicate the available functions for external users.

State diagrams show all possible states and transactions of a system. A change of state caused by an event is called a transition. Figure 8(a) illustrates the states and transitions of the WIP Holon. Please refer to Fig. 7 and Fig. 8 when reading the following explanation.

A user initiates the WIP Holon by invoking the Login() operation of UserInteface. If he passes the security certification, the WIP Holon will activate CORBA services by calling SetInitialService()of HolonKernel. Then, the system is ready to receive WIP object's creating commands.

In fact, the major functions of the WIP holon are how to trace and manage WIP. We define WIP to be temporal objects, as such they have life cycles. Figure 8(b) is the state diagram of WIP life cycle.

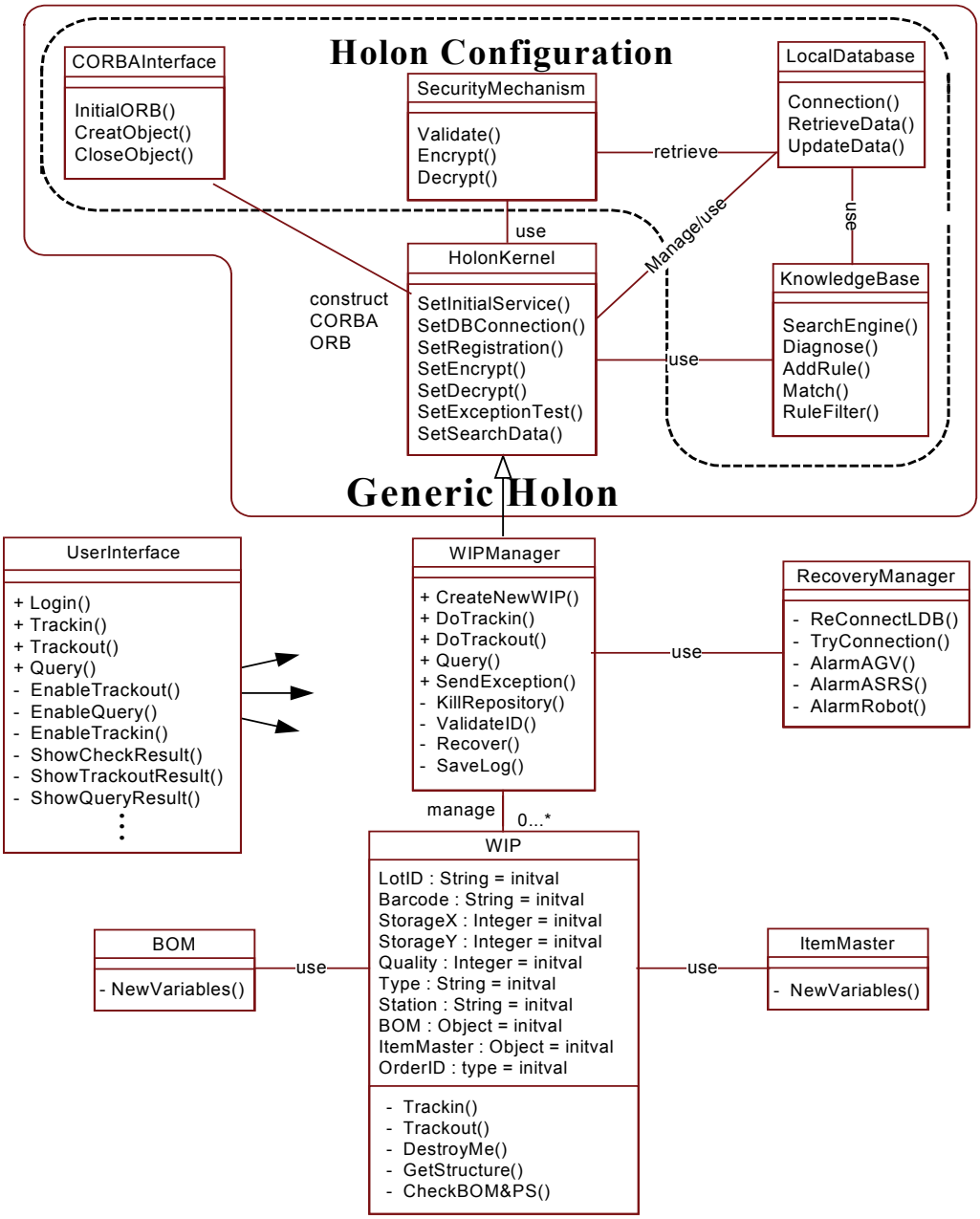


Figure 7. Class Diagram of WIP Holon

When WIPManager gets the message CreateNewWIP() from the Scheduling Holon, a new WIP object is generated based on the data transferred from the Scheduling Holon. WIP object uses NewVariables() operation in BOM to get the contents of BOM. WIP object uses the same approach to obtain ItemMaster information. Then, WIP object gets order status and saves it. Up to this point, initialization of WIP object is completed and it enters Wait for request state.

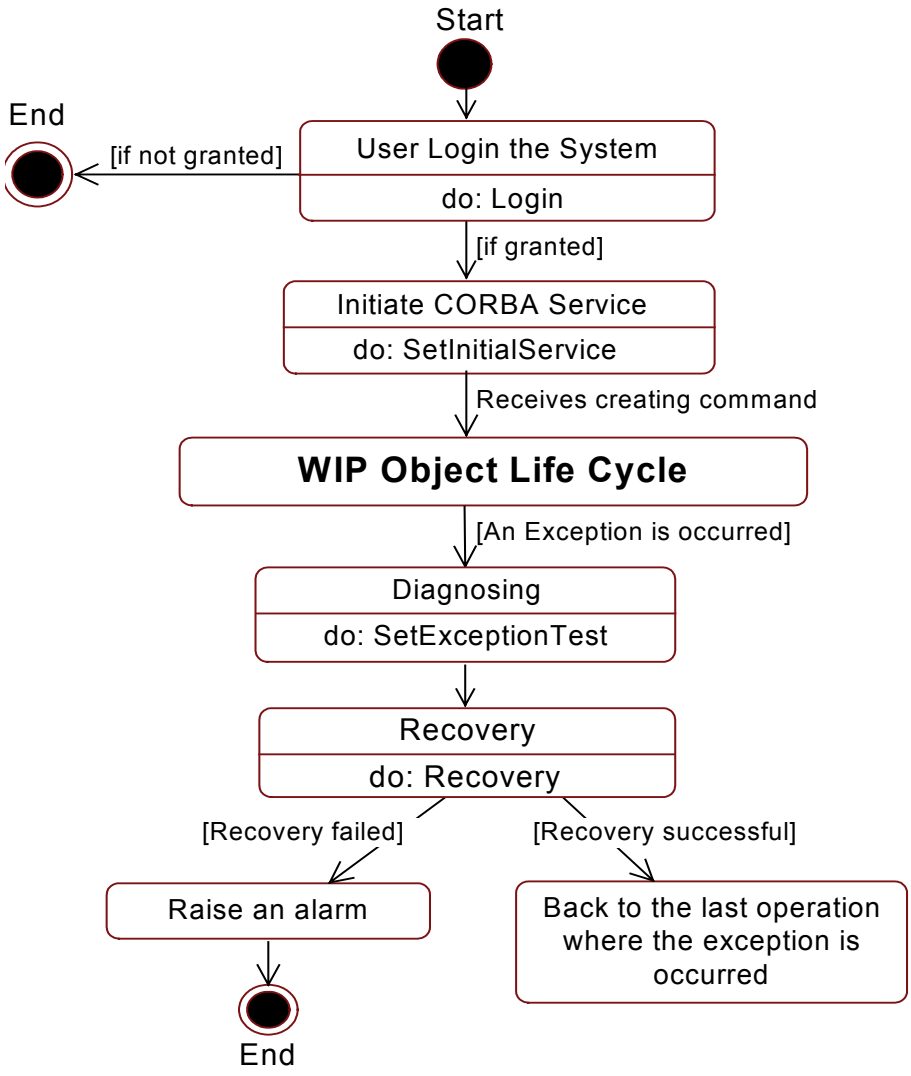
At Wait for request state, the WIP object can take commands, such as track-in, track-out, and query. The query request will bring the WIP object to the Provide WIP status state and the WIP status is then sent to the requester. Track-out and track-in commands will update the WIP status and store it to database. During track-in operation, the WIP object will check if this current process sequence is the last one or not. If it is not, just jumps back to Wait for request state. If it is the last process, this WIP object will be deleted and the memory will be released. It thus completes the life cycle of a WIP object.

Note that, the initial Generic Holon architecture shown in Fig. 4 only specifies the generic skeleton of the intelligence mechanism that consists of KnowledgeBase and LocalDatabase. After inheriting the Generic Holon to become a part of the WIP Holon, its KnowledgeBase and LocalDatabase shall be trained to contain the specific knowledge, information, and rules for WIP holon's exception-diagnosis usage only.

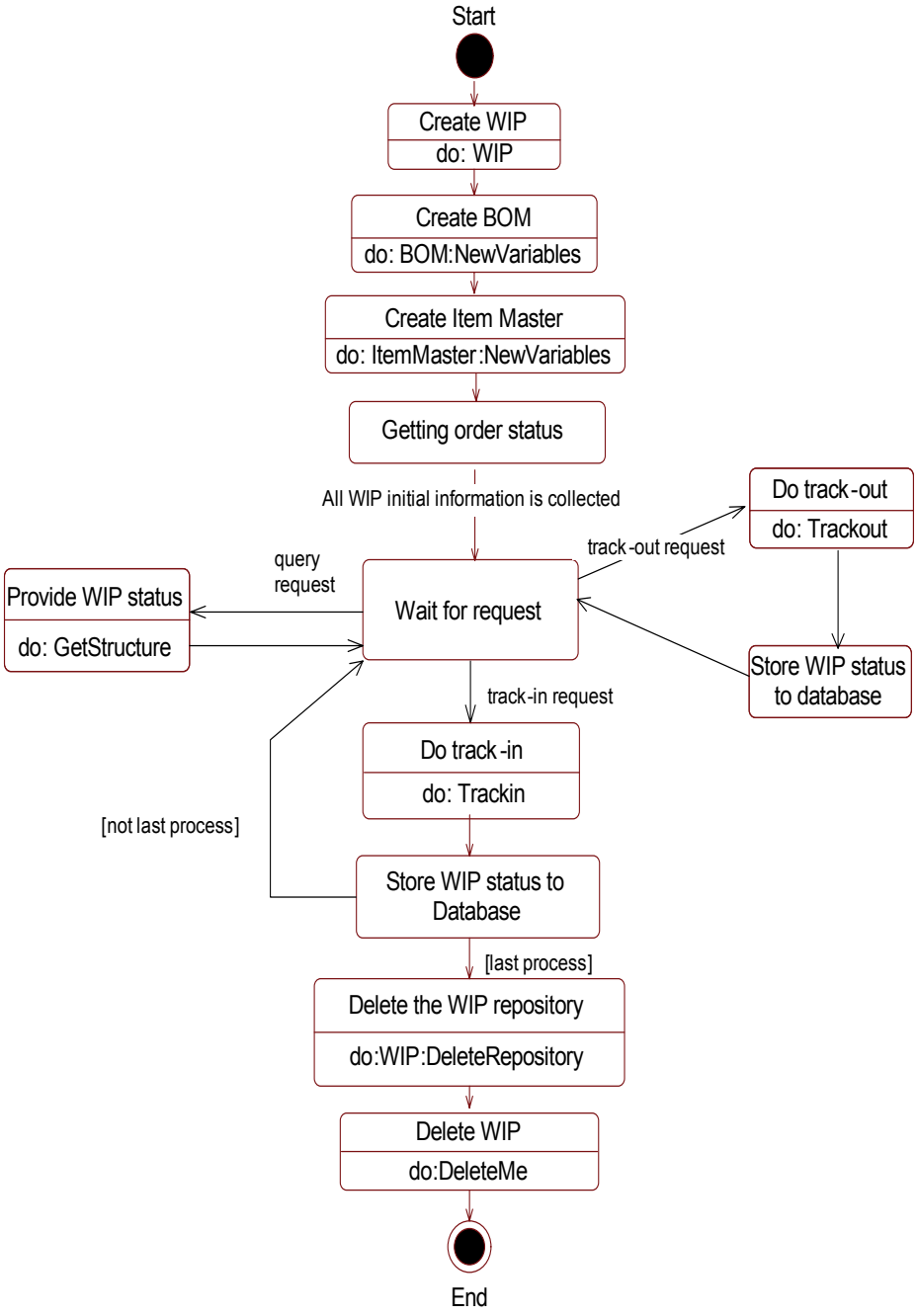
Now, observing Fig. 8(a), if an exception is occurred and detected during the WIP management process, the system will enter the Diagnosing state that invokes SetExceptionTest() of HolonKernel to diagnose the exception.

If the cause is identified by the intelligence mechanism of the Generic Holon, the system will enter the Recovery state that invokes the associated recovery operation implemented in RecoveryManager. If the recovery operation is successful, the system will jump back to the last operational state where the exception was occurred, otherwise the system will raise an alarm and then stop.

After demonstrating how to design functional holons, the holarchy design stage is completed. The following section will explain the application construction and system integration stages.



(a) Entire WIP Holon State Diagram



(b) WIP Object Life Cycle State Diagram

Figure 8. State Diagrams of WIP Holon

6. Application Construction and System Integration

As depicted in Fig. 2, the last two stages are application construction and system integration. Observing the top of Fig. 6, with the advantage of HMES Holarchy, it is obvious that applications can be constructed by invoking operations of associated holons. These holons will cooperate with one another by following the holarchy messages defined in Fig. 5. This meets the characteristics of holon and holarchy. In fact, the deployment diagram, holarchy messages, and a holarchy framework as shown in Figs. 1, 5, and 6, respectively, have been successfully implemented and running at the Factory Automation Laboratory of the Institute of Manufacturing Engineering, National Cheng Kung University, Tainan, Taiwan, Republic of China.

7. Comparisons among Legacy MES, Framework MES, and Holonic MES

The concepts and/or technologies of OOAD, component software, framework, holon, holarchy, security certification, and failure recovery have been taken into account for developing HMES. In this section, characteristic comparisons between Legacy MES, Framework MES, and Holonic MES are presented.

	Legacy MES	Framework MES	Holonic MES
Architecture	Centralization	Distributed OO	Holarchy
Open Interfaces	No	Yes	Yes
Modularization	Low	High	High
Interoperability	Low	High	High
Configurability	Low	High	High
Maintainability	Difficult	Easy	Easy
Security Certification	No	No	Yes
Failure Recovery	No	No	Yes

Table 1. Comparisons between Traditional MES, Framework MES, and Holonic MES

As indicated in Table 1, Legacy MES refers to the commercial products such as Promis, WorkStream, and Poseidon. Framework MES stands for Encore, SiView, and FACTORYWorks. Detailed comparisons are presented below.

7.1. Architecture

Concerning architecture, Legacy MES is a centralized system. All the computations and operations are executed in one large-scale mainframe computer. Framework MES belongs to distributed object-oriented systems that divide all the functions into individual various models. The computations and operations are also distributed into each model. In this way, Framework MES lowers the loading of each mainframe and increases the reliability of the system. Also, Framework MES avoids the malfunction of the entire system due to the breakdown of a single module. Holonic MES is designed with the concepts of holon and holarchy. It has the advantages of distributed object-oriented systems, and also the characteristics of intelligence, autonomy, coordination, and collaboration. Thus, Holonic MES's adaptability can meet the requirements and trends of future manufacturing systems.

7.2. Open Interfaces

When considering interfaces, Legacy MES is a closed system while Framework MES and Holonic MES are open systems. Systems with open interfaces have the advantage of being easy to cooperate and link with other related modules or systems.

7.3. Modularization

Modular design is very important to system software development. With component software, users can apply proper modules based on needs. This is beneficial both for design and maintenance. Both Framework MES and Holonic MES utilize modular design but Legacy MES does not.

7.4. Interoperability

A distributed object-oriented system usually has many functional modules that they need to interoperate with one another. Framework MES and Holonic MES are distributed object-oriented systems, therefore their interoperability with distributed modules is both essential and profuse.

7.5. Configurability

Configurability is important for a manufacturing system to deal with a dynamic, varying and rapidly changing environment. Framework MES and Holonic MES are easier to reconfigure than Legacy MES.

7.6. Maintainability

For Legacy MES, it is not easy to repair and maintain since it is a large-scale and centralized system. For Framework MES and Holonic MES, their maintenance is easier because they are distributed systems and each component of the systems can operate alone and be maintained separately.

7.7. Security Certification

The problem of security is becoming more and more serious. In Holonic MES, the ability of security certification is embedded in the design of the Generic Holon so that it is natural for all the functional holons to possess the capability of security certification.

7.8. Failure Recovery

Reliability is always the most important issue for automatic manufacturing systems. Once there is an exceptional condition that causes the entire production line to shutdown, the loss is beyond evaluation. As a result, a good set of MES needs a failure recovery mechanism so as to minimize the loss caused by occurrences of exceptional conditions. Among those three MES types, only Holonic MES incorporates the capability of failure recovery into the design.

8. Summary and Conclusions

Based on the characteristics of holon and holarchy and by applying distributed object-oriented techniques, this paper proposes a systematic approach for developing Holonic Manufacturing Execution Systems (HMES) with security-certification and failure-recovery considerations. The basic foundations required for developing HMES possessing characteristics of holon and holarchy are summarized. The HMES development procedure that consists of system analysis, holarchy design, application construction, and system integration

and testing stages are proposed. Among these stages, holarchy design is the most important and consists of seven steps: (a) constructing an abstract object model, (b) partitioning the application domain into components, (c) identifying generic functions among the components, (d) developing the Generic Holon, (e) defining holarchy messages, (f) defining the holarchy framework, and (g) designing functional holons. WIP Holon, as an example of a functional holon, is developed for demonstration purposes. Comparisons between Legacy MES, Framework MES, and Holonic MES are made. It reveals that this systematic approach provides a new concept for developing next generation manufacturing execution systems.

Acknowledgments

The authors would like to thank the National Science Council of the Republic of China for financially supporting this research under contracts No. NSC-89-2212-E006-094, NSC-90-2212-E006-026, and NSC-91-2212-E006-062.

9. References

Chang, C.-F. (2000). Development of scheduling holons and WIP holons, Master Thesis of the Institute of Manufacturing Engineering, National Cheng Kung University

Chen, D. J. & Chen, D. T. K. (1994). An experimental study of using reusable software design frameworks to achieve software reuse, Journal of Object Oriented Programming, pp. 56-67

Cheng, F.-T., Shen, E., Deng, J.-Y. & Nguyen, K. (1999). Development of a system framework for the Computer-Integrated Manufacturing Execution System: a Distributed Object-Oriented Approach, International Journal of Computer Integrated Manufacturing, Vol. 12(5), pp. 384-402

Cheng, F.-T., Yang, H.-C. & Huang, E. (2002). Development of an educational supply chain information system using object web technology. Journal of the Chinese Institute of Engineers, Vol. 25(6), pp. 735-752

Eriksson, H.-E. & Penker, M. (1998). UML Toolkit. New York: John Willy & Sons, Inc.

Gamma, E., Helm, R., Johnson, R. & Vlissides, J. (1995). Design Patterns: Elements of Reusable Object-Oriented Software, Addison-Wesley, Greenwich, CT

Hino, R. & Moriwaki, T. (1999). Decentralized Scheduling in Holonic Manufacturing Systems, Proceedings of the Second International Workshop on Intelligent Manufacturing Systems, Leuven, Belgium, pp. 41-47

Holonic Manufacturing System. HMS Introduction and Overview, http://hms.ifw.uni-hannover.de/

Huang, E., Cheng, F.-T. & Yang, H.-C. (1999). Development of a collaborative and event-driven supply chain information system using mobile object technology, in Proceedings of the 1999 IEEE International Conference on Robotics and Automation, Detroit, Michigan, U.S.A., pp. 1776-1781

Kadar, B., Monostori, L.& Szelke, E. (1998). An object-oriented framework for developing distributed manufacturing architectures, Journal of Intelligent Manufacturing, Vol. 9, pp. 73-179

Lin J.-Y. (2000). The development of holonic information coordination systems with security mechanism and error-recovery capability, Master Thesis of the Institute of Manufacturing Engineering, National Cheng Kung University

MacDonald, A. (1993). MESs help drive competitive gains in discrete industries, Instrumentation & Control Systems, pp. 69-72

Markus, A., Vancza, T. Kis & Monostori, L. (1996). A Market Approach to Holonic Manufacturing, Annals of the CIRP, Vol. 45(1), pp. 433-436

McGehee, J., Hebley, J. & Mahaffey, J. (1994). The MMST computer-integrated manufacturing system framework, IEEE Transactions on Semiconductor Manufacturing, Vol. 7(2), pp. 107-115

MESA International (1997). MESA International - White Paper Number 6: MES Explained: A High Level Vision, MESA International, Pittsburgh, PA

Mowbray, Z. (1995). The Essential CORBA: Systems Integration Using Distributed Objects, ISBN0-471-10611-9, New York: John Wiley & Sons

Nguyen, K. (1996). Flexible computer integrated manufacturing systems, in Proceedings of the SEMICON Taiwan 96 IC Seminar, Taipei, Taiwan, R.O.C., pp. 241-247

OMG (1998) The Common Object Request Broker: Architecture and Specification, Revision 2.2. Object Management Group

Orfali, R., Harkey, D. & Edwards, J. (1996). The Essential Distributed Objects Survival Guide, New York: John Willy & Sons

Ramos, C. (1996). A Holonic Approach for Task Scheduling in Manufacturing Syatems, Proceedings of the IEEE International Conference on Robotics and Automation, Minneapolis, U.S.A., pp. 2511-2516

Rumbaugh, J., Blaha, M., Premerlani, W., Eddy, F. & Lorensen, W. (1991). Object-Oriented Modeling and Design, Englewood Cliffs, NJ: Prentice-Hall

Samanich, N. J. (1993). Understand your requirements before choosing an MES, Manufacturing Systems, pp. 34-39

Scott, D. (1996). Comparative advantage through manufacturing execution systems, in Proceedings of the SEMICON Taiwan 96 IC Seminar, Taipei, Taiwan, R.O.C., pp. 227-236

SEMATECH (1998). Computer Integrated Manufacturing (CIM) Framework Specification 2.0 SEMATECH, Austin, Texas

Sparks, S., Benner, K. & Faris, C. (1996). Managing object-oriented framework reuse, IEEE Computer, pp. 52-61

Tonshoff, H. K., Winkler, M., and Aurich, J. C. (1994). Product modeling for holonic manufacturing systems, in Rensselaer's 4th International Conference on Computer Integrated Manufacturing and Automation Technology

Ulieru, M. (1997). Soft computing issues in the intelligent control of holonic manufacturing systems, in Proceedings of the 1997 IEEE Annual Meeting of the North American Fuzzy Information Proceeding Society, NAFIPS'97, Syracuse NY, USA, pp. 323-328

Valckenaers, P., Bonneville, F., Brussel, H. V., Bongaerts, L. & Wyns, J. (1994). Results of the holonic control system benchmark at KULeuven, in Rensselaer's 4th International Conference on Computer Integrated Manufacturing and Automation Technology

Van Leeuwen; E. H. & Norrie, D. (1997). Holons and holarchies, Manufacturing Engineerings, pp. 86-88

4

Bio-inspired approach for autonomous routing in FMS

T. Berger, Y. Sallez and C. Tahon

1. Introduction

Today's mass production strategies are unable to cope with the present needs of the manufacturing industry, which is constantly confronted with changing and increasingly complex product requirements as well as mounting pressure to decrease costs. To meet this challenge, Flexible Manufacturing Systems (FMS) must become more robust, scalable, reconfigurable, dynamic, adaptable and even more flexible. The challenge is even greater given that both production program modifications and resource failures can necessitate a partial or total reconfiguration of the FMS routing. Such reconfiguration is difficult to accomplish both because the problem is huge (combinatory explosion) and because each individual failed resource situation requires anticipating a different partial solution. In response to the FMS challenge, this chapter proposes autonomous routing of physical FMS flows.

Many international research projects focusing on the design of heterarchical (non-hierarchical) architectures have already been completed. Such architectures play a prominent role in the new control systems that dominate the field of FMS research (Duffie & Prabhu, 1996; Pujo & Kieffer, 2002). Our contribution to such decentralized control is inspired by the biological phenomenon called stigmergy, defined as insects' use of chemicals, called pheromones, to organize group activity. For example, foraging ants are known to lay down chemical trails, which are, in turn, followed by other ants who add their own odour to the trail, thus reinforcing it for future use. This stigmergic activity has an optimizing effect on performance. It allows creatures to communicate indirectly by sensing and modifying their local environment, and it is this communication that determines the creature's behaviour.

The next section of this chapter describes the concept of stigmergy. After presenting the different pheromonal characteristics upon which our bio-inspired approach is based, we describe several pheromone emulation mechanisms and some interesting insect-based methods that have been devised for various

manufacturing applications. Section 3 explains the key phases of our approach, focusing successively on the virtual pheromone-based progression of the entities and the updating of the virtual pheromones. Section 4 summarizes the results of a flexible assembly cell simulation that was conducted at the AIP-PRIMECA Center in Valenciennes. Following a description of the cell architecture and its main components and a brief presentation of the Netlogo simulation context, the qualitative and quantitative results are presented.
Section 5 outlines the advantages (adaptability, robustness) and potential disadvantages (stagnation, delay) of the stigmergic approach, and proposes solutions to the problems of stagnation and delay that can occur. Section 6 presents a real-life implementation of our approach, involving the instrumentation of moving entities and their environment. Finally, the last section of this chapter provides a brief overview of our prospective future research on self-organization.

2. An approach based on insect societies

2.1 The stigmergy concept

French entomologist Grassé (1959) introduced the term "stigmergy" to describe the mechanism by which termites coordinate their mound-building activities. In such activities, many individuals participate in a collective task, and the stimuli provided by the emerging structure are used to coordinate the individual actions. A similar mechanism is used by ants laying down pheromone trails between a food source and their nest.
Figure 1 portrays a classic experiment in entomology. Starting in the upper left-hand quadrant of the figure 1a, ants wander around their nest in search of food. Those finding food carry it back to the nest, simultaneously laying down a pheromone trail (figure 1b). Other ants, detecting the pheromones, follow the trails back toward the food. As more ants bring food to the nest, they each reinforce the chemical trail of the path they follow. Since pheromones tend to evaporate over time, the more attractive trails accumulate more pheromones and thus an advantage over the other trails (figure 1c). Over time, due to the natural reinforcement of the ants, only the shortest trail remains (figure 1d). As the experiment illustrates, this stigmergic process has a natural optimizing effect. (For more information about the history of stigmergy in the context of social insects, see (Theraulaz & Bonabeau, 1999))

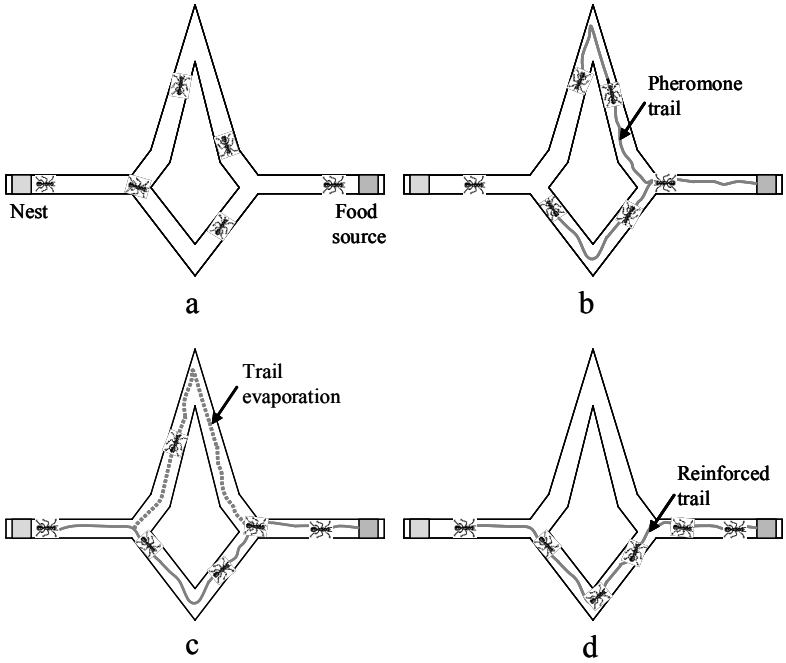


Figure 1. Stigmergy illustration

2.2 Basic operations

In the real world, three basic operations have been associated with the stigmergic process: information fusion, information removal and local information distribution. In the first, deposits from individual entities are aggregated to allow the easy fusion of information. In the second, pheromone evaporation over time is used to remove obsolete or inconsistent information. In the last, information is provided according to pheromone diffusion in the immediate (local) neighbourhood.
In all of these operations, the pheromone field has three main characteristics:

1. Independence: The sender of a pheromone message does not know the identity of the potential receiver and does not wait for any acknowledgement, which makes pheromone use very effective for communication within large populations of simple entities.
2. Local management: Because pheromone diffusion falls off rapidly with distance, pheromone interaction remains local, thus avoiding the need for centralized interaction management.
3. Dynamism: The continuous cycles of reinforcement and evaporation act respectively to integrate new information and to delete obsolete information.

2.3 Pheromone emulation

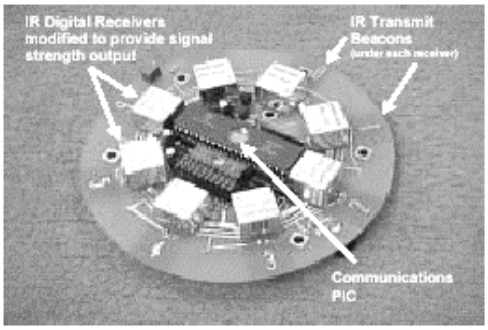
A variety of different approaches to pheromone emulation have been developed. The three most common types are described below:

Common memory In many studies (e.g., Dorigo & Colombetti, 1998), artificial ants cooperate via a common memory that serves the same purpose as the pheromones deposited by real ants. In this common memory, an artificial pheromone is created and accumulated (updated) via a learning mechanism during runtime.

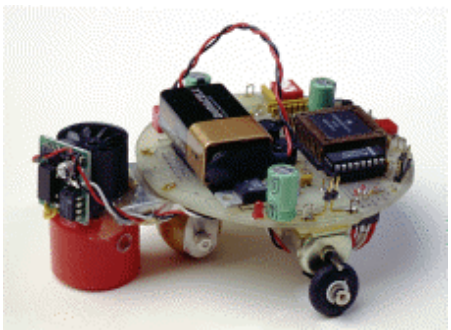
Embedded implementation Using signals transmitted from robot to robot, Payton et al. (2001) have implemented "virtual" pheromones that are sustained on board. Simple beacons and directional Ir sensors are mounted on the robots (figure 2a), with the virtual pheromones attached to the robots rather than laid down in the environment. This particularity is necessitated by the application: guidance of a rescue team in an unfamiliar building.

Direct environmental marking This kind of pheromone emulation can be performed in two ways: using real "pheromones" and using deposit tags.

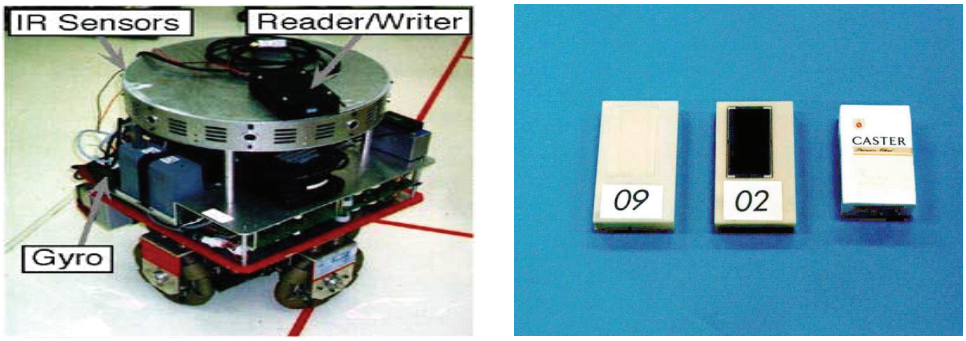
- Real pheromones have been used by researchers in Australia (Russell, 1995) (figure 2b) and in Israël (Wagner et al., 1995) to emulate ant behaviour by creating robots capable of laying down and detecting chemical trails (camphor or thinner).
- Deposit tags, such as the Intelligent Data Carrier (IDC), have been developed by Japanese researchers (Kurabayashi, 1999). The IDC system is composed of reader/writer units attached to mobile robots and tags that are carried and located by the robots. These tags are analogous to pheromones in that they store the information used to guide the robots (figure 2c).



a) Transceiver for virtual pheromones



b) Robot chemical sensing



c) Intelligent Data Carrier (robot and tags)

Figure 2. Examples of existing pheromone emulation hardware

2.4 FMS applications

The first experiments related to the industrial use of stigmergy were conducted in the early 1980s by Deneubourg et al. (1983), who simulated "ant-like robots". Since then, many researchers (e.g., Ferber, 1995; Arkin, 1998; Dorigo & Colombetti, 1998) have applied this concept when studying robot collectives and working to solve optimization problems (e.g., Travelling Salesman Problems, Network Routing for telecommunications and the Internet). Based on the ant foraging analogy, Dorigo et al. (1999) developed the Ant Colony Optimization (ACO) metaheuristic, a population-based approach to solving combinatorial optimization problems. The basic idea behind ACO is that a large number of simple artificial entities can be used to build good solutions to hard combinatorial optimization problems via low-level communications. The ACO approach can be applied to almost any scheduling problem, such as job shop scheduling and vehicle routing, for example.

Researchers have also applied the stigmergy concept to specific situations in manufacturing control systems:

- Parunak et al. (2001) emphasize the importance of the environment in agent systems, in which information flows through the environment complement classic message-based communications between the agents. In this study, the environment is computational, and agents moving over a graph are used to study manufacturing company supply networks. The

authors focus on the dynamics that emerge from the interactions in multi-agent systems; these dynamics are analyzed using methods inspired by statistical mechanics.

- Brückner (2000) applies the stigmergy concept to manufacturing control, and his application is supported by an agent-system approach. He presents an extensive set of guidelines that can be used to design synthetic ecosystems. In this study, different types of agents are used to model the various elements (e.g., resources, part flow and control units) involved in routing a car body through a Mercedes Benz paint shop. Brückner's work is based on the PROSA reference architecture (Wyns, 1999).

- Peeters et al. (1999) and Hadeli et al. (2004) both propose a pheromone-based control algorithm with a bottom-up design. Like Brückner (see above), Peeters et al. based their work on the PROSA reference architecture (Wyns, 1999). (Those interested should consult the Mascada-WP4-Report (1999) for a description of both the agents and the pheromone life-cycle in the routing of a car body through a paint shop, this one at Daimler-Chrysler). Hadeli et al. (2004) emulate a simple flexible manufacturing system characterized by dynamic order arrival, probabilistic processing time, and several disturbances (e.g., machine breakdowns), with the objective of evaluating the possibility of creating short-term forecasts based on agent intentions.

3. Description of our bio-inspired approach

Our approach to FMS is based on the behaviour of biological systems, such as ant colonies (Deneubourg et al., 1983; Di Caro & Dorigo, 1998). An FMS can be seen as a network of nodes that are interconnected by uni/bi-directional paths on which mobile product entities navigate. Each of these entities must obtain a variety of services from resources (service stations) located on the nodes. These autonomous entities move from one node to another until they reach a destination node, where the desired service can be obtained. One after another, the entities choose a destination and the appropriate path to reach it, using the data stored on each node. Typical entity behaviour is portrayed in schematic form below (figure 3).

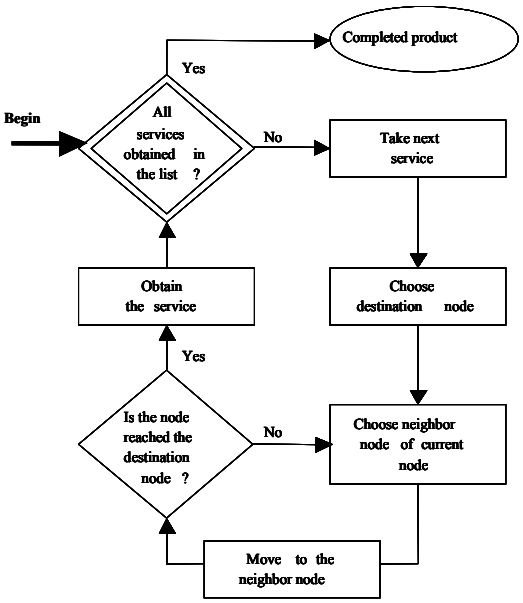


Figure 3. Typical entity behaviour

3.1 Entity progression based on virtual pheromones

In analogy with biological systems, P coefficients characterize the pheromone rate on different trails/paths. Based on the information stored on the current node, which includes the coefficients associated with each destination node, the entity chooses the best route from its current location to the destination node.

Let P_{dn}, on node n_k, represent the preference for the neighbour node n_n that comes with a short traversal time to move from the current node n_k to destination n_d via node n_n, which must belong to the n_k neighbour node set.

In the following notation:

$$P_{dn} \in [0;1],\ P_{dj} > P_{dp} \tag{1}$$

implies n_j allows the entity to reach n_d more rapidly than n_p, where n_j and n_p are neighbours of n_k.

$$Vn_k \Leftrightarrow \{n_k \text{ neighbour}\};\ n_n \in Vn_k \tag{2}$$

After standardization:

$$\sum_i P_{di} = 1; i \in Vn_k \quad (3)$$

The best neighbour is n_c, such that:

$$P_{dc} = \max(P_{di}); i / n_i \in Vnk \quad (4)$$

Still, since both chance and diversity are important adaptation mechanisms in natural biological systems, the choice of n_c is not totally deterministic. In fact, a sort of "roulette wheel", based on the same principles as the one found in a casino, is used to choose the next neighbour. The different P_{dn} coefficients serve as weights to build a 100-sector roulette wheel. Higher P_{di} coefficients increase the likelihood that neighbour n_i will be chosen. To avoid one neighbour becoming predominant, the minimal P_{di} coefficient value is 0.01 (at least one chance in 100). This "roulette wheel" principle is presented in figure 4.

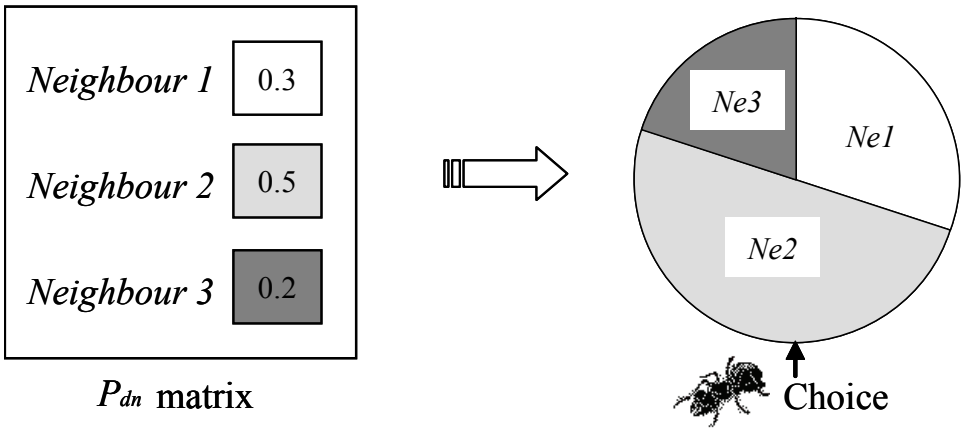


Figure 4. Roulette wheel illustration

3.2 Virtual pheromone updating

While moving, each entity stores data in its embedded memory. The memory records the entity's path through the nodes n_k, including crossing time. Like real ants, which lay down a pheromone trail when returning to the nest, every

time an entity reaches a destination node, a fictitious entity retraces its path virtually, using the information contained in the embedded memory to update coefficients at each node that was crossed.

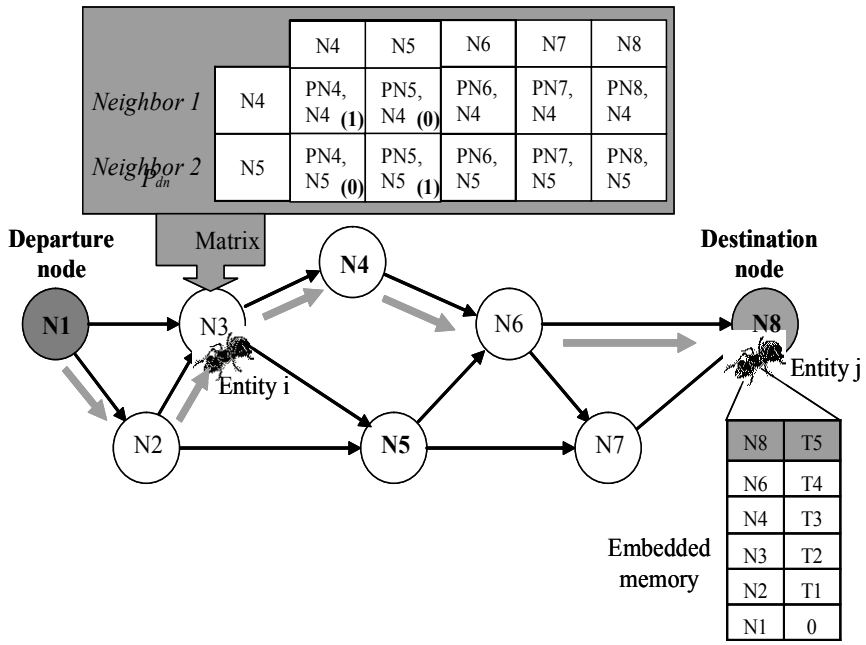


Figure 5. The laying down of pheromone information

Figure 5 shows an example of the embedded memory for an entity j, which has just arrived at a destination node. Each node n_k includes:

- a P_{dn} matrix, whose columns provide all possible destinations n_d and whose rows show all existing neighbours Vn_k; and
- a matrix that contains μ_{dk}, the mean time needed to go from n_k to n_d for all possible destinations (and if used, the standard deviation σ_{dk}).

T_{dk} is the time span needed to go from n_k to n_d. When the fictitious entity arrives at node n_k, T_{dk} and the mean μ_{dk} of the previous T_{dk} are compared. The P_{dn} matrix is then updated by incrementing the coefficient P_{dc} (the possibility of choosing neighbour n_c when the destination is n_d) and decrementing other coefficients P_{do}.

A reinforcement value r (in our case r = 0.2) is used as follows:

$$P_{dc} \Leftarrow P_{dc} + r * (1 - P_{dc}) \quad (5)$$

Coefficients P_{do} for the destination n_d of the other neighbours n_o are negatively reinforced through a process of normalization.

$$P_{do} \Leftarrow P_{do} - r * P_{do}, \; n_o \in Vn_k \text{ with } n_o \neq n_c \quad (6)$$

For a more detailed discussion about how r is chosen, please refer to the literature on "reinforcement learning", specifically the books by Dorigo & Stützle (2004) and Sutton & Barto (1998).
A more sophisticated approach to updating the *P* coefficients could integrate the standard deviation σ_{dk} into the updating process. In such an approach, the comparison of T_{dk} and μ_{dk} would be valid only if μ_{dk} were sufficiently stable in terms of the σ_{dk} of the previous T_{dk}. Three situations are possible:

- given a stable μ_{dk} value and $T_{dk} < \mu_{dk}$, increasing P_{dc} would reinforce n_c,
- given a stable μ_{dk} value and $T_{dk} \geq \mu_{dk}$, decreasing P_{dc} would result in n_c being ignored,
- given an unstable μ_{dk} value, adjustments would need to be made to stabilize it.

4 An FMS routing simulation

4.1 The AIP FMS cell

A flexible assembly cell was simulated at the Valenciennes AIP-PRIMECA Center. This cell was composed of seven workstations *Wi* placed around a flexible conveyor system, which insured a flexible flow of pallets to each workstation. The conveyor system is based on Montech's Montrac system (Montech, 2005). Montrac is a monorail transport system that uses self-propelled shuttles to transport materials on tracks (figure 6). Each shuttle is individually controlled and equipped with a collision-avoidance optical sensor.

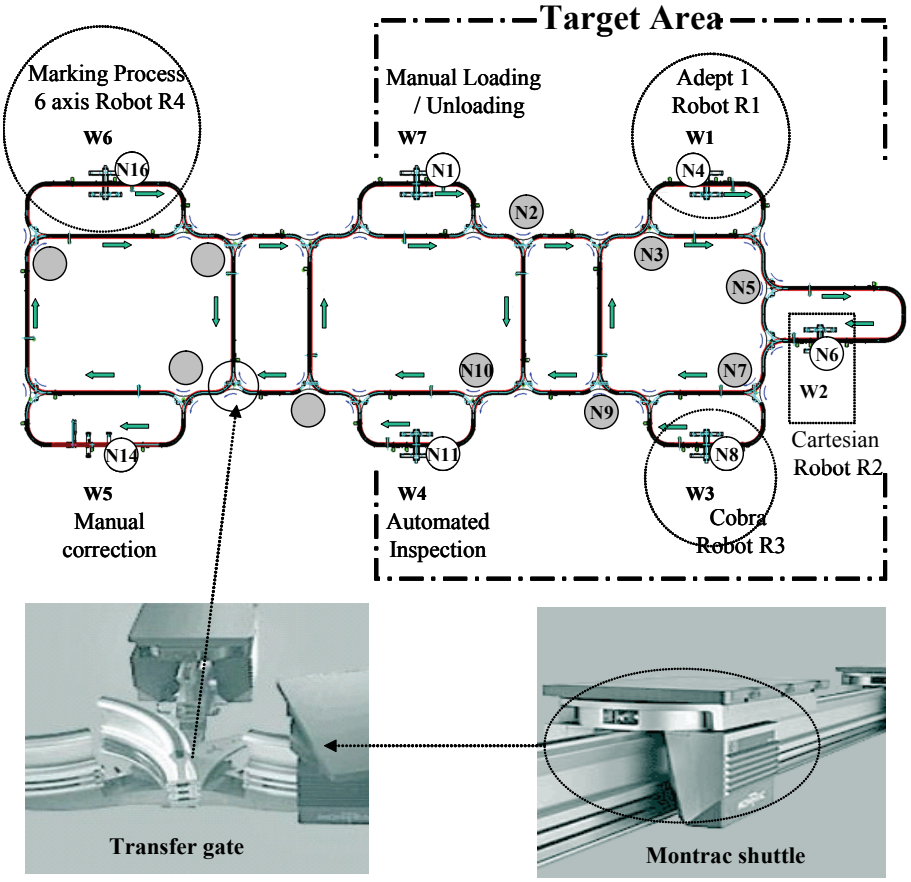


Figure 6. Schematic view of the flexible cell with inset detailed images of the transfer gate and shuttle

Eleven of the major nodes in the cell shown in figure 6 were considered in this simulation:

- The nodes/stations in white (N1, N4, N6, N8 and N11) are possible destination nodes where services can be obtained.
- The nodes/stations in gray (N2, N3, N5, N7, N9 and N10) are divergent transfer gates, from which shuttles can obtain the information available about the destinations in order to make their routing decisions.

The other transfer gates (neither white nor grey) appearing in figure 6 were not taken into account in the simulation. They were only used to connect convergent tracks when no routing decisions were required.

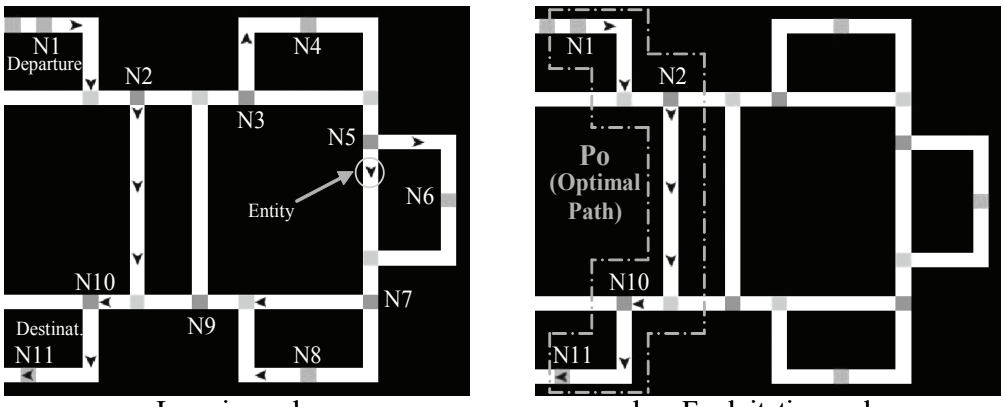
4.2 The context of the simulation

Given the specifications of the cell described in the previous section, we chose the NetLogo platform (NetLogo, 2006) to simulate our approach. The NetLogo platform offers an agent-based parallel modelling and simulation environment. Mainly used to simulate natural and social phenomena, it is particularly well suited to complex system modelling. With Netlogo, each entity can be described as an independent agent interacting with its environment. All agents operate in parallel on a grid of patches (cellular world), and each agent can read and modify some of the attributes linked to the patches in its proximity. The behavioural rules defined for the agents make it possible to describe agent-environment interaction, which is very important when simulating the stigmergic process.

4.3 Results

4.3.1 Qualitative results

The preliminary results of the first simulation clearly illustrate the overall adaptability of the system when faced with small disturbances. Figure 7 shows the right part of the cell as depicted with a NetLogo simulator. In the figure 7a, N1 and N11 are the departure and destination nodes, respectively; and the location of the entities is indicated with small arrows.



a. « Learning» phase b. « Exploitation » phase

Figure 7. Netlogo simulation results

The simulation can be divided in two phases: the "learning" phase and the "exploitation" phase.

- The learning phase is shown in figure 7a. First, the entities traveled randomly along the different paths of the network, and all the *P* coefficients were equal on the different nodes.
- The exploitation phase began when the *P* coefficients were stable, which happened when all the paths to the different destination nodes had been sufficiently explored (figure 7b).

The pivotal point, at which the learning phase became the exploitation phase, occurred when, in accordance with the principle of stigmergy, the optimal path Po (N1-N2-N10-N11) emerged through reinforcement (figure 7b). However, the stigmergic process is not static. Faced with disturbances (e.g., flow reduction) that affected the fluidity of the path N2-N10, the entities travelling on this path began to perform poorly, and the appeal of this path decreased. The decreased appeal of path Po increased the appeal of path NPo (N1-N2-N3-N5-N7-N9-N10-N11), even though this path was originally non-optimal (figure 8). This dynamic response is a classic display of the natural routing reconfiguration capacity of the stigmergic approach.

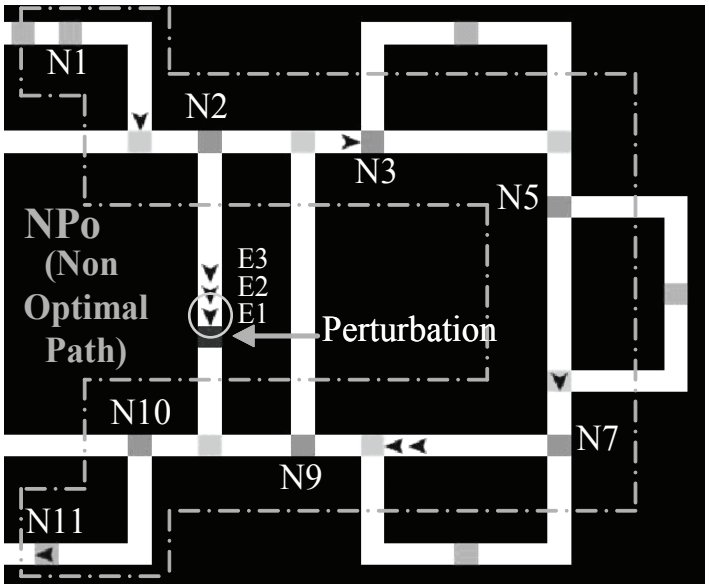


Figure 8. Reconfiguration capacity of the stigmergic approach

4.3.2 Quantitative results for the target area

In this second simulation, the departure and destination nodes were changed to N1 and N8, respectively. Figures 9, 10 and 11 show the evolution of the coefficients, which illustrates the advantage of passing though the current node's neighbours to reach the destination node N8.

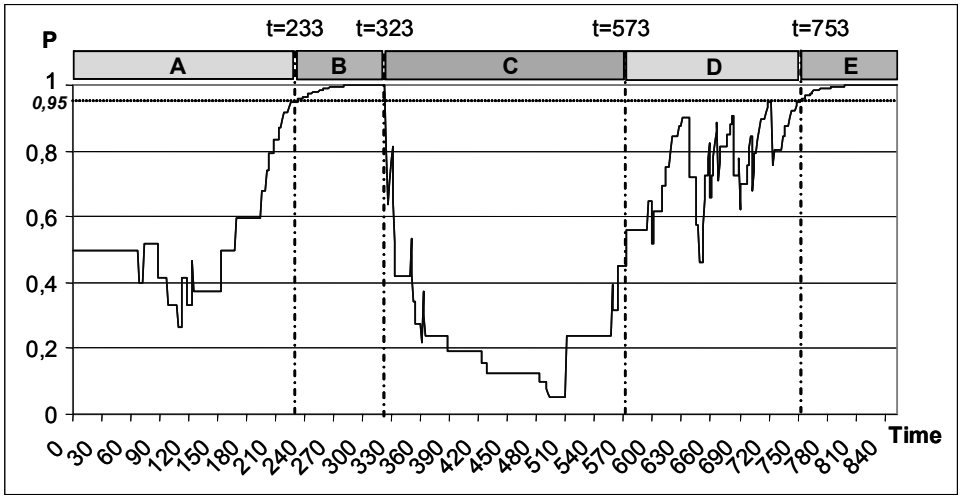


Figure 9. Evolution of coefficient *P* on N2, showing the advantage of going through N3 to reach N8.

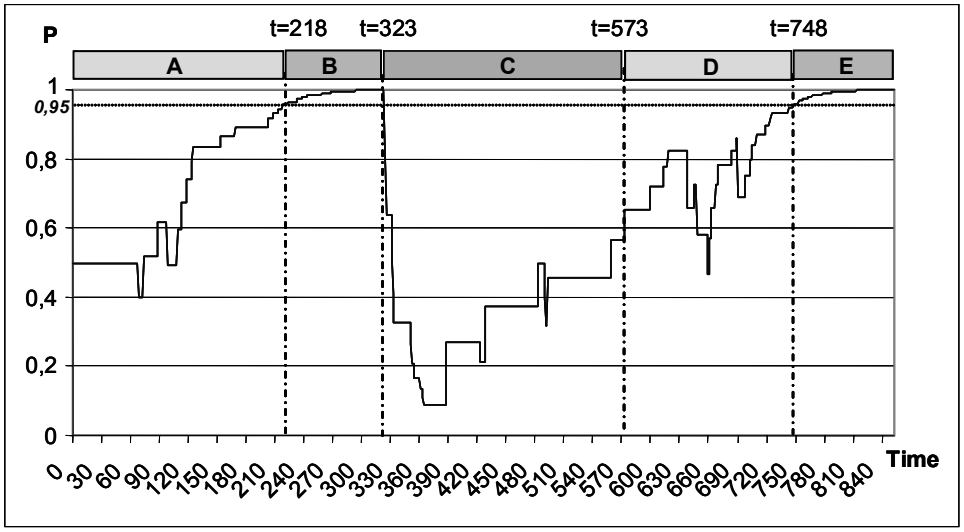


Figure 10. Evolution of coefficient *P* on N3, showing the advantage of going through N5 to reach N8.

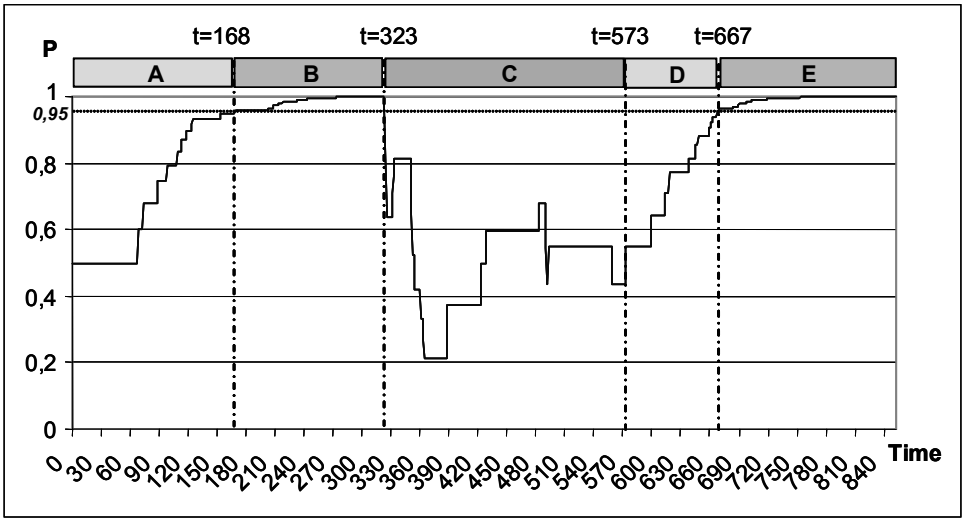


Figure 11. Evolution of coefficient *P* on N5, showing the advantage of going through N7 to reach N8.

The curves depicted in all three figures can be divided into 5 zones:

- Zone A depicts the initial operation mode, in which the system goes through a learning phase.
- Zone B depicts the "exploitation" phase, during which the system can be considered stable.
- Zone C depicts the system during a period of perturbation.
- Zone D depicts the period after the perturbation, in which the system goes through a second learning period in order to readapt/reconfigure itself.
- Zone E depicts the phase that comes about when the system is once again stable.

The system in zone A: During the initialization stage, the coefficients of both destinations are set at 0.5 until they can be updated after the entities have reached the destination node (N8). Starting at node N1, two paths are possible. In figure 9, for instance, looking at time t=60, a change in the coefficient *P* can already be observed at N2. This underlines the advantage of passing through N3 to reach N8, because it means that the first entity reached its destination at t=60, and the entity's embedded memory provided the feedback needed to update the coefficients.

The coefficient *P* characterizes the "optimal" path, deduced from the learning phase in which one of the possible paths passed through N2, N3, N5, N7 and N8. Consequently, there is an increase in coefficients over time. The permanent operating regime is reached at a coefficient value of 0.95. As shown in figure 11, for instance, the closer the entity is to the destination, the more quickly the coefficient reaches the value 0.95. Conversely, the farther away the entity is from the destination, the more slowly the value 0.95 is reached (figures 9 and 10). Figures 9 and 10 also show the values of the coefficients presented tend to oscillate. Depending on the topology of the network, one possible explanation for this oscillation is that the further the entity is from the destination, the larger the number of possible paths. The wider range of choices leads to more attempts to reach N8 along less efficient paths, which hinders the updating process and thus decreases the coefficient's value. This inefficiency is corrected by those entities that reach N8 more quickly, thus reinforcing the shorter paths by increasing the value of the coefficient.

The system in zone B:
Given that all the coefficients for the path from N2 to N8 have reached the value of 0.95, the learning process is presumably over. Therefore, all entities will tend to choose the same path until disturbances occur.

The system in zone C:
The smooth progression on the path from N2 to N8 (passing through N3, N5 and N7) is disrupted by an unexpected flow reduction in the path from N5 to N7, just after N5. In reaction, the system adapts itself, "forgetting" the previously optimal path by decreasing the path's coefficients.

The system in zone D:
The flow reduction has now disappeared, and following a second learning period, the path to N8 through N2, N3, N5 and N7 once again becomes the optimal choice. The closer the entity comes to the destination node, the more quickly the path coefficients stabilize themselves around the value of 0.95. In this way, the coefficients that demonstrate this path's advantage are strengthened.

The system in zone E: Entities can once again use the newly optimal path.

4.3.3 Quantitative results overall

Table 1 shows the results for all destination nodes (N4, N6, N8, N11, N14 and N16) in the cell, given a departure node N1 (see figure 6). These results are based on the average for 100 simulations. The information is presented in 5 columns: Dest Node, N entities, N loops, N straight, and Time.

- "Dest. Node" shows the identifier of the destination node.
- "N entities" gives the average number of entities arriving at destination. These entities update the *P* coefficients and thus help to determine the optimal path.
- "N loops" provides the average number of entities that have travelled across a loop before reaching the destination node.
- "N straight" shows the average number of entities that reached the destination node via the optimal path.
- -"Time" indicates the delay (in units of simulated time) needed for the optimal path to emerge. This delay represents the moment when all the coefficients on the optimal path are greater than 0.95.

Dest. Node	N. entities	N. loops	N. straight	Time
N4	12.00	0.87	11.13	111.61
N6	23.55	4.16	14.93	194.79
N8	38.05	8.70	18.58	278.87
N11	12.01	0.01	10.07	120.27
N14	20.16	3.93	8.71	208.55
N16	38.11	12.12	11.60	325.54

Table 1. Summary of the overall results

Logically, the number of entities and the time needed for the optimal path to emerge increases with the distance between the departure and destination nodes. The main differences in the simulation results are due to the presence of loops in the network (see figure 6). These loops delay the convergence of certain coefficients. Using a 800 MHz PC, the CPU time needed to determine the optimal path ranges from 1 to 6 seconds, depending the destination.

5 Advantages and disadvantages of the stigmergic approach

The results described above demonstrate the advantages of our bio-inspired approach. First, the entities are able to determine the best path from the departure node to the destination node without any centralized control. Second, they are able to surmount disturbances, by seeking out new paths that bypass the disturbance but still lead to the desired destination. These capacities indicate that the approach offers a good level of adaptability and robustness in the face of environmental fluctuations. These qualities are common in classic biological systems based on stigmergic principles (e.g., ant colonies, termite mounds).

However, there are two potential problems, which can be disadvantageous: *stagnation*, once the routing network reaches its convergence point, and the *delays* caused by the traffic jams.

Stagnation occurs when the routing network reaches its convergence point. The optimal path Po is now chosen by all ants, which recursively reinforces the preference for Po. Unfortunately, this reinforcement has two negative consequences, in addition to the positive ones described above: 1) the optimal path becomes congested, while 2) the likelihood that another path will be chosen to avoid the congestion is significantly decreased.
During the simulation, frequent disturbances on the path N2-N10 provoked a decrease in the corresponding coefficient *P* at node N2, resulting in a preference for the path NPo (see figure 8). Little by little, the optimal path Po is "forgotten". However, this problem can be resolved. In their survey of ACO systems, Sim & Sun (2003) describe several mechanisms that can be used:

- Evaporation: This mechanism has been widely used in many studies (e.g., Parunak et al., 2001; Brückner, 2000). Evaporation prevents pheromones from concentrating on certain paths and allows new routing solutions to be explored. This approach is inspired by real ant colonies.
- Aging: With the aging mechanism, older entities deposit less and less of the pheromone as they travel from node to node. This solution, often used in conjunction with evaporation, is based on the fact that "old" entities are less successful at finding optimal paths.
- Limiting: This mechanism limits the quantity of the pheromone that can be amassed on any one path. This upper limit prevents a specific path from becoming "too dominant".

The third mechanism was chosen for our approach, allowing "chance" to play a bigger role in path exploration.

Delays were quite typical in the FMS application simulated in this experiment. When the optimal path Po emerged, it became dominant, and all entities followed this path to N11. If an entity E1 got stuck on the path N2-N10, the entities that followed were also blocked (figure 12). Unfortunately, this kind of disturbance can not be detected immediately because the P coefficients are only updated after entity E1 arrives at its destination, node N11. In nature, real ants bypass the obstacle and continue on the path. However, in our simulation, the entities (shuttles) remained blocked on the tracks and could not bypass the obstacle.
To solve this problem, we introduced a repulsive field, commonly used in mobile robotics (Balch & Parker, 2002). In the real implementation, this repulsive field was created by messages exchanged between nodes. (See section 6.3, below, for more details.) With the use of a repulsive field, the P coefficient corresponding to the blocked path is temporarily ignored, and the other shuttles take a detour that brings them to the destination node. After the problem causing the obstacle has been resolved, the "old" P coefficient is restored for this path.

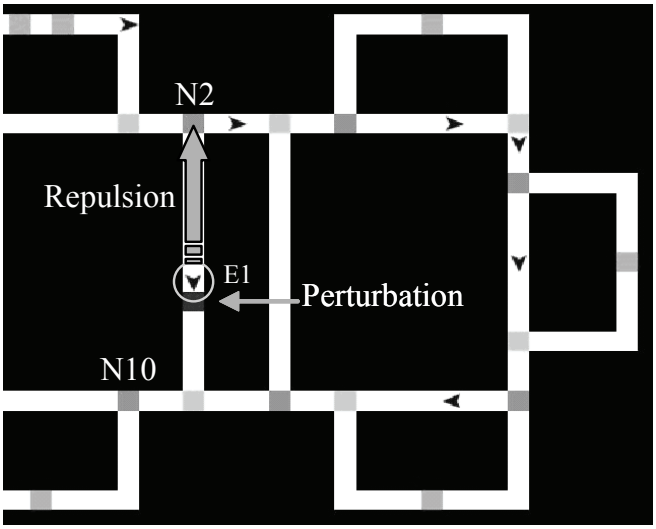


Figure 12. Repulsive effect

6. Details for a real implementation

6.1 The implementation

Following the simulation phase, a real implementation was designed and is currently being tested.

As explained in section 3, a node is assigned to each location where a decision must be made. At the current node n_k, a shuttle queries the node to gain read access to the P coefficients, which are stored in node n_k. To apply our approach, two types of equipment must be installed (see figure 13):

- *Node instrumentation*, including a "gate controller" which works to oversee the transfer gate and to help avoid collisions, a P_{dn} matrix (see figure 5), another matrix containing the mean μ_{dk}, two data communication systems (both Ethernet and Ir), and a data processing system; and
- *Shuttle instrumentation*, including an Ir communication system and a data processing system.

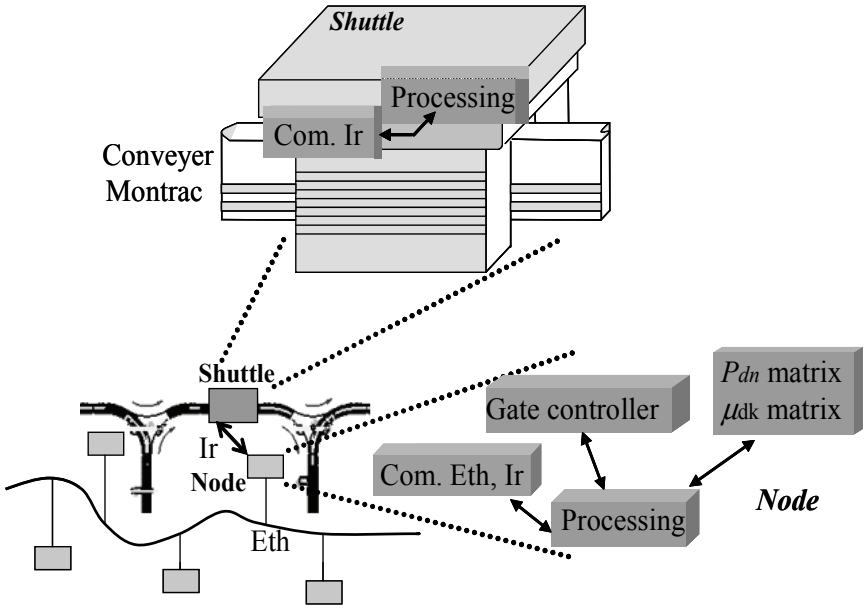


Figure 13. Equipment necessary to implement our approach

6.2 Shuttle progression

In order to identify its position, the shuttle obtains the ID node at each current node n_k through Ir communication. As required in our approach, the shuttle also obtains the *P* coefficients needed to allow it to choose the best neighbour node. These shuttle queries are processed by a microcontroller (Beck-IPC) located on the node. Then, via a microcontroller (Microchip Pic18F) embedded inside the shuttle, the shuttle processes the *P* coefficients in order to choose the best neighbour. Once the processing phase is finished, the shuttle asks to be routed towards the chosen neighbour node. The concrete local routing is performed by the "gate controller" on the node.

6.3 P updating and the implementation of the repulsive field

When a shuttle arrives at the destination node, it uploads its embedded memory. The node controller sends the data to the nodes that the shuttle passed through, and each node updates its P_{dn} matrix and μ_{dk} matrix. This inter-node communication is done through an Ethernet link.
To implement a repulsive field, the node from which a shuttle has just departed launches a timer. When the information from the shuttle arriving at the chosen neighbour is sent back to the previous node via the Ethernet link, this timer is reset. When the time delay exceeds the timeout limit, the previous node considers that the shuttle is blocked on the path to the chosen neighbour, and the *P* coefficient corresponding to this blocked path is temporarily ignored.

7. Perspectives for future research

To solve the delay problem in the first approach, we introduced a solution based on repulsion. A more general and interesting solution might be to couple the stigmergic approach with more reactive methods, such as those based on potential-fields. Potential field-based methods are inspired by electrostatics. They are widely used in swarm robotics to navigate mobile robots in restricted environments (Arkin, 1998; Balch & Parker, 2002). They combine attractive forces (goals) and repulsive forces (obstacles) to guide reactive robots. In fact, all entities (goals, obstacles, others robots in proximity) are considered to be attractive or repulsive forces, with each mobile robot moving under the vector sum of all the previous forces. In classic field-based approaches, the repulsive or attractive forces act only inside an area defined by fixed distances.

In their work on bionic manufacturing systems, Vaario and Ueda (1998) have also used local attraction fields to direct transporters carrying jobs to particular resources, resulting in the emergence of a dynamic schedule of the interactions between the different entities. In the context of FMS routing, the idea of bionic systems would be an interesting avenue to explore.

As seen in section 4, our simulation is bound by the topology of the AIP FMS cell. Shuttles must obligatorily follow the tracks and can be jammed by any failure of the conveyor system. To examine the possible intersection of stigmergy and potential field-based methods, we looked at an experiment similar to the line-less production system, introduced by Ueda et al. (2001).

In a line-less production system, all production entities (products, resources) can move freely on the production floor. Self-organization concepts are used to deal with fluctuations (diversity of the production demand, machine failures). The authors' application context is a car chassis welding process. Car chassis are mounted on automated guided vehicles (agvs) capable of moving around the shop floor. Mobile welding robots move towards the agvs according to the perceived attraction field. Figure 14 shows an agv with 6 sub-parts and 2 types of welding robots (A and B). Depending on the product requirements, different attraction fields can be generated. The results obtained by Ueda et al. suggest that this approach is appropriate for high-variety production and can generate high productivity (Ueda et al., 2001).

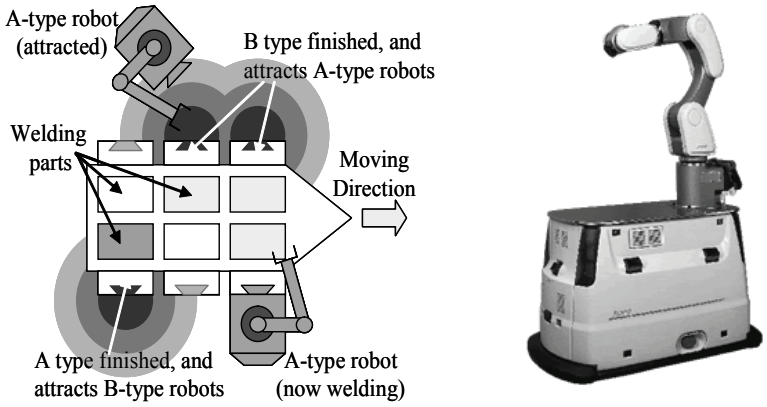


Figure 14. Agv attracting robots (Ueda et al., 2001) and mobile robot

These two complementary approaches—stigmergy and potential field-based mechanism—are dynamic local methods that don't require centralized management; both characteristics are essential when studying emergent phenomena in self-organized processes.

8. Conclusion

The stigmergic approach is an original answer to routing problems in FMS. Applied in many research fields (e.g., robotics, network routing), stigmergy offers robustness and adaptability in the face of environmental variations. The results of our simulation in the NetLogo environment highlight the robustness and adaptability of the approach, and, in fact, these qualities stem directly from our use of virtual pheromones. Although the approach is not perfect, solutions exist to remedy the problems of stagnation and delay. For example, when the stigmergic approach fails to take disturbances into account in real-time (e.g., a blocked shuttle), this lack of reactivity can be resolved by adopting complementary mechanisms, such as a repulsive field.

Our main objective is to develop adequate ways (e.g., stigmergic, field-based or hormone-based mechanisms) to support the interactions between entities in self-organized processes, such as bionic manufacturing systems. The implementation of the bio-inspired method described here is a first step towards the development of more intelligent flows in FMS.

9. References

Arkin, R.C. (1998). *Behaviour-based robotics*, The MIT Press

Balch, T. & Parker, L.E. (2002). *Robot teams: From Diversity to Polymorphism*, Natick, Massachusetts: A K Peters Ltd.

Brückner, S. (2000). Return from the Ant synthetic ecosystems for manufacturing control, Thesis Humboldt-University of Berlin, june 2000

Deneubourg, J.L.; Pasteels, J.M. & Verhaeghe, J.C. (1983). Probabilistic Behaviour in Ants : a Strategy of Errors ? *Journal of Theoretical Biology*, Vol. 105, pp. 259-271, 1983

Di Caro, G. & Dorigo, M. (1998). AntNet : Distributed Stigmergic Control for Communications Networks. *Journal of Intelligence Research*, 9, pp 317-365, 1998

Dorigo, M. & Colombetti, M. (1998). *Robot shaping: An experiment in Behaviour Engineering*, The MIT Press

Dorigo, M.; Di Caro, G. & Gambardella, L.M. (1999). Ant algorithms for discrete optimization. *Artificial Life,* Vol. 5, No 2, pp. 137-172, 1999

Dorigo, M. & Stützle, T. (2004). *Ant Colony optimization*, The MIT Press

Duffie, N.A. & Prabhu, V. (1996). Heterarchical control of highly distributed manufacturing systems, *International Journal of Computer Integrated Manufacturing*, Vol. 9, No. 4, pp. 270-281, 1996

Ferber, J. (1995). *Les systèmes multi-agents*, InterEditions

Grassé, P.P. (1959). La reconstruction du nid et les coordination inter-individuelles chez Bellicositermes natalensis et Cubitermes sp. La théorie de la stigmergie : essai d'interprétation du comportement des termites constructeurs, Insectes Sociaux, Vol. 6, pp. 41-83, 1959

Hadeli, T.; Valckenaers, P.,Kollingbaum, M. & Van Brussel, H. (2004).Multi-agent coordination and control using stigmergy. *Computers in industry*, Elsevier Science, Vol. 53, pp. 75-96, 2004

Kurabayashi, D. (1999). Development of an Intelligent Data Carrier (IDC) system and its applications, *Proceedings of 4th Int. Symp. on Artificial Life and Robotics*, pp. 34-39, 1999

Mascada-WP4 (1999). WP4 Report : ACA (Autonomous Co-operating Agents) Framework for Manufacturing Control Systems, 1999

Montech (2005). website: http://www.montech.ch/montrac/content/

NetLogo (2006). website: http://ccl.northwestern.edu/netlogo/

Parunak, H.V.D.; Brueckner, S. & Sauter, J. (2001). ERIM's Approach to Fine-Grained Agents, *Proceedings of the NASA/JPL Workshop on Radical Agent Concepts (WRAC'2001)*, Greenbelt, MD, Sept. 19-21, 2001

Payton, D.; Daily, M., Estkowski, R., Howard, M. & Lee, C. (2001). Pheromone Robotics. *Autonomous Robots*, Vol. 11, No. 3, pp. 319-324, Kluwer Academic Publishers, Norwell (MA)

Peeters, P.; Van Brussel, H., Valckenaers, P., Wyns, J., Bongaerts, L., Heikkilä, T. & Kollingbaum, M. (1999). Pheromone based emergent shop floor control system for flexible flow shops, *Proceedings of International Workshop on Emergent Synthesis (IWES'99)*, Kobe, Japan, Dec. 6-7, 1999

Pujo, P. & Kieffer, J.P. (2002). *Méthodes du pilotage des systèmes de production*, Traité IC2, série Productique, Hermes Lavoisier

Russell, R.A. (1995). Laying and sensing odor markings as a strategy for assisting mobile robot navigation tasks. IEEE Robotics and Automation Magazine, pp. 3-9

Sim, K.W. & Sun, W.H. (2003). Ant Colony Optimization for Routing and Load-Balancing: Survey and New Directions. *IEEE Trans. On Systems, Man and Cybernetics,* Vol. 3, No. 5, pp. 560-572, sept. 2003.

Sutton, R.S. & Barto, A.G. (1998). *Reinforcement Learning*, The MIT Press.

Theraulaz, G. & Bonabeau, E. (1999). A brief history of stigmergy. *Journal of Artificial Life*, Vol. 5, No 2, pp. 97-116, 1999

Ueda, K.; Hatono, I., Fujii, N. & Vaario, J. (2001). Line-Less Production System Using Self-Organization: A Case Study for BMS. *Annuals of CIRP*, Vol. 50, No. 1, pp. 319-322, 2001

Vaario, J. & Ueda, K. (1998). An emergent modeling method for dynamic scheduling. *Journal of Intelligence Manufacturing*, Vol. 9, pp. 129-140.

Wagner, I.A.; Lindenbaum, M. & Bruckstein, A.M. (1995). Distributed Covering by Ant-Robots Using Evaporating Traces, *IEEE Transactions on Robotics and Automation*, Vol.15, No. 5, pp. 918-933, 1995

Wyns, J. (1999). Reference architecture for Holonic Manufacturing Systems – the key to support evolution and reconfiguration, Ph.D. Thesis K.U. Leuven, 1999

5

Modular Machining Line Design and Reconfiguration: Some Optimization Methods

S. Belmokhtar, A.I. Bratcu and A. Dolgui

1. Introduction

1.1 Machining lines

Automated flow-oriented machining lines are typically encountered in the mechanical industry (Groover, 1987; Hitomi, 1996; Dashchenko, 2003). They are also called transfer (or paced) lines, being preferred mainly for the mass production, as they increase the production rate and minimize the cost of machining parts (Hutchinson, 1976). They consist of a linear sequence of multi-spindle machines (workstations), without buffers in between, arranged along a conveyor belt (transfer system). Each workstation is equipped with several spindle heads, each of these latter being composed of several tools. Each tool executes one or several (for the case of a combined cutting tool) operations. A *block* of operations is defined by the set of the operations executed simultaneously by one spindle head. When all blocks of a workstation have been accomplished, the workstation cycle time is terminated. The cycle time of the line is the longest workstation cycle time; its inverse is the line's production rate.

A machining line designed to produce a single product type is called *dedicated* line; its optimal structure, once found and implemented, is intended for a long exploitation time and needs high investments. The main drawback of such a system is its rigid structure which does not permit any conversion in case of change in product type. Thus, to react to changes effectively an alternative is to design the system from the outset for all the product types intended to be produced. Research has been conducted to an integrated approach of transfer lines design in the context of *flexibility* (Zhang *et al.*, 2002). This is the most important aspect which characterizes the potential of a system for reconfiguration (Koren *et al.*, 1999). The chapter deals with the designing of modular *reconfigurable* transfer lines, where a set of standard spindle heads are used to

produce a *family of similar products*. The objective is to minimise the total investment cost and implicitly minimise the time for reconfiguration. For simplicity, in this chapter, "spindle heads" and "blocks" will have here the same meaning.

Our interest focuses on the configuration/reconfiguration of modular lines. The modularity brought many advantageous: maintenance and overhaul become easier, installation is rapid and reconfiguration becomes possible (Mehrabi et al., 1999). An approach to solve the problem is provided. Such approach is not limited to any specific system. It could be either used to configure a line for one time in case of DML (since the configuration is locked for the whole life time of the system) or to reconfigure the system in case of RMS at each time the demand changes to adapt to the new situation.

The design or configuration of a modular machining lines deals with the selection of modules from a given set and with their assignment to a set of stations. The modules in such lines are the multi-spindle units. Figure 1 illustrates a unit with 2 spindles. When the line has to be configured for the first time, i.e., the line has to be built, the given set of modules is formed on the basis of the following information:

a) The availability on the market, proposed by the manufacturers of spindle units.
b) The knowledge of the engineering team to design and manufacture their own spindle units
c) The already used spindle units which worked on the old lines and are still operational

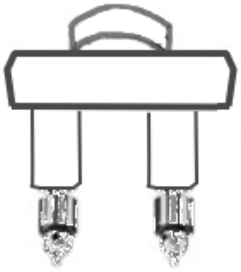


Figure 1. Two-spindle unit

The problem remains in finding an assignment for spindle units such that all operations are performed and all technological within cycle time constraints

are fulfilled. The objective is to minimize the total cost considering the cost of workstations and the costs of spindle units. Depending on the type of system we deal with, the costs can be either the fixed costs in case of dedicated lines or reconfiguration costs considering the amortization of the equipment.

A diagram of the design approach using our IP models is presented in Figure 2. This could be integrated in a holistic approach for line design which is similar to the framework of modular line design suggested by (Zhang *et al.*, 2002).

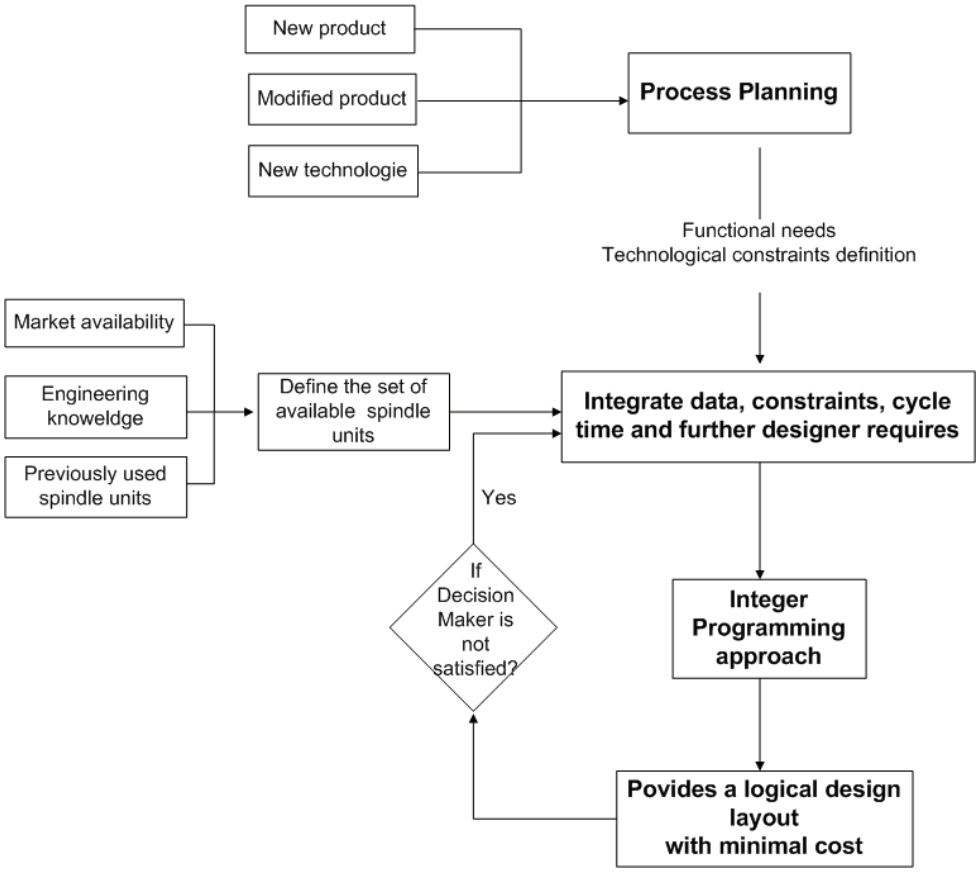


Figure 2. A global conceptual schema

In the next section related works the problem of designing modular machining lines for single product case is firstly addressed. Then, the proposed approach is generalized to the broader case considering a family of products.

1.2 Configuration and reconfiguration – related works

Koren *et al.* (1999) perform a comprehensive analysis of different types of manufacturing systems. Despite of their high cost, the dedicated machining lines (DML) are very effective as long as the demand exceeds the supply. Even these lines could operate at their full capacity; their average utilisation rate does not exceed 53%, as shown in the cited work. Flexible manufacturing systems (FMS), on the other hand, are built with the maximal available flexibility. They are able to respond to market changes, but are very expensive due to the involved CNC technology. The same study shows that 66% of FMS are not exploiting their full flexibility. Consequently, capital lies idle and an important portion is wasted.

A new alternative to the latter systems is brought by the *reconfigurable manufacturing systems* (RMS). The RMS aims to compensate the disadvantages of the last systems. This can be achieved by combining the high productivity of DML and flexibility of FMS, hence, providing a cost-effective and quick response to market changes. The cited authors define the RMS as being "designed at the outset for rapid change in structure, as well as in hardware and software components, in order to quickly adjust production capacity and functionality within a part family in response to sudden changes in market or in regulatory requirements."

Youssef & ElMaraghy (2005) identify two aspects of the reconfiguration, namely the software part and the hardware (physical) part. The effort is placed in the first part – machine re-programming, re-planning, re-scheduling, re-routing – whereas the physical reconfiguration relies upon adding/removing machines and changing the handling material. Son (2000) proposes a genetic algorithm based design methodology of an economic reconfiguration in response to demand variations, by using a configuration similarity index.

RMS must necessarily be based on a *modular* structure to meet the requirements for changeability. To configure machining systems the interfaces between the modules have prior importance, therefore these latter must meet some standard specifications. A first step to reconfigurability and, meanwhile, its strongest justification, is to ensure the possibility of producing a family of products, instead of a single one, such that to enable a smooth re-adaptation of the system to a continuously changing demand. A low cost design of a mixed-model machining line will implicitly ensure the re-adaptation time minimisation also.

A single-part versus multiple-part manufacturing systems (SPMS vs. MSPS) critical analysis is performed by Tang *et al.* (2005a). The SPMS concern the production of a single type of product on a practically rigid line configuration, whereas the MSPS are intended for a product family. The MSPS are obviously more complex, need a more sophisticated transfer system, but conversely offer more flexibility at a lower cost.

In the following, we give a formal description of the designing machining line problem for single product.

1.3 Single product case

1.3.1 Problem description

In order to model the problem we have to understand the mechanism of the machining process. We consider the lines where the activation of the spindle units at workstations is sequential. At the level of a workstation, the spindle units operate one after another on the positioned part to be manufactured. So, each workstation has an execution time equal to the sum total of its spindle times. The cycle time of the line is the elapsed time between the starting machining of the spindle units and their end on all workstations. Thus, the cycle time is determined by the slowest station of the line. At the end of each cycle time, the parts are moved to the next station and another cycle begins. Figure 3 illustrates such a line with 2 workstations. The first workstation is equipped with 2 multi-spindles whereas the second has only one unit. Each unit is composed of two spindles.

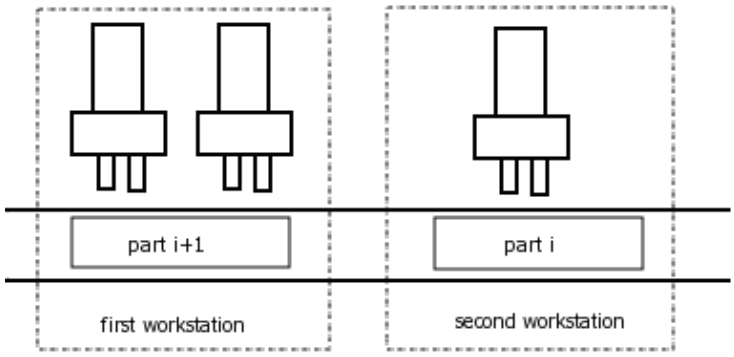


Figure 3. An example of a dedicated line

Notations and assumptions are as follows:

- **N** is the set of operations that have to be performed on each part (drilling, milling, boring, etc.).
- **B** = $\{B_r | B_r \subset \mathbf{N}\}$ corresponds to the set of all available multi-spindle units, where each is defined by the subset of operations it performs. A multi-spindle unit B_r is physically composed of one or several tools performing simultaneously corresponding operations. For the sake of simplicity, the term *block* is used henceforth to refer to a multi-spindle unit. Thus, the block $B_r \subset \mathbf{N}$ is said to contain the operations performed by the corresponding multi-spindle unit.
- q_r is the cost of block B_r,
- t_r is the execution time of block B_r,
- C_s is the average cost for any created workstation,
- C_T is the cycle time to not exceed,
- m_0 and n_0 are the maximum number of workstations which can be created and the maximum number of blocks which can be assigned to any workstation, respectively.

It is assumed that the following constraints are known:

1. cycle time,
2. precedence relations between operations,
3. exclusion conditions for blocks and
4. inclusion constraints for operations to be executed at the same workstation.

The above constraints are defined as follows:

1. An upper bound on the cycle time insures a minimal threshold of throughput.
2. Precedence relations impose an order which should be respected between some operations. For example, before drilling or boring one part a jig boring should be performed. A jig boring consists in making a notch when "true position" locating is required. The order relation over the set **N** can be represented by an acyclic digraph $G^{or} = (\mathbf{N}, D^{or})$. An arc $(i, j) \in \mathbf{N} \times \mathbf{N}$ belongs to the set D^{or} if the operation j must be executed after operation i.

3. Exclusion conditions correspond to the incompatibility between some operations, e.g. it can be the inability to activate some tools on the same workstation. The same kind of constraints have been already studied by Park, Park and Kim (1996) where the assignment of tasks may be restricted by some incompatibilities (minimum or maximum distances in terms of time or space between stations performing a pair of tasks). In our case, this incompatibility is extended to blocks such that blocks involving incompatible operations are not assigned to the same workstation. The constraints are represented by a collection D^{bs} of subsets $D_l \subseteq \mathrm{B}$ such that all blocks from the set D_l cannot be assigned to the same workstation. But any subset strictly included in D_l can be assigned to the same workstation.

4. Restrictions related to operations which have to be executed on the same station are referred to as inclusion relations. For example, if a precise distance is required between two holes, the operations corresponding to their drilling should be performed at the same workstation. If these operations are performed on different workstations, then the impact of moving reduces greatly the chance of successful precision drilling for subsequent holes. The inclusion conditions can be represented by a family D^{os} of subsets $D_t \subseteq \mathrm{N}$ such that all operations of the same subset D_t from D^{os} must be performed at the same workstation. In Pastor and Corominas (2000) similar restrictions are considered, these operations are introduced as one operation. Beyond the possibility of merging the operations, we also consider the case where operations can be performed separately with different spindle units (if such units are available).

1.3.2 The integer linear program

Decision variables are defined as follows:

$$x_{rk} = \begin{cases} 1, & \text{if block } \mathrm{B}_r \text{ is assigned to station } k \\ 0, & \textit{otherwise} \end{cases}$$

$$y_k = \begin{cases} 1, & \text{if station } k \text{ is opened} \\ 0, & \textit{otherwise} \end{cases}$$

Additional parameters have to be defined, they are described as follows.

- $K(r) = [head_r, tail_r] \subseteq [1, m_0]$ is the interval of the workstation indices where block $B_r \in \mathbf{B}$ can be assigned. The $head_r$ is the earliest station where block B_r can be assigned and $tail_r$ is the last; $head_r$ and $tail_r$ values are computed on the basis of problem constraints. Obviously, the number of decision variables is directly proportional to the width of the interval $K(r)$;
- $Q(i) = \{B_r \in \mathbf{B} \mid i \in B_r\}$. Thus, $Q(i)$ contains all blocks from **B** which perform operation $i \in \mathbf{N}$;
- interval $KO(j)$ corresponds to all stations where operation j can be performed:

$$KO(j) = \bigcup_{B_r \in Q(j)} K(r)$$

The objective function is expressed as follows:

$$\text{Minimize} \sum_{k=m^*+1}^{m_0} C_s \cdot y_{k+} \sum_{k=1}^{m_0} \sum_{B_r \in \mathbf{B}} q_r \cdot x_{rk}$$

The following constraints ensure for each operation from set **N** its execution in only one workstation:

$$\sum_{B_r \in Q(i)} \sum_{k \in K(r)} x_{rk} = 1, \quad \forall i \in N$$

The cycle time constraints for each workstation are:

$$\sum_{B_r \in \mathbf{B}} x_{rk} \cdot t_r \le C_T, \quad \forall k = 1, 2, ..., m_0$$

The precedence constraints must not be violated:

$$\sum_{B_r \in Q(i)} \sum_{l=head_r}^{k-1} x_{rl} \ge \sum_{B_r \in Q(j)} x_{rk}, \quad \forall (i,j) \in D^{or}, \quad \forall k \in KO(j)$$

The inclusion constraints for operations are respected with the following constraints:

$$\sum_{B_r \in Q(i)} x_{rk} = \sum_{B_r \in Q(j)} x_{rk}, \quad \forall \{i,j\} \subseteq D_t, \quad \forall D_t \in D^{os}, \quad \forall k \in KO(j)$$

The exclusion constraints for blocks are respected if:

$$\sum_{\mathrm{B}_r \in D_l} x_{rk} \le |D_l| - 1, \quad \forall D_l \in D^{bs}, \quad \forall k \in \bigcap_{\mathrm{B}_r \in D_l} K(r)$$

The maximal number of blocks by workstation is respected by:

$$\sum_{\mathrm{B}_r \in \{\mathrm{B}_s \in \mathbf{B} | k \in K(s)\}} x_{rk} \le n_0, \quad \forall k = 1, 2, ..., m_0$$

The following constraints are added in order to avoid the existence of intermediate empty workstations:

$$y_{k-1} - y_k \ge 0, \quad k = m^* + 2, ..., m_0$$

$$y_k \ge x_{rk}, \quad k \ge m^* + 1, \quad B_r \in \{B_s \in \mathbf{B} | k \in K(s)\}$$

In the next section we show how the model for a single product, above presented, can be extended to a family of products. Many of the assumptions and notation are maintained, and some new assumptions have to be considered, as shown below.

1.4. Family product case

1.4.1 Problem description

The features of the product family are supposed known, each product being described by the corresponding precedence graph and the required cycle time. An admissible resemblance degree must be assumed between products, as to some well known rules of defining a product family (Kamrani & Logendran, 1998) (for example, they must have a minimal number of common operations).

The goal is to design the minimal cost line configuration from the given set of available modules. This configuration must ensure a desired throughput level. By designing, we mean to determine the number of workstation to establish and to equip them with blocks such that all operations are executed only once. We are interested in the best structure of such line with regard to fixed cost point of view. Thus, the objective function is a linear combination of the cost of workstations to be established and the costs of blocks chosen to be assigned to them.

1.4.2 Problem assumptions

For our problem the following assumptions are adopted for the whole family:

- a set of all possible blocks is known and it is authorized that each block may be used only partially; the operations from a block are executed in parallel;
- each operation must be executed only once, by a single block assigned to exactly one workstation;
- the activation of blocks belonging to the same station is sequential;
- station setting cost, blocks' operating times and cost of each block are given.

An admissible assignment of blocks to stations is to find such that all the technological constraints – order, compatibility and desired cycle times – be satisfied and the *total equipment cost* (for all stations and blocks) is minimal. We are not interested to find the blocks activation sequence for each product, but we should ensure for the provided solution that such an order always exists for each product.

1.4.3 Related literature

This kind of problems is known in literature as *Line Balancing* problems (Scholl & Klein, 1998). In case of a single product type and if each block is composed of only one operation, then the problem is reduced to the basic problem, *Simple Assembly Line Balancing* (SALB). The aim is to minimize the unbalance (cost) of the line for a given line cycle time. The unbalance is minimal if and only if the number of stations is minimal.

In general, integer linear programming models for the SALBP are formulated and solved by exact or heuristic methods (Baybars, 1986; Talbot *et al.*, 1986; Johnson, 1988; Erel & Sarin, 1998; Rekiek *et al.*, 2002). The *Mixed-model Assembly Line Balancing* (MALB) problem approaches the optimization of lines with several "versions" of a commodity in an intermixed sequence (Scholl, 1999). The design of such lines – also called *multi-product* or *multi-part* lines – must take into account the possible differences between versions – among others, different precedence relations, different task times, etc. By enriching the basic assumptions of SALB, the *Generalized* ALB (GALB) problems have also been stated, in order to solve more realistic problems – a comprehensive survey may be found in Becker & Scholl (2006).

The balancing problems with *equipment selection* have some common features with the studied problem. In this context, a recent approach proposes the use of a genetic algorithm for configuring a multi-part optimal line, having the maximal efficiency (minimal ratio of cost to throughput) as criterion for the fitness function (Tang *et al.*, 2005b). Our problem differs essentially from the balancing and equipment selection problems because the operations are simultaneous into blocks. This feature makes it impossible to directly solve by the known methods.

The closest problems are studied in Dolgui *et al.* (1999), Dolgui *et al.* (2000), Dolgui *et al.* (2001), Dolgui *et al.* (2005) and Dolgui *et al.* (2006b) where all blocks at the same station are executed sequentially (block by block) and any alternative variants of blocks are not given beforehand (any subset of the given operation set is a potential block). In Dolgui *et al.* (2004), Belmokhtar *et al.* (2004), Dolgui *et al.* (2006a) the blocks are known and are executed in parallel. All these papers concern the case of a single product, for which three solving approaches have been proposed: a constrained shortest path; mixed integer programming (MIP); heuristics. A generalization of the linear programming approach to the case of a product family and sequential activation of blocks was for the first time presented by Bratcu *et al.* (2005).

The rest of this chapter is organised as follows. In Section 2 a detailed formal description of the problem is presented, along with the needed notations and explanations on how the constraints' aggregation is made. In Section 3 the proposed solving procedure is discussed, based upon a linear programming model, possibly to improve by some reductions. Section 4 is dedicated to some concluding remarks and to perspectives.

2. Formal statement of the problem

2.1 Input data and notations

The problem is identified by answering to the following questions:

a) *what* must be produced? – this is the set of features characterizing the product family (number of products, set of operations and precedence constraints for each product);
b) *how* should them be produced? – these are the blocks' characteristics (cost and operating time of each block);

c) production conditions (*external* environment – like demand, for example – and *internal constraints* – for example, maximal number of stations or maximal number of stations on the line).

As consequence, the following input data are given for each instance of the problem:

a) - p is the number of product types to manufacture;
   - $\mathbf{N}_i$ is the set of operations corresponding to product i, i=1,2,…p;
   - $\mathbf{N} = \bigcup_{i=1}^{p} \mathbf{N}_i$ is set of operations of the whole family;
b) - **B** is the set of blocks for realizing the operations from **N**, with **R** being the set of **B**'s indices;
   - Cb_r is the cost of block r and tb_r is its operating time;
c) - Cs0 is the cost of setting a new station;
   - m0 is the maximal number of workstations and n0 is the maximal num ber of blocks assigned to a station;
   - Δt is the time interval in which the demand of product i is ni, i=1,2,…p.

From quantitative information about the demand, that is, from Δt and n_i, the imposed cycle times for each product, Tc_i, may be computed. It is assumed that **B** contains only blocks having operating times smaller than the smallest cycle time of the products: $tb_r \leq \min_{i=1,2,\ldots p} \{Tc_i\}$ for all $r \in \mathbf{R}$.

Three types of technological constraints are considered:

1. the precedence constraints
2. the inclusion constraints
3. the exclusion constraints.

Their meanings and formalizations are detailed hereafter:

1. The precedence relation is a partial order relation over each set $\mathbf{N}_i$. It is represented by an acyclic digraph $G_i=(\mathbf{N}_i, Dor_i)$. One should notice that the precedence relation is here taken in the non strict sense: a vertex $(j,k) \in \mathbf{N}_i \times \mathbf{N}_i$ belongs to the set Dor_i if *either* operation k must be executed *after* operation j, *or* the two operations are performed *in parallel* (in the same time).

2. Exclusion conditions correspond to incompatibility between some operations and have the same meaning like in the single product case (see section 1.3.1).
3. Restrictions related to operations which have to be executed on the same station are referred to as inclusion relations. These also have the same meaning like in the single product case.

The inclusion conditions can be represented by a family D_i^{os} of subsets $D_t \subseteq \mathbf{N}_i$ such that all operations of the same subset D_t from D_i^{os} should be performed at the same workstation. One can note that each D_i^{os} is at most a partition over the set $\mathbf{N}_i$. *Remark*: In the general case, where the blocks of operations are not known, there exist inclusion and exclusion constraints of assigning operations to the same block, that is, sets of operations forbidden to be assigned to a block all together, respectively sets of operations which are mandatory to be assigned to a same block. For our problem, these constraints are taken into account while forming the block set **B**, therefore it is supposed that all the blocks of **B** already meet these constraints.

2.2 Aggregated constraints

As all the products should be produced on the same line, using the same equipment, an initial phase in dealing with a reconfigurable line optimization is the aggregation of the constraints concerning the individual products. Due to the assumption on the same characteristics for products belonging to the same family, the constraints should not be contradictory. In case where it happens, there will be no feasible solution to the design problem: a line for machining the given product family cannot be designed under the given constraints.

In particular, there should normally not be contradictory precedence constraints between operations common to several products of the family. But if this however happens, one must first make the aggregation of the precedence constraints to obtain a single precedence graph for the whole family. This operation will influence also the other two types of constraints, as shown later. The *aggregated* (or *total*) *precedence graph* is obtained by merging together the sets of individual precedence relations, according to the following steps:

- represent all graphs superposed and merge the multiple vertices in the same sense;
- delete redundant arc (*i,j*), i.e., if there is a path from *i* to *j* (containing several transitive vertices), then the arc (*i,j*) is said to be redundant and consequently should be deleted;
- identify the *circuits* due to contrary precedence between operations in different individual graphs; the nodes from these circuits correspond to operations that cannot be separated without violating the precedence constraints, therefore, such operations are merged together into the newly introduced *macro-operations*;
- redraw the total acyclic graph, where the macro-operations are represented by ordinary nodes.
- The definition of the macro-operations will consequently induce changes in all the operation sets, $\mathbf{N}_i$, i=1,2…p, and also in the total set, **N**, as well as in the set of both inclusion and exclusion constraints.

Concerning the *inclusion constraints*, each element of each D_i^{os}, i=1,2,…p, containing only some operations of a macro-operation, is extended with the absent operations. These sets are then united and the elements having non empty intersection are merged together. Furthermore, the *blocks* which execute just a part of the macro-operations should be eliminated; the final, aggregated set of inclusion constraints is denoted by *Dos*. Next, the sets of *exclusion constraints* (elements of *Dbs*) containing these blocks are to be eliminated too. These two latter actions may also be viewed as part of the model reduction (detailed in Section 3.2).

2.3 Example

Here below is detailed an example to illustrate the aggregated constraints.

Let a product family be composed of p=3 products, given by their precedence graphs, as in Figure 4. The corresponding sets of operations are:

$$\mathbf{N}_1=\{1,2,3,4,5,6,7,8,9,10,11\},\ |\mathbf{N}_1|=11,$$
$$\mathbf{N}_2=\{1,2,3,4,5,6,7,8,9,10,11,12\},\ |\mathbf{N}_2|=12,$$
$$\mathbf{N}_3=\{1,2,4,5,7,8,9,10,11,12,13\},\ |\mathbf{N}_3|=11.$$

The total operation set is therefore:

$$\mathbf{N} = \bigcup_{i=1}^{p} \mathbf{N}_i = \{1,2,3,4,5,6,7,8,9,10,11,12,13\},\ |\mathbf{N}|=13,$$

and there are 9 common operations for all products:

$$\mathbf{N_b} = \bigcap_{i=1}^{\mathbf{p}} \mathbf{N}_i = \{1,2,4,5,7,8,9,10,11\}$$

Ellipses in Figure 4 represent the inclusion constraints (defining which operations must be performed on the same workstation) for product *i*, D_i^{os}. The corresponding sets are:

$$D_1^{os} = \left\{ \underbrace{\{7,8\}}_{D_{11}^{os}}, \underbrace{\{9,10\}}_{D_{12}^{os}} \right\},\ D_2^{os} = \left\{ \underbrace{\{7,8\}}_{D_{21}^{os}}, \underbrace{\{9,10,12\}}_{D_{22}^{os}} \right\} \text{ and } D_3^{os} = \left\{ \underbrace{\{7,8\}}_{D_{31}^{os}}, \underbrace{\{9,10\}}_{D_{32}^{os}} \right\};$$

to these ones a supplementary set of inclusion constraints is added, $D_s^{os} = \{\{3,13\}\}$, concerning two operations belonging to different products, which have to be together when the products are realized on the same line, This latter set is suggested by dotted rectangle in Figure 4. Taking into account the aggregation of constraints, this set will be treated just like any other D_i^{os}.

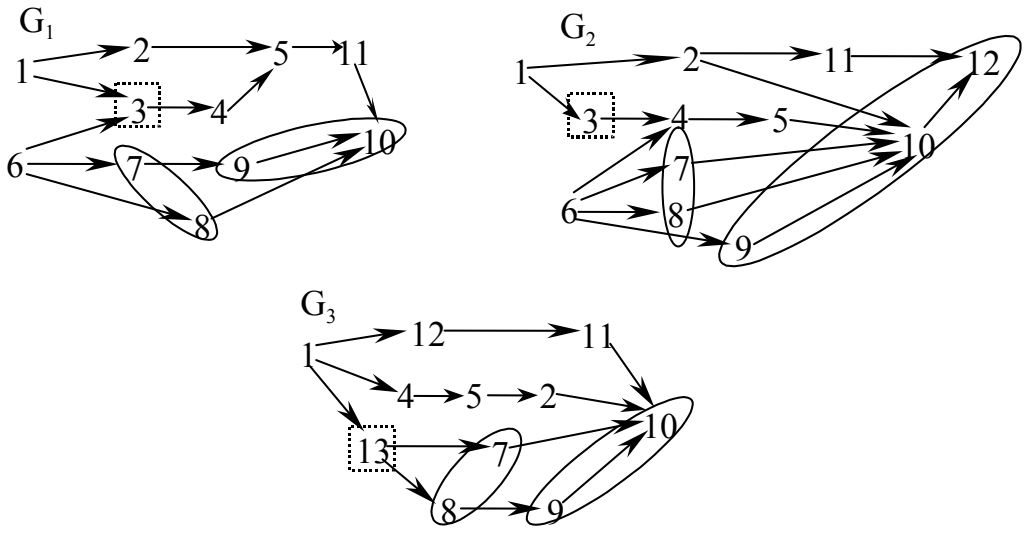


Figure 4. Precedence graphs and inclusion constraints for a family of 3 products.

Next, suppose that the total set **N** is to be executed by a set **B** of 12 multifunctional tools (blocks), whose features – operations to perform, operating times and costs – are provided in Table 1 (abbreviations t.u. and m.u. denote respectively time units and monetary units).

Block r	Operations	Operating time, tb_r [t.u.]	Cost, Cb_r [m.u.]
B_1	{1,3,6,13}	9	250
B_2	{1,3,13}	8	170
B_3	{1,2,7,8}	6	281
B_4	{2,5,9}	10	150
B_5	{2,5,7,8,11}	9	275
B_6	{2,6,9,10}	11	230
B_7	{4,6,8,10}	13	211
B_8	{4,7,8}	9	160
B_9	{4,7,8,9}	10	215
B_{10}	{5,12,13}	6	158
B_{11}	{10,11,12,13}	12	230
B_{12}	{2,5,10,11,12}	11	260

Table 1. Set of blocks to manufacture the product family described in Figure 4.

The fixed cost of setting a new station is Cs_0=350 m.u., the maximal number of stations is m_0=5 and the maximal number of blocks per station is n_0=3.

Next, suppose that the following constraints related to minimal line throughput are imposed: in a period of Δt=48300 t.u. n_1=2100 pieces of the first product , n_2=1932 pieces of the second product and n_3=2300 pieces of the third product must be manufactured. Computing the required cycle with the expression $Tc_i=\Delta t/n_i$ for time for product i, one obtains Tc_1=23 t.u., Tc_2=25 t.u. and Tc_3=21 t.u. respectively.

The exclusion constraints are provided by a unique block set for all products. Suppose that the sets of blocks forbidden to be assigned to the same station are:

$$D^{bs}=\left\{\underbrace{\{B_1,B_6,B_7,B_{10}\}}_{D_1^{bs}},\underbrace{\{B_1,B_9\}}_{D_2^{bs}},\underbrace{\{B_3,B_5,B_{11}\}}_{D_3^{bs}}\right\}$$

The constraints aggregation is part of a pre-processing performed on the initial data about the problem; it will lead to a single set of each type of constraints, distinguished by the exponent "*pp*" (acronym corresponding to *p*re-*p*rocessing). The generation of the total precedence graph – a single one for the whole product family – by merging together the individual precedence graphs, is the starting point in aggregating constraints. In the considered case, this operation allows to identifying two circuits, 2→5→2 and 11→10→12→11, due to contrary precedence relations between same operations in different individual graphs. Therefore, two macro-operations are formed, denoted by a={2,5} and b={10,11,12}. The total precedence graph, G, defining a partial order relation on the new set of operations, $\mathbf{N}^{pp}$={1,3,4,*a*,*b*,5,6,7,8,13}, is shown in Figure 5.

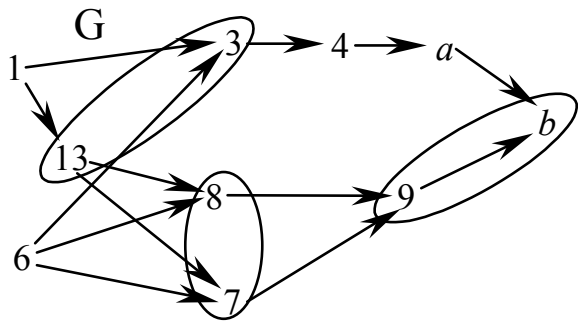


Figure 5. Precedence graph of the whole product family.

As the macro-operations have been introduced, the *operation set* for each product has also changed:

$$\mathbf{N}^{pp1}=\{1,a,3,4,6,7,8,9,b\},\ |\mathbf{N}^{pp1}|=9,$$
$$\mathbf{N}^{pp2}=\{1,a,3,4,6,7,8,9,b\},\ |\mathbf{N}^{pp2}|=9,$$
$$\mathbf{N}^{pp3}=\{1,a,4,7,8,9,b,13\},\ |\mathbf{N}^{pp3}|=8.$$

Also, those *blocks* which cannot execute but only some operations from a macro-operation must be eliminated. These are B_3, B_5, B_6, B_7 and B_{10}. Therefore, the new set of blocks is $\mathbf{B}^{pp}$={B_1,B_2,B_4,B_8,B_9,B_{11},B_{12}}, having noted that now B_4={*a*,9}, B_{11}={*b*,13} and B_{12}={*a*,*b*}.

Next, one must perform the aggregation of the *inclusion constraints*, taking into account the existence of macro-operations. Thus, each element of each D_i^{os}, i=1,2,3, and each element of D_s^{os} containing only some of the operations included in macro-operations is first extended with the absent operations:

$$D_1^{os\,pp} = \left\{ \underbrace{\{7,8\}}_{D_{11}^{os\,pp}}, \underbrace{\left\{ \underbrace{9,10,11,12}_{b} \right\}}_{D_{12}^{os\,pp}} \right\}$$

$$D_2^{os\,pp} = \left\{ \underbrace{\{7,8\}}_{D_{21}^{os\,pp}}, \underbrace{\left\{ \underbrace{9,10,11,12}_{b} \right\}}_{D_{22}^{os\,pp}} \right\}$$

$$D_3^{os\,pp} = \left\{ \underbrace{\{7,8\}}_{D_{31}^{os\,pp}}, \underbrace{\left\{ \underbrace{9,10,11,12}_{b} \right\}}_{D_{32}^{os\,pp}} \right\}$$

Set $D_s^{os\,pp} = \left\{ \underbrace{\{3,13\}}_{D_{s1}^{os\,pp}} \right\}$ remains unchanged. Then, all these 4 sets are united and the elements having non empty intersection are merged together. The aggregated set of inclusion constraints is:

$$D^{os\,pp} = \left\{ \underbrace{\{3,13\}}_{D_1^{os\,pp}}, \underbrace{\{7,8\}}_{D_2^{os\,pp}}, \underbrace{\{9,b\}}_{D_3^{os\,pp}} \right\}$$

The set of *exclusion constraints*, *Dbs*, must be changed because of the elimination of some blocks in the previous aggregation steps. Each element of *Dbs* has the meaning of forbidding the blocks to be *all together* on the same station (note that the global exclusion relation does not necessarily mean mutually exclu-

sion). Hence, if a block happens to be eliminated, then all the exclusion constraints containing it will also be eliminated. In the considered example, sets D_1^{bs} and D_3^{bs} are those to be eliminated. Therefore, the final exclusion constraints set is:

$$D^{bs\,pp} = \left\{ \underbrace{\{B_1, B_9\}}_{D_1^{bs\,pp}} \right\}$$

3. Solving by integer linear programming (IP)

3.1 IP formulation

The cost optimization of a reconfigurable machining line admits a IP formulation. The presented model is an extension of the one built for the single product case (Dolgui *et al.,* 2004; Belmokhtar *et al.*, 2004). The main difference is that individual constraints are aggregated for all products, the blocks work sequentially in each station and the precedence relation is not strict (see above).

The model needs that the following variables and additional parameters be introduced:

- binary decision variables x_{rk}, with x_{rk}=1 if block r is assigned to station k and x_{rk}=0 otherwise, k=1,…, m_0;
- $y \geq 0$ to denote the number of stations;
- m^* to denote a lower bound of the number of stations;
- for each block r, the interval $K(r)$=[$head(r)$,$tail(r)$], with $head(r)$ being the earliest station and $tail(r)$ being the latest station where block r can be assigned;
- family F_s={F_1,…F_v} of pairs of blocks having common operations: F_q={r,t} such that $B_r \cap B_t \neq \varnothing$ for any $q \in V$={1,…,v} – i.e., only one block from each pair of F_s can be used in a decision; F_s is called the subset of (pairs of) *alternative* blocks;
- F_0=$\mathbf{B} \setminus \bigcup_{q \in V} F_q$ – i.e., F_0 is the set of blocks that will surely appear in the solution;

- $w_{rt}=|B_r \cap D_t|$ and $W_t=\{r \in \mathbf{R} |\ w_{rt}>0\}$ for any block r and any $D_t \in Dos$, that is, W_t are the blocks able to execute the operations belonging to subset D_t of aggregated inclusion constraints;
- $U_t=\{r \in \mathbf{R} |\ i_t \in B_r\}$, where it is a given operation from the set $D_t \in Dos$;
- for each block B_r, the set $M(r)$ of operations not belonging to B_r which directly precede the operations of B_r;
- for each block r, the set $H(r)=\{t \in \mathbf{R} |\ B_t \cap M(r) \neq \varnothing\}$, containing the blocks capable of performing the operations from $M(r)$;
- $\mathbf{H}=\{r \in \mathbf{R} |\ M(r) \neq \varnothing\}$, i.e., the set of operations having predecessors;
- $h_{tr}=|B_t \cap M(r)|$ for any $r \in \mathbf{H}$ and any $t \in H(r)$;
- R^* to denote an upper bound of the set of blocks to be assigned to the last station of the line.

The *objective function* corresponds to the line total investment cost minimization:

$$Cs_0 \cdot y + \sum_{r=1}^{|\mathbf{R}|} \sum_{k=1}^{m_0} Cb_r \cdot x_{rk} \rightarrow \min \qquad (1)$$

which, for reasons of speeding up computation, can be also expressed as:

$$Cs_0 \cdot y + \sum_{r=1}^{|\mathbf{R}|} \sum_{k=1}^{m_0} (Cb_r + \varepsilon_r k) x_{rk} \rightarrow \min \qquad (1.1)$$

where ε_r is a sufficiently small nonnegative value. The optimization is subject to a set of constraints, whose mathematical forms are given and explained hereafter.

The first constraints ensures the execution of every operation from the aggregate operation set, **N**, in exactly one station. Both cases are considered: either choosing blocks without intersection with the others (from F_0), or choosing alternative blocks (from elements F_q of F_s):

$$\sum_{k \in K(r)} x_{rk} \leq 1, \quad r \in F_0 \qquad (2)$$

$$\sum_{r \in F_q} \sum_{k \in K(r)} x_{rk} \leq 1, \quad q \in V \qquad (3)$$

As all the operations from the total set **N** must be executed, it holds that:

$$\sum_{r\in\mathbf{R}}\ \sum_{k\in K(r)} |B_r\cap\mathbf{N}|\cdot x_{rk}=|\mathbf{N}| \tag{4}$$

The aggregate precedence constraints on set **N** impose that:

$$\sum_{t\in H(r)}\ \sum_{s\in K(t),s\le k} h_{tr}\cdot x_{ts}\ge |M(r)|\cdot x_{rk},\quad r\in\mathbf{H},\quad k\in K(r) \tag{5}$$

The aggregate inclusion constraints for the stations are met if:

$$\sum_{r\in W_t} w_{rt}\cdot x_{rk}=|D_t|\cdot\sum_{s\in U_t} x_{sk},\quad D_t\in Dos,\quad k\in\bigcup_{s\in U_t} K(s) \tag{6}$$

Respect of the aggregate exclusion constraints for assigning blocks to the same station writes as:

$$\sum_{r\in D_t} x_{rk}\le |D_t|-1,\quad D_t\in Dbs,\quad k\in\bigcap_{s\in D_t} K(s) \tag{7}$$

As n_0 is the maximal number of blocks to be allocated to a workstation, then:

$$\sum_{r\in\{t\in\mathbf{R}|k\in K(t)\}} x_{rk}\le n_0,\quad k=1,2\ldots m_0 \tag{8}$$

The constraints concerning the number of stations require that:

$$y\ge k\cdot x_{rk},\quad r\in R^*,\quad k\in K(r),\quad k\ge m^* \tag{9}$$

The last constraints impose that the cycle time requirements be met:

$$\sum_{r\in\{t\in\mathbf{R}|k\in K(t)\}} t_r\cdot x_{rk}\le Tc_i,\quad i=1,2,\ldots p,\quad k\ge m^* \tag{10}$$

In the above model one can note the dependence of the number of stations, y, on the variables x_{rk}. The model does not explicitly claim the integrality constraint on y, but constraint (9) and the objective function (1) implicitly force it.

Some possible model reductions may be performed, in order to minimize the number of decision variables, as proposed below.

3.2 Reduction of model and computation of bounds

In order to reduce computation time, an analysis of the block set after performing the aggregation of constraints – that is, after identifying the macro-operations – can allow some supplementary block eliminations. The steps presented hereafter are not mandatory, but can contribute to avoid useless computation.

The first action is to check if situations like the one described in Figure 6a) happen. In this figure, the precedence relations between two operations from different blocks are such that a "block circuit" appears. Obviously, a solution cannot contain all the blocks involved in such a circuit, but it is however sufficient that a single block be deleted. It is proposed that a heuristic elimination rule be used in this case, namely the most expensive – as cost per operation – block be eliminated, which is consistent with the goal of the total investment cost minimization. Note that such eliminations must start from the maximal circuits identified.

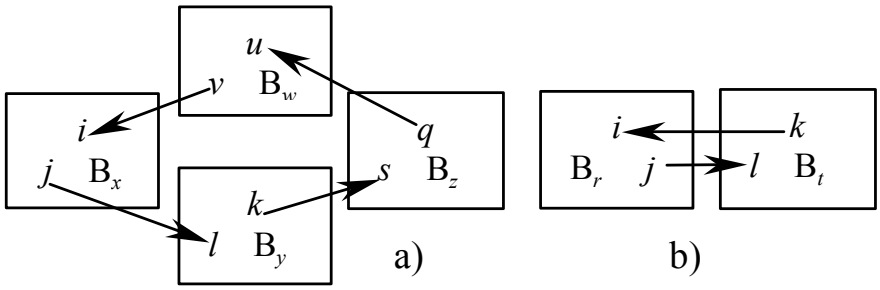


Figure 6. Example of blocks forming circuits and loops.

A special case is that of two vertices circuits (loops). If two blocks r and t form a loop (like in Figure 6b)), it is not necessary that one of them be deleted, but certainly only one will appear in a solution. It is therefore sufficient to treat them as alternative blocks (i.e., the pair (r,t) be an element of F_s).

The second step of the model reduction concerns also the set of blocks. Remember that this set, **B**, will have already undertaken some changes due to the constraints aggregation, as above mentioned.

Thus, for each block B_r from the last block set **B**, a subset B' is searched, such that:

- operations from B′ give the total set, **N**;
- $|B'| \leq m_0 \cdot n_0$;
- all blocks from B′ are mutually disjoint.

Each block for which such a subset, B′, does not exist must be eliminated.

Even if these reductions are not performed before the optimization phase, the optimizer will implicitly make them. But in large scale problems this could negatively affect the computation time.

Hereafter are presented the reductions possible for the example considered in Section 2.3.

The first reduction step is to check the existence of "*circuits*" on subsets of $\mathbf{B}^{pp}$. It is said that a precedence relation exists between two blocks if and only if all the operations from a block precede all the operations from the other block. In Figure 7 precedence relations between two blocks have been represented by thick arrows, whereas the thin arrows denote precedence relation between operations.

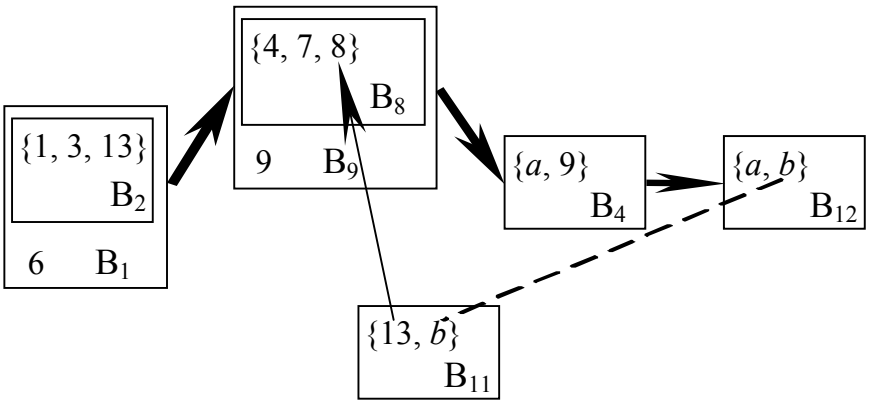


Figure 7. Circuits on the set of blocks after the constraints aggregation, B^{pp}.

One can remark the existence of a circuit of blocks, that is, $B_8 \rightarrow B_4 \rightarrow B_{12} \rightarrow B_{11} \rightarrow B_8$. According to the heuristic of eliminating the most expensive block as cost per operation, block B_{11} (57.5 m.u. per operation) must be eliminated, as to data provided in Table 1, Hence, the reduced block set is:

$\mathbf{B}^{pp}=\{B_1,B_2,B_4,B_8,B_9,B_{12}\}$,

which still contains all the operations of $\mathbf{N}^{pp}$ (if this were not be the case, then the problem would not have any solution).

Concerning the second reduction step, this is not very important in small scale problems. But for large scale problems, it may be useful to make it in the pre-processing phase. In the analyzed case, one can verify that blocks B_2, B_4 and B_8 may be eliminated.

As for the computation of bounds, we briefly present here below how the intervals $K(r)$ are computed for any block r, as well as the minimal number of stations, m^*, and the maximal block set, R^*, to be assigned to the last station.

Intervals $K(r)$ and m^* are computed based upon the algorithm proposed by Dolgui *et al.* (2000), using the notion of *distance* between two operations. In the general case, this distance takes one of three values (0, 1 or 2) – in our case, it can take only two values: either 2, if the two operations can only be performed by blocks forbidden to be on the same station (i.e., belonging to elements of *Dbs*), or 0, otherwise.

In the cited work, the blocks are not a priori known. Therefore, the problem is solved in two steps: first determining the bounds of assigning operations to blocks, then for allocating blocks to workstations.

Thus, the algorithm begins with computing the values $q^-(i)$ and $q^+(i)$ for any operation i, which denote the earliest and respectively the latest block where operation i can be assigned. In our case, to compute values $q^-(i)$, the algorithm needs as input data the total precedence graph, G, the aggregated inclusion constraints, *Dos*, and the distance matrix, d ($|\mathbf{N}|\times|\mathbf{N}|$). Values $q^+(i)$ result from the same algorithm, but entering the reversed precedence graph, G^r.

In the second step, values $k^-(i)$ and $k^+(i)$ of the earliest and the latest station where operation i can be assigned are computed, using the relation:

$$k^{+/-}(i) = [q^{+/-}(i)/n_0],$$

with $[\cdot]$ denoting the smallest integer larger or equal with the argument.

For any block r, there are finally computed:

$$\begin{cases} head(r) = \max\{k^-(i) \mid i \in B_r\} \\ tail(r) = \min\{k^+(i) \mid i \in B_r\} \end{cases} \tag{11}$$

The lower bound on the number of stations results as:

$$m^* = \max\{k^-(i) \mid i \in \mathbf{N}\} \qquad (12)$$

Having computed intervals $K(r)$ for any block r, a sufficiently good value of R^* may result from the following algorithm.

1. $tail\_max \leftarrow \max\{tail(r) \mid B_r \in \mathbf{B}\}$
2. Find the minimum head of the blocks having the tail equal to *tail_max*. Let be *head_min* this minimum.
3. Form the subset of blocks having the tail strictly larger than *head_min*. This subset is R^*.

An immediate goal aimed at in the near future is to improve the value of R^*.

4. Conclusion and perspectives

This chapter has approached the problem of optimizing the investment cost of modular machining lines (also called transfer lines) aimed at producing a family of products. The possibility of allowing variations over the set of products is the most important step for a manufacturing system to become reconfigurable. The specificity of the lines analyzed here is the parallelization of the operations' execution by the same spindle head. Due to the important investment cost required to build such lines, the search of an optimal design decision for the whole family appears as necessary. The potential economic benefits achieved are not negligible and is one of the motivating reasons to propose such an approach. The powerful mathematical programming tools make it possible to solve exactly and efficiently such problem, providing cost effective solutions. However, searching for the optimal solution may be prohibitively time-consuming, as much as the scale problem is larger.

The cost optimization of this kind of machining lines is a new and poorly studied problem, different from the classical SALB problem, but also NP-hard. This work presented a complete mathematical formulation of the problem as a linear program and proposed a procedure to follow for obtaining an exact (optimal) solution. An important phase of the solving procedure is the aggregation of constraints, which practically allows that the studied problem be treated like the single product one. Some proposals of model reduction have also been

presented, to avoid running time exhaustion.

A particular attention should be focused on improving the bounds. Due to the exponential complexity of the integer linear programming solving algorithm, a bad behavior when increasing the problem's dimension is highly possible. The large number of constraints is a feature that will potentially allow the coupling of the presented exact method with different types of heuristics, able to provide good bounds to exact methods. We consider that, for applying the proposed method in real life environments, this coupling is definitely necessary.

5. References

Baybars, I. (1986). A survey of exact algorithms for the simple line balancing problem, *Management Science*, Vol. 32, pp. 909-932, ISSN 0025-1909

Becker, C. & Scholl, A. (2006). A survey on problems and methods in generalized assembly line balancing, *European Journal of Operational Research*, Vol. 168, No. 3, (1 February 2006), pp. 694-715, ISSN 0377-2217

Belmokhtar, S.; Dolgui, A.; Guschinsky, N.; Ihnatsenka, I. & Levin, G. (2004). Optimization of transfer line by constraint programming approach, *Proceedings of Computer and Industrial Engineering Conference* (*CD-ROM*), November 13-16 2004, San Francisco, U.S.A.

Bratcu, A. I.; Dolgui, A. & Belmokhtar, S. (2005). Reconfigurable Transfer Lines Cost Optimization – A Linear Programming Approach, *Proceedings of the 10th IEEE International Conference on Emerging Technologies and Factory Automation – ETFA 2005*, Lo Bello, L. & Sauter, T. (Eds.), pp. 625-632, ISBN 0-7803-9402-X, Catania, Italy, September 19-22 2005, IEEE, Piscataway, NJ, U.S.A.

Dashchenko A. I. (Ed) (2003). *Manufacturing Technologies for Machines of the Future 21st Century Technologies*, Springer.

Dolgui, A.; Guschinsky, N. & Levin, G. (1999). On problem of optimal design of transfer lines with parallel and sequential operations, *Proceedings of the 7th International Conference on Emerging Technologies and Factory Automation – ETFA'99*, J. M. Fuertes (Ed.), pp. 329-334, ISBN 0-7803-5670-5, Barcelona, Spain, October 18-20 1999, IEEE, Piscataway, NJ, U.S.A.

Dolgui, A.; Guschinsky, N. & Levin, G. (2000). *Approaches for transfer lines balancing*, Preprint No. 8, Institute of Engineering Cybernetics/University of Technology of Troyes

Dolgui, A.; Guschinsky, N.; Levin, G. & Harrath, Y. (2001b). Optimal design of a class of transfer lines with blocks of parallel operations, *Proceedings of the IFAC Symposium on Manufacturing, Modeling, Management and Control MIM 2000*, P. P. Groumpos, A. P. Tzes (Eds.), pp. 36-41, ISBN 0080435548, Patras, Greece, July 12-14, 2000, Elsevier.

Dolgui, A.; Guschinsky, N.; Levin, G.; Louly, M. & Belmokhtar, S. (2004). Balancing of Transfer Lines with Simultaneously Activated Spindles, *Preprints of the IFAC Symposium on Information Control Problems in Manufacturing – INCOM'04* (*CD-ROM*), April 5-7 2004, Salvador da Bahia, Brazil (to appear also in the *IFAC Proceedings Volume*, Elsevier)

Dolgui, A.; Finel, B.; Guschinski, N.; Levin, G. & Vernadat, F. (2005). A heuristic approach for transfer lines balancing. *Journal of Intelligent Manufacturing*, Vol. 16, No 2, pp. 159-171, ISSN 0956-5515

Dolgui, A.; Guschinsky, N. & Levin, G. (2006a). A special case of transfer lines balancing by graph approach. *European Journal of Operational Research* Vol. 168, No. 3, (1 February 2006), pp. 732-746, ISSN 0377-2217

Dolgui, A.; Finel, B.; Guschinski, N.; Levin, G. & Vernadat, F. (2006b). MIP Approach to Balancing Transfer Lines with Blocks of Parallel Operations, *IIE Transactions*, 2006, (In Press)

Erel, E. & Sarin, C. (1998). A survey of the assembly line balancing procedures, *Production Planning and Control*, Vol. 9, pp. 414-434, ISSN 0953-7287

Groover, M. P. (1987). *Automation, production systems and computer integrated manufacturing*, Prentice Hall, Englewood Cliffs, New Jersey

Hitomi, K. (1996). *Manufacturing system engineering*, Taylor & Francis

Hutchinson, G. (1976). Production Capacity: CAM vs. transfer line, *IE*, September 1976, pp. 30-35

Johnson, J. R. (1988). Optimally balancing large assembly lines with FABLE, *Management Science*, Vol. 34, pp. 240-253, ISSN 0025-1909

Kamrani, A. K. & Logendran, R. (1998). *Group technology and cellular manufacturing: methodologies and applications*, Gordon and Breach

Koren, Y.; Heisel, U.; Jovane, F.; Moriwaki, T.; Pritschow, G.; Van Brussel, H. & Ulsoy, G. (1999). Reconfigurable Manufacturing Systems. *CIRP Annals*, Vol. 48, pp. 527-598, ISSN 0007-8506

Park, K.; Park, S. & Kim, W. (1996). A heuristic for an assembly line balancing problem with incompatibility, range and partial precedence constraints. *Computers and Industrial Engineering*, Vol. 32, No 2, pp. 321-332, ISSN 0360-8352

Pastor, R. & Corominas, A. (2000). Assembly line balancing with incompatibilities and bounded workstation loads. *Ricerca Operativa*, Vol. 30, pp. 23-45, ISSN 0390-8127

Rekiek, B.; Dolgui, A.; Dechambre, A. & Bratcu, A. (2002). State of art of assembly lines design optimization, *Annual Reviews in Control,* Vol. 26, No. 2, pp. 163-174, ISSN 1367-5788

Scholl, A. & Klein, R. (1998). Balancing assembly lines effectively: a computational comparison. *European Journal of Operational Research*, Vol. 114, pp. 51-60, ISSN 0377-2217

Scholl, A. (1999). *Balancing and sequencing assembly lines*, 2nd edition, Physica, Heidelberg

Son, S. Y., (2000). *Design Principles and Methodologies for Reconfigurable Machining Systems*, Ph.D. Dissertation, University of Michigan

Talbot, F. B.; Paterson, J. H. & Gehrlein, W. V. (1986). A comparative evaluation of heuristic line balancing techniques, *Management Science*, Vol. 32, pp. 430-454

Tang, L.; Yip-Hoi, D. M.; Wang, W. & Koren, Y. (2005a). Selection Principles on Manufacturing System for Part Family, *Proceedings of the 2005 CIRP International Conference on Reconfigurable Manufacturing* (CD-ROM)

Tang, L.; Yip-Hoi, D.; Wang, W. and Koren, Y. (2004). Concurrent line-balancing, equipment selection and throughput analysis for multi-part optimal line design. *The International Journal for Manufacturing Science & Production*, Vol. 6, Nos. 1-2, pp.71-81, ISSN 0793-6648

Youssef, A. Y. M. & ElMaraghy, H. A. (2005). A New Approach for RMS Configuration Selection, *Proceedings of the 2005 CIRP International Conference on Reconfigurable Manufacturing* (CD-ROM)

Zhang, G. W.; Zhang S. C. & Xu, Y. S. (2002). Research on flexible transfer line schematic design using hierarchical process planning. *Journal of Materials Processing Technology*, Vol. 129, pp. 629-633, ISSN 0924-0136

6

Flexible Manufacturing System Simulation Using Petri Nets

Carlos Mireles, Alfonso Noriega and Gerardo Leyva

1. Introduction to Petri Nets

In 1962 German mathematician Kart Adam on his Ph.D. work "Kommunikation mit automaten", he proposed a new way for modelling & representating systems where events and transitions are present. This new representation is now known as Petri Nets (PN).
Since then, several researchers have used this tool. In the USA and in Europe several developments have been done. In 1980, in Europe, a work table was controlled using PN and the related work was presented at one International Conference. Researchers in France have done great contributions to the development and the applications of PN. Some of them used the PN to describe Programmable Logic Controllers (PLC) and this application had a big influence in the development of the Grafcet.

2. Petri Nets Theory

2.1 Basic Concepts

2.1.1 Petri Net

Graphic and executable technique to specify and analize dynamic & concurrent discrete event systems.

2.1.2 Formal

Petri Nets analysis is a mathematical technique well defined. Many static and dynamic properties of a PN (and therefore for a system represented by a PN) can be mathematically proved.

2.1.3 Graphics

This technique belongs to the area of mathematics known as "Graph Theory", which makes a PN being able to represented both by a graphic and by mathematic expressions. This graphic property provides a better understanding of the system which is represented by the PN.

2.1.4 Executable

A PN can be executed and then the dynamics of the system can be observed.

2.1.5 Concurrency

Multiple independent entities can be represented and supported by PN.

2.1.6 Discrete event dynamic systems

A system that can change its current state based in both its current state and the transition conditions between states.

2.2 Structure of a Petri Net

2.2.1 Formal Definition

A Petri Net has a set of Places, a set of Transitions, an Input Function and an Output Function.
The structure of a Petri Net is the array (L, T, E, S, m_0) where:

- L is the set of places in the graph.
- T is the set of transitions in the graph. L & T satisfy the following conditions $L \cup T = \varnothing$ $L \cap T = \varnothing$.
- $E: LxT \rightarrow \{0,1\}$ is the input function which specify the connectivity between Places & Transitions.
- $S: LxT \rightarrow \{0,1\}$ is the output function which specify the connectivity between Transitions & Places.

2.2.2 Graphic Representation

A Petri Net is an oriented graph that contains two types of nodes: Places and Transitions which are connected by oriented arcs that connect Places with Transitions –connectivity between nodes of the same kind is not allowed. In the graphic representation, Places are shown as circles, Transitions are shown

as bars and the arcs are shown as arrows. Places represent conditions and Transitions represent events. A Transition has certain number of places either as inputs or outputs which are pre and post conditions. See figure 1.

2.2.3 Marking a Petri Net

The marking of a Petri Net is a positive integer μ_i for every place L_i. A mark is represented as a dot within the circle for a given place. These marks move between the places which provide the dynamic feature of the Petri Net.
A Petri Net is considered marked when at least one place has a mark.One place can have N marks, where N is a positive integer. If N = 0 then the place has no marks. A marker M is a function $\mathrm{M}: \mathrm{L} \rightarrow \mathrm{N}$ that can be expressed by:

$$\mathrm{M} = \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_N \end{bmatrix} \qquad (1)$$

m_i is the number of dots for place L_i.

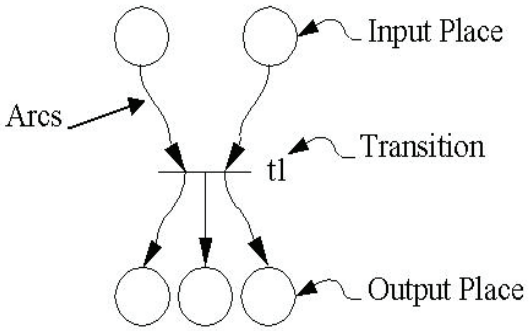


Figure 1. Graphic representation of a PN showing the Input & Outputs places as well as the Transition.

2.2.4 Interpretation of a Petri Net

Marks are the resources. Resources can be physical entites or no physical such as messages or information.
Places are the locations where the resources are stored.
Transitions are actions that transfer the resources to generate new resources.

The weight of each arrow (a number on each arrow) is the minimum number of resources needed in a place in order to get the action indicated by the transition.

2.2.5 Triggering a Transition in a Petri Net

To trigger a Transition, it has to be validated which means that every Place connected to this Transition has to have at least one mark. The execution of a Petri Net is controlled by the number and distribution of the marks on their Places.

When a Transition is triggered, a change in the marking of the Petri Net is performed and every input place lose the number of marks indicate by the weight in the arc connecting the Transition with that input place. On the other hand, every output place gain the number of places inidicated by the weight in the arc connecting the Transition with that output place. In figure 2 are shown different cases for the triggering of a transition.

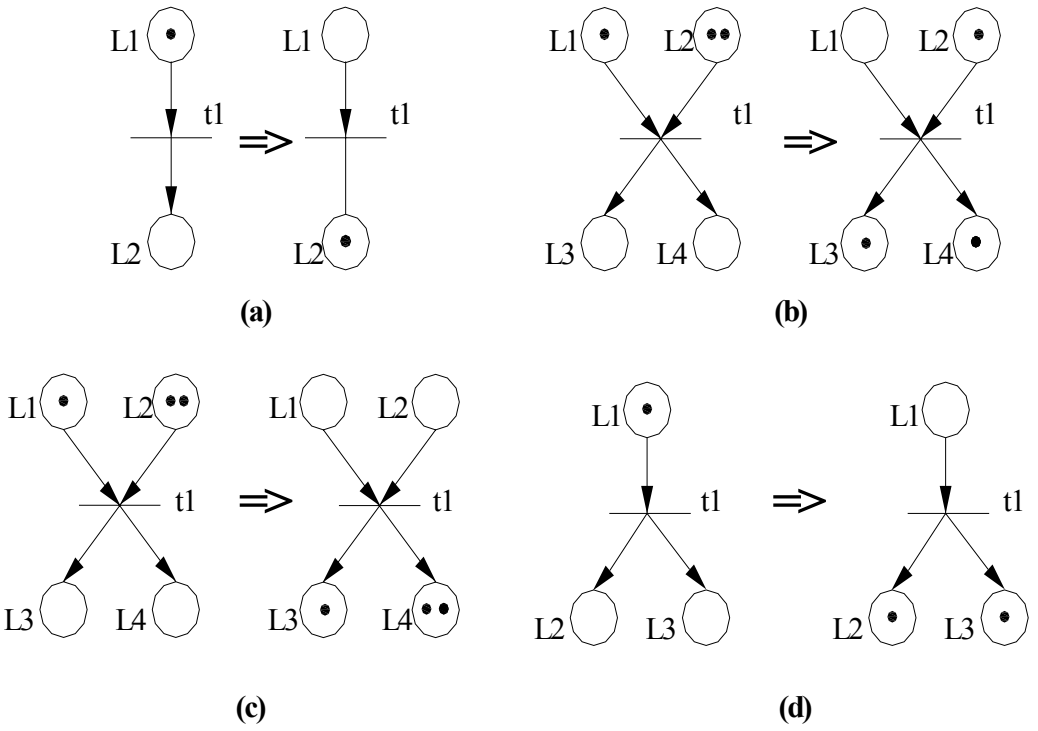


Figure 2. Different triggering cases for some Petri Nets.

3. Petri Nets showing relationships amog sub-processes

3.1 Task Sequencing

Figure 3 shows how different task can be done in sequential way.

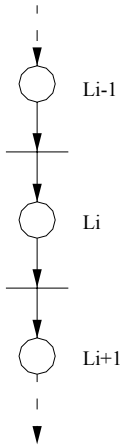


Figure 3. Task Sequencing

3.2 Task Selector

Figure 4 shows a task selector. This graph is useful when we need in a system an EXOR function between transitions t_1 & t_2.

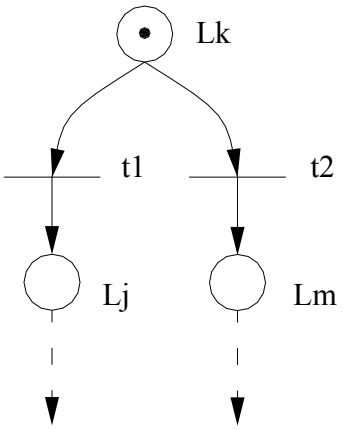


Figure 4. Task Selector

3.3 Task Synchronization

This is a very useful concept to synchronize two or more ub-processes assuming that they may have different running times or different evolution in time. Any of the processes can be out of phase respect to the others, however this model provides a wait state for any of the processes until the triggering of the transition. Synchronization is achieved when, being validated by the proper marking on the input places, the event occurs triggering the transition.

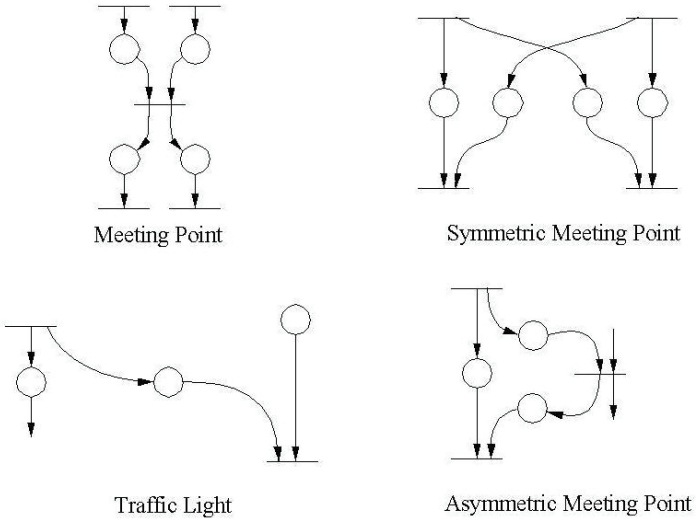


Figure 5. Different models to synchronize processes

3.4 Tasks Concurrency

Figure 6 shows how concurrent processes are represented.

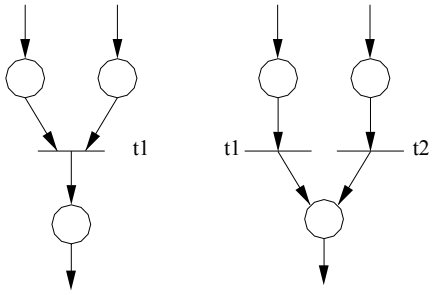


Figure 6. Tasks Concurrency

3.5 Sharing Resources

Sharing resources is done when two or more sub-processes are using a unique resource in the system –such a robot to load/unload two conveyors. The objective is to avoid the malfunction of one of the processes because the resource is not available.

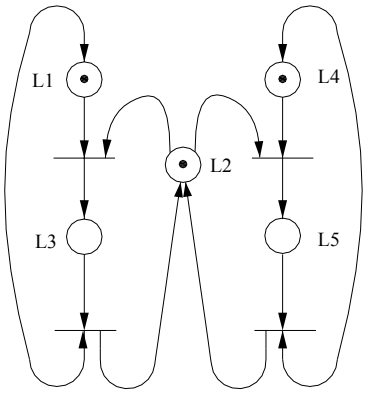


Figure 7. Sharing Resources

4. Equation of a Petri Net.

Figure 8 shows a Petri Net which will be used to explain the Petri Net equation.

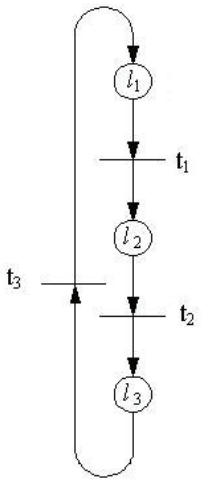


Figure 8. Sharing Resources

4.1 Input Matrix

This matrix provides the connectivity information from Places to Transitions. In this case Place l_1 goes to Transtion t_2, Place l_2 goes to Transition t_2 and so on. See figure 8.

$$E = \begin{matrix} \\ t_1 \\ t_2 \\ t_3 \end{matrix} \begin{matrix} \begin{matrix} l_1 & l_2 & l_3 \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{matrix} \tag{2}$$

4.2 Output Matrix

This matrix provides the connectivity information from Transitions to Places. In this case Transtion t_2 goes to Place l_1, Transtion t_2 goes to Place l_2 and so on. See figure 8.

$$S = \begin{matrix} \\ l_1 \\ l_2 \\ l_3 \end{matrix} \begin{matrix} \begin{matrix} t_1 & t_2 & t_3 \end{matrix} \\ \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \end{matrix} \tag{3}$$

4.3 General Matrix of a Petri Net

The two last matrixes are used to define the Net Matrix given by:

$$A = S - E \tag{4}$$

4.4 Evolution of a Petri Net

The evolution of the Petri Net is related with the "movement" of the marks in the graph from one place to another. The recursive equation is given by:

$$\mu_{k+1} = \mu_k + AV_k \tag{5}$$

where:

- μ_k represents the marking vector before the evolution of the net.
- μ_{k+1} represents the marking vector after the evolution of the net.
- A represents the General Matrix of the net.
- V_k represents the Transitions that are been triggered in the current evolution step of the net.

4.5 A working Petri Net

Figure 9 shows a Petri Net before and after the evolution step (trigger of t_1).

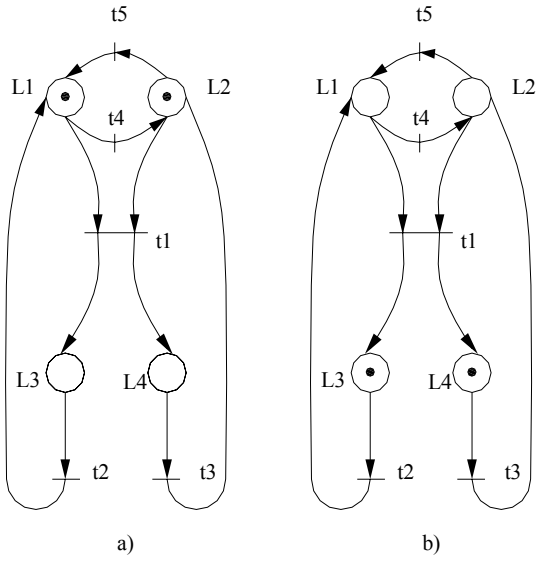


Figure 9. A Petri Net before and after the trigger of transition t_1

5. Application Examples

5.1 Wagon Control

Figure 10 shows a wagon that moves along a rail by the action of signals R (right) and L (left). Extreme positions are detected by sensors A & B. Button M starts the sequence when wagon is in initial position.

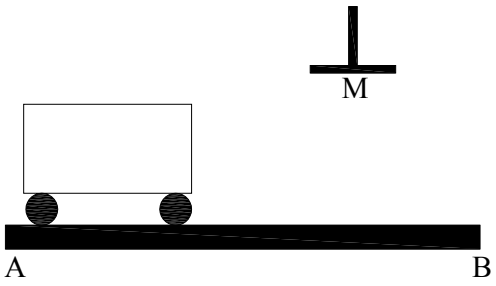


Figure 10. Wagon Control

There are three places (l_1, l_2 & l_3): Idle, Moving from A to B & Moving from B to A. There are three transitions (t_1, t_2 & t_3): Switch M, wagon detected by sensor A & wagon detetcted by sensor B. When M is press, the wagon moves from A to B, once reaches this point goes back to A. Sequence can only starts when wagon is in Idle again.

Figure 11 shows the Petri Net for this system.

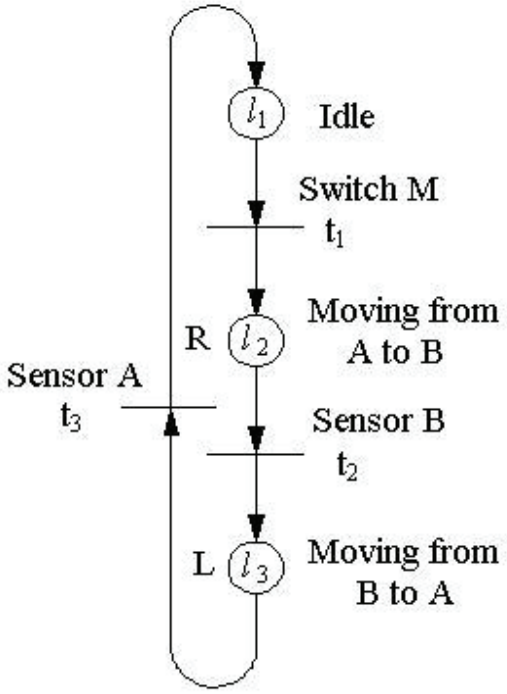


Figure 11. Petri Net for the Wagon Control system

Following the matrixes for this system:

$$E = \begin{matrix} \\ l_1 \\ l_2 \\ l_3 \end{matrix}\!\!\begin{matrix} \begin{matrix} t_1 & t_2 & t_3 \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{matrix} \qquad S = \begin{matrix} \\ l_1 \\ l_2 \\ l_3 \end{matrix}\!\!\begin{matrix} \begin{matrix} t_1 & t_2 & t_3 \end{matrix} \\ \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \end{matrix} \qquad V_0 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \qquad \mu_0 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \tag{6}$$

$$A = S - E$$

$$A = \begin{bmatrix} -1 & 0 & 1 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \tag{7}$$

Where $\mu_{k+1} = \mu_k + AV_k$ then:

$$\mu_1 = \mu_0 + AV_0 : \mu_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} -1 & 0 & 1 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \tag{8}$$

$$\mu_1 = \mu_0 + AV_0 : \mu_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} -1 & 0 & 1 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \tag{9}$$

$$\mu_1 = \mu_0 + AV_0 : \mu_1 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} + \begin{bmatrix} -1 & 0 & 1 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \tag{10}$$

6. Simulation Software

In order to simulate the Petri Nets, a software kit was developed. Although there are some simulators available, we took the decision to develop our own as the main goal and to have a software system able to be interfaced with the real world. The software was developed in Lab Windows CVI, having a Graphic User Interface which allows the user to create easily the net that represents the system that is needed to simulate. Figure 12 shows the tipycal interface of the system.

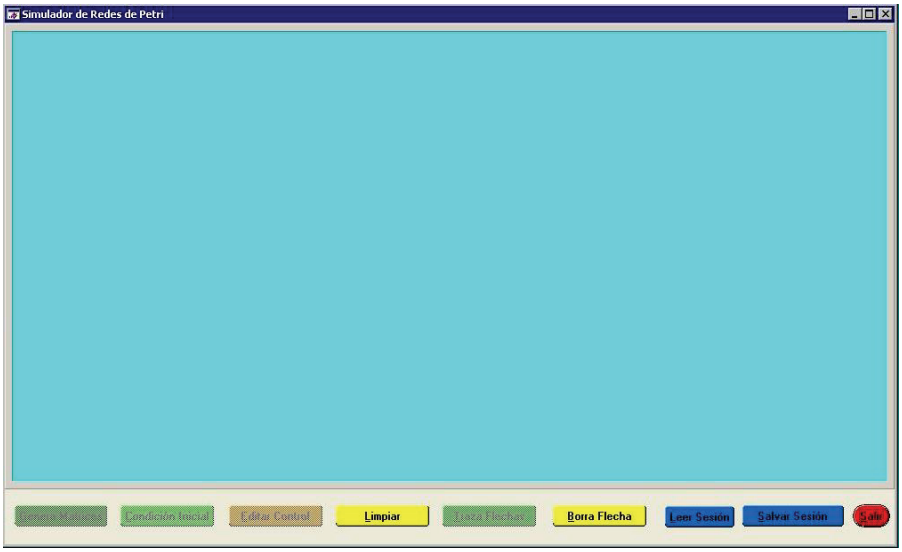


Figure 12. Main view of the simulator

The operator inputs the data related with the PN structure, the information is then interpreted by the simulator which shows the development of the PN. On figure 13, the Petri Net shown on figure 11 was introduced to the simulator.

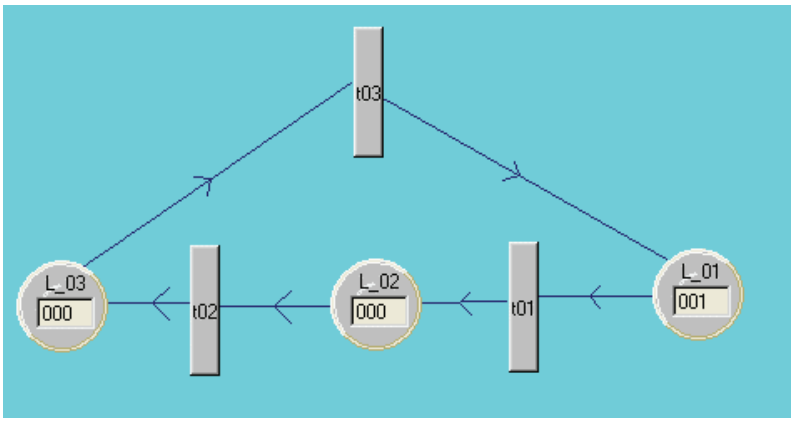


Figure 13. Petri Net for the Wagon Control system

Figures 14, 15 & 16 shows the status of the Petri Net when transitions t_1, t_2 & t_3 are triggered.

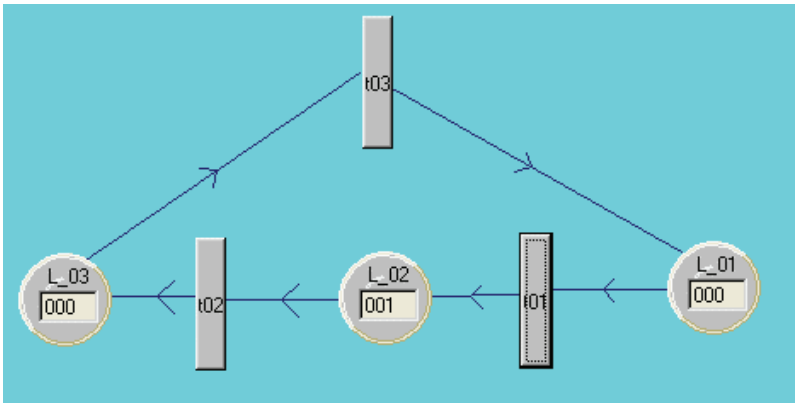


Figure 14. Petri Net for the Wagon Control system after t_1 is triggered

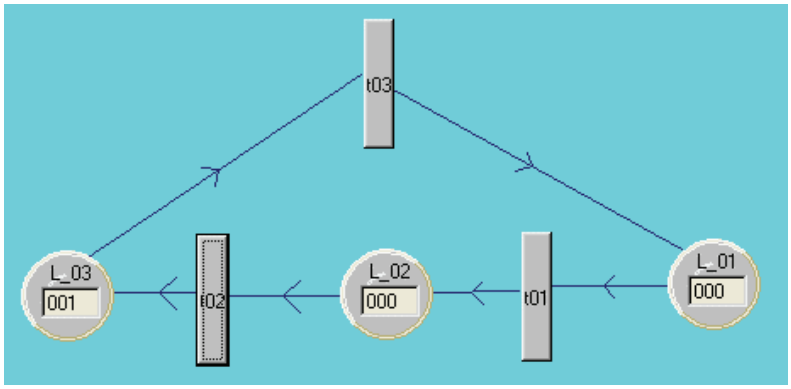


Figure 15. Petri Net for the Wagon Control system after t_2 is triggered

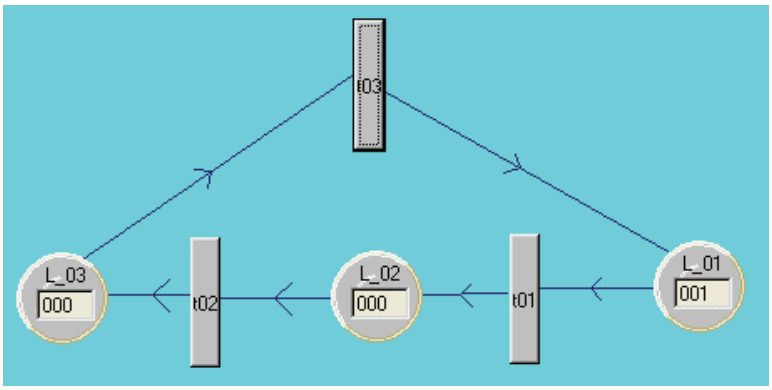


Figure 16. Petri Net for the Wagon Control system after t_3 is triggered

Inside the simulator each circle represents a Place where the wagon is moving to the right, to the left, or simply is on the idle state waiting for the button M to be pushed (transition t_1) to start a sequence.

It is worth to say that this simulator can easily be used as a Control System just by adding an Acquisition Board for sampling the actual signals which in turn can activate the transitions. The purpose with this was to develop a non expensive control system which can be used by very small companies that do not have the capital to buy very expensive equipment. In comparison with a PLC controller the cost involved when implementing this approach is very low and also offers an intuitive way of developing the control system.

7. Benefits of a Petri Net

Petri Nets have the following benefits:

a) They have a graphic representation which allows a simply and secure way to observe how the system is working.
b) They have semantics well defined that specify the net.
c) They can be used to represent several types of systems.
d) They have an explicit description for states and actions.
e) They provide interactive simulations where the results are shown directly on the graphic diagram of the net.

In future research this system can be used on the control of production lines where series of events (discrete events) need to be controlled, such as a robot feeding a conveyor, a conveyor moving cans or packages into storage areas, etc.

8. References

Mireles, C. et al. (2004), *Flexible Manufacturing System Simulation using Petri Nets*
http://www.lurpa.ens-cachan.fr/grafcet/generalites/presentation_uk.html

7

Applications of Petri Nets to Human-in-the-Loop Control for Discrete Automation Systems

Jin-Shyan Lee and Pau-Lo Hsu

1. Introduction

For discrete automation systems, certain human operations may violate desired safety requirements and result in catastrophic failure. For such human-in-the-loop systems, this paper proposes a systematic approach to developing supervisory agents which guarantee that manual operations meet required safety specifications. In the present approach, Petri nets (PN) are applied to construct a system model and synthesize a desired supervisor. Applications to 1) a rapid thermal process (RTP) in semiconductor manufacturing controlled over the Internet (Lee and Hsu, 2003) and 2) a two-robot remote surveillance system (Lee et al., 2005) are provided to demonstrate the practicability of the developed supervisory control approach. The technique developed in this paper is significant in industrial applications.

1.1 General Review

Basically, an automated process is inherently a discrete event system (DES). The Petri net (PN) has been developed as a powerful tool for modelling, analysis, simulation, and control of DES. PN was named after Carl A. Petri (1962), who created a net-like mathematical tool for describing relations between the conditions and the events. PN was further developed to meet the need in specifying process synchronization, asynchronous events, concurrent operations, and conflicts or resource sharing for a variety of industrial automated systems at the discrete-event level. Starting in the late of 1970's, researchers investigated possible industrial applications of PN in discrete-event systems as in the survey/tutorial papers of Murata (1989), Zurawski and Zhou (1994), David and Alla (1994), and Zhou and Jeng (1998).

Recently, due to the rapid development of Internet technology, system monitoring and control no longer needs to be conducted within a local area. Several

remote approaches have been proposed which allow people to monitor the automated processes from great distances (Weaver et al., 1999; Yang et al., 2002; Kress et al., 2001; Lee and Hsu, 2004; Huang and Mak, 2001).

Practically, to perform maintenance functions in hazardous environments without their exposure to dangers is a unique application of the remote technology. By conducting remote access using IP-based networks, an entire Internet-based control system is inherently a DES and its state change is driven by occurrences of individual events. The supervisory control theory provides a suitable framework for analyzing DES (Ramadge and Wonham, 1987, 1989; Balemi et al., 1993) and most existing methods are based on automata models. The calculus of communicating systems (CCS), which was invented by Robin Milner (1989), is another classical formalism for representing systems of concurrent processes. However, these available methods often involve exhaustive searches of overall system behaviours and result in state-space explosion design as system becomes more complex. On the other hand, PN is an efficient approach to model the DES and its models are normally more compact than the automata models. Also, PN is better suitable for modelling systems with parallel and concurrent activities. In addition, PN has an appealing graphical representation with a powerful algebraic formulation for supervisory control design (Giua and DiCesare, 1991; Moody and Antsaklis, 1998; Uzam et al., 2000).

1.2 Problem Statement

Typically, an Internet-based control system (remote access using IP-based networks) is a "human-in-the-loop" system since people use a general web browser or specific software to monitor and control remotely located systems. As shown in Figure 1 (a), the human operator is involved in the loop and sends control commands according to the observed status displayed by the state and/or image feedback. Research results indicate that approximately 80% of industrial accidents are attributed to human errors, such as omitting a step, falling asleep and improper control of the system (Rasmussen et al., 1994). However, the Internet-based control literature provides few solutions for reducing or eliminating the possibility of human errors. Therefore, solutions to reduce or eliminate the possibility of human errors are required in the Internet-based control systems.

1.2 Proposed Approach

In this chapter, we propose applying a supervisory design to the present remotely-controlled and human-in-the-loop system so as to prevent abnormal operations from being carried out.

As shown in Figure 1 (b), the supervisory agent acquires the system status and makes the decision to enable/disable associated events to meet the required specifications, typically safety requirements. The human operator is then only allowed to perform the enabled events to control the system. The role of a supervisory agent is to interact with the human operator and the controlled system so that the closed human-in-the-loop system meets the required specifications and to guarantee that undesirable executions do not occur.

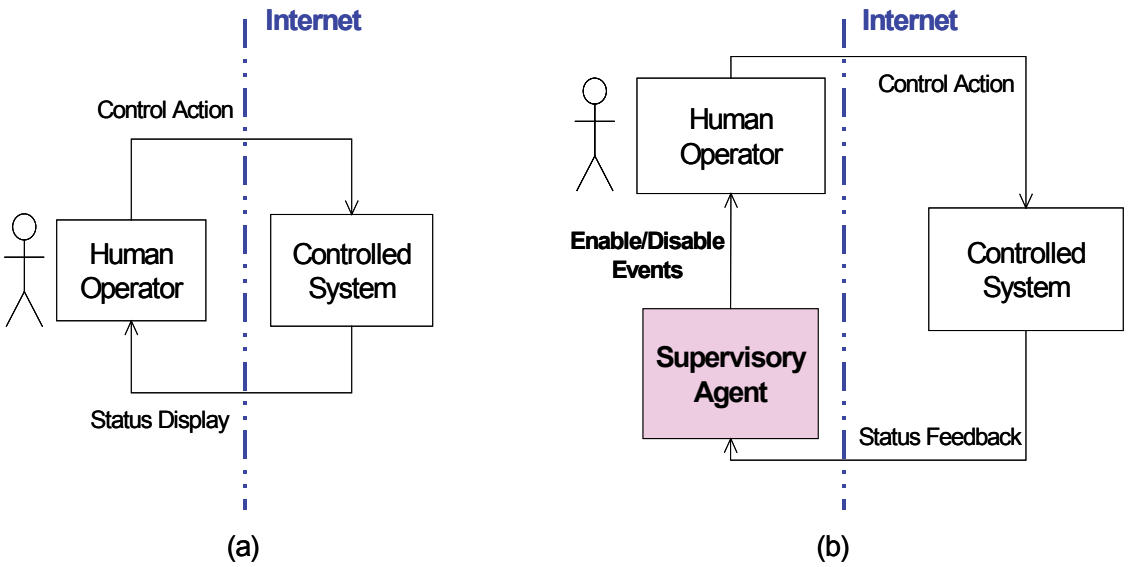


Figure 1. (a) Typical remote control system with the human in the loop. (b) The proposed remote supervisory control scheme

2. PN-Based Modelling

2.1 Basic PN Concepts

A PN is identified as a particular kind of bipartite directed graph populated by three types of objects. They are places, transitions, and directed arcs connecting places and transitions. Formally, a PN can be defined as

$$G = (P, T, I, O) \tag{1}$$

where,

$P = \{p_1, p_2, \ldots, p_m\}$ is a finite set of places, where $m > 0$;
$T = \{t_1, t_2, \ldots, t_n\}$ is a finite set of transitions with $P \cup T \neq \varnothing$ and $P \cap T = \varnothing$, where $n > 0$;
$I : P \times T \rightarrow N$ is an input function that defines a set of directed arcs from P to T, where $N = \{0, 1, 2, \ldots\}$;
$O : T \times P \rightarrow N$ is an output function that defines a set of directed arcs from T to P.

A marked PN is denoted as (G, M_0), where M_0: $P \rightarrow N$ is the initial marking. A transition t is enabled if each input place p of t contains at least the number of tokens equal to the weight of the directed arc connecting p to t. When an enabled transition fires, it removes the tokens from its input places and deposits them on its output places. PN models are suitable to represent the systems that exhibit concurrency, conflict, and synchronization.
Some important PN properties in manufacturing systems include boundedness (no capacity overflow), liveness (freedom from deadlock), conservativeness (conservation of non-consumable resources), and reversibility (cyclic behavior). The concept of liveness is closely related to the complete absence of deadlocks. A PN is said to be live if, no matter what marking has been reached from the initial marking, it is possible to ultimately fire any transition of the net by progressing through some further firing sequences. This means that a live PN guarantees deadlock-free operation, no matter what firing sequence is chosen. Validation methods of these properties include reachability analysis, invariant analysis, reduction method, siphons/traps-based approach, and simulation (Zhou and Jeng, 1998).

2.2 Elementary PN Models

At the modelling stage, one needs to focus on the major operations and their sequential or precedent, concurrent, or conflicting relationships. The basic relations among these processes or operations can be classified as follows.

1. *Sequential*: As shown in Figure 2 (a), if one operation follows the other, then the places and transitions representing them should form a cascade or sequential relation in PNs.
2. *Concurrent*: If two or more operations are initiated by an event, they form a parallel structure starting with a transition, i.e., two or more places are the outputs of a same transition. An example is shown in Figure 2 (b). The pipeline concurrent operations can be represented with a sequentially-connected series of places/transitions in which multiple places can be marked simultaneously or multiple transitions are enabled at certain markings.
3. *Cyclic*: As shown in Figure 2 (c), if a sequence of operations follow one after another and the completion of the last one initiates the first one, then a cyclic structure is formed among these operations.
4. *Conflicting*: As shown in Figure 2 (d), if either of two or more operations can follow an operation, then two or more transitions form the outputs from the same place.
5. *Mutually Exclusive*: As shown in Figure 2 (e), two processes are mutually exclusive if they cannot be performed at the same time due to constraints on the usage of shared resources. A structure to realize this is through a common place marked with one token plus multiple output and input arcs to activate these processes.

In this chapter, PN models of the human behaviours will be constructed based on these elementary models.

2.3 System Modeling

The human behaviors can be modeled using the command/response concept. As shown in Figure 3, each human operation is modeled as a task with a start transition, end transition, progressive place and completed place. Transitions drawn with dark symbols are events that are controllable by the remote-located human through the network. Note that the start transition is a control-

lable event as "command" input, while the end transition is an uncontrollable event as "response" output.

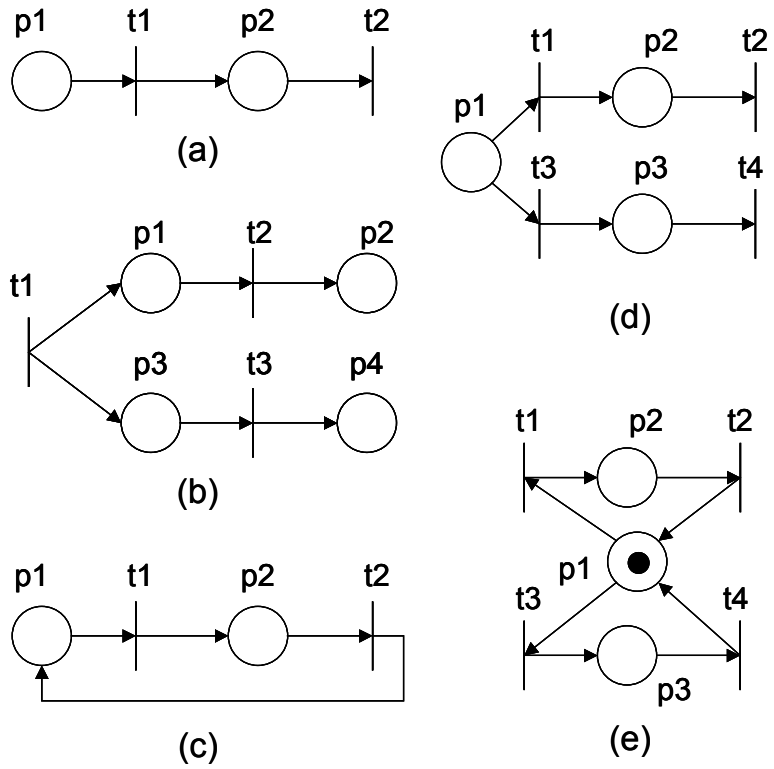


Figure 2. Basic PN models for (a) sequential, (b) concurrent, (c) cyclic, (d) conflicting, and (e) mutually exclusive relations

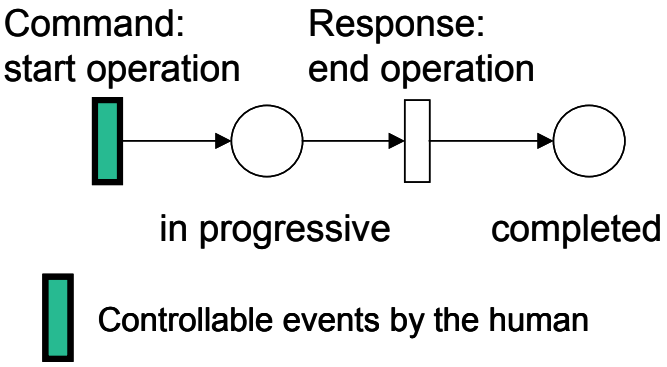


Figure 3. Modeling of human behavior using the command/response concept

3. PN-Based Supervisor Synthesis

3.1 Control Modes

For remote control via the Internet, we are interested in the following two control modes:

1. *Automatic mode*: When the system is in automatic control mode, the automatic controller autonomously controls the manufacturing process without user intervention (the human operator only needs to push a button to start the control cycle). Generally, an active sequence controller is used to automatically complete several operations in a certain order.
2. *Manual mode*: A system often must be open to manual control for various purposes, such as for test runs and fault diagnosis. Here, we examine the case in which the user can directly perform each operation. To ensure that safety constraints are not violated, the supervisory agent is on-line executed to acquire the system status and decide to either enable or disable specific operations.

3.2 Specification Types

The objective of the supervisor is to restrict the behavior of the system so that it is contained within the set of admissible states, called the specification. Two types of specifications are classified as follows:

1. *Explicit specifications for control sequences*: Generally, these specifications are "recipe-dependent". They are enforced by a sequence controller in automatic mode or by a human operator in manual mode so as to accomplish certain tasks in a desired logical order.
2. *Implicit specifications for safety requirements*: These specifications are "recipe-independent" and thus must always be obeyed throughout the whole operation of the system. Basically, these specifications are required to satisfy safety and liveness constraints. The safety specification prevents the system from performing undesirable actions, while the liveness specification ensures that a given behavior is repeatable. In automatic mode, these specifications can be effectively dealt with by the sequence controller. In manual mode, the supervisor enforces these specifications by restricting the commands available to human operators.

3.3 Supervisor Synthesis

PNs have been used to model, analyze, and synthesize control laws for DES. Zhou and DiCesare (1991), moreover, addressing the shared resource problem recognized that mutual exclusion theory plays a key role in synthesizing a bounded, live, and reversible PN. In mutual exclusion theory, parallel mutual exclusion consists of a place marked initially with one token to model a single shared resource, and a set of pairs of transitions. Each pair of transitions models a unique operation that requires the use of the shared resource.

Definition 1: Given two nets $G_1 = (P_1, T_1, I_1, O_1)$ and $G_2 = (P_2, T_2, I_2, O_2)$ with initial marking $M_{0,1}$ and $M_{0,2}$, respectively. The synchronous composition of G_1 and G_2 is a net $G = (P, T, I, O)$ with initial marking M_0:

$$G = G_1 \,||\, G_2 \tag{2}$$

where,

$P = P_1 \cup P_2$;

$T = T_1 \cup T_2$;

$I(p,t) = I_i(p,t)$ if $(\exists i \in \{1,2\})[p \in P_i \wedge t \in T_i]$, else $I(p,t) = 0$;

$O(p,t) = O_i(p,t)$ if $(\exists i \in \{1,2\})[p \in P_i \wedge t \in T_i]$, else $O(p,t) = 0$;

$M_0(p) = M_{0,1}(p)$ if $p \in P_1$, else $M_0(p) = M_{0,2}(p)$.

An agent that specifies which events are to be enabled and disabled when the system is in a given state is called a supervisor. For a system with plant model G and specification model H, the supervisor can be obtained by synchronous composition of the plant and the specification models:

$$S_G = G \,||\, H \tag{3}$$

where the transitions of H are a subset of the transitions of G, i.e. $T_H \in T_G$. Note that S_G obtained through the above construction, in the general case, does not represent a proper supervisor, since it may contain deadlock states from which a final state cannot be reached. Thus, the behavior of S should be further refined and restricted by PN analysis.

In this chapter, we adopt mutual exclusion concept to build the PN specification model and then compose it with the plant model to design the supervisor. The supervisor design procedure consists of the following steps:

Step 1) Construct the PN model of the human behaviors for system plants.
Step 2) Construct the PN model of specifications using the mutual exclusion concept for shared resources.
Step 3) Compose the behavior and specification models to synthesize the preliminary supervisor model.
Step 4) Analyze and verify the properties of the composed model.
Step 5) Refine the model to obtain a deadlock-free, bounded, and reversible model.

4. Agent-based Implementation

4.1 Agent Technology

The agent technology is a new and important technique in recent novel researches of the artificial intelligence. Using agent technology leads to a number of advantages such as scalability, event-driven actions, task-orientation, and adaptivity (Bradshaw, 1997). The concept of an agent as a computing entity is very dependent on the application domain in which it operates. As a result, there exists many definitions and theories on what actually constitutes an agent and the sufficient and necessary conditions for agency. Wooldridge and Jennings (1995) depicts an agent as a computer system that is situated in some environment, and that is capable of autonomous actions in this environment in order to meet its design objectives. From a software technology point of view, agents are similar to software objects, which however run upon call by other higher-level objects in a hierarchical structure. On the contrary, in the narrow sense, agents must run continuously and autonomously. In addition, the distributed multiagent coordination system is defined as the agents that share the desired tasks in a cooperative point of view, and they are autonomously executing at different sites. For our purposes, we have adopted the description of an agent as a software program associated to the specific function of remote supervision for the manufacturing system. A supervisory agent is implemented to acquire the system status and then enable and disable associated tasks so as to advise and guide the manager in issuing commands.

4.2 Client/Server Architecture

Figure 4 shows the client/server architecture for implementing the remote supervisory control system. On the remote client, the human operator uses a Java-capable web browser, such as Netscape Navigator or Microsoft Internet Explorer, to connect to the web server through the Internet. On the web server side, a Java servlet handles user authentication, while a Java applet is provides a graphical human/machine interface (HMI) and invokes the supervisory agent. In this chapter, we use Java technology to implement the supervisory agent on an industrial PLC, with a built-in Java-capable web server assigned to handle the client requests.

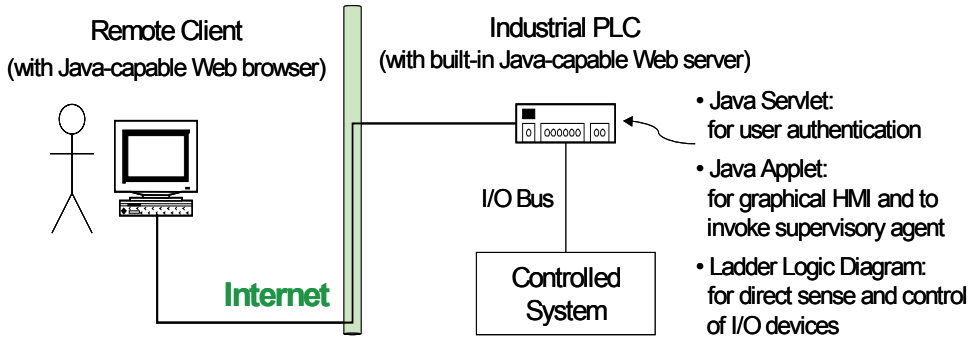


Figure 4. Implementation architecture of the supervisory control system

4.3 Interactive Modeling

A sequence diagram of the UML (Booch et al., 1999) is applied to model client/server interaction in the remote control system. Within a sequence diagram, an object is shown as a box at the top of a vertical dashed line, called the object's lifeline and representing the life of the object during the interaction. Messages are represented by horizontal arrows and are drawn chronologically from the top of the diagram to the bottom.

Figure 5 shows the sequence diagram of the implemented remote supervisory control system. At the first stage, the *Remote Client* sends a hypertext transfer protocol (HTTP) request to the *Web Server*. Next, the *Web Server* sends an HTTP response with an authentication web page, on which the *Remote Client* can

login to the system by sending a request with user/password. The *Web Server* then invokes a Java servlet to authenticate the user. If the authentication fails, the Java servlet will respond with the authentication web page again. On the other hand, if the authentication succeeds, the response of the Java servlet will be a control web page with a Java applet. The Java applet first builds a graphical HMI and constructs a socket on the specified port to maintain continuous communication with the server. Then, the Java applet acquires the system status through the constructed socket and displays it on the control web page iteratively by invoking the *Device Handler* to fetch the sensor states of *Device* objects. Finally, the supervisory agent called by the Java applet determines enable/disable control buttons on the HMI according to the current system status so as to meet the required specifications. Thus, the *Remote Client* can send an action command by pushing an enabled button to control the remote system through the constructed socket.

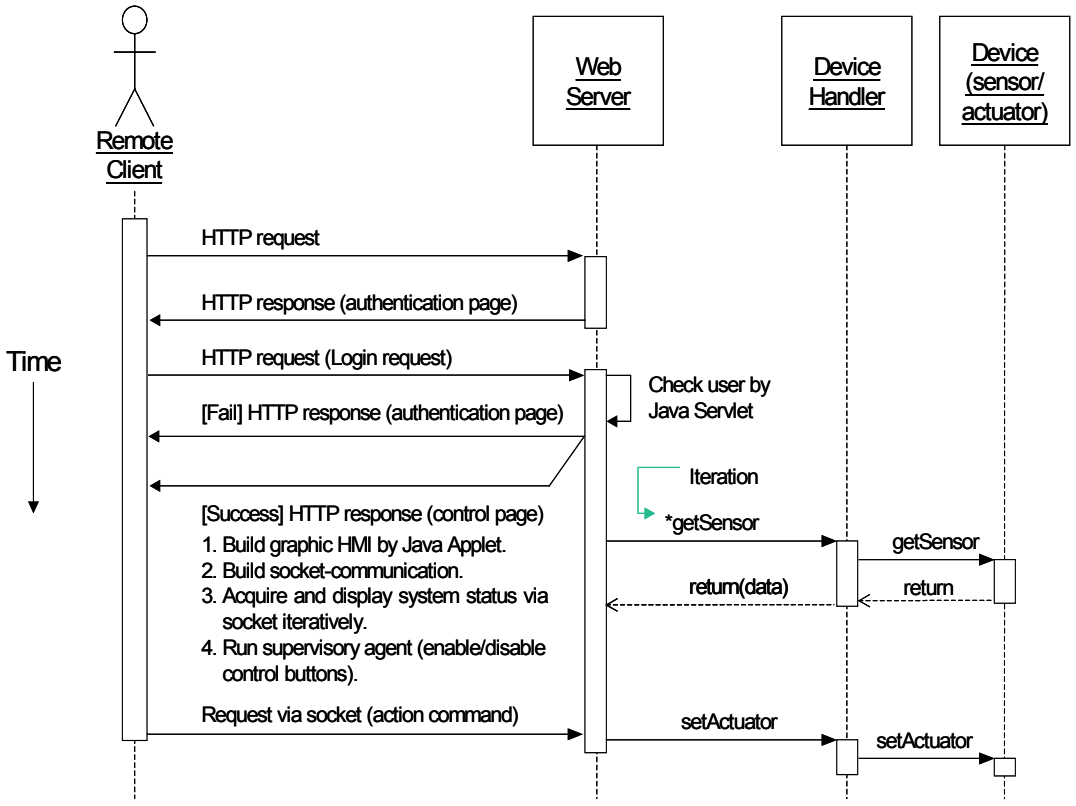


Figure 5. Interactive modeling with sequence diagram

5. Application 1: A Rapid Thermal Process

5.1 System Description

A rapid thermal processor is a relatively new semiconductor manufacturing device (Fair, 1993). A schematic diagram of the RTP system is shown in Figure 6, which is composed of 1) a reaction chamber with a door, 2) a robot arm for wafer loading/unloading, 3) a gas supply module with a mass flow controller and pressure controller-I, 4) a heating lamp module with a temperature controller, and 5) a flush pumping system with a pressure controller-II.

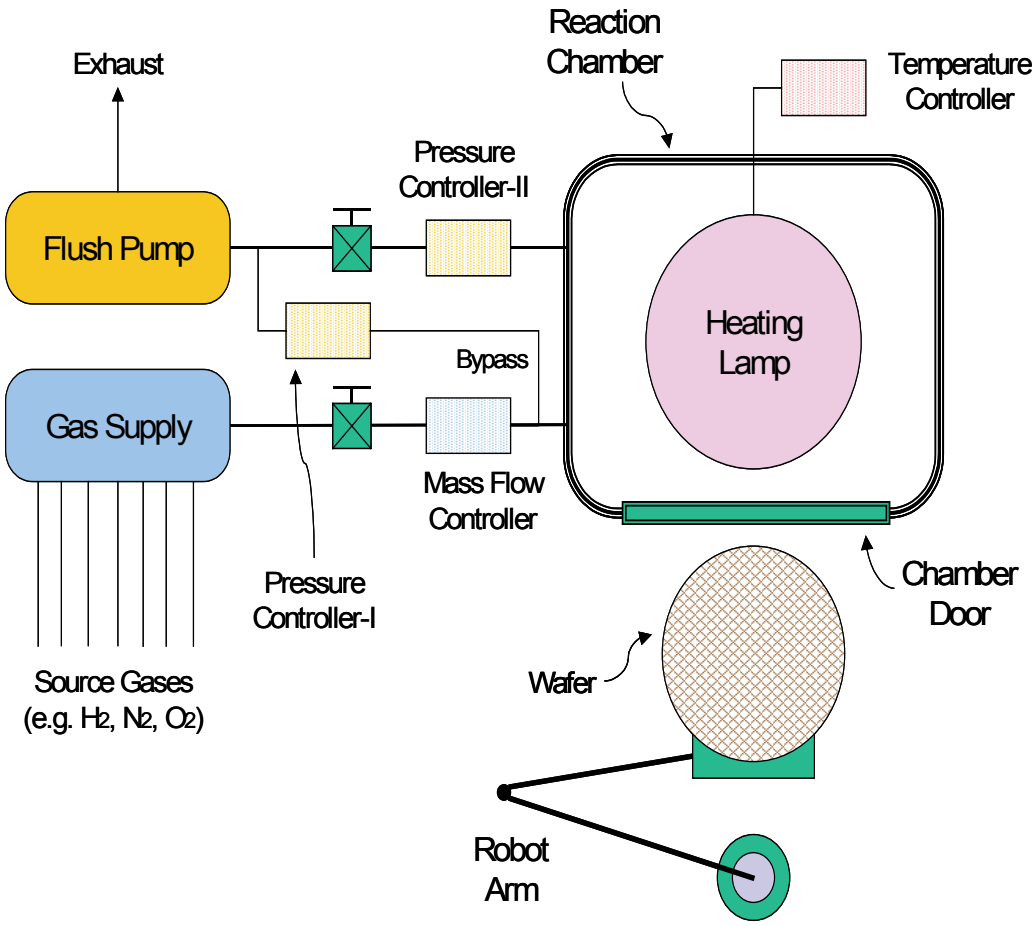


Figure 6. Schematic diagram of the RTP system

A realistic "recipe" of the hydrogen baking process, i.e. the explicit specification as mentioned in Section 3.2, is as follows:

Step 6) Load the raw wafer.

Step 7) Close the chamber door.

Step 8) Open the gas valve to supply gases with a desired gas flow rate and pressure of 2.8 liters per minute (lpm) and 0.5 Torr, respectively.

Step 9) Close the gas valve.

Step 10) Turn on the heating lamp to bake the wafer with a desired baking temperature and duration of 1000 $^{\circ}C$ and 4 seconds, respectively.

Step 11) Turn off the heating lamp to cool down the chamber to a desired temperature of less than 20 $^{\circ}C$.

Step 12) Turn on the flush pump with a desired pressure of less than 0.05 Torr.

Step 13) Turn off the flush pump.

Step 14) Open the chamber door.

Step 15) Unload the processed wafer.

The initial state of the components in the RTP is either closed or off, except that the door is open. The following safety specifications, i.e. the implicit specification mentioned in Section 3.2, must be enforced throughout system operation.

Spec-1: Wafer Loading is allowed only when no wafer is in the chamber.

Spec-2: Wafer Loading/unloading is allowed only when the door is open.

Spec-3: The gas valve must be closed when the flush pump is applied to the chamber.

Spec-4: The gas valve, heating lamp, and flush pump cannot be started when the door is open.

5.2 Automatic Controller Design

The specifications can be satisfied and involved in the sequence controller in the present automatic control mode. By applying the task-oriented concept, the PN model for the automatic control mode of the RTP is constructed as shown in Figure 7, which consists of 26 places and 20 transitions, respectively.

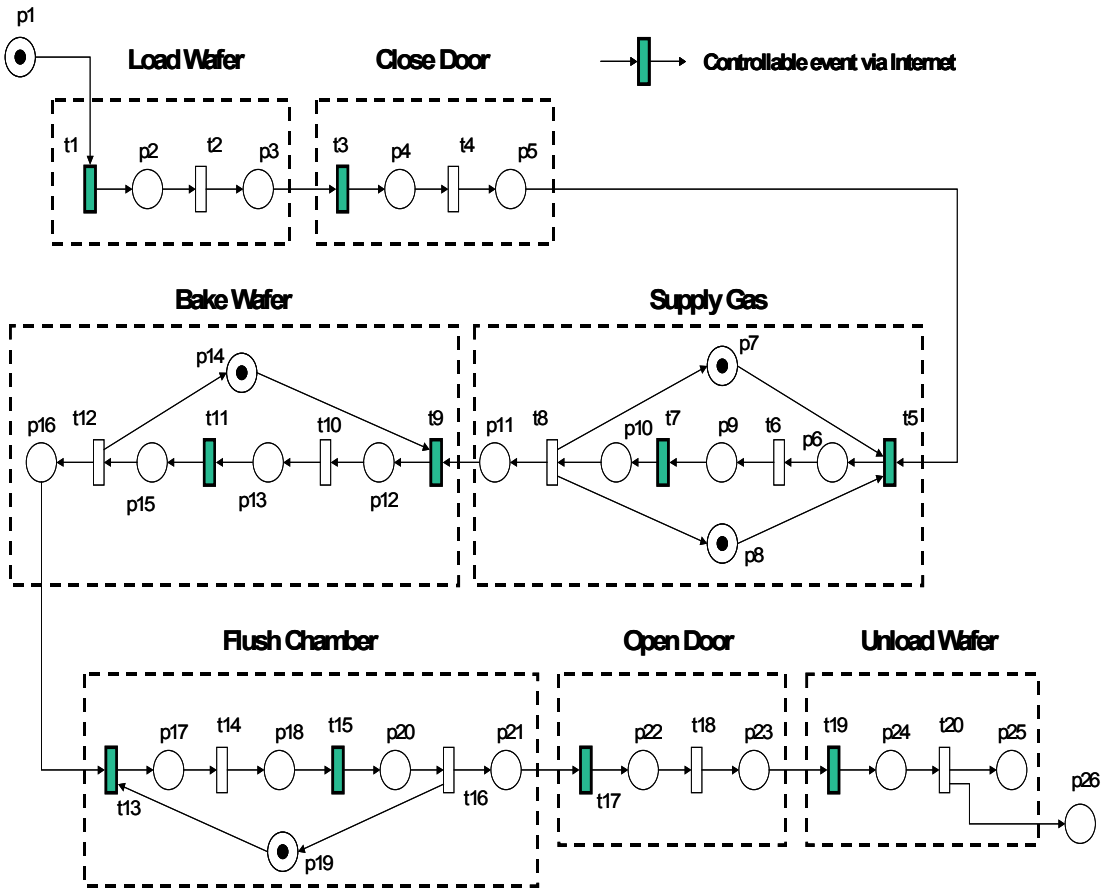


Figure 7. The PN model for automatic control of the RTP system

Corresponding notations are described in Table 1. Transitions drawn with dark symbols are events that are controllable by remote clients via the Internet.

Place	Description	Transition	Description
p1	Raw wafer buffer	t1	Cmd: start loading wafer
p2	Loading wafer	t2	Re: end loading wafer
p3	Loading wafer completed	t3	Cmd: start closing chamber door
p4	Closing chamber door	t4	Re: end closing chamber door
p5	Closing chamber door completed	t5	Cmd: start opening gas valve
p6	Opening gas valve	t6	Re: end opening gas valve
p7	Mass flow controller ready	t7	Cmd: start closing gas valve
p8	Pressure controller-I ready	t8	Re: end closing gas valve
p9	Opening gas valve completed	t9	Cmd: start turning on heating lamp
p10	Closing gas valve	t10	Re: end turning on heating lamp
p11	Closing gas valve completed	t11	Cmd: start turning off heating lamp
p12	Turning on heating lamp	t12	Re: end turning off heating lamp
p13	Turning on heating lamp completed	t13	Cmd: start turning on flush pump
p14	Temperature controller ready	t14	Re: end turning on flush pump
p15	Turning off heating lamp	t15	Cmd: start turning off flush pump
p16	Turning off heating lamp completed	t16	Re: end turning off flush pump
p17	Turning on flush pump	t17	Cmd: start opening chamber door
p18	Turning on flush pump completed	t18	Re: end opening chamber door
p19	Pressure controller-II ready	t19	Cmd: start unloading wafer
p20	Turning off flush pump	t20	Re: end unloading wafer
p21	Turning off flush pump completed		
p22	Opening chamber door		
p23	Opening chamber door completed		
p24	Unloading wafer		
p25	Unloading wafer completed		
p26	Processed wafer buffer		

Table 1. Notations for the PN of the RTP system in Figure 7

5.3 Supervisor Synthesis

For manual control mode, the plant model is formed by unconnecting each pair of transitions for the tasks in Figure 7. In the specification model, Spec-1 and Spec-2 are modeled as the pre-conditions of the associated operations, while Spec-3 and Spec-4 are built by using the mutual exclusion concept. The

composed PN model of both the plant and specifications is shown in Figure 8, where A-J represent ten remote controllable tasks for the RTP system. The supervisory places **ps1-7** (**ps1** for Spec-1, **ps2-3** for Spec-2, **ps4** for Spec-3, **ps5-7** for Spec-4) are used to prevent undesired and unsafe operations on the part of the human operator. Corresponding notations for the supervisory places are described in Table 2. At this stage, the software package ARP (Maziero, 1990) is chosen to verify the behavioral properties of the composed PN model due to its graphical representation, ease of manipulation, and ability to perform structural and performance analyses. The ARP uses the reachability analysis to validate the PN properties. Results reveal that the present PN model is live and bounded. The liveness property means that the system can be executed properly without deadlocks, while the boundedness property means that the system can be executed with limited resources (e.g., limited buffer sizes).

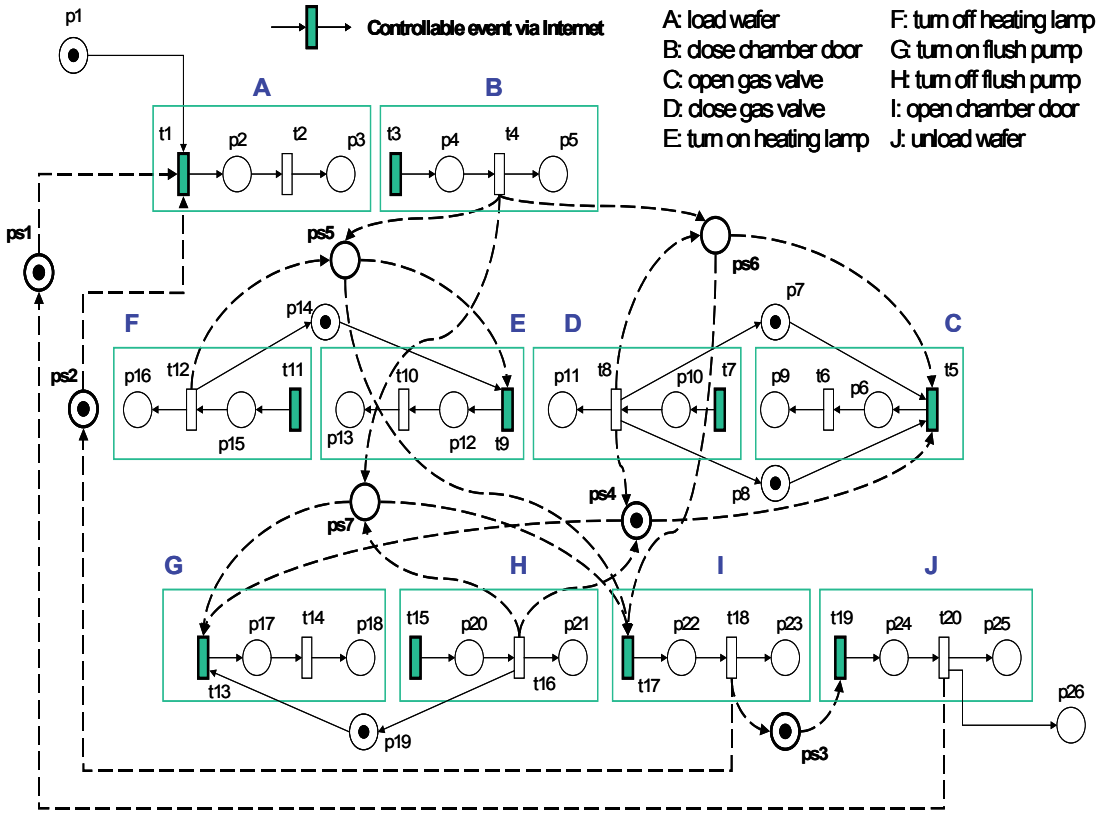


Figure 8. The composed PN model for manual control of the RTP system

Place	Description
ps1	Spec-1: chamber is empty
ps2	Spec-2: chamber door is open
ps3	Spec-2: chamber door is open
ps4	Spec-3: gas is closed/pump is off
ps5	Spec-4: door is closed/lamp is off
ps6	Spec-4: door is closed/gas is closed
ps7	Spec-4: door is closed/pump is off

Table 2. Notations for supervisory places of PN in Figure 8

5.4 Implementation with Agent Technology

The system modeling and design developed in previous stages provide supervisory control models for implementation with agent technology. The developed supervisory agent is implemented on the Mirle SoftPLC (80486-100 CPU), an advanced industrial PLC with built-in Web server and Java virtual machine so that it can interpret the LLD, HTTP requests, and Java programs (Mirle Automation Corporation, 1999; SoftPLC Corporation, 1999).

The developed HMI, shown in Figure 9, is carefully designed to make its web pages more user-friendly and also to increase download speed by avoiding unnecessary images. Since the client users will be mainly operators and engineers, they will want effective information delivery and will not be interested in flashy graphics (Shikli, 1997).

The current system status is placed on the left, the system message is in the center, and the button control area is on the right. Figure 9 also shows the web pages for manual control mode after the **Open Valve** button has just been pushed (Step 3 in Section 5.1). In this situation, since one wafer is already in the chamber and the door is closed, the **Load Wafer** and **Unload Wafer** buttons are both disabled by the supervisory agent to meet Spec-1 and Spec-2. Moreover, the **Turn_On Pump** and **Open Door** buttons are disabled to meet Spec-3 and Spec-4, respectively. Thus, the safety requirements of the RTP processing are guaranteed as human operations are conducted.

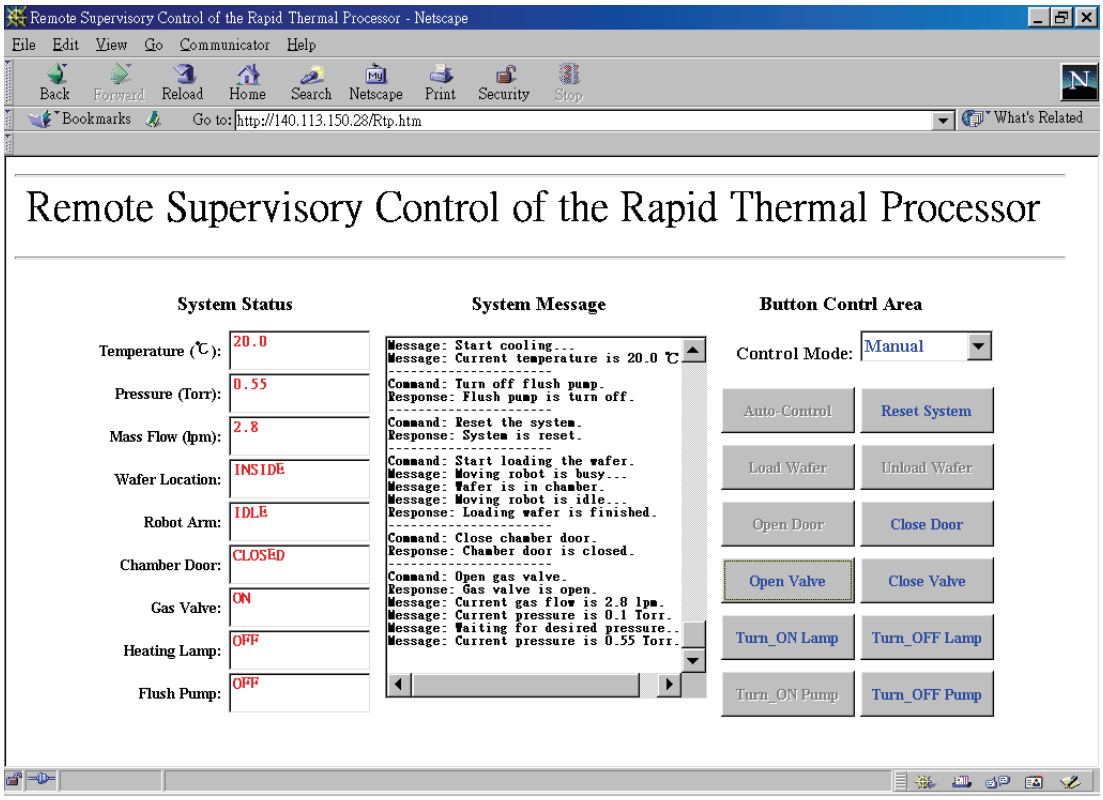


Figure 9. Interactive web page in manual control mode at Step 3 of RTP processing (seven buttons are enabled)

6. Application 2: A Two-Robot Remote Surveillance System

6.1 System Description

Figure 10 shows a human-computer interactive system (HCIS), in which a human operator issues a command to trigger a human-controlled (semi-autonomous) robot and a computer controller automatically regulates a computer-controlled (fullly autonomous) robot both with the status feedback from the overall controlled system (i.e. both robots) through a network. Such HCIS can be applied as a remote surveillance system, which is composed of one human-controlled robot (simplified as Robot-h) and one computer-controlled robot (simplified as Robot-c). These two robots are placed on a floor with five rooms, and the moving directions for each robot are shown in Figure 11, respectively.

The Robot-h and Robot-c must traverse each doorway in the direction indicated. Moreover, in order to avoid possible collisions, Robot-h and Robot-c are not allowed simultaneously in the same room during the surveillance period. The initial states of the Robot-h and Robot-c are in R5 and R2, respectively.

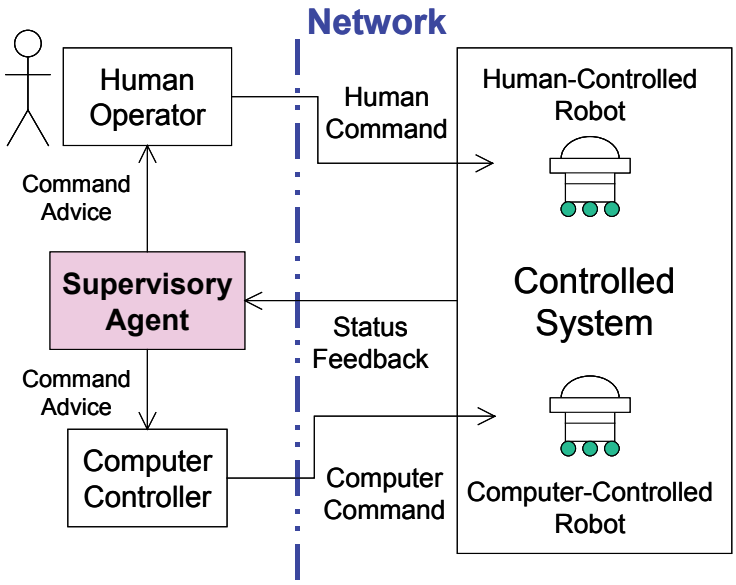


Figure 10. A two-robot human-computer interactive system

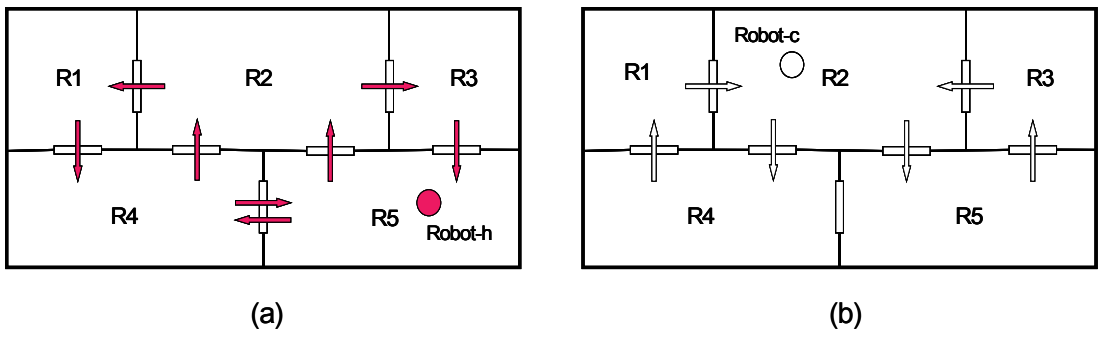


Figure 11. The schematic diagram of the two-robot remote surveillance system with the moving directions for (a) human-controlled robot, and (b) computer-controlled robot.

6.2 PN-Based System Modeling

By applying the command/response concept and based on the system description, the PN model for the human-controlled robot is constructed as shown in Figure 12 (a). It consists of 13 places and 16 transitions, respectively. On the other hand, for the computer-controlled robot, the PN model is directly built according to its located room, as shown in Figure 12 (b), which respectively consists of 5 places and 6 transitions. Corresponding notation of both the PN models is described in Table 3.

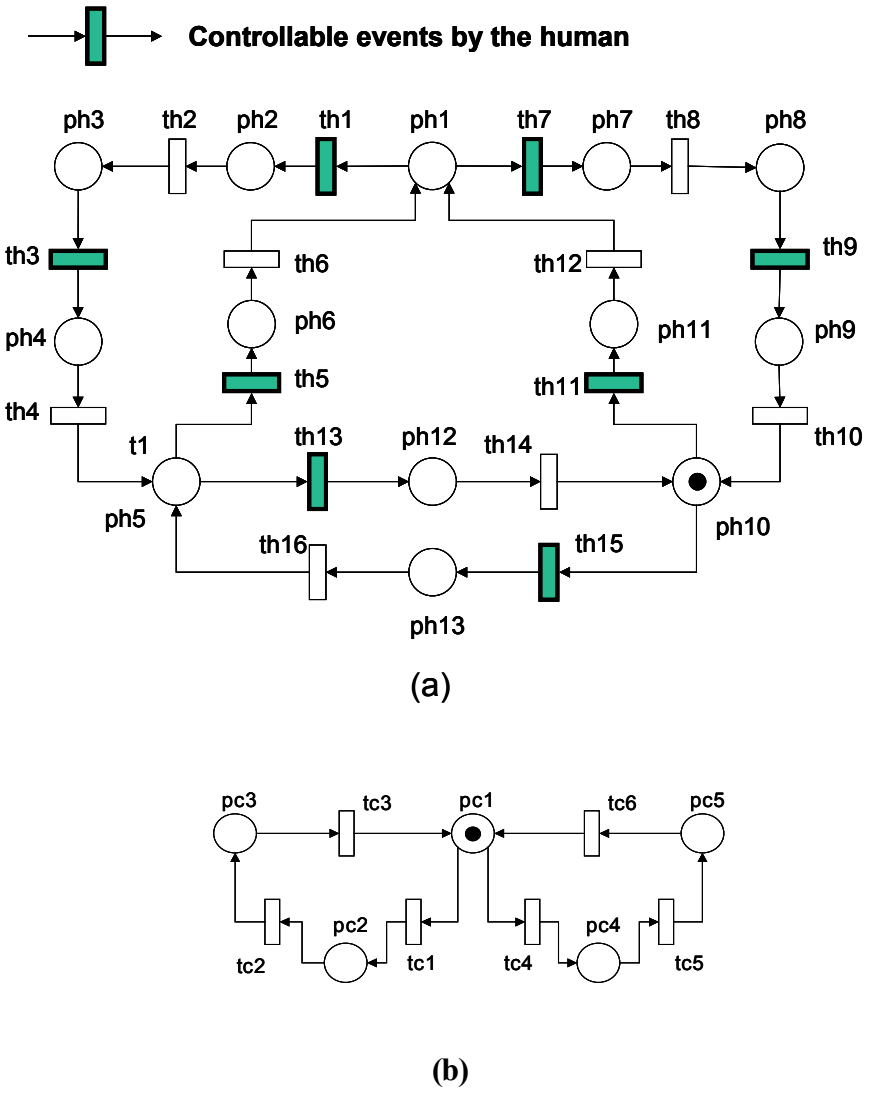


Figure 12. PN models of (a) human-controlled robot, and (b) computer-controlled robot

Place	Description	Transition	Description
ph1	Robot-h is in R2	th1	Cmd: start moving to R1
ph2	Moving to R1	th2	Re: end moving to R1
ph3	Robot-h is in R1	th3	Cmd: start moving to R4
ph4	Moving to R4	th4	Re: end moving to R4
ph5	Robot-h is in R4	th5	Cmd: start moving to R2
ph6	Moving to R2	th6	Re: end moving to R2
ph7	Moving to R3	th7	Cmd: start moving to R3
ph8	Robot-h is in R3	th8	Re: end moving to R3
ph9	Moving to R5	th9	Cmd: start moving to R5
ph10	Robot-h is in R5	th10	Re: end moving to R5
ph11	Moving to R2	th11	Cmd: start moving to R2
ph12	Moving to R5	th12	Re: end moving to R2
ph13	Moving to R4	th13	Cmd: start moving to R5
		th14	Re: end moving to R5
		th15	Cmd: start moving to R4
		th16	Re: end moving to R4
pc1	Robot-c is in R2	tc1	Move to R4
pc2	Robot-c is in R4	tc2	Move to R1
pc3	Robot-c is in R1	tc3	Move to R2
pc4	Robot-c is in R5	tc4	Move to R5
pc5	Robot-c is in R3	tc5	Move to R3
		tc6	Move to R2

Table 3. Notation for the Petri nets in Figure 12

6.3 PN-Based Supervisor Synthesis

The five rooms represent the resources shared by the two robots. Since more than one robot may require access to the same room, but in order to avoid collisions, each room can only be allowed to have one robot at a time, operations with collisions and deadlocks may thus occur. Hence, the objective is to design a supervisor to insure the whole system against these undesired situations. The required two main specifications are formulated as follows:

Spec-1: *Collision-free motions*: Robot-h or Robot-c moving to Room i is allowed only when Room i is available, where i = 1, 2, …, 5. Thus, we have five sub-specifications denoted as Spec-1.1 to Spec-1.5.

Spec-2: *Deadlock-free operations*: No deadlock states occur throughout system operation.

In the specification models, Spec-1.1 to Spec-1.5 are enforced by using the mutual exclusion concept. The composed PN model of both the systems and specifications is shown in Figure 13. The supervisory arcs are shown with dashed lines and the places showing the supervisory positions are drawn thicker than those showing the system positions. A supervisory place is modeled as an input place of the transitions that need such a resource, and as an output place of those that release this resource. Take an example of **ps1** that physically means Room 1 being available. It makes two transitions *th1* and *tc2* mutually exclusive. Intuitively, performance of *th1* is only allowed if Room 1 is available and *tc2* has not yet been fired. If *tc2* has been fired, *th1* cannot be executed until *tc3* is given to signal that Room 1 is available again. Thus, only one robot is allowed to be in Room 1 at any time, thereby avoiding the collision there.

The supervisory places **ps1** to **ps5** (for Spec-1.1 to Spec-1.5, respectively) are used to prevent the remote human operator and computer controller from issuing undesired commands leading to resource conflicts on the part of the system. The corresponding notation for the supervisory places (**ps1**-**ps5**) is described in Table 4.

Place	Description
ps1	Spec-1.1: R1 is available.
ps2	Spec-1.2: R2 is available.
ps3	Spec-1.3: R3 is available.
ps4	Spec-1.4: R4 is available.
ps5	Spec-1.5: R5 is available.
ps6	Spec-2.1: Robot-h is admitted into R1. Robot-c is admitted into R4.
ps7	Spec-2.2: Robot-h is admitted into R3. Robot-c is admitted into R5.

Table 4. Notation for the supervisory places in Figure 13

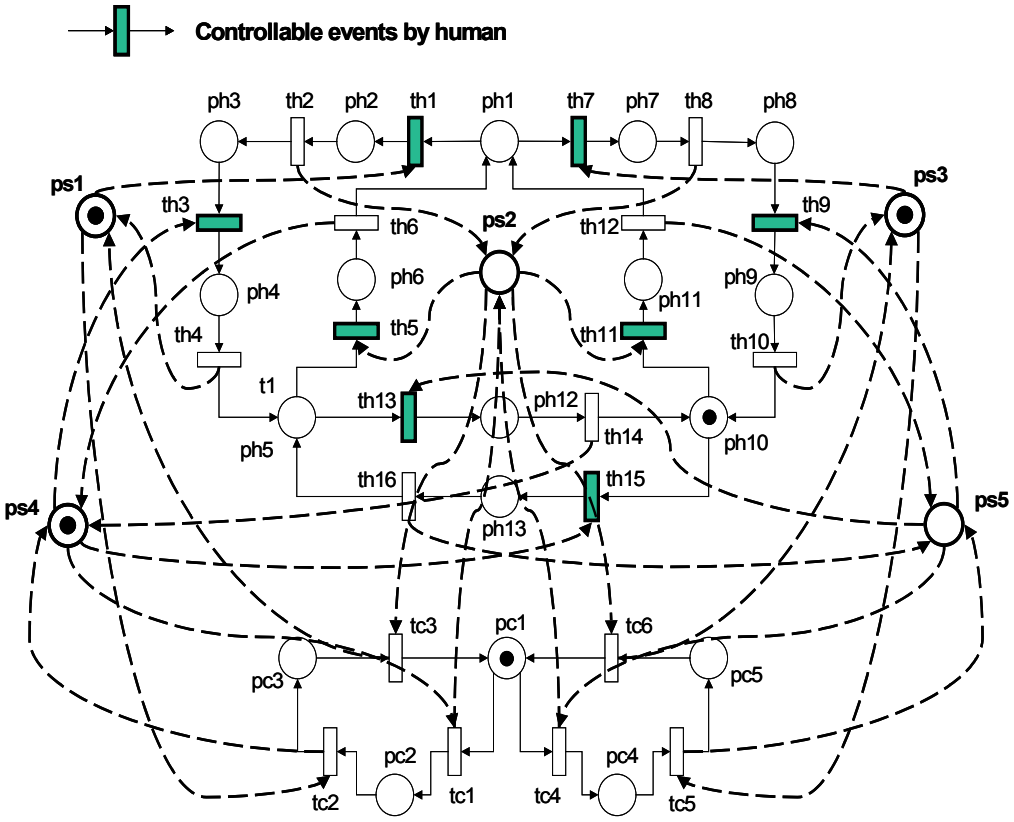


Figure 13. Preliminary composed PN model of the remote surveillance system

6.4 System Verification and Deadlock Resolution

Again, the software package ARP (Maziero, 1990) is used to verify the behavioral properties of the composed PN model using the reachability analysis. The validation result shows that two deadlocks occur with the marked places {ph3, pc2, **ps2**, **ps3**, **ps5**} and {ph8, pc4, **ps1**, **ps2**, **ps4**}, respectively. Figure 14 shows the real situations of the two deadlock states, of which the physical meaning is that if Room 1 (or Room 3) is occupied with Robot-h and Room 4 (or Room 5) is held by Robot-c, respectively, then no new events can be fired by the human or computer, and the system is deadlocked. Hence, for deadlock-free requirements, Spec-2 has two sub-specifications as follows:

Spec-2.1: Robot-h is allowed to enter Room 1 only when Robot-c is not in Room 4, and vice versa.

Spec-2.2: Robot-h is allowed to enter Room 3 only when Robot-c is not in Room 5, and vice versa.

As shown in Figure 15, **ps6** and **ps7** are further designed by using the mutual exclusion concept and then combined with the PN model in Figure 13. Take an example of **ps6**. It makes transitions *th1* and *tc1* mutually exclusive. That means either Robot-h moving to R1 or Robot-c moving to R4 is allowed to perform at a time. If *tc1* has been fired, *th1* cannot be executed until *tc2* is given to signal that Robot-c is not in R4.

Validation results (with **ps6** and **ps7**) reveal that the present PN model is deadlock-free, bounded, and reversible. The deadlock-free property means that the system can be executed properly without deadlocks, while boundedness indicates that the system can be executed with limited resources, and reversibility implies that the initial system configuration is always reachable. The corresponding notation for the supervisory places (**ps6** and **ps7**) is described in Table 4.

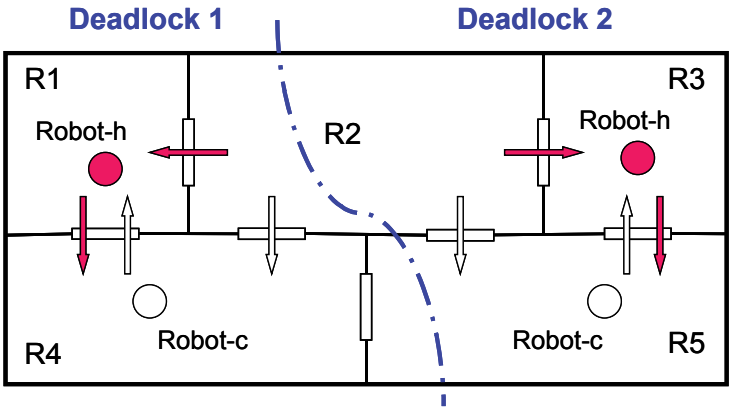


Figure 14. Two deadlock states of the PN model in the Figure 13

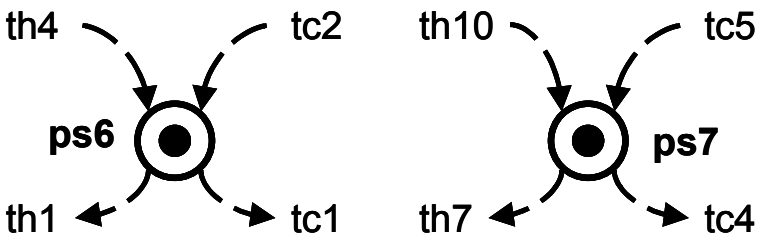


Figure 15. Supervisory places for the deadlock resolution

6.5 Discussions

On the part of the human-controlled robot, in the proposed supervisory framework, the human behavior is advised and restricted to satisfy the specifications so that the collision and deadlock are avoid during the surveillance period. As shown in Table 5, without supervisory control, the state space is 65, including the undesired collision and deadlock states. By using our proposed approach, in the preliminary supervision, i.e., only the collision-free specification (Spec-1.1 to Spec-1.5) is enforced, the state space reduces to 44. Finally, with the deadlock resolution, the state space is limited to 40 only. That means the undesired collision and deadlock states will be successfully avoided during the surveillance period. In this approach, the supervisor only consists of places and arcs, and its size is proportional to the number of specifications that must be satisfied.

Petri net models	Unsupervised system	Preliminary supervision (with deadlocks)	Complete supervision (deadlock-free)
Places	18	23	25
Transitions	22	22	22
State space	65	44	40

Table 5. Comparison between unsupervised and supervised systems

7. Conclusion

This chapter has presented a PN-based framework to design supervisors for human-in-the-loop systems. The supervisor is systematically synthesized to enforce the requirements. To demonstrate the practicability of the proposed supervisory approach, an application to 1) the RTP system in semiconductor manufacturing controlled over the Internet and 2) the two-robot remote surveillance system are provided. According to the feedback status of the remotely located system, the designed supervisory agent guarantees that all requested commands satisfy the desired specifications. On the part of human-

controlled systems, the developed supervisor can be implemented as an intelligent agent to advise and guide the human operator in issuing commands by enabling or disabling the associated human-controlled buttons. Hence, for human-in-the-loop systems, the proposed approach would be also beneficial to the human-machine interface design.

Future work includes the extension of specifications to timing constraints, the multiple-operator access, and error recovery functions. Moreover, constructive definition of the synthesis algorithm should be investigated. Also, for the scalability of the supervisor synthesis, the hierarchical design can be further applied to more complex and large-scale systems.

8. References

Balemi, S.; Hoffmann, G. J.; Gyugyi, P.; Wong-Toi, H. & Franklin, G. F. (1993). Supervisory control of a rapid thermal multiprocessor. *IEEE Trans. Automat. Contr.*, Vol. 38, No. 7, pp. 1040-1059.

Booch, G.; Rumbaugh, J. & Jacobson, I. (1999). *The Unified Modeling Language User Guide*, Addison-Wesley, Reading, MA.

Bradshaw, J. M. (1997), Introduction to software agents, *Software Agents*, Bradshaw, J. M. Ed., Cambridge, MA: AAAI Press/MIT Press.

David, R. & Alla, H. (1994), Petri nets for modeling of dynamics systems– A survey, *Automatica*, Vol. 30, No. 2, pp. 175-202.

Fair, R. B. (1993), *Rapid Thermal Processing: Science and Technology*, New York: Academic.

Giua, A. & DiCesare, F. (1991), Supervisory design using Petri nets, *Proceedings of IEEE Int. Conf. Decision Contr.*, pp. 92-97, Brighton, England.

Huang, G. Q. & Mak, K. L. (2001), Web-integrated manufacturing: recent developments and emerging issues, *Int. J. Comput. Integrated Manuf.*, Vol. 14, No. 1, pp. 3-13, (Special issue on Web-integrated manufacturing).

Kress, R. L., Hamel, W. R., Murray, P. & Bills, K. (2001), Control strategies for teleoperated Internet assembly, *IEEE/ASME Trans. Mechatronics*, Vol. 6, No. 4, pp. 410-416, (Focused section on Internet-based manufacturing systems).

Lee, J. S. & Hsu, P. L. (2003), Remote supervisory control of the human-in-the-loop system by using Petri nets and Java, *IEEE Trans. Indu. Electron.*, Vol. 50, No. 3, pp. 431-439.

Lee, J. S. & Hsu, P. L. (2004), Design and implementation of the SNMP agents for remote monitoring and control via UML and Petri nets, *IEEE Trans. Contr. Syst. Technol.*, Vol. 12, No. 2, pp. 293-302.

Lee, J. S.; Zhou M. C. & Hsu P. L. (2005), An application of Petri nets to supervisory control for human-computer interactive systems, *IEEE Transactions on Industrial Electronics*, Vol. 52, No. 5, pp. 1220-1226.

Maziero, C. A. (1990), *ARP: Petri Net Analyzer*. Control and Microinformatic Laboratory, Federal University of Santa Catarina, Brazil.

Milner R. (1989), *Communication and Concurrency*. Englewood Cliffs, NJ: Prentice Hall.

Mirle Automation Corporation (1999), *SoftPLC Controller User's Manual Version 1.2*. Hsinchu, Taiwan.

Moody, J. O. & Antsaklis, P. J. (1998), *Supervisory Control of Discrete Event systems Using Petri Nets.* Boston, MA: Kluwer.

Murata, T. (1989), Petri nets: Properties, analysis, and applications, *Proc. IEEE*, Vol. 77, No. 4, pp. 541-580.

Petri, C. A. (1962), *Kommunikation mit Automaten*. Bonn: Institut für Instrumentelle Mathematik, Schriften des IIM Nr. 2. English translation, *Communication with Automata*. New York: Griffiss Air Force Base, Tech.l Rep. RADC-TR-65--377, Vol. 1, pages 1-Suppl. 1. 1966.

Ramadge, P. J. & Wonham, W. M. (1987), Supervisory control of a class of discrete event processes, *SIAM J. Contr. Optimiz.*, Vol. 25, No. 1, pp. 206-230.

Ramadge, P. J. & Wonham, W. M. (1989), The control of discrete event systems, *Proc. IEEE*, Vol. 77, No. 1, pp. 81-98.

Rasmussen, J., Pejtersen, A. M. & Goodstein, L. P. (1994), *Cognitive Systems Engineering*. New York, NY: John Wiley and Sons.

Shikli, P. (1997), Designing winning Web sites for engineers, *Machine Design*, Vol. 69, No. 21, pp. 30-40.

SoftPLC Corporation (1999), *SoftPLC-Java Programmer's Toolkit*. Spicewood, TX.

Uzam, M., Jones, A. H. & Yücel, I. (2000), Using a Petri-net-based approach for the real-time supervisory control of an experimental manufacturing system, *Int. J. Adv. Manuf. Tech.*, Vol. 16, No. 7, pp. 498-515.

Weaver, A., Luo, J. & Zhang, X. (1999), Monitoring and control using the Internet and Java, *Proceedings of IEEE Int. Conf. Industrial Electronics*, pp. 1152-1158, San Jose, CA.

Wooldridge, M. & Jenkins, M. R. (1995), Intelligent agents: theory and practice, *Knowledge Engineering Review*, Vol. 10, No. 2, pp. 115–152.

Yang, S. H., Chen, X. & Alty, J. L. (2002), Design issues and implementation of Internet-based process control systems, *Contr. Engin. Pract.*, Vol. 11, No. 6, pp. 709-720.

Zhou, M. C. & DiCesare, F. (1991), Parallel and sequential mutual exclusions for Petri net modeling for manufacturing systems, *IEEE Trans. Robot. Automat.*, Vol. 7, No. 4, pp. 515-527.

Zhou, M. C. & Jeng, M. D. (1998), Modeling, analysis, simulation, scheduling, and control of semiconductor manufacturing systems: A Petri net approach, *IEEE Trans. Semicond. Manuf.*, Vol. 11, No. 3, pp. 333-357, (Special section on Petri nets in semiconductor manufacturing).

Zurawski, R. & Zhou, M. C. (1994), Petri nets and industrial applications: a tutorial, *IEEE Trans. Ind. Electron.*, Vol. 41, No. 6, pp. 567-583, (Special section on Petri nets in manufacturing).

8

Application Similarity Coefficient Method to Cellular Manufacturing

Yong Yin

1. Introduction

Group technology (GT) is a manufacturing philosophy that has attracted a lot of attention because of its positive impacts in the batch-type production. Cellular manufacturing (CM) is one of the applications of GT principles to manufacturing. In the design of a CM system, similar parts are groups into families and associated machines into groups so that one or more part families can be processed within a single machine group. The process of determining part families and machine groups is referred to as the cell formation (CF) problem.
CM has been considered as an alternative to conventional batch-type manufacturing where different products are produced intermittently in small lot sizes. For batch manufacturing, the volume of any particular part may not be enough to require a dedicated production line for that part. Alternatively, the total volume for a family of similar parts may be enough to efficiently utilize a machine-cell (Miltenburg and Zhang, 1991).
It has been reported (Seifoddini, 1989a) that employing CM may help overcome major problems of batch-type manufacturing including frequent setups, excessive in-process inventories, long through-put times, complex planning and control functions, and provides the basis for implementation of manufacturing techniques such as just-in-time (JIT) and flexible manufacturing systems (FMS).

A large number of studies related to GT/CM have been performed both in academia and industry. Reisman *et al*. (1997) gave a statistical review of 235 articles dealing with GT and CM over the years 1965 through 1995. They reported that the early (1966-1975) literature dealing with GT/CM appeared predominantly in book form. The first written material on GT was Mitrofanov (1966) and the first journal paper that clearly belonged to CM appeared in 1969 (Optiz *et al*., 1969). Reisman *et al*. (1997) also reviewed and classified these 235 articles on a five-point scale, ranging from pure theory to bona fide applications.

In addition, they analyzed seven types of research processes used by authors. There are many researchable topics related to cellular manufacturing. Wemmerlöv and Hyer (1987) presented four important decision areas for group technology adoption – applicability, justification, system design, and implementation. A list of some critical questions was given for each area.

Applicability, in a narrow sense, can be understood as feasibility (Wemmerlöv and Hyer, 1987). Shafer *et al*. (1995) developed a taxonomy to categorize manufacturing cells. They suggested three general cell types: process cells, product cells, and other types of cells. They also defined four shop layout types: product cell layouts, process cell layouts, hybrid layouts, and mixture layouts. Despite the growing attraction of cellular manufacturing, most manufacturing systems are hybrid systems (Wemmerlöv and Hyer, 1987; Shambu and Suresh, 2000). A hybrid CM system is a combination of both a functional layout and a cellular layout. Some hybrid CM systems are unavoidable, since some processes such as painting or heat treatment are frequently more efficient and economic to keep the manufacturing facilities in a functional layout.
Implementation of a CM system contains various aspects such as human, education, environment, technology, organization, management, evaluation and even culture. Unfortunately, only a few papers have been published related to these areas. Researches reported on the human aspect can be found in Fazakerley (1976), Burbidge *et al*. (1991), Beatty (1992), and Sevier (1992). Some recent studies on implementation of CM systems are Silveira (1999), and Wemmerlöv and Johnson (1997; 2000).

The problem involved in justification of cellular manufacturing systems has received a lot of attention. Much of the research was focused on the performance comparison between cellular layout and functional layout. A number of researchers support the relative performance supremacy of cellular layout over functional layout, while others doubt this supremacy. Agarwal and Sarkis (1998) gave a review and analysis of comparative performance studies on functional and CM layouts. Shambu and Suresh (2000) studied the performance of hybrid CM systems through a computer simulation investigation.

System design is the most researched area related to CM. Research topics in this area include cell formation (CF), cell layout (Kusiak and Heragu, 1987; Balakrishnan and Cheng; 1998; Liggett, 2000), production planning (Mosier and Taube, 1985a; Singh, 1996), and others (Lashkari *et al*, 2004; Solimanpur *et al*, 2004). CF is the first, most researched topic in designing a CM system. Many approaches and methods have been proposed to solve the CF problem. Among

these methods, Production flow analysis (PFA) is the first one which was used by Burbidge (1971) to rearrange a machine part incidence matrix on trial and error until an acceptable solution is found. Several review papers have been published to classify and evaluate various approaches for CF, some of them will be discussed in this paper. Among various cell formation models, those based on the similarity coefficient method (SCM) are more flexible in incorporating manufacturing data into the machine-cells formation process (Seifoddini, 1989a). In this paper, an attempt has been made to develop a taxonomy for a comprehensive review of almost all similarity coefficients used for solving the cell formation problem.

Although numerous CF methods have been proposed, fewer comparative studies have been done to evaluate the robustness of various methods. Part reason is that different CF methods include different production factors, such as machine requirement, setup times, utilization, workload, setup cost, capacity, part alternative routings, and operation sequences. Selim, Askin and Vakharia (1998) emphasized the necessity to evaluate and compare different CF methods based on the applicability, availability, and practicability. Previous comparative studies include Mosier (1989), Chu and Tsai (1990), Shafer and Meredith (1990), Miltenburg and Zhang (1991), Shafer and Rogers (1993), Seifoddini and Hsu (1994), and Vakharia and Wemmerlöv (1995).

Among the above seven comparative studies, Chu and Tsai (1990) examined three array-based clustering algorithms: rank order clustering (ROC) (King, 1980), direct clustering analysis (DCA) (Chan & Milner, 1982), and bond energy analysis (BEA) (McCormick, Schweitzer & White, 1972); Shafer and Meredith (1990) investigated six cell formation procedures: ROC, DCA, cluster identification algorithm (CIA) (Kusiak & Chow, 1987), single linkage clustering (SLC), average linkage clustering (ALC), and an operation sequences based similarity coefficient (Vakharia & Wemmerlöv, 1990); Miltenburg and Zhang (1991) compared nine cell formation procedures. Some of the compared procedures are combinations of two different algorithms *A1*/*A2*. *A1*/*A2* denotes using *A1* (algorithm 1) to group machines and using *A2* (algorithm 2) to group parts. The nine procedures include: ROC, SLC/ROC, SLC/SLC, ALC/ROC, ALC/ALC, modified ROC (MODROC) (Chandrasekharan & Rajagopalan, 1986b), ideal seed non-hierarchical clustering (ISNC) (Chandrasekharan & Rajagopalan, 1986a), SLC/ISNC, and BEA.

The other four comparative studies evaluated several similarity coefficients. We will discuss them in the later section.

2. Background

This section gives a general background of machine-part CF models and detailed algorithmic procedures of the similarity coefficient methods.

2.1 Machine-part cell formation

The CF problem can be defined as: "If the number, types, and capacities of production machines, the number and types of parts to be manufactured, and the routing plans and machine standards for each part are known, which machines and their associated parts should be grouped together to form cell?" (Wei and Gaither, 1990). Numerous algorithms, heuristic or non-heuristic, have emerged to solve the cell formation problem. A number of researchers have published review studies for existing CF literature (refer to King and Nakornchai, 1982; Kumar and Vannelli, 1983; Mosier and Taube, 1985a; Wemmerlöv and Hyer, 1986; Chu and Pan, 1988; Chu, 1989; Lashkari and Gunasingh, 1990; Kamrani *et al*., 1993; Singh, 1993; Offodile *et al*., 1994; Reisman *et al*., 1997; Selim *et al*., 1998; Mansouri *et al*., 2000). Some timely reviews are summarized as follows.

Singh (1993) categorized numerous CF methods into the following sub-groups: part coding and classifications, machine-component group analysis, similarity coefficients, knowledge-based, mathematical programming, fuzzy clustering, neural networks, and heuristics.

Offodile *et al*. (1994) employed a taxonomy to review the machine-part CF models in CM. The taxonomy is based on Mehrez *et al*. (1988)'s five-level conceptual scheme for knowledge representation. Three classes of machine-part grouping techniques have been identified: visual inspection, part coding and classification, and analysis of the production flow. They used the production flow analysis segment to discuss various proposed CF models.

Reisman *et al*. (1997) gave a most comprehensive survey. A total of 235 CM papers were classified based on seven alternatives, but not mutually exclusive, strategies used in Reisman and Kirshnick (1995).

Selim *et al*. (1998) developed a mathematical formulation and a methodology-based classification to review the literature on the CF problem. The objective function of the mathematical model is to minimize the sum of costs for purchasing machines, variable cost of using machines, tooling cost, material handling cost, and amortized worker training cost per period. The model is combinatorially complex and will not be solvable for any real problem. The

classification used in this paper is based on the type of general solution methodology. More than 150 works have been reviewed and listed in the reference.

2. Similarity coefficient methods (SCM)

A large number of similarity coefficients have been proposed in the literature. Some of them have been utilized in connection with CM. SCM based methods rely on similarity measures in conjunction with clustering algorithms. It usually follows a prescribed set of steps (Romesburg, 1984), the main ones being:

Step (1). Form the initial machine part incidence matrix, whose rows are machines and columns stand for parts. The entries in the matrix are 0s or 1s, which indicate a part need or need not a machine for a production. An entry a_{ik} is defined as follows.

$$a_{ik} = \begin{cases} 1 \text{ if part } k \text{ visits machine } i, \\ 0 \text{ otherwise.} \end{cases} \quad (1)$$

where

i -- machine index (i=1,…, M)
k -- part index (k=1,…, P)
M --number of machines
P -- number of parts

Step (2). Select a similarity coefficient and compute similarity values between machine (part) pairs and construct a similarity matrix. An element in the matrix represents the sameness between two machines (parts).

Step (3). Use a clustering algorithm to process the values in the similarity matrix, which results in a diagram called a tree, or dendrogram, that shows the hierarchy of similarities among all pairs of machines (parts). Find the machines groups (part families) from the tree or dendrogram, check all predefined constraints such as the number of cells, cell size, etc.

3. Why present a taxonomy on similarity coefficients?

Before answer the question “Why present a taxonomy on similarity coefficients?”, we need to answer the following question firstly “Why similarity co-

efficient methods are more flexible than other cell formation methods?".
In this section, we present past review studies on similarity coefficients, discuss their weaknesses and confirm the need of a new review study from the viewpoint of the flexibility of similarity coefficients methods.

3.1 Past review studies on similarity coefficients

Although a large number of similarity coefficients exist in the literature, very few review studies have been performed on similarity coefficients. Three review papers on similarity coefficients (Shafer and Rogers, 1993a; Sarker, 1996; Mosier *et al*., 1997) are available in the literature.
Shafer and Rogers (1993a) provided an overview of similarity and dissimilarity measures applicable to cellular manufacturing. They introduced general measures of association firstly, then similarity and distance measures for determining part families or clustering machine types are discussed. Finally, they concluded the paper with a discussion of the evolution of similarity measures applicable to cellular manufacturing.
Sarker (1996) reviewed a number of commonly used similarity and dissimilarity coefficients. In order to assess the quality of solutions to the cell formation problem, several different performance measures are enumerated, some experimental results provided by earlier researchers are used to evaluate the performance of reviewed similarity coefficients.
Mosier *et al*. (1997) presented an impressive survey of similarity coefficients in terms of structural form, and in terms of the form and levels of the information required for computation. They particularly emphasized the structural forms of various similarity coefficients and made an effort for developing a uniform notation to convert the originally published mathematical expression of reviewed similarity coefficients into a standard form.

3.2 Objective of this study

The three previous review studies provide important insights from different viewpoints. However, we still need an updated and more comprehensive review to achieve the following objectives.

- Develop an explicit taxonomy
  To the best of our knowledge, none of the previous articles has developed or employed an explicit taxonomy to categorize various similarity coefficients.

We discuss in detail the important role of taxonomy in the section 3.3.
Neither Shafer and Rogers (1993a) nor Sarker (1996) provided a taxonomic review framework. Sarker (1996) enumerated a number of commonly used similarity and dissimilarity coefficients; Shafer and Rogers (1993a) classified similarity coefficients into two groups based on measuring the resemblance between: (1) part pairs, or (2) machine pairs.

- Give a more comprehensive review
  Only a few similarity coefficients related studies have been reviewed by previous articles.
  Shafer and Rogers (1993a) summarized 20 or more similarity coefficients related researches; Most of the similarity coefficients reviewed in Sarker (1996)'s paper need prior experimental data; Mosier et al. (1997) made some efforts to abstract the intrinsic nature inherent in different similarity coefficients, Only a few similarity coefficients related studies have been cited in their paper.

Owing to the accelerated growth of the amount of research reported on similarity coefficients subsequently, and owing to the discussed objectives above, there is a need for a more comprehensive review research to categorize and summarize various similarity coefficients that have been developed in the past years.

3.3 Why similarity coefficient methods are more flexible

The cell formation problem can be extraordinarily complex, because of various different production factors, such as alternative process routings, operational sequences, production volumes, machine capacities, tooling times and others, need to be considered. Numerous cell formation approaches have been developed, these approaches can be classified into following three groups:

1. Mathematical Programming (MP) models.
2. (meta-)Heurestic Algorithms (HA).
3. Similarity Coefficient Methods (SCM).

Among these approaches, SCM is the application of cluster analysis to cell formation procedures. Since the basic idea of GT depends on the estimation of the similarities between part pairs and cluster analysis is the most basic

method for estimating similarities, it is concluded that SCM based method is one of the most basic methods for solving CF problems.
Despite previous studies (Seifoddini, 1989a) indicated that SCM based approaches are more flexible in incorporating manufacturing data into the machine-cells formation process, none of the previous articles has explained the reason why SCM based methods are more flexible than other approaches such as MP and HA. We try to explain the reason as follows.
For any concrete cell formation problem, there is generally no "correct" approach. The choice of the approach is usually based on the tool availability, analytical tractability, or simply personal preference. There are, however, two effective principles that are considered reasonable and generally accepted for large and complex problems. They are as follows.

- Principle 1:
  Decompose the complex problem into several small conquerable problems. Solve small problems, and then reconstitute the solutions.

  All three groups of cell formation approaches (MP, HA, SCM) mentioned above can use principle 1 for solving complex cell formation problems. However, the difficulty for this principle is that a systematic mean must be found for dividing one complex problem into many small conquerable problems, and then reconstituting the solutions. It is usually not easy to find such systematic means.

- Principle 2:
  It usually needs a complicated solution procedure to solve a complex cell formation problem. The second principle is to decompose the complicated solution procedure into several small tractable stages.

Comparing with MP, HA based methods, the SCM based method is more suitable for principle 1. We use a concrete cell formation model to explain this conclusion. Assume there is a cell formation problem that incorporates two production factors: production volume and operation time of parts.

(1). MP, HA:
By using MP, HA based methods, the general way is to construct a mathematical or non-mathematical model that takes into account production volume and operation time, and then the model is analyzed, optimal or heuristic solution

procedure is developed to solve the problem. The advantage of this way is that the developed model and solution procedure are usually unique for the original problem. So, even if they are not the "best" solutions, they are usually "very good" solutions for the original problem. However, there are two disadvantages inherent in the MP, HA based methods.

• Firstly, extension of an existing model is usually a difficult work. For example, if we want to extend the above problem to incorporate other production factors such as alternative process routings and operational sequences of parts, what we need to do is to extend the old model to incorporate additional production factors or construct a new model to incorporate all required production factors: production volumes, operation times, alternative process routings and operational sequences. Without further information, we do not know which one is better, in some cases extend the old one is more efficient and economical, in other cases construct a new one is more efficient and economical. However, in most cases both extension and construction are difficult and cost works.

• Secondly, no common or standard ways exist for MP, HA to decompose a complicated solution procedure into several small tractable stages. To solve a complex problem, some researchers decompose the solution procedure into several small stages. However, the decomposition is usually based on the experience, ability and preference of the researchers. There are, however, no common or standard ways exist for decomposition.

(2). SCM:
SCM is more flexible than MP, HA based methods, because it overcomes the two mentioned disadvantages of MP, HA. We have introduced in section 2.2 that the solution procedure of SCM usually follows a prescribed set of steps:

Step 1. Get input data;
Step 2. Select a similarity coefficient;
Step 3. Select a clustering algorithm to get machine cells.

Thus, the solution procedure is composed of three steps, this overcomes the second disadvantage of MP, HA. We show how to use SCM to overcome the first disadvantage of MP, HA as follows.
An important characteristic of SCM is that the three steps are independent

with each other. That means the choice of the similarity coefficient in step2 does not influence the choice of the clustering algorithm in step3. For example, if we want to solve the production volumes and operation times considered cell formation problem mentioned before, after getting the input data; we select a similarity coefficient that incorporates production volumes and operation times of parts; finally we select a clustering algorithm (for example ALC algorithm) to get machine cells. Now we want to extend the problem to incorporate additional production factors: alternative process routings and operational sequences. We re-select a similarity coefficient that incorporates all required 4 production factors to process the input data, and since step2 is independent from step3, we can easily use the ALC algorithm selected before to get new machine cells. Thus, comparing with MP, HA based methods, SCM is very easy to extend a cell formation model.
Therefore, according above analysis, SCM based methods are more flexible than MP, HA based methods for dealing with various cell formation problems. To take full advantage of the flexibility of SCM and to facilitate the selection of similarity coefficients in step2, we need an explicit taxonomy to clarify and classify the definition and usage of various similarity coefficients. Unfortunately, none of such taxonomies has been developed in the literature, so in the next section we will develop a taxonomy to summarize various similarity coefficients.

4. A taxonomy for similarity coefficients employed in cellular manufacturing

Different similarity coefficients have been proposed by researchers in different fields. A similarity coefficient indicates the degree of similarity between object pairs. A tutorial of various similarity coefficients and related clustering algorithms are available in the literature (Anderberg, 1973; Bijnen, 1973; Sneath and Sokal, 1973; Arthanari and Dodge, 1981; Romesburg, 1984; Gordon, 1999). In order to classify similarity coefficients applied in CM, a taxonomy is developed and shown in figure 1. The objective of the taxonomy is to clarify the definition and usage of various similarity or dissimilarity coefficients in designing CM systems. The taxonomy is a 5-level framework numbered from level 0 to 4. Level 0 represents the root of the taxonomy. The detail of each level is described as follows.

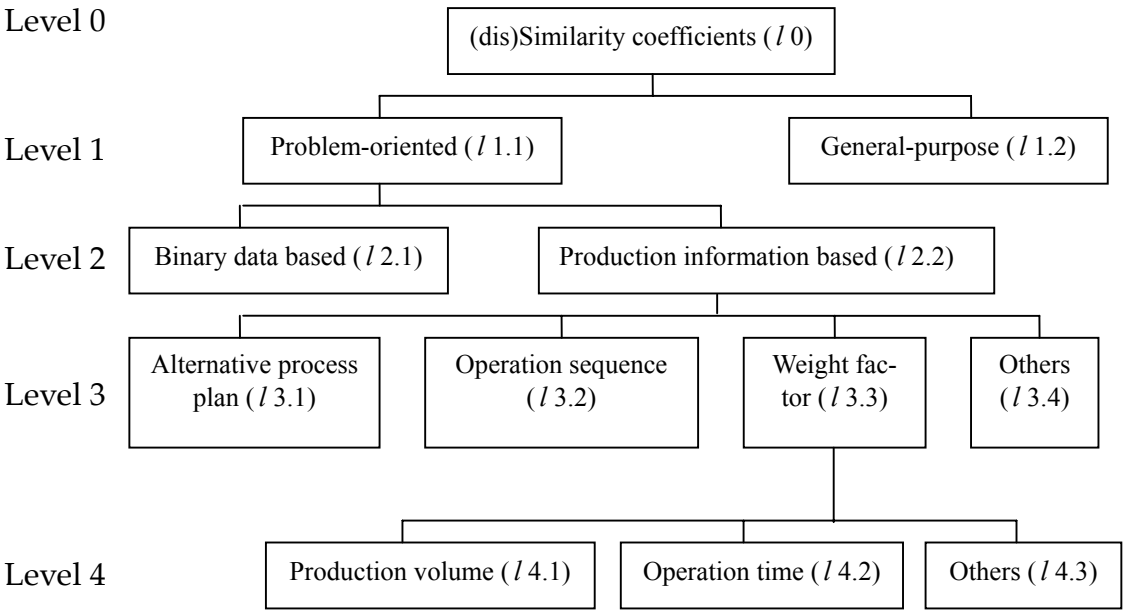


Figure 1. A taxonomy for similarity coefficients

Level 1.

l1 categorizes existing similarity coefficients into two distinct groups: problem-oriented similarity coefficients (l1.1) and general-purpose similarity coefficients (l1.2). Most of the similarity coefficients introduced in the field of numerical taxonomy are classified in l1.2 (general-purpose), which are widely used in a number of disciplines, such as psychology, psychiatry, biology, sociology, the medical sciences, economics, archeology and engineering. The characteristic of this type of similarity coefficients is that they always maximize similarity value when two objects are perfectly similar.

On the other hand, problem-oriented (l1.1) similarity coefficients aim at evaluating the predefined specific "appropriateness" between object pairs. This type of similarity coefficient is designed specially to solve specific problems, such as CF. They usually include additional information and do not need to produce maximum similarity value even if the two objects are perfectly similar. Two less similar objects can produce a higher similarity value due to their "appropriateness" and more similar objects may produce a lower similarity value due to their "inappropriateness".

We use three similarity coefficients to illustrate the difference between the problem-oriented and general-purpose similarity coefficients. Jaccard is the most commonly used general-purpose similarity coefficient in the literature, Jaccard similarity coefficient between machine i and machine j is defined as follows:

$$s_{ij} = \frac{a}{a+b+c}, \qquad 0 \le s_{ij} \le 1 \tag{2}$$

where

a : the number of parts visit both machines,

b : the number of parts visit machine i but not j,

c : the number of parts visit machine j but not i,

Two problem-oriented similarity coefficients, MaxSC (Shafer and Rogers, 1993b) and Commonality score (CS, Wei and Kern, 1989), are used to illustrate this comparison. MaxSC between machine i and machine j is defined as follows:

$$ms_{ij} = max[\frac{a}{a+b}, \frac{a}{a+c}], \qquad 0 \le ms_{ij} \le 1 \tag{3}$$

and CS between machine i and machine j is calculated as follows:

$$c_{ij} = \sum_{k=1}^{P} \delta(a_{ik}, a_{jk}) \tag{4}$$

Where

$$\delta(a_{ik}, a_{jk}) = \begin{cases} (P-1), & \text{if } a_{ik} = a_{jk} = 1 \\ 1, & \text{if } a_{ik} = a_{jk} = 0 \\ 0, & \text{if } a_{ik} \neq a_{jk}. \end{cases} \tag{5}$$

$$a_{ik} = \begin{cases} 1, & \text{if machine } i \text{ uses part } k, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

k : part index (k=1,… P), is the k th part in the machine-part matrix.

We use figure 2 and figure 3 to illustrate the "appropriateness" of problem-oriented similarity coefficients. Figure 2 is a machine-part incidence matrix whose rows represent machines and columns represent parts. The Jaccard coefficient s_{ij}, MaxSC coefficient ms_{ij} and commonality score c_{ij} of machine pairs in figure 2 are calculated and given in figure 3.

The characteristic of general-purpose similarity coefficients is that they always maximize similarity value when two objects are perfectly similar. Among the four machines in figure 2, we find that machine 2 is a perfect copy of machine

1, they should have the highest value of similarity. We also find that the degree of similarity between machines 3 and 4 is lower than that of machines 1 and 2. The results of Jaccard in figure 3 reflect our finds straightly. That is, $max(s_{ij})=s_{12}=1$, and $s_{12}>s_{34}$.

		Part													
		p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12	p13	p14
Machine	m1	1	1	1											
	m2	1	1	1											
	m3	1	1	1	1										
	m4	1	1	1	1	1	1	1							

Figure 2. Illustrative machine-part matrix for the "appropriateness"

	Similarity values, s_{ij}, ms_{ij} and c_{ij}			
	$i=1, j=2$	$i=3, j=4$	$i=1$or$2, j=3$	$i=1$or$2, j=4$
Jaccard s_{ij}	1	4/7	3/4	3/7
MaxSC ms_{ij}	1	1	1	1
CS c_{ij}	50	59	49	46

Figure 3. Similarity values of Jaccard, MaxSC and CS of figure 2

Problem-oriented similarity coefficients are designed specially to solve CF problems. CF problems are multi-objective decision problems. We define the "appropriateness" of two objects as the degree of possibility to achieve the objectives of CF models by grouping the objects into the same cell. Two objects will obtain a higher degree of "appropriateness" if they facilitate achieving the predefined objectives, and vice versa. As a result, two less similar objects can produce a higher similarity value due to their "appropriateness" and more similar objects may produce a lower similarity value due to their "inappropriateness". Since different CF models aim at different objectives, the criteria of "appropriateness" are also varied. In short, for problem-oriented similarity coefficients, rather than evaluating the similarity between two objects, they evaluate the "appropriateness" between them.

MaxSC is a problem-oriented similarity coefficient (Shafer and Rogers, 1993b). The highest value of MaxSC is given to two machines if the machines process exactly the same set of parts or if one machine processes a subset of the parts processed by the other machine. In figure 3, all machine pairs obtain the highest MaxSC value even if not all of them are perfectly similar. Thus, in the procedure of cell formation, no difference can be identified from the four machines by MaxSC.

CS is another problem-oriented similarity coefficient (Wei and Kern, 1989). The objective of CS is to recognize not only the parts that need both machines, but also the parts on which the machines both do not process. Some characteristics of CS have been discussed by Yasuda and Yin (2001). In figure 3, the highest CS is produced between machine 3 and machine 4, even if the degree of similarity between them is lower and even if machines 1 and 2 are perfectly similar. The result $s_{34} > s_{12}$ illustrates that two less similar machines can obtain a higher similarity value due to the higher "appropriateness" between them.

Therefore, it is concluded that the definition of "appropriateness" is very important for every problem-oriented similarity coefficient, it determines the quality of CF solutions by using these similarity coefficients.

Level 2.

In figure 1, problem-oriented similarity coefficients can be further classified into binary data based (l2.1) and production information based (l2.2) similarity coefficients. Similarity coefficients in l2.1 only consider assignment information, that is, a part need or need not a machine to perform an operation. The assignment information is usually given in a machine-part incidence matrix, such as figure 2. An entry of "1" in the matrix indicates that the part needs a operation by the corresponding machine. The characteristic of l2.1 is similar to l1.2, which also uses binary input data. However, as we mentioned above, they are essentially different in the definition for assessing the similarity between object pairs.

Level 3.

In the design of CM systems, many manufacturing factors should be involved when the cells are created, e.g. machine requirement, machine setup times, utilization, workload, alternative routings, machine capacities, operation sequences, setup cost and cell layout (Wu and Salvendy, 1993). Choobineh and Nare (1999) described a sensitivity analysis for examining the impact of ignored manufacturing factors on a CMS design. Due to the complexity of CF

problems, it is impossible to take into consideration all of the real-life production factors by a single approach. A number of similarity coefficients have been developed in the literature to incorporate different production factors. In this paper, we use three most researched manufacturing factors (alternative process routing *l* 3.1, operation sequence *l* 3.2 and weighted factors *l* 3.3) as the base to perform the taxonomic review study.

Level 4.
Weighted similarity coefficient is a logical extension or expansion of the binary data based similarity coefficient. Merits of the weighted factor based similarity coefficients have been reported by previous studies (Mosier and Taube, 1985b; Mosier, 1989; Seifoddini and Djassemi, 1995). This kind of similarity coefficient attempts to adjust the strength of matches or misses between object pairs to reflect the resemblance value more realistically and accurately by incorporating object attributes.
The taxonomy can be used as an aid to identify and clarify the definition of various similarity coefficients. In the next section, we will review and map similarity coefficients related researches based on this taxonomy.

5. Mapping SCM studies onto the taxonomy

In this section, we map existing similarity coefficients onto the developed taxonomy and review academic studies through 5 tables. Tables 1 and 2 are general-purpose (*l* 1.2) similarity/dissimilarity coefficients, respectively. Table 3 gives expressions of some binary data based (*l* 2.1) similarity coefficients, while table 4 summarizes problem-oriented (*l* 1.1) similarity coefficients. Finally, SCM related academic researches are illustrated in table 5.
Among the similarity coefficients in table 1, eleven of them have been selected by Sarker and Islam (1999) to address the issues relating to the performance of them along with their important characteristics, appropriateness and applications to manufacturing and other related fields. They also presented numerical results to demonstrate the closeness of the eleven similarity and eight dissimilarity coefficients that is presented in table 2. Romesburg (1984) and Sarker (1996) provided detailed definitions and characteristics of these eleven similarity coefficients, namely Jaccard (Romesburg, 1984), Hamann (Holley and Guilford, 1964), Yule (Bishop *et al*., 1975), Simple matching (Sokal and Michener, 1958), Sorenson (Romesburg, 1984), Rogers and Tanimoto (1960), Sokal and

Sneath (Romesburg, 1984), Rusell and Rao (Romesburg, 1984), Baroni-Urbani and Buser (1976), Phi (Romesburg, 1984), Ochiai (Romesburg, 1984). In addition to these eleven similarity coefficients, table 1 also introduces several other similarity coefficients, namely PSC (Waghodekar and Sahu, 1984), Dot-product, Kulczynski, Sokal and Sneath 2, Sokal and Sneath 4, Relative matching (Islam and Sarker, 2000). Relative matching coefficient is developed recently which considers a set of similarity properties such as no mismatch, minimum match, no match, complete match and maximum match. Table 2 shows eight most commonly used general-purpose (^{l}1.2) dissimilarity coefficients.

Similarity Coefficient	*Definition* s_{ij}	*Range*
1. Jaccard	$a/(a+b+c)$	0-1
2. Hamann	$[(a+d)-(b+c)]/[(a+d)+(b+c)]$	-1-1
3. Yule	$(ad-bc)/(ad+bc)$	-1-1
4. Simple matching	$(a+d)/(a+b+c+d)$	0-1
5. Sorenson	$2a/(2a+b+c)$	0-1
6. Rogers and Tanimoto	$(a+d)/[a+2(b+c)+d]$	0-1
7. Sokal and Sneath	$2(a+d)/[2(a+d)+b+c]$	0-1
8. Rusell and Rao	$a/(a+b+c+d)$	0-1
9. Baroni-Urbani and Buser	$[a+(ad)^{1/2}]/[a+b+c+(ad)^{1/2}]$	0-1
10. Phi	$(ad-bc)/[(a+b)(a+c)(b+d)(c+d)]^{1/2}$	-1-1
11. Ochiai	$a/[(a+b)(a+c)]^{1/2}$	0-1
12. PSC	$a^2/[(b+a)*(c+a)]$	0-1
13. Dot-product	$a/(b+c+2a)$	0-1
14. Kulczynski	$1/2[a/(a+b)+a/(a+c)]$	0-1
15. Sokal and Sneath 2	$a/[a+2(b+c)]$	0-1
16. Sokal and Sneath 4	$1/4[a/(a+b)+a/(a+c)+d/(b+d)+d/(c+d)]$	0-1
17. Relative matching	$[a+(ad)^{1/2}]/[a+b+c+d+(ad)^{1/2}]$	0-1

Table 1. Definitions and ranges of some selected general-purpose similarity coefficients (***l* 1.2**). a: the number of parts visit both machines; b: the number of parts visit machine i but not j; c: the number of parts visit machine j but not i; d: the number of parts visit neither machine

The dissimilarity coefficient does reverse to those similarity coefficients in table 1. In table 2, d_{ij} is the original definition of these coefficients, in order to

show the comparison more explicitly, we modify these dissimilarity coefficients and use binary data to express them. The binary data based definition is represented by d_{ij}

Dissamilarity Coefficient	*Definition* d_{ij}	*Range Definition*	d'_{ij}	*Range*
1. Minkowski	$\left(\sum_{k=1}^{M}\left\|a_{ki}-a_{kj}\right\|^{r}\right)^{1/r}$	Real	$(b+c)^{1/r}$	Real
2. Euclidean	$\left(\sum_{k=1}^{M}\left\|a_{ki}-a_{kj}\right\|^{2}\right)^{1/2}$	Real	$(b+c)^{1/2}$	Real
3. Manhattan (City Block)	$\sum_{k=1}^{M}\left\|a_{ki}-a_{kj}\right\|$	Real	$b+c$	0-M
4. Average Euclidean	$\left(\sum_{k=1}^{M}\left\|a_{ki}-a_{kj}\right\|^{2}/M\right)^{1/2}$	Real	$\left(\frac{b+c}{a+b+c+d}\right)^{1/2}$	Real
5. Weighted Minkowski	$\left(\sum_{k=1}^{M}w_k\left\|a_{ki}-a_{kj}\right\|^{r}\right)^{1/r}$	Real	$[w_k(b+c)]^{1/r}$	Real
6. Bray-Curtis	$\sum_{k=1}^{M}\left\|a_{ki}-a_{kj}\right\| / \sum_{k=1}^{M}\left\|a_{ki}+a_{kj}\right\|$	0-1	$\frac{b+c}{2a+b+c}$	0-1
7. Canberra Metric	$\frac{1}{M}\sum_{k=1}^{M}\left(\frac{\left\|a_{ki}-a_{kj}\right\|}{a_{ki}+a_{kj}}\right)$	0-1	$\frac{b+c}{a+b+c+d}$	0-1
8. Hamming	$\sum_{k=1}^{M}\delta(a_{kl},a_{kj})$	0-M	$b+c$	0-M

Table 2. Definitions and ranges of some selected general-purpose dissimilarity coefficients. ($\boldsymbol{l}$ 1.2) $\delta(a_{kl},a_{kj})=\begin{cases}1, \text{if } a_{kl}\neq a_{kj};\\ 0, \text{otherwise.}\end{cases}$; r : a positive integer; d_{ij} : dissimilarity between i *and* j; d'_{ij} : dissimilarity by using binary data; k : attribute index (k =1,…, M).

Table 3 presents some selected similarity coefficients in group l 2.1. The expressions in table 3 are similar to that of table 1. However, rather than judging the similarity between two objects, problem-oriented similarity coefficients evaluate a predetermined "appropriateness" between two objects. Two objects

that have the highest "appropriateness" maximize similarity value even if they are less similar than some other object pairs.

Coefficient/Resource	*Definition* S_{ij}	*Range*
1. Chandrasekharan & Rajagopalan (1986b)	$a / Min[(a+b),(a+c)]$	0-1
2. Kusiak *et al.* (1986)	a	integer
3. Kusiak (1987)	$a+d$	integer
4. Kaparthi *et al.* (1993)	$a^{'}/(a+b)^{'}$	0-1
5. MaxSC / Shafer & Rogers (1993b)	$max[a/(a+b), a/(a+c)]$	0-1
6. Baker & Maropoulos (1997)	$a / Max[(a+b),(a+c)]$	0-1

Table 3. Definitions and ranges of some selected problem-oriented binary data based similarity coefficients (*I* 2.1). $a^{'}$ is the number of matching ones between the matching exemplar and the input vector; $(a+b)^{'}$ is the number of ones in the input vector

Table 4 is a summary of problem-oriented (*l* 1.1) similarity coefficients developed so far for dealing with CF problems. This table is the tabulated expression of the proposed taxonomy. Previously developed similarity coefficients are mapped into the table, additional information such as solution procedures, novel characteristics are also listed in the "*Notes/KeyWords*" column.

Finally, table 5 is a brief description of the published CF studies in conjunction with similarity coefficients. Most studies listed in this table do not develop new similarity coefficients. However, all of them use similarity coefficients as a powerful tool for coping with cell formation problems under various manufacturing situations. This table also shows the broad range of applications of similarity coefficient based methods.

No	Resource/Coefficient		Binary data based (*l2.1*)	Production Information (*l2.2*)						Notes/KeyWords
				Alternative Proc. (*l3.1*)	Operation sequ. (*l3.2*)	Weights (*l3.3*)			Others (*l3.4*)	
	Author(s)/(SC)	Year				Prod. Vol.(*l4.1*)	Oper. Time(*l4.2*)	Others (*l4.3*)		
1	De Witte	1980				Y			MM	3 SC created; Graph theory
2	Waghodekar & Sahu (PSC & SCTF)	1984	Y							*l* 1.2; 2 SC created
3	Mosier & Taube	1985b				Y				2 SC created
4	Selvam & Balasubramanian	1985			Y	Y				Heuristic
5	Chandrasekharan & Rajagopalan	1986b	Y							*l* 2.1; hierarchical algorithm
6	Dutta *et al.*	1986							CS; NC	5 D developed;
7	Faber & Carter (MaxSC)	1986	Y							*l* 2.1; Graph
8	Kusiak *et al.*	1986	Y							*l* 2.1; 3 distinct integer models
9	Kusiak	1987	Y							*l* 2.1; APR by p-median
10	Seifoddini	87/88			Y	Y				
11	Steudel & Ballakur	1987					Y			Dynamic programming
12	Choobineh	1988			Y					Mathematical model
13	Gunasingh & Lashkari	1989						T		Math.; Compatibility index
14	Wei & Kern	1989	Y							*l* 2.1; Heuristic
15	Gupta & Seifoddini	1990			Y	Y	Y			Heuristic

Table 4. Summary of developed problem-oriented (dis)similarity coefficients (SC) for cell formation (***l* 1.1**)

No	Resource/Coefficient		Binary data based (**l2.1**)	Production Information (**l2.2**)						Notes/KeyWords
				Alternative Proc. (**l3.1**)	Operation sequ. (**l3.2**)	Weights (**l3.3**)				
	Author(s)/(SC)	Year				Prod. Vol.(*l4.1*)	Oper. Time(*l4.2*)	Others (**l4.3**)	Others (**l3.4**)	
16	Tam	1990			Y					*k* Nearest Neighbour
17	Vakharia & Wemmerlöv	1987 ; 1990			Y					Heuristic
18	Offodile	1991						Y		Parts coding and classification
19	Kusiak & Cho	1992	Y							*l* 2.1; 2 SC proposed
20	Zhang & Wang	1992							Y	Combine SC with fuzziness
21	Balasubramanian & Panneerselvam	1993			Y	Y			M H C	D; covering technique
22	Ho *et al.*	1993			Y					Compliant index
23	Gupta	1993		Y	Y	Y	Y			Heuristic
24	Kaparthi *et al.*	1993	Y							*l* 2.1; Improved neural network
25	Luong	1993							C S	Heuristic
26	Ribeiro & Pradin	1993	Y							D, *l* 1.2; Knapsack
27	Seifoddini & Hsu	1994						Y		Comparative study
28	Akturk & Balkose	1996			Y					D; multi objective model
29	Ho & Moodie (POSC)	1996							F P R	Heuristic; Mathematical
30	Ho & Moodie (GOSC)	1996				Y				SC between two part groups
31	Suer & Cedeno	1996						C		
32	Viswanathan	1996	Y							*l* 2.1; modify p-median

Table 4 (continued)

No	Resource/Coefficient		Binary data based (**l2.1**)	Production Information (**l2.2**)						Notes/KeyWords
						Weights (**l3.3**)				
	Author(s)/(SC)	Year		Alternative Proc. (**l3.1**)	Operation sequ. (**l3.2**)	Prod. volu. (**l4.1**)	Oper. Time (**l4.2**)	Others (**l4.3**)	Others (**l3.4**)	
33	Baker & Maropoulos	1997	Y							*l* 2.1; Black box algorithm
34	Lee *et al.*	1997			Y	Y				APR by genetic algorithm
35	Won & Kim	1997		Y						Heuristic
36	Askin & Zhou	1998			Y					Shortest path
37	Nair & Narendran	1998			Y					Non-hierarchical
38	Jeon *et al.*	1998 b		Y						Mathematical
39	Kitaoka *et al.* (Double Centering)	1999	Y							*l* 2.1; quantification model
40	Nair & Narendran	1999							W L	Mathematical; Non-hierarchical
41	Nair & Narendran	1999			Y	Y			W L	Mathematical; Non-hierarchical
42	Seifoddini & Tjahjana	1999							B S	
43	Sarker & Xu	2000			Y					3 phases algorithm
44	Won	2000 a		Y						Modify p-median
45	Yasuda & Yin	2001							C S	D; Heuristic

Table 4 (continued). Summary of developed problem-oriented (dis)similarity coefficients (SC) for cell formation (***l* 1.1**)

APR: Alternative process routings; *BS*: Batch size; *C*: Cost of unit part, *CS*: cell size; *D*: dissimilarity coefficient; *FPR*: Flexible processing routing, *MHC*: Material handling cost; *MM*: Multiple machines available for a machine type, *NC*: number of cell; *SC*:

Similarity coefficient; T: Tooling requirements of parts, *WL*: Workload

Articles		***Similarity coefficients (SC) used***	***Description/Keywords***
Author(s)	*Year*		
McAuley	1972	Jaccard	First study of SC on cell formation
Carrie	1973	Jaccard	Apply SC on forming part families
Rajagopalan & Batra	1975	Jaccard	Graph theory
Waghodekar & Sahu	1984	Jaccard; PSC; SCTF	Propose MCSE method
Kusiak	1985	Minkowski (D)	p-median; heuristics
Chandrasekharan & Rajagopalan	1986a	Minkowski (D)	Non-hierarchical algorithm
Han & Ham	1986	Manhattan (D)	Classification and coding system
Seifoddini & Wolfe	1986	Jaccard	Bit-level data storage technique
Chandrasekharan & Rajagopalan	1987	Manhattan (D)	Develop ZODIAC algorithm
Marcotorchino	1987	Jaccard; Sorenson	Create a block seriation model
Seifoddini & Wolfe	1987	Jaccard	Select threshold on material handling cost
Chandrasekharan& Rajagopalan	1989	Jaccard; Simple matching; Manhattan (D)	An analysis of the properties of data sets
Mosier	1989	7 similarity coefficients	Comparative study
Seifoddini	1989a	Jaccard	SLC vs. ALC
Seifoddini	1989b	Jaccard	Improper machine assignment
Srinivasan *et al.*	1990	Kusiak (1987)	An assignment model
Askin *et al.*	1991	Jaccard	Hamiltonian path; TSP
Chow	1991	CS	Unjustified claims of LCC
Gongaware & Ham	1991	---*	Classification & coding; multi-objective model
Gupta	1991	Gupta & Seifoddini (1990)	Comparative study on chaining effect
Logendran	1991	Jaccard; Kusiak (1987)	Identification of key machine

Srinivasan & Narendran	1991	Kusiak (1987)	A nonhierarchical clustering algorithm
Wei & Kern	1991	CS	Reply to Chow (1991)
Chow & Hawaleshka	1992	CS	Define machine unit concept
Shiko	1992	Jaccard	Constrained hierarchical
Chow & Hawaleshka	1993a	CS	Define machine unit concept
Chow & Hawaleshka	1993b	CS	A knowledge-based approach
Kang & Wemmerlöv	1993	Vakharia & Wemmerlov (87,90)	Heuristic; Alternative operations of parts
Kusiak *et al.*	1993	Hamming (D)	Branch-Bound & A* approaches
Offodile	1993	Offodile (1991)	Survey of robotics & GT; robot selection model
Shafer & Rogers	1993a	Many	Review of similarity coefficients
Shafer & Rogers	1993b	16 similarity coefficients	Comparative study
Vakharia & Kaku	1993	Kulczynski	Long-term demand change
Ben-Arieh & Chang	1994	Manhattan (D)	Modify p-median algorithm
Srinivasan	1994	Manhattan (D)	Minimum spanning trees
Balakrishnan & Jog	1995	Jaccard	TSP algorithm
Cheng *et al.*	1995	Hamming (D)	Quadratic model; A* algorithm
Kulkarni & Kiang	1995	Euclidean (D)	Self-organizing neural network
Murthy & Srinivasan	1995	Manhattan (D)	Heuristic; Consider fractional cell formation
Seifoddini & Djassemi	1995	Jaccard	Merits of production volume consideration
Vakharia & Wemmerlöv	1995	8 dissimilarity coefficients	Comparative study
Wang & Roze	1995	Jaccard, Kusiak (1987), CS	An experimental study
Balakrishnan	1996	Jaccard	CRAFT
Cheng *et al.*	1996	Hamming (D)	Truncated tree search algorithm

Hwang & Ree	1996	Jaccard	Define compatibility coefficient
Lee & Garcia-Diaz	1996	Hamming (D)	Use a 3-phase network-flow approach
Leem & Chen	1996	Jaccard	Fuzzy set theory
Lin *et al.*	1996	Bray-Curtis (D)	Heuristic; workload balance within cells
Sarker	1996	Many	Review of similarity coefficient
Al-sultan & Fedjki	1997	Hamming (D)	Genetic algorithm
Askin *et al*	1997	MaxSC	Consider flexibility of routing and demand
Baker & Maropoulos	1997	Jaccard, Baker & Maropoulos (1997)	Black Box clustering algorithm
Cedeno & Suer	1997	---	Approach to "remainder clusters"
Masnata & Settineri	1997	Euclidean (D)	Fuzzy clustering theory
Mosier *et al.*	1997	Many	Review of similarity coefficients
Offodile & Grznar	1997	Offodile (1991)	Parts coding and classification analysis
Wang & Roze	1997	Jaccard, Kusiak (1987), CS	Modify p-median model
Cheng *et al.*	1998	Manhattan (D)	TSP by genetic algorithm
Jeon *et al.*	1998a	Jeon *et al.* (1998b)	p-median
Onwubolu & Mlilo	1998	Jaccard	A new algorithm (SCDM)
Srinivasan & Zimmers	1998	Manhattan (D)	Fractional cell formation problem
Wang	1998	---	A linear assignment model
Ben-Arieh & Sreenivasan	1999	Euclidean (D)	A distributed dynamic algorithm
Lozano *et al.*	1999	Jaccard	Tabu search
Sarker & Islam	1999	Many	Performance study
Baykasoglu & Gindy	2000	Jaccard	Tabu search
Chang & Lee	2000	Kusiak (1987)	Multi-solution heuristic
Josien & Liao	2000	Euclidean (D)	Fuzzy set theory
Lee-post	2000	Offodile (1991)	Use a simple genetic algorithm
Won	2000a	Won & Kim(1997)	Alternative process plan with p-median model

Won	2000b	Jaccard, Kusiak (1987)	Two-phase p-median model
Dimopoulos & Mort	2001	Jaccard	Genetic algorithm
Samatova *et al.*	2001	5 dissimilarity coefficients	Vector perturbation approach

Table 5. Literature of cell formation research in conjunction with similarity coefficients (SC). *: no specific SC mentioned

6. General discussion

We give a general discussion of production information based similarity coefficients (***I* 2.2**) and an evolutionary timeline in this section.

6.1. Production information based similarity coefficients

6.1.1 Alternative process routings

In most cell formation methods, parts are assumed to have a unique part process plan. However, it is well known that alternatives may exist in any level of a process plan. In some cases, there may be many alternative process plans for making a specific part, especially when the part is complex (Qiao *et al*. 1994). Explicit consideration of alternative process plans invoke changes in the composition of all manufacturing cells so that lower capital investment in machines, more independent manufacturing cells and higher machine utilization can be achieved (Hwang and Ree 1996).

Gupta (1993) is the first person who incorporated alternative process routings into similarity coefficient. His similarity coefficient also includes other production information such as operation sequences, production volumes and operation times. The similarity coefficient assigns pairwise similarity among machines with usage factors of all alternative process routings. The usage factors are determined by satisfying production and capacity constraints. The production volumes and operation times are assumed to be known with certainty.

An alternative process routings considered similarity coefficient was developed by Won and Kim (1997) and slightly modified by Won (2000a). In the definition of the similarity coefficient, if machine i is used by some process routing of part j, then the number of parts processed by machine i is counted as one for that part even if the remaining process routings of part j also use

machine i. The basic idea is that in the final solution only one process routing is selected for each part. p-median approach was used by Won (2000a) to associate the modified similarity coefficient.
A similarity coefficient that considers the number of alternative process routings when available during machine failure is proposed by Jeon *et al.* (1998b). The main characteristic of the proposed similarity coefficient is that it draws on the number of alternative process routings during machine failure when alternative process routings are available instead of drawing on operations, sequence, machine capabilities, production volumes, processing requirements and operational times. Based on the proposed similarity coefficient, p-median approach was used to form part families.

6.1.2 Operation sequences

The operation sequence is defined as an ordering of the machines on which the part is sequentially processed (Vakharia and Wemmerlov 1990). A lot of similarity coefficients have been developed to consider operation sequence.
Selvam and Balasubramanian (1985) are the first persons who incorporated alternative process routings into similarity coefficient. Their similarity coefficient is very simple and intuitive. The value of similarity coefficient is determined directly by the production volume of parts moves between machines.
Seifoddini (1987/1988) modified Jaccard similarity coefficient to take into account the production volume of parts moves between machine pairs. A simple heuristic algorithm was used by the author to form machine cells. Choobineh (1988) gave a similarity coefficient between parts j and k which is based on the common sequences of length 1 through L between the two parts. To select the value L, one has to balance the need to uncover the natural strength of the relationships among the parts and the computational efforts necessary to calculate the sequences of length 1 through L. In general, the higher the value of L, the more discriminating power similarity coefficient will have.Gupta and Seifoddini (1990) proposed a similarity coefficient incorporating operation sequence, production volume and operation time simultaneously. From the definition, each part that is processed by at least one machine from a pair of machines contributes towards their similarity coefficient value. A part that is processed by both machines increases the coefficient value for the two machines whereas, a part that is processed on one machine tends to reduce it. The similarity coefficient developed by Tam (1990) is based on Levenshtein's distance measure of two sentences. The distance between two sentences is defined

as the minimum number of transformations required to derive one sentence from the other. Three transformations are defined. The similarity coefficient between two operation sequences x and y, is defined as the smallest number of transformations required to derive y from x. Vakharia and Wemmerlov (1990) proposed a similarity coefficient based on operation sequences to integrate the intracell flow with the cell formation problem by using clustering methodology. The similarity coefficient measures the proportion of machine types used by two part families in the same order.

Balasubramanian and Panneerselvam (1993) developed a similarity coefficient which needs following input data: (1) operation sequences of parts; (2) additional cell arrangements; (3) production volume per day and the bulk factor; (4) guidelines for computing excess moves; (5) actual cost per move.

Ho *et al*. (1993)'s similarity coefficient calculates a compliant index firstly. The compliant index of the sequence of a part compared with a flow path is determined by the number of operations in the sequence of the part that have either "in-sequence" or "by-passing" relationship with the sequence of the flow path. There are two kinds of compliant indexes: forward compliant index and backward index. These two compliant indexes can be calculated by comparing the operation sequence of the part with the sequence of the flow path forwards and backwards.As mentioned in 6.1.1, Gupta (1993) proposed a similarity coefficient which incorporates several production factors such as operation sequences, production volumes, alternative process routings.

Akturk and Balkose (1996) revised the Levenshtein distance measure to penalize the backtracking parts neither does award the commonality. If two parts have no common operations, then a dissimilarity value is found by using the penalizing factor.Lee *et al.* (1997)'s similarity coefficient takes the direct and indirect relations between the machines into consideration. The direct relation indicates that two machines are connected directly by parts; whereas the indirect relation indicates that two machines are connected indirectly by other machines. Askin and Zhou (1998) proposed a similarity coefficient which is based on the longest common operation subsequence between part types and used to group parts into independent, flow-line families.

Nair and Narendran (1998) gave a similarity coefficient as the ratio of the sum of the moves common to a pair of machines and the sum of the total number of moves to and from the two machines. Latterly, They extended the coefficient to incorporate the production volume of each part (Nair and Narendran, 1999). Sarker and Xu (2000) developed an operation sequence-based similarity coeffi-

cient. The similarity coefficient was applied in a *p*-median model to group the parts to form part families with similar operation sequences.

6.1.3 Weight factors

Weighted similarity coefficient is a logical extension or expansion of the binary data based similarity coefficient. Two most researched weight factors are production volume and operation time.

De Witte (1980) is the first person who incorporated production volume into similarity coefficient. In order to analyse the relations between machine types three different similarity coefficients has be used by the author. Absolute relations, mutual interdependence relations and relative single interdependence relations between machine pairs are defined by similarity coefficients *SA*, *SM* and *SS*, respectively.

Mosier and Taube (1985b)'s similarity coefficient is a simple weighted adaptation of McAuley's Jaccard similarity coefficient with an additional term whose purpose is to trap the coefficient between -1.0 and +1.0. Production volumes of parts have been incorporated into the proposed similarity coefficient.

Ho and Moodie (1996) developed a similarity coefficient, namely group-operation similarity coefficient (*GOSC*) to measure the degree of similarity between two part groups. The calculation of *GOSC* considers the demand quantities of parts. A part with a larger amount of demand will have a heavier weight. This is reasonable since if a part comprises the majority of a part group, then it should contribute more in the characterization of the part group it belongs to.

The operation time is considered firstly by Steudel and Ballakur (1987). Their similarity coefficient is based on the Jaccard similarity coefficient and calculates the operation time by multiplying each part's operation time by the production requirements for the part over a given period of time. Operation set-up time is ignored in the calculation since set-up times can usually be reduced after the cells are implemented. Hence set-up time should not be a factor in defining the cells initially.

Other production volume / operation time considered studies include Selvam and Balasubramanian (1985), Seifoddini (1987/1988), Gupta and Seifoddini (1990), Balasubramanian and Panneerselvam (1993), Gupta (1993), Lee *et al*. (1997) and Nair and Narendran (1999). Their characteristics have been discussed in sections 6.1.1 and 6.1.2.

6.2 Historical evolution of similarity coefficients

Shafer and Rogers (1993a) delineated the evolution of similarity coefficients until early 1990s. Based on their work and table 4, we depict the historical evolution of similarity coefficients over the last three decades.

McAuley (1972) was the first person who used the Jaccard similarity coefficient to form machine cells. The first weighted factor that was considered by researchers is the production volume of parts (De Witte, 1980; Mosier and Taube, 1985b). Operation sequences, one of the most important manufacturing factors, was incorporated in 1985 (Selvam and Balasubramanian). In the late 1980s and early 1990s, other weighted manufacturing factors such as tooling requirements (Gunasingh and Lashkari, 1989) and operation times (Gupta and Seifoddini, 1990) were taken into consideration. Alternative process routings of parts is another important manufacturing factor in the design of a CF system. Although it was firstly studied by Kusiak (1987), it was not combined into the similarity coefficient definition until Gupta (1993).

Material handling cost was also considered in the early 1990s (Balasubramanian and Panneerselvam, 1993). In the middle of 1990s, flexible processing routings (Ho and Moodie, 1996) and unit cost of parts (Sure and Cedeno, 1996) were incorporated.

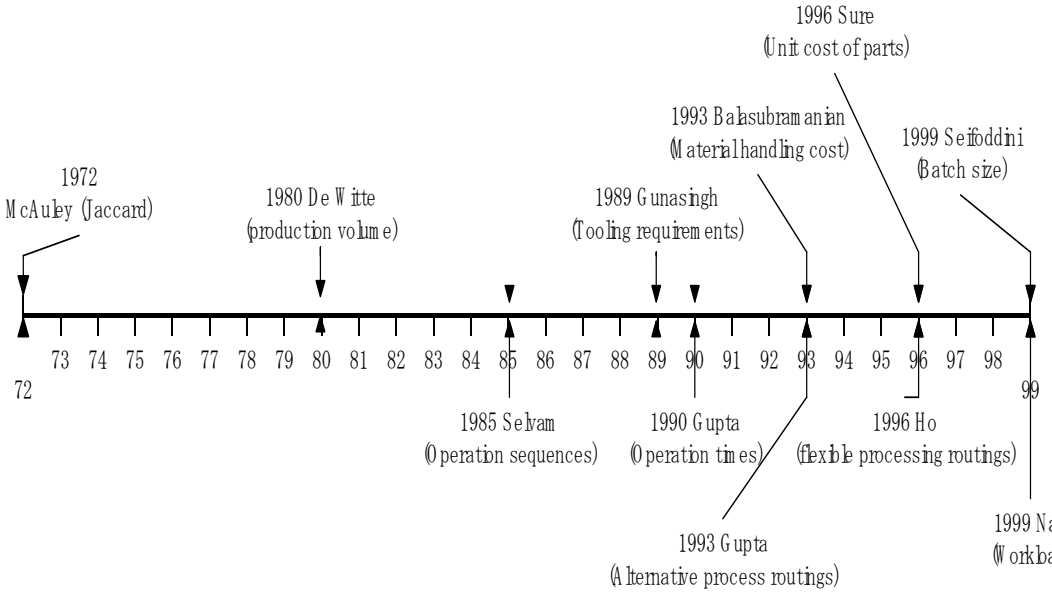


Figure 4. The similarity coefficient's evolutionary timeline

Finally, some impressive progresses that have been achieved in the late 1990s were workload (Nair and Narendran, 1999) and batch size (Seifoddini and Tjahjana, 1999) consideration in the definition of similarity coefficients.
The similarity coefficient's evolutionary timeline is given in figure 4.

7. Comparative study

7.1 The objective of the comparison

Although a large number of similarity coefficients exist in the literature, only a handful has been used for solving CF problems. Among various similarity coefficients, Jaccard similarity coefficient (Jaccard, 1908) was the most used similarity coefficient in the literature (Table 5). However, contradictory viewpoints among researchers have been found in the previous studies: some researchers advocated the dominant power of Jaccard similarity coefficient; whereas some other researchers emphasized the drawbacks of Jaccard similarity coefficient and recommended other similarity coefficients; moreover, several researchers believed that there is no difference between Jaccard and other similarity coefficients, they considered that none of the similarity coefficients seems to perform always well under various cell formation situations.
Therefore, a comparative research is crucially necessary to evaluate various similarity coefficients. Based on the comparative study, even if we cannot find a dominant similarity coefficient for all cell formation situations, at least we need to know which similarity coefficient is more efficient and more appropriate for some specific cell formation situations.
In this paper, we investigate the performance of twenty well-known similarity coefficients. A large number of numerical data sets, which are taken from the open literature or generated specifically, are tested on nine performance measures.

7.2 Previous comparative studies

Four studies that have focused on comparing various similarity coefficients and related cell formation procedures have been published in the literature.
Mosier (1989) applied a mixture model experimental approach to compare seven similarity coefficients and four clustering algorithms. Four performance measures were used to judge the goodness of solutions: simple matching

measure, generalized matching measure, product moment measure and intercellular transfer measure. As pointed out by Shafer and Rogers (1993), the limitation of this study is that three of the four performance measures are for measuring how closely the solution generated by the cell formation procedures matched the original machine-part matrix. However, the original machine-part matrix is not necessarily the best or even a good configuration. Only the last performance measure, intercellular transfer measure is for considering specific objectives associated with the CF problem.

Shafer and Rogers (1993) compared sixteen similarity coefficients and four clustering algorithms. Four performance measures were used to evaluate the solutions. Eleven small, binary machine-part group technology data sets mostly from the literature were used for the purpose of comparison. However, small and/or "well-structured" data sets may not have sufficient discriminatory power to separate "good" from "inferior" techniques. Further, results based on a small number of data sets may have little general reliability due to clustering results' strong dependency on the input data (Vakharia & Wemmerlöv, 1995; Milligan & Cooper, 1987; Anderberg, 1973).

Seifoddini and Hsu (1994) introduced a new performance measure: grouping capability index (GCI). The measure is based on exceptional elements and has been widely used in the subsequent researches. However, only three similarity coefficients have been tested in their study.

Vakharia and Wemmerlöv (1995) studied the impact of dissimilarity measures and clustering techniques on the quality of solutions in the context of cell formation. Twenty-four binary data sets were solved to evaluate eight dissimilarity measures and seven clustering algorithms. Some important insights have been provided by this study, such as data set characteristics, stopping parameters for clustering, performance measures, and the interaction between dissimilarity coefficients and clustering procedures. Unfortunately, similarity coefficients have not been discussed in this research.

8. Experimental design

8.1 Tested similarity coefficients

*T*wenty well-known similarity coefficients (Table 6) are compared in this paper. Among these similarity coefficients, several of them have never been studied by previous comparative researches.

Coefficient	Definition S_{ij}	Range
1. Jaccard	$a/(a+b+c)$	0-1
2. Hamann	$[(a+d)-(b+c)]/[(a+d)+(b+c)]$	-1-1
3. Yule	$(ad-bc)/(ad+bc)$	-1-1
4. Simple matching	$(a+d)/(a+b+c+d)$	0-1
5. Sorenson	$2a/(2a+b+c)$	0-1
6. Rogers and Tanimoto	$(a+d)/[a+2(b+c)+d]$	0-1
7. Sokal and Sneath	$2(a+d)/[2(a+d)+b+c]$	0-1
8. Rusell and Rao	$a/(a+b+c+d)$	0-1
9. Baroni-Urbani and Buser	$[a+(ad)^{1/2}]/[a+b+c+(ad)^{1/2}]$	0-1
10. Phi	$(ad-bc)/[(a+b)(a+c)(b+d)(c+d)]^{1/2}$	-1-1
11. Ochiai	$a/[(a+b)(a+c)]^{1/2}$	0-1
12. PSC	$a^2/[(b+a)*(c+a)]$	0-1
13. Dot-product	$a/(b+c+2a)$	0-1
14. Kulczynski	$1/2[a/(a+b)+a/(a+c)]$	0-1
15. Sokal and Sneath 2	$a/[a+2(b+c)]$	0-1
16. Sokal and Sneath 4	$1/4[a/(a+b)+a/(a+c)+d/(b+d)+d/(c+d)]$	0-1
17. Relative matching	$[a+(ad)^{1/2}]/[a+b+c+d+(ad)^{1/2}]$	0-1
18. Chandrasekharan & Rajagopalan (1986b)	$a/Min[(a+b),(a+c)]$	0-1
19. MaxSC	$Max[a/(a+b),a/(a+c)]$	0-1
20. Baker & Maropoulos (1997)	$a/Max[(a+b),(a+c)]$	0-1

Table 6: Definitions and ranges of selected similarity coefficients a: the number of parts visit both machines; b: the number of parts visit machine i but not j; c: the number of parts visit machine j but not i; d: the number of parts visit neither machine.

8.2 Data sets

It is desirable to judge the effectiveness of various similarity coefficients under varying data sets conditions. The tested data sets are classified into two distinct groups: selected from the literature and generated deliberately. Previous comparative studies used either of them to evaluate the performance of various similarity coefficients. Unlike those studies, this paper uses both types of the data sets to evaluate twenty similarity coefficients.

8.2.1 Data sets selected from the literature

In the previous comparative studies, Shafer and Rogers (1993), and Vakharia and Wemmerlöv (1995) took 11 and 24 binary data sets from the literature, respectively. The advantage of the data sets from the literature is that they stand for a variety of CF situations. In this paper, 70 data sets are selected from the literature. Table 7 shows the details of the 70 data sets.

8.2.2 Data sets generated deliberately

From the computational experience with a wide variety of CF data sets, one finds that it may not always be possible to obtain a good GT solution, if the original CF problem is not amenable to well-structural data set (Chandrasekharan & Rajagopalan, 1989). Hence, it is important to evaluate the quality of solutions of various structural data sets. Using data sets that are generated deliberately is a shortcut to evaluate the GT solutions obtained by various similarity coefficients. The generation process of data sets is often controlled by using experimental factors. In this paper, we use two experimental factors to generate data sets.

- Ratio of non-zero Element in Cells (REC)

Density is one of the most used experimental factors (Miltenburg & Zhang, 1991). However, in our opinion, density is an inappropriate factor for being used to control the generation process of cell formation data sets. We use following Fig.5 to illustrate this problem.

Cell formation data are usually presented in a machine-part incidence matrix such as Fig.5a. The matrix contains 0s and 1s elements that indicate the machine requirements of parts (to show the matrix clearly, 0s are usually unshown). Rows represent machines and columns represent parts.

A '1' in the i^{th} row and j^{th} column represents that the j^{th} part needs an operation on the i^{th} machine; similarly, a '0' in the i^{th} row and j^{th} column represents that the i^{th} machine is not needed to process the j^{th} part.

For Fig.5a, we assume that two machine-cells exist. The first cell is constructed by machines 2 ,4, 1 and parts 1, 3, 7, 6, 10; The second cell is constructed by machines 3, 5 and parts 2, 4, 8, 9, 5, 11. Without loss of generality, we use Fig.5b to represent Fig.5a. The two cells in Fig.5a are now shown as capital letter 'A', we call 'A' as the inside cell region. Similarly, we call 'B' as the outside cell region.

There are three densities that are called Problem Density (PD), non-zero elements Inside cells Density (ID) and non-zero elements Outside cells Density (OD). The calculations of these densities are as follows:

$$PD=\frac{total\ number\ of\ non\text{-}zero\ elements\ in\ regions\ A+B}{total\ number\ of\ elements\ in\ regions\ A+B} \quad (7)$$

$$ID=\frac{total\ number\ of\ non\text{-}zero\ elements\ in\ regions\ A}{total\ number\ of\ elements\ in\ regions\ A} \quad (8)$$

$$OD=\frac{total\ number\ of\ non\text{-}zero\ elements\ in\ regions\ B}{total\ number\ of\ elements\ in\ regions\ B} \quad (9)$$

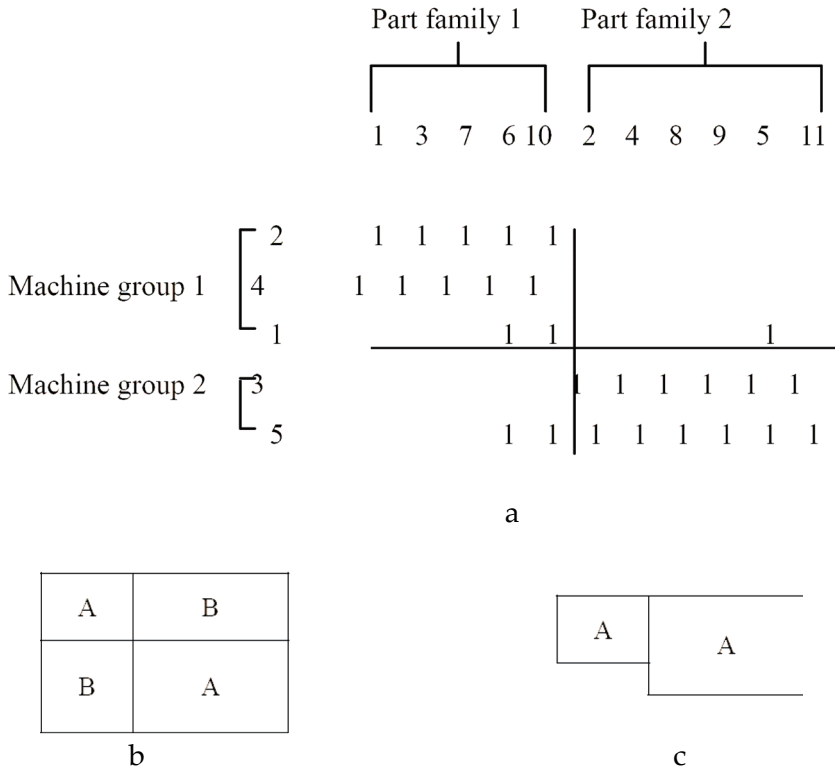


Figure 5. Illustration of three densities used by previous studies

In the design of cellular manufacturing systems, what we concerned is to find out appropriate machine-part cells – the region A. In practice, region B is only

a virtual region that does not exist in the real job shops. For example, if Fig.5a is applied to a real-life job shop, Fig.5c is a possible layout. There is no region B exists in the real-life job shop. Therefore, we conclude that region B based densities are meaningless. Since PD and OD are based on B, this drawback weakens the quality of generated data sets in the previous comparative studies.
To overcome the above shortcoming, we introduce a ratio to replace the density used by previous researchers. The ratio is called as Ratio of non-zero Element in Cells (REC) and is defined as follows:

$$REC = \frac{total\ number\ of\ non\text{-}zero\ elements}{total\ number\ of\ elements\ in\ region\ A} \qquad (10)$$

The definition is intuitive. REC can also be used to estimates the productive capacity of machines. If REC is bigger than 1, current machine capacity cannot response to the productive requirements of parts. Thus, additional machines need to be considered. Therefore, REC can be used as a sensor to assess the capacity of machines.

- Radio of Exceptions (RE)

The second experimental factor is Radio of Exceptions (RE). An exception is defined as a '1' in the region B (an operation outside the cell). We define RE as follows:

$$RE = \frac{total\ number\ of\ non\text{-}zero\ elements\ in\ region\ B}{total\ number\ of\ non\text{-}zero\ elements} \qquad (11)$$

RE is used to judge the "goodness" of machine-part cells and distinguish well-structured problems from ill-structured problems.
In this paper, 3 levels of REC, from sparse cells (0.70) to dense cells (0.90), and 8 levels of RE, from well-structured cells (0.05) to ill-structured cells (0.40), are examined. 24 (3*8) combinations exist for all levels of the two experimental factors. For each combination, five 30*60-sized (30 machines by 60 parts) problems are generated. The generation process of the five problems is similar by using the random number. Therefore, a total of 120 test problems for all 24 combines are generated, each problem is made up of 6 equally sized cells. The levels of REC and RE are shown in Table 8.

	Data set	Size	NC
1.	Singh & Rajamani	1996	4*4 2
2.	Singh & Rajamani	1996	4*5 2
3.	Singh & Rajamani	1996	5*6 2
4.	Waghodekar& Sahu	1984	5*7 2
5.	Waghodekar& Sahu	1984	5*7 2
6.	Chow & Hawaleshka	1992	5*11 2
7.	Chow & Hawaleshka	1993a	5*13 2
8.	Chow & Hawaleshka	1993b	5*13 2
9.	Seifoddini	1989b	5*18 2
10.	Seifoddini	1989b	5*18 2
11.	Singh & Rajamani	1996	6*8 2
12.	Chen *et al*.	1996	7*8 3
13.	Boctor	1991	7*11 3
14.	Islam & Sarker	2000	8*10 3
15.	Seifoddini & Wolfe	1986	8*12 3
16.	Chandrasekharan & Rajagopalan	1986a	8*20 2, 3
17.	Chandrasekharan & Rajagopalan	1986b	8*20 2,3
18.	Faber & Carter	1986	9*9 2
19.	Seifoddini & Wolfe	1986	9*12 3
20.	Chen *et al*.	1996	9*12 3
21.	Hon & Chi	1994	9*15 3
22.	Selvam & Balasubramanian	1985	10*5 2
23.	Mosier & Taube	1985a	10*10 3
24.	Seifoddini & Wolfe	1986	10*12 3
25.	McAuley	1972	12*10 3
26.	Seifoddini	1989a	11*22 3
27.	Hon & Chi	1994	11*22 3
28.	De Witte	1980	12*19 2, 3
29.	Irani & Khator	1986	14*24 4
30.	Askin & Subramanian	1987	14*24 4
31.	King	1980(machine 6, 8removed)	14*43 4, 5
32.	Chan & Milner	1982	15*10 3
33.	Faber & Carter	1986	16*16 2, 3
34.	Sofianopoulou	1997	16*30 2, 3
35.	Sofianopoulou	1997	16*30 2, 3
36.	Sofianopoulou	1997	16*30 2, 3
37.	Sofianopoulou	1997	16*30 2, 3

38.	Sofianopoulou	1997	16*30 2, 3
39.	Sofianopoulou	1997	16*30 2, 3
40.	Sofianopoulou	1997	16*30 2, 3
41.	Sofianopoulou	1997	16*30 2, 3
42.	Sofianopoulou	1997	16*30 2, 3
43.	Sofianopoulou	1997	16*30 2, 3
44.	King	1980	16*43 4, 5
45.	Boe & Cheng	1991 (mach 1 removed)	19*35 4
46.	Shafer & Rogers	1993	20*20 4
47.	Shafer & Rogers	1993	20*20 4
48.	Shafer & Rogers	1993	20*20 4
49.	Mosier & Taube	1985b	20*20 3, 4
50.	Boe & Cheng	1991	20*35 4
51.	Ng	1993	20*35 4
52.	Kumar & Kusiak	1986	23*20 2, 3
53.	McCormick *et al.* 1	972	24*16 6
54.	Carrie	1973	24*18 3
55.	Chandrasekharan & Rajagopalan	1989	24*4 7
56.	Chandrasekharan & Rajagopalan 1	989	24*40 7
57.	Chandrasekharan & Rajagopalan	1989	24*40 7
58.	Chandrasekharan & Rajagopalan	1989	24*40 7
59.	Chandrasekharan & Rajagopalan	1989	24*40 7
60.	Chandrasekharan & Rajagopalan	1989	24*40 7
61.	Chandrasekharan & Rajagopalan	1989	24*40 7
62.	McCormick *et al.*	1972	27*27 8
63.	Carrie	1973	28*46 3, 4
64.	Lee *et al.*	1997	30*40 6
65.	Kumar & Vannelli	1987	30*41 2,3,9
66.	Balasubramanian & Panneerselvam	1993	36*21 7
67.	King & Nakornchai	1982	36*90 4, 5
68.	McCormick *et al.*	1972	37*53 4,5,6
69.	Chandrasekharan & Rajagopalan	1987	40*100 10
70.	Seifoddini & Tjahjana	1999	50*22 14

Table 7. Data sets from literature

Level	1	2	3	4	5	6	7	8
REC	0.70	0.80	0.90	--	--	--	--	--
RE	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40

Table 8. Test levels of REC and RE

8.3 Clustering procedure

The most well-known clustering procedures that have been applied to cell formation are single linkage clustering (SLC) algorithm, complete linkage clustering (CLC) algorithm and average linkage clustering (ALC) algorithm. These three procedures have been investigated by lots of studies. A summary of the past comparative results is shown in Table 9.
Due to that ALC has the advantage of showing the greatest robustness regardless of similarity coefficients, in this paper, we select ALC as the clustering algorithm to evaluate the twenty similarity coefficients (Table 6).

Procedure	**Advantage**	**Drawback**
SLC	Simplicity; Minimal computational requirement; Tends to minimize the degree of adjusted machine duplication. (Vakharia & Wemmerlöv, 1995).	Largest tendency to chain; Leads to the lowest densities and the highest degree of single part cells (Seifoddini, 1989a; Gupta, 1991; Vakharia & Wemmerlöv, 1995).
CLC	Simplicity; Minimal computational requirement (does the reverse of SLC)	Performed as the worst procedure (Vakharia & Wemmerlöv, 1995; Yasuda & Yin, 2001).
ALC	Performed as the best procedure; Produces the lowest degree of chaining; Leads to the highest cell densities; Indifferent to choice of similarity coefficients; Few single part cells (Tarsuslugil & Bloor, 1979; Seifoddini, 1989a; Vakharia & Wemmerlöv, 1995; Yasuda & Yin, 2001).	Requires the highest degree of machine duplication; Requires more computation (Vakharia & Wemmerlöv, 1995).

Table 9. Comparative results of SLC, ALC and CLC

The ALC algorithm usually works as follows:

Step (1). Compute similarity coefficients for all machine pairs and store the values in a similarity matrix.

Step (2). Join the two most similar objects (two machines, a machine and a machine group or two machine groups) to form a new machine group.

Step (3). Evaluate the similarity coefficient between the new machine group and other remaining machine groups (machines) as follows:

$$S_{tv} = \frac{\sum_{i \in t} \sum_{j \in v} S_{ij}}{N_t N_v} \qquad (12)$$

where i is the machine in the machine group t; j is the machine in the machine group v. And N_t is the number of machines in group t; N_v is the number of machines in group v.

Step (4). When all machines are grouped into a single machine group, or pre determined number of machine groups has obtained, go to step 5; otherwise, go back to step 2.

Step (5). Assign each part to the cell, in which the total number of exceptions is minimum.

8.4 Performance measures

A number of quantitative performance measures have been developed to evaluate the final cell formation solutions. Sarker and Mondal (1999) reviewed and compared various performance measures.

Nine performance measures are used in this study to judge final solutions. These measures provide different viewpoints by judging solutions from different aspects.

8.4.1 Number of exceptional elements (EE)

Exceptional elements are the source of inter-cell movements of parts. One objective of cell formation is to reduce the total cost of material handling. Therefore, EE is the most simple and intuitive measure for evaluating the cell formation solution.

8.4.2 Grouping efficiency

Grouping efficiency is one of the first measures developed by Chandrasekharan and Rajagopalan (1986a, b). Grouping efficiency is defined as a weighted average of two efficiencies η_1 and η_2:

$$\eta = w\eta_1 + (1-w)\eta_2 \tag{13}$$

where

$$\eta_1 = \frac{o-e}{o-e+v}$$

$$\eta_2 = \frac{MP-o-v}{MP-o-v+e}$$

M number of machines
P number of parts
o number of operations (1s) in the machine-part matrix $\{a_{ik}\}$
e number of exceptional elements in the solution
v number of voids in the solution

A value of 0.5 is recommended for w. η_1 is defined as the ratio of the number of 1s in the region A (Fig.5b) to the total number of elements in the region A (both 0s and 1s). Similarly, η_2 is the ratio of the number of 0s in the region B to the total number of elements in the region B (both 0s and 1s). The weighting factor allows the designer to alter the emphasis between utilization and intercell movement. The efficiency ranges from 0 to 1.
Group efficiency has been reported has a lower discriminating power (Chandrasekharan & Rajagopalan, 1987). Even an extremely bad solution with large number of exceptional elements has an efficiency value as high as 0.77.

8.4.3 Group efficacy

To overcome the problem of group efficiency, Kumar and Chandrasekharan (1990) introduced a new measure, group efficacy.

$$\tau = (1-\varphi)/(1+\phi) \tag{14}$$

where φ is the ratio of the number of exceptional elements to the total number of elements; ϕ is the ratio of the number of 0s in the region A to the total number of elements.

8.4.4 Machine utilization index (Grouping measure, GM)

Proposed by Miltenburg and Zhang (1991), which is used to measure machine utilization in a cell. The index is defined as follows:

$$\eta_g = \eta_u - \eta_m \tag{15}$$

where $\eta_u = d/(d+v)$ and $\eta_m = 1-(d/o)$. d is the number of 1s in the region A, η_u is the measure of utilization of machines in a cell and η_m is the measure of inter-cell movements of parts. η_g ranges form –1 to 1, η_u and η_m range from 0 to 1. A bigger value of machine utilization index η_g is desired.

8.4.5 Clustering measure (CM)

This measure tests how closely the 1s gather around the diagonal of the solution matrix, the definition of the measure is as follows (Singh & Rajamani, 1996).

$$\eta_c = \frac{\left\{\sum_{i=1}^{M}\sum_{k=1}^{P}\left(\sqrt{\delta_h^2(a_{ik})+\delta_v^2(a_{ik})}\right)\right\}}{\sum_{i=1}^{M}\sum_{k=1}^{P}a_{ik}} \tag{16}$$

where $\delta_h(a_{ik})$ and $\delta_v(a_{ik})$ are horizontal and vertical distances between a non-zero entry a_{ik} and the diagonal, respectively.

$$\delta_h = i - \frac{k(M-1)}{(P-1)} - \frac{(P-M)}{(P-1)} \tag{17}$$

$$\delta_v = k - \frac{i(P-1)}{(M-1)} - \frac{(P-M)}{(M-1)} \tag{18}$$

8.4.6 Grouping index (GI)

Nair and Narendran (1996) indicated that a good performance measure should be defined with reference to the block diagonal space. And the definition should ensure equal weightage to voids (0s in the region A) and exceptional elements. They introduced a measure, incorporating the block diagonal space, weighting factor and correction factor.

$$\gamma = \frac{1 - \frac{qv + (1-q)(e-A)}{B}}{1 + \frac{qv + (1-q)(e-A)}{B}} \tag{19}$$

where B is the block diagonal space and q is a weighting factor ranges between 0 and 1. A=0 for $e \le B$ and $A = e - B$ for $e > B$. For convenience, equation (19) could be written as follows:

$$\gamma = \frac{1-\alpha}{1+\alpha} \tag{20}$$

where

$$\alpha = \frac{qv + (1-q)(e-A)}{B} \tag{21}$$

Both α and γ range from 0 to 1.

8.4.7 Bond energy measure (BEM)

McCormick *et al*. (1972) used the BEM to convert a binary matrix into a block diagonal form. This measure is defined as follows:

$$\eta_{BE} = \frac{\sum_{i=1}^{M}\sum_{k=1}^{P-1} a_{ik}a_{i(k+1)} + \sum_{i=1}^{M-1}\sum_{k=1}^{P} a_{ik}a_{(i+1)k}}{\sum_{i=1}^{M}\sum_{k=1}^{P} a_{ik}} \tag{22}$$

Bond energy is used to measure the relative clumpiness of a clustered matrix. Therefore, the more close the 1s are, the larger the bond energy measure will be.

8.4.8 Grouping capability index (GCI)

Hsu (1990) showed that neither group efficiency nor group efficacy is consistent in predicting the performance of a cellular manufacturing system based on the structure of the corresponding machine-part matrix (Seifoddini & Djassemi, 1996). Hsu (1990) considered the GCI as follows:

$$GCI = 1 - \frac{e}{o} \tag{23}$$

Unlike group efficiency and group efficacy, GCI excludes zero entries from the calculation of grouping efficacy.

8.4.9. Alternative routeing grouping efficiency (ARG efficiency)

ARG was propose by Sarker and Li (1998). ARG evaluates the grouping effect in the presence of alternative routings of parts. The efficiency is defined as follows:

$$\eta_{ARG} = \frac{(1-\frac{e}{o'})(1-\frac{v}{z'})}{(1+\frac{e}{o'})(1+\frac{v}{z'})} = \left(\frac{o'-e}{o'+e}\right)\left(\frac{z'-v}{z'+v}\right) \qquad (24)$$

where o' is the total number of 1s in the original machine-part incidence matrix with multiple process routings, z' is the total number of 0s in the original machine-part incidence matrix with multiple process routings. ARG efficiency can also be used to evaluate CF problems that have no multiple process routings of parts. The efficiency ranges from 0 to 1 and is independent of the size of the problem.

9. Comparison and results

Two key characteristics of similarity coefficients are tested in this study, discriminability and stability. In this study, we compare the similarity coefficients by using following steps.

Comparative steps

1. Computation.
   - 1.1. At first, solve each problem in the data sets by using 20 similarity coefficients; compute performance values by 9 performance measures. Thus, we obtain at least a total of δ*20*9 solutions. δ is the number of the problems (some data sets from literature are multi-problems due to the different number of cells, see the item NC of Table 7).
   - 1.2. Average performance values matrix: create a matrix whose rows are problems and columns are 9 performance measures. An element in row i and column j indicates, for problem i and performance measure j, the average performance value produced by 20 similarity coefficients.

2. Based on the results of step 1, construct two matrixes whose rows are 20 similarity coefficients and columns are 9 performance measures, an entry SM_{ij} in the matrixes indicates:
   2.1. Discriminability matrix: the number of problems to which the similarity coefficient i gives the best performance value for measure j.
   2.2. Stability matrix: the number of problems to which the similarity coefficient i gives the performance value of measure j with at least average value (better or equal than the value in the matrix of step 1.2).
3. For each performance measure, find the top 5 values in the above two matrixes. The similarity coefficients correspond to these values are considered to be the most discriminable/stable similarity coefficients for this performance measure.
4. Based on the results of step 3, for each similarity coefficient, find the number of times that it has been selected as the most discriminable/stable coefficient for the total 9 performance measures.

We use small examples here to show the comparative steps.

Step 1.1: a total of 214 problems were solved. 120 problems were deliberately generated; 94 problems were from literature, see Table 2 (some data sets were multi-problems due to the different number of cells). A to tal of 38,520 (214*20*9) performance values were gotten by using 20 similarity coefficients and 9 performance measures. For example, by using Jaccard similarity coefficient, the 9 performance values of the problem McCormick *et al*. (no.62 in Table 7) are as follows (Table 10):

	EE	*Grouping efficiency*	*Group efficacy*	*GM*	*CM*	*GI*	*BEM*	*GCI*	ARG
Jaccard	87	0.74	0.45	0.25	7.85	0.44	1.07	0.6	0.32

Table 10: The performance values of McCormick et al. by using Jaccard similarity coefficient

Step 1.2: The average performance values matrix contained 214 problems (rows) and 9 performance measures (columns). An example of row (problem McCormick *et al*.) is as follows (Table 11):

	EE	*Grouping efficiency*	*Group efficacy*	*GM*	*CM*	*GI*	*BEM*	*GCI*	**ARG**
Average values	94.7	0.77	0.45	0.28	8.06	0.4	1.06	0.57	0.31

Table 11. The average performance values of 20 similarity coefficients, for the problem McCormick *et al*

We use Jaccard similarity coefficient and the 94 problems from literature to explain following steps 2, 3, and 4.

Step 2.1 (discriminability matrix): among the 94 problems and for each performance measure, the numbers of problems to which Jaccard gave the best values are shown in Table 12. For example, the 60 in the column EE means that comparing with other 19 similarity coefficients, Jaccard produced minimum exceptional elements to 60 problems.

	EE	*Grouping efficiency*	*Group efficacy*	*GM*	*CM*	*GI*	*BEM*	*GCI*	**ARG**
Jaccard	60	51	55	62	33	65	41	60	57

Table 12. The number of problems to which Jaccard gave the best performance values

Step 2.2 (stability matrix): among the 94 problems and for each performance measure, the numbers of problems to which Jaccard gave the value with at least average value (matrix of step 1.2) are shown in Table 13. For example, the meaning of 85 in the column EE is as follows: comparing with the average exceptional elements of 94 problems in the matrix of step 1.2, the numbers of problems to which Jaccard produced a fewer exceptional elements are 85.

	EE	*Grouping efficiency*	*Group efficacy*	*GM*	*CM*	*GI*	*BEM*	*GCI*	**ARG**
Jaccard	85	85	85	89	69	91	75	88	73

Table 13. The number of problems to which Jaccard gave the best performance values

Step 3: For example, for the exceptional elements, the similarity coefficients that corresponded to the top 5 values in the discriminability matrix are Jaccard, Sorenson, Rusell and Rao, Dot-product, Sokal and Sneath 2, Relative matching, and Baker and Maropoulos. These similarity coefficients are considered as the most discriminable coefficients for the performance measure – exceptional elements. The same procedures are conducted to the other performance measures and stability matrix.

Step 4: Using the results of step 3, Jaccard have been selected 5/6 times as the most discriminable/stable similarity coefficient. That means, among 9 performance measures, Jaccard is the most discriminable/stable similarity coefficient for 5/6 performance measures. The result is shown in the column – literature of Table 14.
The results are shown in Table 14 and Figs. 6-8 (in the figures, the horizontal axes are 20 similarity coefficients and the vertical axes are 9 performance measures).

The tables and figures show the number of performance measures for which these 20 similarity coefficients have been regarded as the most discriminable/stable coefficients. The columns of the table represent different conditions of data sets. The column – literature includes all 94 problems from literature; the column – all random includes all 120 deliberately generated problems. The deliberately generated problems are further investigated by using different levels of REC and RE.

"Literature" and "All random" columns in Table 14 (also Fig.6) give the performance results of all 214 tested problems. We can find that Jaccard and Sorenson are two best coefficients. On the other hand, four similarity coefficients: Hamann, Simple matching, Rogers & Tanimoto, and Sokal & Sneath are inefficient in both discriminability and stability.

"REC" columns in table 9 (also Fig.3) show the performance results under the condition of different REC ratios. We can find that almost all similarity coefficients perform well under a high REC ratio. However, four similarity coefficients: Hamann, Simple matching, Rogers & Tanimoto, and Sokal & Sneath, again produce bad results under the low REC ratio.
"RE" columns in Table 14 (also Fig.8) give the performance results under the

condition of different RE ratios. All similarity coefficients perform best under a low RE ratio (data sets are well-structured). Only a few of similarity coefficients perform well under a high RE ratio (data sets are ill-structured), Sokal & Sneath 2 is very good for all RE ratios. Again, the four similarity coefficients: Hamann, Simple matching, Rogers & Tanimoto, and Sokal & Sneath, perform badly under high RE ratios.

No.	**Similarity Coefficient**	*Literature*		*All random*		*REC*						*RE*					
						0.7		*0.8*		*0.9*		*0.05-0.15*		*0.2-0.3*		*0.35-0.4*	
		D	*S*	*D*	*S*	*D*	*S*	*D*	*S*	*D*	*S*	*D*	*S*	*D*	*S*	*D*	*S*
1	Jaccard	5	6	6	9	8	9	8	9	9	9	9	9	9	9	8	9
2	Hamann	0	0	2	1	1	1	2	3	7	7	9	9	1	0	2	2
3	Yule	4	4	2	6	3	7	5	7	7	8	9	9	2	6	6	7
4	Simple matching	0	0	2	0	1	0	3	5	6	8	9	9	0	0	2	2
5	Sorenson	6	4	9	8	7	9	8	9	9	9	9	9	9	9	7	7
6	Rogers & Tanimoto	0	0	2	1	2	2	4	4	6	7	9	9	1	2	2	2
7	Sokal & Sneath	0	0	0	0	2	1	5	6	6	8	9	9	1	1	2	2
8	Rusell & Rao	4	4	5	3	5	5	9	8	8	6	9	9	9	8	6	6
9	Baoroni-Urban & Buser	5	6	1	3	3	7	9	7	7	8	9	9	4	7	2	6
10	Phi	5	5	6	6	9	7	8	8	7	8	9	9	9	8	7	7
11	Ochiai	1	4	8	7	9	7	8	8	9	9	9	9	9	9	7	7
12	PSC	2	2	9	8	9	9	9	8	9	9	9	9	9	9	8	9
13	Dot-product	3	5	9	8	7	9	8	9	9	9	9	9	9	9	7	7
14	Kulczynski	2	5	8	7	8	8	8	8	9	9	9	9	9	9	7	7
15	Sokal & Sneath 2	4	5	6	8	9	9	7	9	9	9	9	9	9	9	9	9
16	Sokal & Sneath 4	5	5	7	6	8	7	8	8	7	8	9	9	8	8	7	7
17	Relative matching	5	4	4	8	7	9	9	9	9	9	9	9	5	9	6	8
18	Chandraseharan & Rajagopalan	2	5	8	6	9	8	8	8	7	7	9	9	9	9	6	7
19	MaxSC	1	4	8	6	9	8	8	8	7	7	9	9	9	9	6	7
20	Baker & Maropoulos	5	3	6	9	7	9	8	9	9	9	9	9	6	9	6	8

Table 14. Comparative results under various conditions. *D:* discriminability; *S:* stability

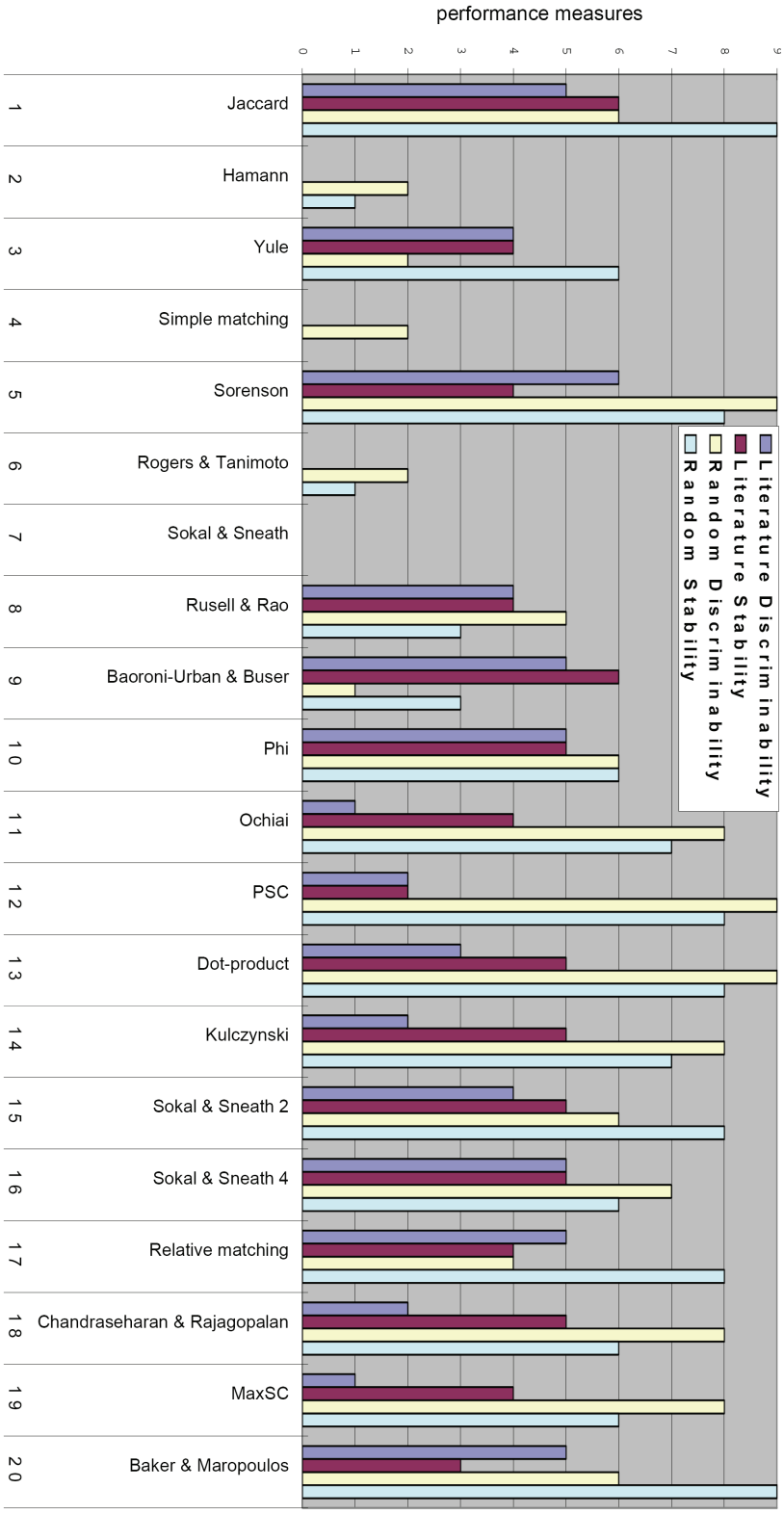


Figure 6. Performance for all tested problems

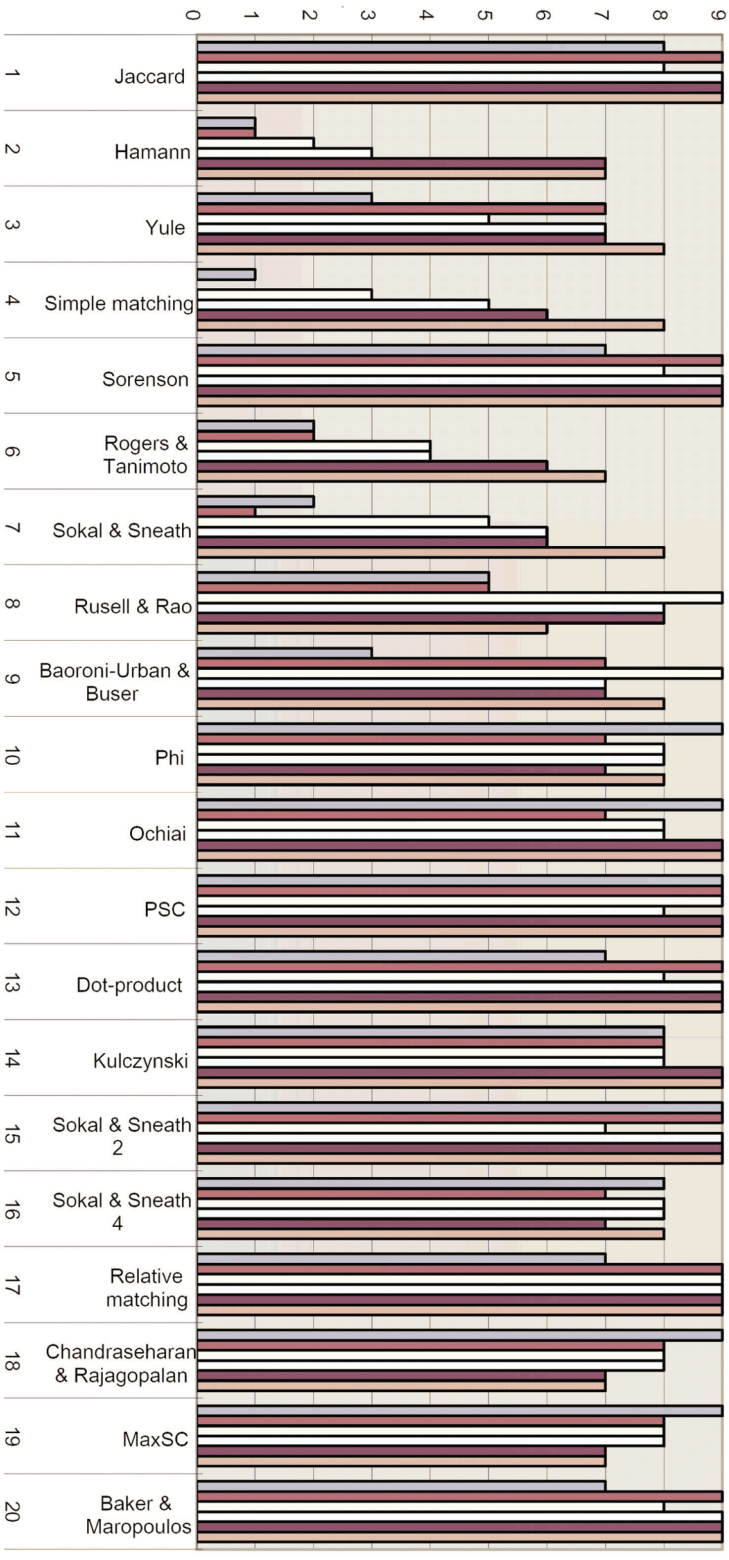


Figure 7. Performance under different REC

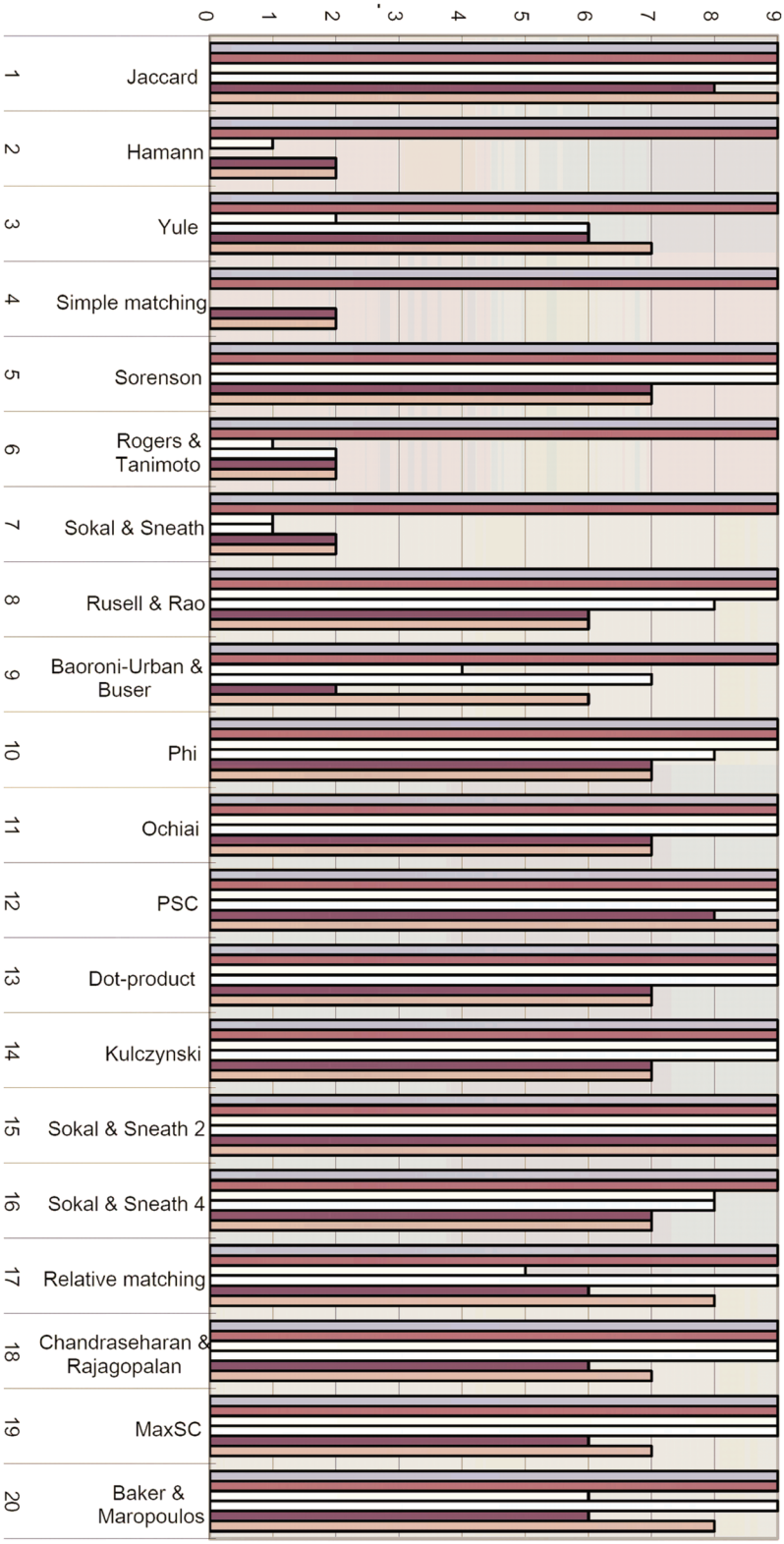


Figure 8. Performance under different RE

In summary, three similarity coefficients: Jaccard, Sorenson, and Sokal & Sneath 2 perform best among twenty tested similarity coefficients. Jaccard emerges from the twenty similarity coefficients for its stability. For all problems, from literature or deliberately generated; and for all levels of both REC and RE ratios, Jaccard similarity coefficient is constantly the most stable coefficient among all twenty similarity coefficients. Another finding in this study is four similarity coefficients: Hamann, Simple matching, Rogers & Tanimoto, and Sokal & Sneath are inefficient under all conditions. So, these similarity coefficients are not recommendable for using in cell formation applications.

9. Conclusions

In this paper various similarity coefficients to the cell formation problem were investigated and reviewed. Previous review studies were discussed and the need for this review was identified. The reason why the similarity coefficient based methods (SCM) is more flexible than other cell formation methods were explained through a simple example. We also proposed a taxonomy which is combined by two distinct dimensions. The first dimension is the general-purpose similarity coefficients and the second is the problem-oriented similarity coefficients. The difference between two dimensions is discussed through three similarity coefficients. Based on the framework of the proposed taxonomy, existing similarity (dissimilarity) coefficients developed so far were reviewed and mapped onto the taxonomy. The details of each production information based similarity coefficient were simply discussed and a evolutionary timeline was drawn based on reviewed similarity coefficients. Although a number of similarity coefficients have been proposed, very fewer comparative studies have been done to evaluate the performance of various similarity coefficients. This paper evaluated the performance of twenty well-known similarity coefficients. 94 problems from literature and 120 problems generated deliberately were solved by using the twenty similarity coefficients. To control the generation process of data sets, experimental factors have been discussed. Two experimental factors were proposed and used for generating experimental problems. Nine performance measures were used to judge the solutions of the tested problems. The numerical results showed that three similarity coefficients are more efficient and four similarity coefficients are inefficient for solving the cell formation problems. Another finding is that Jaccard similarity coefficient is the most stable similarity coefficient. For the further studies, we

suggest comparative studies in consideration of some production factors, such as production volumes, operation sequences, etc. of parts.

7. References

Agarwal, A., Sarkis, J., 1998. A review and analysis of comparative performance studies on functional and cellular manufacturing layouts. Computers and Industrial Engineering 34, 77-89.

Akturk, M.S., Balkose, H.O., 1996. Part-machine grouping using a multi-objective cluster analysis. International Journal of Production Research 34, 2299-2315.

Al-Sultan, K.S., Fedjki, C.A., 1997. A genetic algorithm for the part family formation problem. Production Planning & Control 8, 788-796.

Anderberg, M.R., 1973. Cluster analysis for applications (New York: Academic Press).

Arthanari, T.S., Dodge, Y., 1981. Mathematical programming in statistics (New York: John Wiley & Sons, Inc).

Askin, R.G., Cresswell, S.H., Goldberg, J.B., Vakharia, A.J., 1991. A Hamiltonian path approach to reordering the part-machine matrix for cellular manufacturing. International Journal of Production Research 29, 1081-1100.

Askin, R.G., Selim, H.M., Vakharia, A.J., 1997. A methodology for designing flexible cellular manufacturing systems. IIE Transaction 29, 599-610.

Askin, R.G., & Subramanian, S.P., 1987. A cost-based heuristic for group technology configuration. *International Journal of Production Research, 25*(1), 101-113.

Askin, R.G., Zhou, M., 1998. Formation of independent flow-line cells based on operation requirements and machine capabilities. IIE Transactions 30, 319-329.

Baker, R.P., Maropoulos, P.G., 1997. An automatic clustering algorithm suitable for use by a computer-based tool for the design, management and continuous improvement of cellular manufacturing systems. Computers Integrated Manufacturing Systems 10, 217-230.

Balakrishnan, J., 1996. Manufacturing cell formation using similarity coefficients and pair-wise interchange: formation and comparison. Production Planning & Control 7, 11-21.

Balakrishnan, J., Cheng, C. H., 1998. Dynamic layout algorithms: a state-of-the-art survey. Omega 26, 507-521.

Balakrishnan, J., Jog, P.D., 1995. Manufacturing cell formation using similarity coefficients and a parallel genetic TSP algorithm: formulation and comparison. Mathematical and Computer Modelling 21, 61-73.

Balasubramanian, K.N., Panneerselvam, R., 1993. Covering technique-based algorithm for machine grouping to form manufacturing cells. International Journal of Production Research 31, 1479-1504.

Baroni-Urbani, C., Buser, M.W., 1976. Similarity of binary data. Systematic Zoology 25, 251-259.

Baykasoglu, A., Gindy, N.N.Z., 2000. MOCACEF 1.0: multiple objective capability based approach to form part-machine groups for cellular manufacturing applications. International Journal of Production Research 38, 1133-1161.

Beatty, C.A., 1992. Implementing advanced manufacturing technologies: rules of the road. Sloan Management Review Summer, 49-60.

Ben-Arieh, D., Chang, P.T., 1994. An extension to the p-median group technology algorithm. Computers and Operations Research 21, 119-125.

Ben-Arieh, D., Sreenivasan, R., 1999. Information analysis in a distributed dynamic group technology method. International Journal of Production Economics 60-61, 427-432.

Bijnen, E.J., 1973. Cluster analysis (The Netherlands: Tilburg University Press).

Bishop, Y.M.M., Fienberg, S.E., Holland, P.W., 1975. Discrete multivariate analysis: theory and practice (MA: MIT Press Cambridge).

Boctor, F.F., 1991. A linear formulation of the machine-part cell formation problem. *International Journal of Production Research, 29*(2), 343-356.

Boe, W.J., & Cheng, C.H., 1991. A close neighbour algorithm for designing cellular manufacturing systems. *International Journal of Production Research, 29*(10), 2097-2116.

Burbidge, J.L., 1971. Production flow analysis. Production Engineer 50, 139-152.

Burbidge, J.L., Falster, P., Rhs, J.O., 1991. Why is it difficult to sell group technology and just-in-time to industry? Production Planning & Control 2, 160-166.

Carrie, A.S., 1973. Numerical taxonomy applied to group technology and plant layout. International Journal of Production Research 11, 399-416.

Cedeno, A.A., Suer, G.A., 1997. The use of a similarity coefficient-based method to perform clustering analysis to a large set of data with dissimilar parts. Computers and Industrial Engineering 33, 225-228.

Chan, H.M., & Milner, D.A., 1982. Direct clustering algorithm for group formation in cellular manufacture. *Journal of Manufacturing Systems, 1*(1), 65-75.

Chandrasekharan, M.P., Rajagopalan, R., 1986a. An ideal seed non-hierarchical clustering algorithm for cellular manufacturing. International Journal of Production Research 24, 451-464.

Chandrasekharan, M.P., Rajagopalan, R., 1986b. MODROC: an extension of rank order clustering for group technology. International Journal of Production Research 24, 1221-1233.

Chandrasekharan, M.P., Rajagopalan, R., 1987. ZODIAC: an algorithm for concurrent formation of part families and machine cells. International Journal of Production Research 25, 451-464.

Chandrasekharan, M.P., Rajagopalan, R., 1989. GROUPABILITY: an analysis of the properties of binary data matrices for group technology. International Journal of Production Research. 27, 1035-1052.

Chang, P.T., Lee, E.S., 2000. A multisolution method for cell formation – exploring practical alternatives in group technology manufacturing. Computers and Mathematics with Applications 40, 1285-1296.

Chen, D.S., Chen H.C., & Part, J.M., 1996. An improved ART neural net for machine cell formation. *Journal of Materials Processing Technology, 61*, 1-6.

Cheng, C.H., Goh, C.H., Lee, A., 1995. A two-stage procedure for designing a group technology system. International Journal of Operations & Production Management 15, 41-50.

Cheng, C.H., Gupta, Y.P., Lee, W.H., Wong, K.F., 1998. A TSP-based heuristic for forming machine groups and part families. International Journal of Production Research 36, 1325-1337.

Cheng, C.H., Madan, M.S., Motwani, J., 1996. Designing cellular manufacturing systems by a truncated tree search. International Journal of Production Research 34, 349-361.

Choobineh, F., 1988. A framework for the design of cellular manufacturing systems. International Journal of Production Research 26, 1161-1172.

Choobineh, F., Nare, A., 1999. The impact of ignored attributes on a CMS design. International Journal of Production Research 37, 3231-3245.

Chow, W.S., 1991. Discussion: a note on a linear cell clustering algorithm. International Journal of Production Research 29, 215-216.

Chow, W.S., Hawaleshka, O., 1992. An efficient algorithm for solving the machine chaining problem in cellular manufacturing. Computers and Industrial Engineering 22, 95-100.

Chow, W.S., Hawaleshka, O., 1993a. Minimizing intercellular part movements in manufacturing cell formation. International Journal of Production Research 31, 2161-2170.

Chow, W.S., Hawaleshka, O., 1993b. A novel machine grouping and knowledge-based approach for cellular manufacturing. European Journal of Operational Research 69, 357-372.

Chu, C.H., 1989. Cluster analysis in manufacturing cellular formation. Omega 17, 289-295.

Chu, C.H., Pan, P., 1988. The use of clustering techniques in manufacturing cellular formation. Proceedings of International Industrial Engineering Confer-

ence, Orlando, Florida, pp. 495-500.

Chu, C.H., & Tsai, M., 1990. A comparison of three array-based clustering techniques for manufacturing cell formation. *International Journal of Production Research, 28*(8), 1417-1433.

De Witte, J., 1980. The use of similarity coefficients in production flow analysis. International Journal of Production Research 18, 503-514.

Dimopoulos, C., Mort, N., 2001. A hierarchical clustering methodology based on genetic programming for the solution of simple cell-formation problems. International Journal of Production Research 39, 1-19.

Dutta, S.P., Lashkari, R.S., Nadoli, G., Ravi, T., 1986. A heuristic procedure for determining manufacturing families from design-based grouping for flexible manufacturing systems. Computers and Industrial Engineering 10, 193-201.

Faber, Z., Carter, M.W., 1986. A new graph theory approach for forming machine cells in cellular production systems. In A. Kusiak (ed), Flexible Manufacturing Systems: Methods and Studies (North-Holland: Elsevier Science Publishers B.V), pp. 301-315.

Fazakerley, G.M., 1976. A research report on the human aspects of group technology and cellular manufacture. International Journal of Production Research 14, 123-134.

Gongaware, T.A., Ham, I., 1991. Cluster analysis applications for group technology manufacturing systems. Proceedings Ninth North American Manufacturing Research Conference, pp. 503-508.

Gordon, A.D., 1999. Classification, 2nd edition (US: Chapman & Hall).

Gunasingh, K.R., Lashkari, R.S., 1989. The cell formation problem in cellular manufacturing systems – a sequential modeling approach. Computers and Industrial Engineering 16, 469-476.

Gupta, T., 1991. Clustering algorithms for the design of a cellular manufacturing system – an analysis of their performance. Computers and Industrial Engineering 20, 461-468.

Gupta, T., 1993. Design of manufacturing cells for flexible environment considering alternative routeing. International Journal of Production Research 31, 1259-1273.

Gupta, T., Seifoddini, H., 1990. Production data based similarity coefficient for machine-component grouping decisions in the design of a cellular manufacturing system. International Journal of Production Research 28, 1247-1269.

Han, C., Ham, I., 1986. Multiobjective cluster analysis for part family formations. Journal of Manufacturing Systems 5, 223-230.

Ho, Y.C., Lee, C., Moodie, C.L., 1993. Two sequence-pattern, matching-based, flow analysis methods for multi-flowlines layout design. International Journal of

Production Research 31, 1557-1578.

Ho, Y.C., Moodie, C.L., 1996. Solving cell formation problems in a manufacturing environment with flexible processing and routeing capabilities. International Journal of Production Research 34, 2901-2923.

Holley, J.W., Guilford, J.P., 1964. A note on the G index of agreement. Educational and Psychological Measurement 24, 749-753.

Hon, K.K.B, & Chi, H., 1994. A new approach of group technology part families optimization. *Annals of the CIRP, 43*(1), 425-428.

Hsu, C.P., 1990. *Similarity coefficient approaches to machine-component cell formation in cellular manufacturing: a comparative study*. Ph.D. thesis. Department of Industrial and Manufacturing Engineering, University of Wisconsin-Milwaukee.

Hwang, H., Ree, P., 1996. Routes selection for the cell formation problem with alternative part process plans. Computers and Industrial Engineering 30, 423-431.

Irani, S.A.,& Khator, S.K., 1986. A microcomputer-based design of a cellular manufacturing system. In: *Proceedings of the 8th Annual Conference on Computers and Industrial Engineering*, *11*, 68-72.

Islam, K.M.S., Sarker, B.R., 2000. A similarity coefficient measure and machine-parts grouping in cellular manufacturing systems. International Journal of Production Research 38, 699-720.

Jaccard, P., 1908. Novelles recgerches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.*, *44*, 223-270.

Jeon, G., Broering, M., Leep, H.R., Parsaei, H.R., Wong, J.P., 1998a. Part family formation based on alternative routes during machine failure. Computers and Industrial Engineering 35, 73-76.

Jeon, G., Leep, H.R., Parsaei, H.R., 1998b. A cellular manufacturing system based on new similarity coefficient which considers alternative routes during machine failure. Computers and Industrial Engineering 34, 21-36.

Josien, K., Liao, T.W., 2000. Integrated use of fuzzy c-means and fuzzy KNN for GT part family and machine cell formation. International Journal of Production Research 38, 3513-3536.

Kamrani, A.K., Parsaei, H.R., Chaudhry, M.A., 1993. A survey of design methods for manufacturing cells. Computers and Industrial Engineering 25, 487-490.

Kang, S.L., Wemmerlöv, U., 1993. A work load-oriented heuristic methodology for manufacturing cell formation allowing reallocation of operations. European Journal of Operational Research 69, 292-311.

Kaparthi, S., Suresh, N.C., Cerveny, R.P., 1993. An improved neural network leader algorithm for part-machine grouping in group technology. European Journal of Operational Research 69, 342-356.

King, J.R., 1980. Machine-component grouping in production flow analysis: an approach using a rank order clustering algorithm. *International Journal of Production Research, 18*(2), 213-232.

King, J.R., Nakornchai, V., 1982. Machine component group formation in group technology: review and extension. International Journal of Production Research 20, 117-133.

Kitaoka, M., Nakamura, R., Serizawa, S., Usuki, J., 1999. Multivariate analysis model for machine-part cell formation problem in group technology. International Journal of Production Economics 60-61, 433-438.

Kulkarni, U.R., Kiang, M. Y., 1995. Dynamic grouping of parts in flexible manufacturing systems – a self-organizing neural networks approach. European Journal of Operational Research 84, 192-212.

Kumar, C.S., Chandrasekharan, M. P., 1990. Grouping efficacy: a quantitative criterion for goodness of block diagonal forms of binary matrices in group technology. International Journal of Production Research 28, 233-243.

Kumar, K.R., Kusiak, A., & Vannelli, A., 1986. Grouping of parts and components in flexible manufacturing systems. *European Journal of Operational Research, 24*, 387-397.

Kumar, K.R., Vannelli, A., 1987. Strategic subcontracting for efficient disaggregated manufacturing. International Journal of Production Research 25, 1715-1728.

Kusiak, A., 1985. The part families problem in flexible manufacturing systems. Annals of Operations Research 3, 279-300.

Kusiak, A., 1987. The generalized group technology concept. International Journal of Production Research 25, 561-569.

Kusiak, A., Boe, W.J., Cheng, C., 1993. Designing cellular manufacturing systems: branch-and-bound and A* approaches. IIE Transactions 25, 46-56.

Kusiak, A., Cho, M., 1992. Similarity coefficient algorithms for solving the group technology problem. International Journal of Production Research 30, 2633-2646.

Kusiak, A., & Chow, W.S., 1987. Efficient solving of the group technology problem. *Journal of Manufacturing Systems, 6*, 117-124.

Kusiak, A., Heragu, S.S., 1987. The facility layout problem. European Journal of Operational Research 29, 229-251.

Kusiak, A., Vannelli, A., Kumar, K.R., 1986. Clustering analysis: models and algorithms. Control and Cybernetics 15, 139-154.

Lashkari, R.S., Boparai, R., Paulo, J., 2004. Towards an integrated model of operation allocation and material handling selection in cellular manufacturing systems. International Journal of Production Economics 87, 115-139.

Lashkari, R.S., Gunasingh, K.R., 1990. A Lagrangian relaxation approach to machine allocation in cellular manufacturing systems. Computers and Industrial Engineering 19, 442-446.

Lee, H., Garcia-Diaz, A., 1996. Network flow procedures for the analysis of cellular manufacturing systems. IIE Transactions 28, 333-345.

Lee, M.K., Luong, H.S., Abhary, K., 1997. A genetic algorithm based cell design considering alternative routing. Computers Integrated Manufacturing Systems 10, 93-107.

Leem, C.W., Chen, J. J.G., 1996. Fuzzy-set-based machine-cell formation in cellular manufacturing. Journal of Intelligent Manufacturing 7, 355-364.

Lee-post, A., 2000. Part family identification using a simple genetic algorithm. International Journal of Production Research 38, 793-810.

Liggett, R.S., 2000. Automated facilities layout: past, present and future. Automation in Construction 9, 197-215.

Lin, T.L., Dessouky, M.M., Kumar, K.R., Ng, S.M., 1996. A heuristic-based procedure for the weighted production-cell formation problem. IIE Transactions 28, 579-589.

Logendran, R., 1991. Effect of the identification of key machines in the cell formation problem of cellular manufacturing systems. Computers and Industrial Engineering 20, 439-449.

Lozano, S., Adenso-Diaz, B., Eguia, I., Onieva, L., 1999. A one-step tabu search algorithm for manufacturing cell design. Journal of the Operational Research Society 50, 509-516.

Luong, L.H.S., 1993. A cellular similarity coefficient algorithm for the design of manufacturing cells. International Journal of Production Research 31, 1757-1766.

Mansouri, S.A., Husseini S.M.M., Newman, S.T., 2000. A review of the modern approaches to multi-criteria cell design. International Journal of Production Research 38, 1201-1218.

Marcotorchino, F., 1987. Block seriation problems: a unified approach. Applied Stochastic Models and Data Analysis 3, 73-91.

Masnata, A., Settineri, L., 1997. An application of fuzzy clustering to cellular manufacturing. International Journal of Production Research 35, 1077-1094.

McAuley, J., 1972, Machine grouping for efficient production. The Production Engineer, 51, 53-57.

McCormick, W.T., Schweitzer P.J., & White, T.W., 1972. Problem decomposition and data reorganization by a clustering technique. *Operations Research, 20*(5), 993-1009.

Mehrez, A., Rabinowitz, G., Reisman, A., 1988. A conceptual scheme of knowledge

systems for MS/OR. Omega 16, 421-428.

Milligan, G.W., & Cooper, S.C., 1987. Methodology review: clustering methods. *Applied Psychological Measurement, 11*(4), 329-354.

Miltenburg, J., Zhang, W., 1991. A comparative evaluation of nine well-known algorithms for solving the cell formation problem in group technology. Journal of Operations Management 10, 44-72.

Mitrofanov, S.P., 1966. Scientific principles of group technology, Part® (MA, Boston: National Lending Library of Science and Technology).

Mosier, C.T., 1989. An experiment investigating the application of clustering procedures and similarity coefficients to the GT machine cell formation problem. International Journal of Production Research 27, 1811-1835.

Mosier, C.T., Taube, L., 1985a. The facets of group technology and their impacts on implementation – a state of the art survey. Omega 13, 381-391.

Mosier, C.T., Taube, L., 1985b. Weighted similarity measure heuristics for the group technology machine clustering problem. Omega 13, 577-583.

Mosier, C.T., Yelle, J., Walker, G., 1997. Survey of similarity coefficient based methods as applied to the group technology configuration problem. Omega 25, 65-79.

Murthy, Ch.V.R., Srinivasan, G., 1995. Fractional cell formation in group technology. International Journal of Production Research 33, 1323-1337.

Nair, G.J.K., Narendran, T.T., 1996. Grouping index: a new quantitative criterion for goodness of block-diagonal forms in group technology. *International Journal of Production Research, 34*(10), 2767-2782.

Nair, G.J.K., Narendran, T.T., 1998. CASE: A clustering algorithm for cell formation with sequence data. International Journal of Production Research 36, 157-179.

Nair, G.J.K., Narendran, T.T., 1999. ACCORD: A bicriterion algorithm for cell formation using ordinal and ratio-level data. International Journal of Production Research 37, 539-556.

Ng, S.M., 1993. Worst-case analysis of an algorithm for cellular manufacturing. *European Journal of Operational Research, 69*, 384-398.

Offodile, O.F., 1991. Application of similarity coefficient method to parts coding and classification analysis in group technology. Journal of Manufacturing Systems 10, 442-448.

Offodile, O.F., 1993. Machine grouping in cellular manufacturing. Omega 21, 35-52.

Offodile, O.F., Mehrez, A., Grznar, J., 1994. Cellular manufacturing: a taxonomic review framework. Journal of Manufacturing Systems 13, 196-220.

Offodile, O.F., Grznar, J., 1997. Part family formation for variety reduction in flexi-

ble manufacturing systems. International Journal of Operations & Production Management 17, 291-304.

Onwubolu, G.C., Mlilo, P.T., 1998. Manufacturing cell grouping using similarity coefficient-distance measure. Production Planning & Control 9, 489-493.

Opitz, H., Eversheim, W., Wienhal, H.P., 1969. Work-piece classification and its industrial applications. International Journal of Machine Tool Design and Research 9, 39-50.

Qiao, L.H., Yang, Z.B., Wang, H.P., 1994. A computer-aided process planning methodology. Computers in Industry 255, 83-94.

Rajagopalan, R., Batra, J.L., 1975. Design of cellular production system: a graph theoretic approach. International Journal of Production Research 13, 567-579.

Reisman, A., Kirshnick, F., 1995. Research strategies used by OR/MS workers as shown by an analysis of papers in flagship journals. Operations Research 43, 731-739.

Reisman, A., Kumar, A., Motwani, J., Cheng, C.H., 1997. Cellular manufacturing: a statistical review of the literature (1965-1995). Operations Research 45, 508-520.

Ribeiro, J.F.F., Pradin, B., 1993. A methodology for cellular manufacturing design. International Journal of Production Research 31, 235-250.

Rogers, D.J., Tanimoto, T.T., 1960. A computer program for classifying plants. Science 132, 1115-1118.

Romesburg, H.C., 1984. Cluster analysis for researchers (CA: Lifetime Learning Publications (Wadsworth Inc.), Belmont).

Samatova, N.F., Potok, T.E., Leuze, M.R., 2001. Vector space model for the generalized parts grouping problem. Robotics and Computer Integrated Manufacturing 17, 73-80.

Sarker, B.R., 1996. The resemblance coefficients in group technology: a survey and comparative study of relational metrics. Computers and Industrial Engineering 30, 103-116.

Sarker, B.R., Islam, K.M.S., 1999. Relative performances of similarity and dissimilarity measures. Computers and Industrial Engineering 37, 769-807.

Sarker, B.R., Li, Z., 1998. Measuring matrix-based cell formation considering alternative routings. *Journal of the Operational Research Society, 49*(9), 953-965.

Sarker, B.R., Mondal, S., 1999. Grouping efficiency measures in cellular manufacturing: a survey and critical review. *International Journal of Production Research, 37*(2), 285-314.

Sarker, B.R., Xu, Y., 2000. Designing multi-product lines: job routing in cellular manufacturing systems. IIE Transactions 32, 219-235.

Seifoddini, H., 1987. Incorporation of the production volume in machine cells

formation in group technology applications. Proceedings of the ⊚th ICPR, October, pp. 2348-2356; or In A. Mital (ed), 1988, Recent Developments in Production Research (The Netherlands: Elsevier Science Publishers B.V.), pp. 562-570.

Seifoddini, H., 1989a. Single linkage versus average linkage clustering in machine cells formation applications. Computers and Industrial Engineering 16, 419-426.

Seifoddini, H., 1989b. A note on the similarity coefficient method and the problem of improper machine assignment in group technology applications. International Journal of Production Research 27, 1161-1165.

Seifoddini, H., Djassemi, M., 1995. Merits of the production volume based similarity coefficient in machine cell formation. Journal of Manufacturing Systems 14, 35-44.

Seifoddini, H., Djassemi, M., 1996. A new grouping measure for evaluation of machine-component matrices. *International Journal of Production Research, 34*(5), 1179-1193.

Seifoddini, H., Hsu, C.P., 1994. Comparative study of similarity coefficients and clustering algorithms in cellular manufacturing. Journal of Manufacturing Systems 13, 119-127.

Seifoddini, H., Tjahjana, B., 1999. Part-family formation for cellular manufacturing: a case study at Harnischfeger. International Journal of Production Research, 37 3263-3273.

Seifoddini, H., Wolfe, P.M., 1986 Application of the similarity coefficient method in group technology. IIE Transactions 18, 271-277.

Seifoddini, H., Wolfe, P.M., 1987. Selection of a threshold value based on material handling cost in machine-component grouping. IIE Transactions 19, 266-270.

Selim H.M., Askin, R.G., Vakharia, A.J., 1998. Cell formation in group technology: review, evaluation and directions for future research. Computers and Industrial Engineering 34, 3-20.

Selvam, R.P., Balasubramanian, K.N., 1985. Algorithmic grouping of operation sequences. Engineering Cost and Production Economics 9, 125-134.

Sevier, A.J., 1992. Managing employee resistance to just-in-time: creating an atmosphere that facilitates implementation. Production and Inventory Management Journal 33, 83-87.

Shafer, S.M., Meredith, J.R., 1990. A comparison of selected manufacturing cell formation techniques. *International Journal of Production Research, 28*(4), 661-673.

Shafer, S.M., Meredith, J.R., Marsh, R.F., 1995. A taxonomy for alternative equipment groupings in batch environments. Omega 23, 361-376.

Shafer, S.M., Rogers, D.F., 1993a. Similarity and distance measures for cellular manufacturing. Part®. A survey. International Journal of Production Research 31, 1133-1142.

Shafer, S.M., Rogers, D.F., 1993b. Similarity and distance measures for cellular manufacturing. Part®. An extension and comparison. International Journal of Production Research 31, 1315-1326.

Shambu, G., Suresh, N.C., 2000. Performance of hybrid cellular manufacturing systems: a computer simulation investigation. European Journal of Operational Research 120, 436-458.

Shiko, G., 1992. A process planning-orientated approach to part family formation problem in group technology applications. International Journal of Production Research 30, 1739-1752.

Silveira, G.D., 1999 A methodology of implementation of cellular manufacturing. International Journal of Production Research 37, 467-479.

Singh, N., 1993. Design of cellular manufacturing systems: an invited review. European Journal of Operational Research 69, 284-291.

Singh, N., 1996. Systems Approach to Computer-Integrated Design and Manufacturing (New York: Wiley).

Singh, N., Rajamani, D., 1996. *Cellular Manufacturing Systems: Design, planning and control*. London: Chapman & Hall.

Sneath, P.H.A., Sokal, R.R., 1973. Numerical taxonomy (San Francisco: W.H. Freeman).

Sofianopoulou, S., 1997. Application of simulated annealing to a linear model for the formulation of machine cells in group technology. *International Journal of Production Research, 35*(2), 501-511.

Sokal, R.R., Michener, C.D., 1958. A statistical method for evaluating systematic relationships. The University of Kansas Science Bulletin 38, 1409-1438.

Solimanpur, M., Vrat, P., Shankar, R., 2004. A heuristic to minimize makespan of cell scheduling problem. International Journal of Production Economics 88, 231-241.

Srinivasan, G., 1994. A clustering algorithm for machine cell formation in group technology using minimum spanning trees. International Journal of Production Research 32, 2149-2158.

Srinivasan, G., Narendran, T.T., Mahadevan, B., 1990. An assignment model for the part-families problem in group technology. International Journal of Production Research 28, 145-152.

Srinivasan, G., Narendran, T.T., 1991. GRAFICS – a nonhierarchical clustering algorithm for group technology. International Journal of Production Research 29, 463-478.

Srinivasan, G., Zimmers, E.W., 1998. Fractional cell formation – issues and approaches. International Journal of Industrial Engineering 5, 257-264.

Steudel, H.J., Ballakur, A., 1987. A dynamic programming based heuristic for machine grouping in manufacturing cell formation. Computers and Industrial Engineering 12, 215-222.

Suer, G.A., Cedeno, A.A., 1996. A configuration-based clustering algorithm for family formation. Computers and Industrial Engineering 31, 147-150.

Tam, K.Y., 1990. An operation sequence based similarity coefficient for part families formations. Journal of Manufacturing Systems 9, 55-68.

Tarsuslugil, M., Bloor, J., 1979. The use of similarity coefficients and cluster analysis in production flow analysis. In: *Proceedings 20th International Machine Tool Design and Research Conference, Birmingham, UK, September*, 525-532.

Vakharia, A.J., Kaku, B.K., 1993. Redesigning a cellular manufacturing system to handle long-term demand changes: a methodology and investigation. Decision Sciences 24, 909-930.

Vakharia, A.J., Wemmerlöv, U., 1987. A new design method for cellular manufacturing systems. Proceedings of the IXth ICPR, Cincinnati, Ohio, pp. 2357-2363.

Vakharia, A.J., Wemmerlöv, U., 1990. Designing a cellular manufacturing system: a materials flow approach based on operation sequences. IIE Transactions 22, 84-97.

Vakharia, A.J., Wemmerlöv, U., 1995. A comparative investigation of hierarchical clustering techniques and dissimilarity measures applied to the cell formation problem. Journal of Operations Management 13, 117-138.

Viswanathan, S., 1996. A new approach for solving the p-median problem in group technology. International Journal of Production Research 34, 2691-2700.

Waghodekar, P.H., Sahu, S., 1984. Machine-component cell formation in group technology: MACE. International Journal of Production Research 22, 937-948.

Wang, J., 1998. A linear assignment algorithm for formation of machine cells and part families in cellular manufacturing. Computers and Industrial Engineering 35, 81-84.

Wang, J., Roze, C., 1995. Formation of machine cells and part families in cellular manufacturing: an experimental study. Computers and Industrial Engineering 29, 567-571.

Wang, J., Roze, C., 1997. Formation of machine cells and part families: a modified p-median model and a comparative study. International Journal of Production Research 35, 1259-1286.

Wei, J.C., Gaither, N., 1990. A capacity constrained multiobjective cell formation method. Journal of Manufacturing Systems 9, 222-232.

Wei, J.C., Kern, G.M., 1989. Commonality analysis: a linear cell clustering algorithm for group technology. International Journal of Production Research 27, 2053-2062.

Wei, J.C., Kern, G.M., 1991. Discussion: reply to 'A note on a linear cell clustering algorithm'. International Journal of Production Research 29, 217-218.

Wemmerlöv, U., Hyer, N.L., 1986. Procedures for the part family/machine group identification problem in cellular manufacturing. Journal of Operations Management 6, 125-147.

Wemmerlöv, U., Hyer, N.L., 1987. Research issues in cellular manufacturing. International Journal of Production Research 25, 413-431.

Wemmerlöv, U., Johnson, D.J., 1997. Cellular manufacturing at 46 user plants: implementation experiences and performance improvements. International Journal of Production Research 35, 29-49.

Wemmerlöv, U., Johnson, D.J., 2000. Empirical findings on manufacturing cell design. International Journal of Production Research 38, 481-507.

Won, Y.K., 2000a. New p-median approach to cell formation with alternative process plans. International Journal of Production Research 38, 229-240.

Won, Y.K., 2000b. Two-phase approach to GT cell formation using efficient p-median formulation. International Journal of Production Research 38, 1601-1613.

Won, Y.K., Kim, S.H., 1997. Multiple criteria clustering algorithm for solving the group technology problem with multiple process routings. Computers and Industrial Engineering 32, 207-220.

Wu, N., Salvendy, G., 1993. A modified network approach for the design of cellular manufacturing systems. International Journal of Production Research 31, 1409-1421.

Yasuda, K., Yin, Y., 2001. A dissimilarity measure for solving the cell formation problem in cellular manufacturing. Computers and Industrial Engineering 39, 1-17.

Zhang, C., Wang, H.P., 1992. Concurrent formation of part families and machine cells based on the fuzzy set theory. Journal of Manufacturing Systems 11, 61-67.

9

Maintenance Management and Modeling in Modern Manufacturing Systems

Mehmet Savsar

1. Introduction

The cost of maintenance in industrial facilities has been estimated as 15-40% (an average of 28%) of total production costs (Mobley, 1990; Sheu and Krajewski, 1994). The amount of money that companies spent yearly on maintenance can be as large as the net income earned (McKone and Wiess, 1998). Modern manufacturing systems generally consist of automated and flexible machines, which operate at much higher rates than the traditional or conventional machines. While the traditional machining systems operate at as low as 20% utilization rates, automated and Flexible Manufacturing Systems (FMS) can operate at 70-80% utilization rates (Vineyard and Meredith, 1992). As a result of this higher utilization rates, automated manufacturing systems may incur four times more wear and tear than traditional manufacturing systems. The effect of such an accelerated usage on system performance is not well studied. However, the accelerated usage of an automated system would result in higher failure rates, which in turn would increase the importance of maintenance and maintenance-related activities as well as effective maintenance management. While maintenance actions can reduce the effects of breakdowns due to wear-outs, random failures are still unavoidable. Therefore, it is important to understand the implications of a given maintenance plan on a system before the implementation of such a plan.

Modern manufacturing systems are built according to the volume/variety ratio of production. A facility may be constructed either for high variety of products, each with low volume of production, or for a special product with high volume of production. In the first case, flexible machines are utilized in a job shop environment to produce a variety of products, while in the second case special purpose machinery are serially linked to form transfer lines for high production rates and volumes. In any case, the importance of maintenance function has increased due to its role in keeping and improving the equipment

availability, product quality, safety requirements, and plant cost-effectiveness levels since maintenance costs constitute an important part of the operating budget of manufacturing firms (Al-Najjar and Alsyouf, 2003).

Without a rigorous understanding of their maintenance requirements, many machines are either under-maintained due to reliance on reactive procedures in case of breakdown, or over-maintained by keeping the machines off line more than necessary for preventive measures. Furthermore, since industrial systems evolve rapidly, the maintenance concepts will also have to be reviewed periodically in order to take into account the changes in systems and the environment. This calls for implementation of flexible maintenance methods with feedback and improvement (Waeyenbergh and Pintelon, 2004).
Maintenance activities have been organized under different classifications. In the broadest way, three classes are specified as (Creehan, 2005):

1. Reactive: Maintenance activities are performed when the machine or a function of the machine becomes inoperable. Reactive maintenance is also referred to as corrective maintenance (CM).
2. Preventive: Maintenance activities are performed in advance of machine failures according to a predetermined time schedule. This is referred to as preventive maintenance (PM).
3. Predictive/Condition-Based: Maintenance activities are performed in advance of machine failure when instructed by an established condition monitoring and diagnostic system.

Several other classifications, as well as different names for the same classifications, have been stated in the literature. While CM is an essential repair activity as a result of equipment failure, the voluntary PM activity was a concept adapted in Japan in 1951. It was later extended by Nippon Denso Co. in 1971 to a new program called Total Productive Maintenance (TPM), which assures effective PM implementation by total employee participation. TPM includes Maintenance Prevention (MP) and Maintainability Improvement (MI), as well as PM. This also refers to "maintenance-free" design through the incorporation of reliability, maintainability, and supportability characteristics into the equipment design. Total employee participation includes Autonomous Maintenance (AM) by operators through group activities and team efforts, with operators being held responsible for the ultimate care of their equipments (Chan et al., 2005).

The existing body of theory on system reliability and maintenance is scattered over a large number of scholarly journals belonging to a diverse variety of disciplines. In particular, mathematical sophistication of preventive maintenance models has increased in parallel to the growth in the complexity of modern manufacturing systems. Extensive research has been published in the areas of maintenance modeling, optimization, and management. Excellent reviews of maintenance and related optimization models can be seen in (Valdez-Flores and Feldman, 1989; Cho and Parlar, 1991; Pintelon and Gelders, 1992; and Dekker, 1996).

Limited research studies have been carried out on the maintenance related issues of FMS (Kennedy, 1987; Gupta et al., 1988; Lin et al., 1994; Sun, 1994). Related analysis include effects of downtimes on uptimes of CNC machines, effects of various maintenance policies on FMS failures, condition monitoring system to increase FMS and stand-alone flexible machine availabilities, automatic data collection, statistical data analysis, advanced user interface, expert system in maintenance planning, and closed queuing network models to optimize the number of standby machines and the repair capacity for FMS. Recent studies related to FMS maintenance include, stochastic models for FMS availability and productivity under CM operations (Savsar, 1997a; Savsar, 2000) and under PM operations (Savsar, 2005a; Savsar, 2006).

In case of serial production flow lines, literature abounds with models and techniques for analyzing production lines under various failure and maintenance activities. These models range from relatively straight-forward to extremely complex, depending on the conditions prevailing and the assumptions made. Particularly over the past three decades a large amount of research has been devoted to the analysis and modeling of production flow line systems under equipment failures (Savsar and Biles, 1984; Boukas and Hourie, 1990; Papadopoulos and Heavey, 1996; Vatn et al., 1996; Ben-Daya and Makhdoum, 1998; Vouros et al., 2000; Levitin and Meizin, 2001; Savsar and Youssef, 2004; Castro and Cavalca, 2006; Kyriakidis and Dimitrakos, 2006). These models consider the production equipment as part of a serial system with various other operational conditions such as random part flows, operation times, intermediate buffers with limited capacity, and different types of maintenance activities on each equipment. Modeling of equipment failures with more than one type of maintenance on a serial production flow line with limited buffers is relatively complicated and need special attention. A comprehensive model and an iterative computational procedure has been developed (Savsar, 2005b)

to study the effects of different types of maintenance activities and policies on productivity of serial lines under different operational conditions, such as finite buffer capacities and equipment failures. Effects of maintenance policies on system performance when applied during an opportunity are discussed by (Dekker and Smeitnik, 1994). Maintenance policy models for just-in-time production control systems are discussed by (Albino, et al., 1992 and Savsar, 1997b).

In this chapter, procedures that combine analytical and simulation models to analyze the effects of corrective, preventive, opportunistic, and other maintenance policies on the performance of modern manufacturing systems are presented. In particular, models and results are provided for the FMS and automated Transfer Lines. Such performance measures as system availability, production rate, and equipment utilization are evaluated as functions of different failure/repair conditions and various maintenance policies.

2. Maintenance Modeling in Modern Manufacturing Systems

It is known that the probability of failure increases as an equipment is aged, and that failure rates decrease as a result of PM and TPM implementation. However, the amount of reduction in failure rate, from the introduction of PM activities, has not been studied well. In particular, it is desirable to know the performance of a manufacturing system before and after the introduction of PM. It is also desirable to know the type and the rate at which preventive maintenance should be scheduled. Most of the previous studies, which deal with maintenance modeling and optimization, have concentrated on finding an optimum balance between the costs and benefits of preventive maintenance. The implementation of PM could be at scheduled times (*scheduled PM*) or at other times, which arise when the equipment is stopped because of other reasons (*opportunistic PM*). Corrective maintenance (CM) policy is adapted if equipment is to be maintained only when it fails. The best policy has to be selected for a given system with respect to its failure, repair, and maintenance characteristics.

Two well-known preventive maintenance models originating from the past research are called *age-based* and *block-based replacement* models. In both models, PM is scheduled to be carried out on the equipment. The difference is in the timing of consecutive PM activities. In the aged-based model, if a failure occurs before the scheduled PM, PM is rescheduled from the time the corrective

maintenance is completed on the equipment. In the block-based model, on the other hand, PM is always carried out at scheduled times regardless of the time of equipment failures and the time that corrective maintenance is carried out. Several other maintenance models, based on the above two concepts, have been discussed in the literature as listed above.

One of the main concerns in PM scheduling is the determination of its effects on time between failures (TBF). Thus, the basic question is to figure out the amount of increase in TBF due to implementation of a PM. As mentioned above, introduction of PM reduces failure rates by eliminating the failures due to wear outs. It turns out that in some cases, we can theoretically determine the amount of reduction in total failure rate achieved by separating failures due to wear outs from the failures due to random causes.

2.1 Mathematical Modeling for Failure Rates Partitioning

Following is a mathematical procedure to separate random failures from wear-out failures. This separation is needed in order to be able to see the effects of maintenance on the productivity and operational availability of an equipment or a system. The procedure outlined here can be utilized in modeling and simulating maintenance operations in a system.

Let f(t) = Probability distribution function (pdf) of time between failures.

F(t) = Cumulative distribution function (cdf) of time between failures.

R(t) = Reliability function (probability of equipment survival by time t).

h(t) = Hazard rate (or instantaneous failure rate of the equipment).

Hazard rate $h(t)$ can be considered as consisting of two components, the first from random failures and the second from wear-out failures, as follows:

$$h(t) = h_1(t) + h_2(t) \qquad (1)$$

Since failures are from both, chance causes (unavoidable) and wear-outs (avoidable), reliability of the equipment by time t, can be expressed as follows:

$$R(t) = R1(t)\ R2(t) \qquad (2)$$

Where, $R_1(t)$ = Reliability due to chance causes or random failures and $R_2(t)$ = Reliability from wear-outs, $h_1(t)$ = Hazard rate from random failures, and $h_2(t)$

= Hazard rate from wear-out failures. Since the hazard rate from random failures is independent of aging and therefore constant over time, we let $h_1(t) = \lambda$. Thus, the reliability of the equipment from random failures with constant hazard rate:

$$R_1(t) = e^{-\lambda t} \text{ and } h(t) = \lambda + h_2(t) \quad (3)$$

It is known that:

$$h(t) = f(t)/R(t) = f(t)/[1-F(t)] = \lambda + h_2(t) \quad (4)$$

$$h_2(t) = h(t) - h_1(t) = f(t)/[1-F(t)] - \lambda \quad (5)$$

$$R_2(t) = R(t)/R_1(t) = [1-F(t)]/ e^{-\lambda t} \quad (6)$$

$$h_2(t) = f_2(t)/R_2(t) \quad (7)$$

$$f_2(t) = h_2(t)R_2(t) = [\frac{f(t)}{1-F(t)} - \lambda][\frac{1-F(t)}{e^{-\lambda t}}] = \frac{f(t)}{e^{-\lambda t}} - \frac{\lambda}{e^{-\lambda t}}[1-F(t)]$$

where

$$f_2(t) = \frac{dF_2(t)}{dt} \qquad F_2(t) = 1 - R_2(t) = 1 - \frac{1-F(t)}{e^{-\lambda t}} = \frac{e^{-\lambda t} - R(t)}{e^{-\lambda t}} \quad (8)$$

Equation (8) can be used to determine $f_2(t)$. These equations show that total time between failures, f(t), can be separated into two distributions, time between failures from random causes, with pdf given by $f_1(t)$, and time between failures from wear-outs, with pdf given by $f_2(t)$. Since the failures from random causes could not be eliminated, we concentrate on decreasing the failures from wear-outs by using appropriate maintenance policies. By the procedure described above, it is possible to separate the two types of failures and develop the best maintenance policy to eliminate wear-out failures. It turns out that this separation is analytically possible for uniform distribution. However, it is not possible for other distributions. Another approach is used for other distribu-

tions when analyzing and implementing PM operations. Separation of failure rates is particularly important in simulation modeling and analysis of maintenance operations. Failures from random causes are assumed to follow an exponential distribution with constant hazard rate since they are unpredictable and do not depend on operation time of equipment. Exponential distribution is the type of distribution that has memoryless property; a property that results in constant failure rates over time regardless of aging and wear outs due to usage. Following section describes maintenance modeling for different types of distributions.

2.2 Uniform Time to Failure Distribution

For uniformly-distributed time between failures, t, in the interval $0 < t < \mu$, the pdf of time between failures without introduction of PM is given by: $f(t)=1/\mu$. If we let $\alpha = 1/\mu$, then $F(t)= \alpha t$ and reliability is given as $R(t)=1-\alpha t$ and the total failure rate is given as $h(t)=f(t)/R(t)=\alpha/(1-\alpha t)$. If we assume that the hazard rate from random failures is a constant given by $h_1(t)=\alpha$, then the hazard rate from wear-out failures can be determined by $h_2(t)=h(t)-h_1(t)=\alpha/(1-\alpha t)-\alpha=\alpha^2 t/(1-\alpha t)$. The corresponding time to failure pdf for each type of failure rate is as follows:

$$f_1(t) = \alpha \times e^{(-\alpha t)}, \qquad 0 < t < \mu \tag{9}$$

$$f_2(t) = \alpha^2 \times t \times e^{(\alpha t)}, \qquad 0 < t < \mu \tag{10}$$

The reliability function for each component is as follows:

$$R_1(t) = e^{(-\alpha t)}, \qquad 0 < t < \mu \tag{11}$$

$$R_2(t) = (1 - \alpha t) \times e^{\alpha t}, \qquad 0 < t < \mu \tag{12}$$

$$R(t) = R_1(t) \times R_2(t) \tag{13}$$

When PM is introduced, failures from wear-outs are eliminated and thus the machines fail only from random failures, which are exponentially distributed as given by $f_1(t)$. Sampling for the time to failures in simulations is then based on an exponential distribution with mean μ and a constant failure rate of $\alpha=1/\mu$. In case of CM without PM, in addition to the random failures, wear-out failures are also present and thus the time between equipment failures is uniformly distributed between zero and μ as given by f(t). The justification behind this assumption is that uniform distribution implies an increasing failure rate with two components, namely, failure rate from random causes and failure rate from wear-out causes as given by $h_1(t)$ and $h_2(t)$, respectively. Initially when t = 0, failures are from random causes with a constant rate $\alpha=1/\mu$. As the equipment operates, wear-out failures occur and thus the total failure rate h(t) increases with time t. Sampling for the time between failures in modeling and simulation is based on uniform distribution with mean $\mu/2$ and increasing failure rate, h(t).

2.3. Non-uniform time to failure distributions

2.3.1 Normal distribution:

If the times between failures are normally distributed, it is not possible to separate the two types of failures analytically. However, the following procedure can be implemented in simulation models:
When no preventive maintenance is implemented, times between failures are sampled from a normal distribution with mean μ and standard deviation σ. When PM is implemented, wear-out failures are eliminated and the remaining random failures follow an exponential distribution with constant failure rate with extended mean time between failures. It is assumed that mean time between equipment failures after introduction of PM extends from μ to $k\mu$, where k is a constant greater than 1.

2.3.2 Gamma Distribution:

For a gamma distribution, which is Erlang when its shape parameter α is integer and exponential when $\alpha=1$, the expected value of random variable T is defined by $E(T) = \alpha\beta$. Thus, by changing α and β values, mean time between failures can be specified as required. When no PM is introduced, times between failures are sampled from a gamma distribution with mean time between fail-

ures of αβ. If PM is introduced and wear-out failures are eliminated, times between failures are extended by a constant k. Therefore, sampling is made from an exponential distribution with mean k(αβ).

2.3.3 Weibull Distribution:

For the Weibull distribution, α is a shape parameter and β is a scale parameter. The expected value of time between failures, $E(T)=MTBF=\beta\Gamma(1/\alpha)/\alpha$, and its variance is $V(T)= \beta^2[2\Gamma(2/\alpha)-\{\Gamma(1/\alpha)\}^2/\alpha]$. For a given value of α, $\beta=\alpha(MTBF)/\Gamma(1/\alpha)$. When there is no PM, times between failures are sampled from Weibull with parameters α and β in simulation models. When PM is introduced, wear-out failures are eliminated and the random failures are sampled in simulation from an exponential distribution with mean=$k[\beta\Gamma(1/\alpha)/\alpha]$, where α and β are the parameters of the Weibull distribution and k is a constant greater than 1.

2.3.4 Triangular Distribution:

The triangular distribution is described by the parameters a, m, and b (i.e., minimum, mode, and maximum). Its mean is given by E(T)=(a+m+b)/3 and variance by $V(T) = (a^2+m^2+b^2-ma-ab-mb)/18$. Since the times between failures can be any value starting from zero, we let a=0 and thus m=b/3 from the property of a triangular distribution. Mean time between failures is E(T)=(m+b)/3=[b+b/3]/3=4b/9=4m/3. If no PM is introduced, time between failures are sampled in simulation from a triangular distribution with parameters a, m, b or 0, b/3, b. If PM is introduced, again wear-out failures are eliminated and the random failures are sampled from an exponential distribution with an extended mean of k[a+m+b]/3, where a, m, and b are parameters of the triangular distribution that describe the time between failures. The multiplier k is a constant greater than 1.

3. Analysis of the Effects of Maintenance Policies on FMS Availability

Equipment in FMS systems can be subject to corrective maintenance; corrective maintenance combined with a preventive maintenance; and preventive maintenance implemented at different opportunities. FMS operates with an increasing failure rate due to random causes and wear-outs. The stream of mixed failures during system operation is separated into two types: (i) random failures due to chance causes; (ii) time dependent failures due to equipment usage

and wear-outs. The effects of preventive maintenance policies (scheduled and opportunistic), which are introduced to eliminate wear-out failures of an FMS, can be investigated by analytical and simulation models. In particular, effects of various maintenance policies on system performance can be investigated under various time between failure distributions, including uniform, normal, gamma, triangular, and Weibull failure time distributions, as well as different repair and maintenance parameters.

3.1 Types of Maintenance Policies

In this section, five types maintenance policies, which resulted in six distinct cases, and their effects on FMS availability are described.

i) No Maintenance Policy:
In this case, a fully reliable FMS with no failures and no maintenance is considered.

ii) Corrective Maintenance Policy (CM):
The FMS receives corrective maintenance only when equipment fails. Time between equipment failures can follow a certain type of distribution. In case of uniform distribution, two different types of failures can be separated in modeling and analysis.

iii) Block-Based PM with CM Policy (BB):
In this case, the equipment is subject to preventive maintenance at the end of each shift to eliminate the wear out failures during the shift. However, regardless of any CM operations between the two scheduled PMs, the PM operations are always carried out as scheduled at the end of the shifts without affecting the production schedule. This policy is evaluated under various time between failure distributions. Figure 1 illustrates this maintenance process.

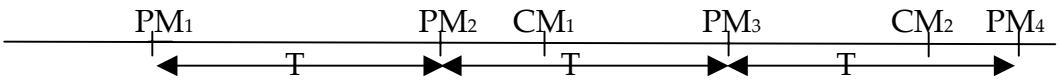


Figure 1. Illustration of PM operations under a block-based policy

iv) Age-Based PM with CM Policy (AB):
In this policy, preventive maintenance is scheduled at the end of a shift, but the PM time changes as the equipment undergoes corrective maintenance. Suppose that the time between PM operations is fixed as T hours and before performing a particular PM operation the equipment fails. Then the CM operation is carried out and the next PM is rescheduled T hours from the time the repair for the CM is completed. CM has eliminated the need for the next PM. If the scheduled PM arrives before a failure occurs, PM is carried out as scheduled. Figure 2 illustrates this process.

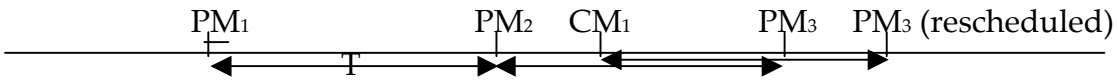


Figure 2. Illustration of PM operations under age-based policy.

v) Opportunity-Triggered PM with CM Policy (OT):
In this policy, PM operations are carried out only when they are triggered by failure. In other words, if a failure that requires CM occurs, it also triggers PM. Thus, corrective maintenance as well as preventive maintenance is applied to the machine together at the time of a failure. This is called triggered preventive maintenance. Since the equipment is already stopped and some parts are already maintained for the CM, it is expected that the PM time would be reduced in this policy. We assign a certain percentage of reduction in the PM operation. A 50% reduction was assumed reasonable in the analysis below.

vi) Conditional Opportunity-Triggered PM with CM Policy (CO): In this policy, PM is performed on each machine at either scheduled times or when a specified opportunistic condition based on the occurrence of a CM arises. The maintenance management can define the specified condition. In our study, a specific condition is defined as follows: if a machine fails within the last quarter of a shift, before the time of next PM, the next PM will be combined with CM for this machine. In this case, PM scheduled at the end of the shift would be skipped. On the other hand, if a machine failure occurs before the last quarter of the shift, only CM is introduced and the PM is performed at the end of the shift as it was scheduled. This means that the scheduled PM is performed only for those machines that did not fail during the last quarter of the shift.

The maintenance policies described above are compared under similar operating conditions by using simulation models with analytical formulas incorporated into the model as described in section 2. The FMS production rate is first determined under each policy. Then, using the production rate of a fully reliable FMS as a basis, an index, called Operational Availability Index (OAI_i) of the FMS under each policy i, is developed: $OAI_i=P_i/P_1$, where P_1 = production rate of the reliable FMS and P_i = production rate of the FMS operated under maintenance policy i (i=2, 3, 4, 5, and 6). General formulation is described in section 2 for five different times between failure distributions and their implementation with respect to the maintenance policies. The following section presents a maintenance simulation case example for an FMS system.

3.2 Simulation Modeling of FMS Maintenance Operations

In order to analyze the performance measures of FMS operations under different maintenance policies, simulation models are developed for the fully reliable FMS and for each of the five maintenance related policies described above. Simulation models are based on the SIMAN language (Pegden et al., 1995). In order to experiment with different maintenance policies and to illustrates their effects on FMS performance, a case problem, as in figure 3 is considered. Table 1 shows the distance matrix for the FMS layout and Table 2 shows mixture of three different types of parts arriving on a cart, the sequence of operations, and the processing times on each machine. An automated guided vehicle (AGV) selects the parts and transports them to the machines according to processing requirements and the sequence. Each part type is operated on by a different sequence of machines. Completed parts are placed on a pallet and moved out of the system. The speed of the AGV is set at 175 feet/minute. Parts arrive to the system on pallets containing 4 parts of type 1, 2 parts of type 2, and 2 parts of type 3 every 2 hours. This combination was fixed in all simulation cases to eliminate the compounding effects of randomness in arriving parts on the comparisons of different maintenance policies. The FMS parameters are set based on values from an experimental system and previous studies.

One simulation model was developed for each of the six cases as: i) A fully reliable FMS (denoted by FR); ii) FMS with corrective maintenance policy only (CM); iii) FMS with block-based policy (BB); iv) FMS with age-based policy (AB); v) FMS with opportunity-triggered maintenance policy (OP); and vi) FMS with conditional opportunity-triggered maintenance policy (CO). Each

simulation experiment was carried out for the operation of the FMS over a period of one month (20 working days or 9600 minutes). In the case of PM, it was assumed that a PM time of 30 minutes (or 15 minutes when combined with CM) is added to 480 minutes at the end of each shift. Twenty simulation replicates are made and the average output rate during one month is determined. The output rate is then used to determine the FMS operational availability index for each policy. The output rate is calculated as the average of the sum of all parts of all types produced during the month. The fully reliable FMS demonstrates maximum possible output (P_i) and is used as a base to compare other maintenance policies with $OAI_i = P_i/P_1$.

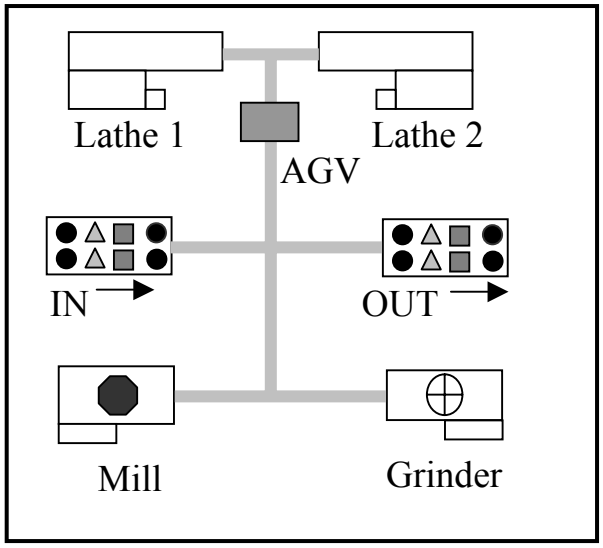


Figure 3. A flexible manufacturing system

	In	Lathe	Mill	Grind	Out
In	-	100	75	100	40
Lathe	-	-	150	175	155
Mill	-	-	-	50	90
Grind	-	-	-	-	115
Out	-	-	-	-	-

Table 1. Distance matrix (in feet).

Part Type	Lathe(L)	Milling(M)	Grinding(G)
1 (L-M-G)	Norm(30,5)	Norm(15,3)	Unif(10,15)
2 (M-G-L)	Norm(25,8)	Tria(2,10,15)	Norm(10,2)
3 (G-L)	Unif (5,10)		Norm(15,3)

Table 2. Processing time and operation sequences.

In *the first simulation experiment*, times between failures are assumed to be uniformly distributed between 0 and T for all machines with MTBF of T/2. Uniform distribution permits theoretical separation of chance-caused failures from wear-out failures. In the absence of any preventive maintenance, a machine can fail anytime from 0 to T. However, when PM is introduced, wear-out failures are eliminated; only the failures from chance causes remain, which have a constant hazard rate and exponential distribution with MTBF of T. In this experiment, the value of T is varied from 500 to 4000 minutes, in increments of 500 minutes. Repair time is assumed to be normal with mean 100 and standard deviation of 10 minutes for all machines. If PM is introduced on a machine, it is assumed that the PM is done at the end of each shift and it takes 30 minutes for each machine. If PM is triggered by the CM and done at this opportunity, PM time reduces to half, i.e., 15 minutes, since it is combined with the CM tasks. Mean production rate values are normalized with respect to fully reliable (FR) FMS values and converted into OAIs. These results are shown in figure 4. As it is seen from figure 4, performing CM alone without any PM is the worst policy of all. Observing all the policies in the figure, the best policy appears to be the opportunity triggered maintenance policy (OT). Between the age and block-based policies, the age-based policy (AB) performed better. Among all the policies with PM, block-based policy (BB) appears to be the worst policy.

As the MTBF increases, all the policies reach a steady state level with respect to operational availability, but the gap between them is almost the same at all levels of MTBF. In the case of CM only policy, the operational availability index sharply increases with the initial increase in MTBF from 500 to 1000.

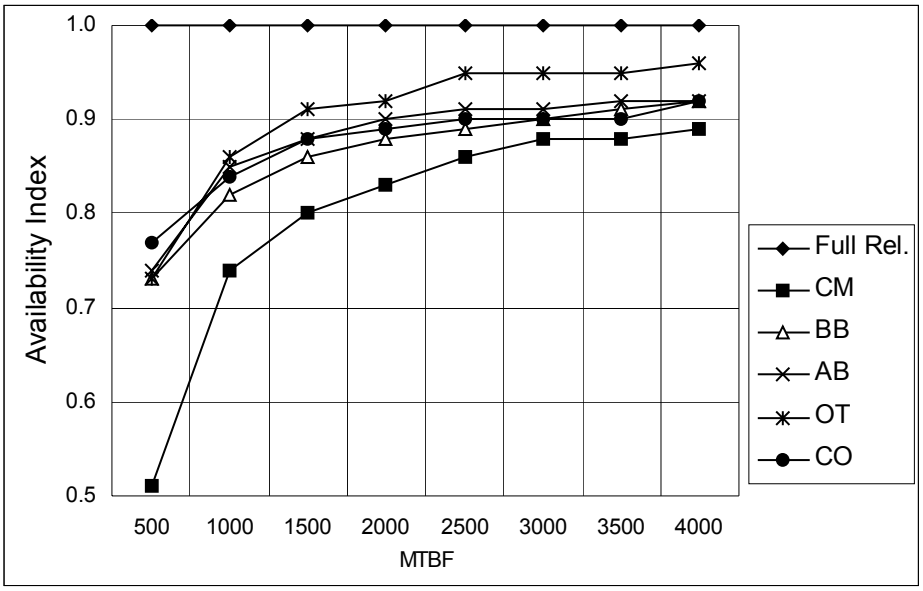


Figure 4. Operational availability index under different maintenance policies.

As indicated above, when PM is introduced, time between failures become exponential regardless of the type of initial distribution. Experiments with different distributions show that all distributions give the same performance results under the last four maintenance policies, which include some form of PM. However, FMS performance would differ under different failure distributions when a CM policy is implemented. This is investigated in *the second experiment*, which compares the effects of various time to failure distributions, including uniform, normal, gamma, Weibull, and triangular distributions, on FMS performance under the CM policy only. All of the FMS parameters related to operation times, repair, and PM times were kept the same as given in the first experiment. Only time to failure distributions and related parameters were changed such that MTBF was varied between 500 and 4000.

In the case of the gamma distribution, $E(T) = \alpha\beta$. Thus, $\alpha = 250$ and $\beta = 2$ resulted in a MTBF of 500; $\alpha = 750$ and $\beta = 2$ resulted in a MTBF=1500; $\alpha = 1250$ and $\beta = 2$ resulted in a MTBF=2500; and $\alpha = 2000$ and $\beta = 2$ resulted in a MTBF=4000, which are the same values specified in the second experiment for the normal distribution. For the Weibull distribution, which has $MTBF=E(T)=\beta\Gamma(1/\alpha)/\alpha$, two the parameters α (shape parameter) and β (scale parameter) have to be defined. For example, if MTBF=500 and $\alpha=2$, then, $500=\beta\Gamma(1/\alpha)/\alpha$

=βΓ(1/2)/2. Since Γ(1/2)=√π, β=1000/√π. Thus, for MTBF=500, β=564.2. Similarly, for MTBF=1500, β=1692.2, for MTBF=2500, β=2820.95, and for MTBF=4000, β=4513.5 are used. Triangular distribution parameters are also determined similarly as follows: E(T) = (a+m+b)/3 and V(T)= (a^2+m^2+b^2-ma-ab-mb)/18. Since the times between failures can be any value starting from zero, we let a=0 and m=b/3 from the property of triangular distribution. E(T)= (m+b)/3=[b+b/3]/3=4b/9=4m/3. In order to determine values of the parameters, we utilize these formula. For example, if MTBF =500, then 500=4b/9 and thus b=4500/4 = 1125 and m=b/3=1500/4=375. Similarly, for MTBF=1500, b=3375 and m=1125. For MTBF=2500, b=5625 and m=1875. For MTBF=4000, b=9000 and m=3000. Table 3 presents a summary of the related parameters.

Distribution	MTBF	Parameters that result in the specified MTBF		
Gamma		α	β	
	500	250	2	
	1500	750	2	
	2500	1250	2	
	4000	2000	2	
Weibull	500	2	564.2	
	1500	2	1692.2	
	2500	2	2820.95	
	4000	2	4513.5	
Triangular	MTBF	a	b	m
	500	0	1125	375
	1500	0	3375	1125
	2500	0	5625	1875
	4000	0	9000	3000

Table 3. Parameters of the distributions used in simulation.

Comparisons of five distributions, uniform, normal, gamma, Weibull, and triangular, with respect to CM are illustrated in figure 5, which plots the OAI values normalized with respect to fully reliable system using production rates. All of the distributions show the same trend of increasing OAI values, and thus production rates, with respect to increasing MTBF values. As it seen in figure 5, uniformly distributed time between failures resulted in significantly

different FMS availability index as compared the other four distributions. This is because in a uniform distribution, which is structurally different from other distributions, probability of failure is equally likely at all possible values that the random variable can take, while in other distributions probability concentration is around the central value. The FMS performance was almost the same under the other four distributions investigated. This indicates that the type of distribution has no critical effects on FMS performance under CM policy if the distribution shapes are relatively similar.

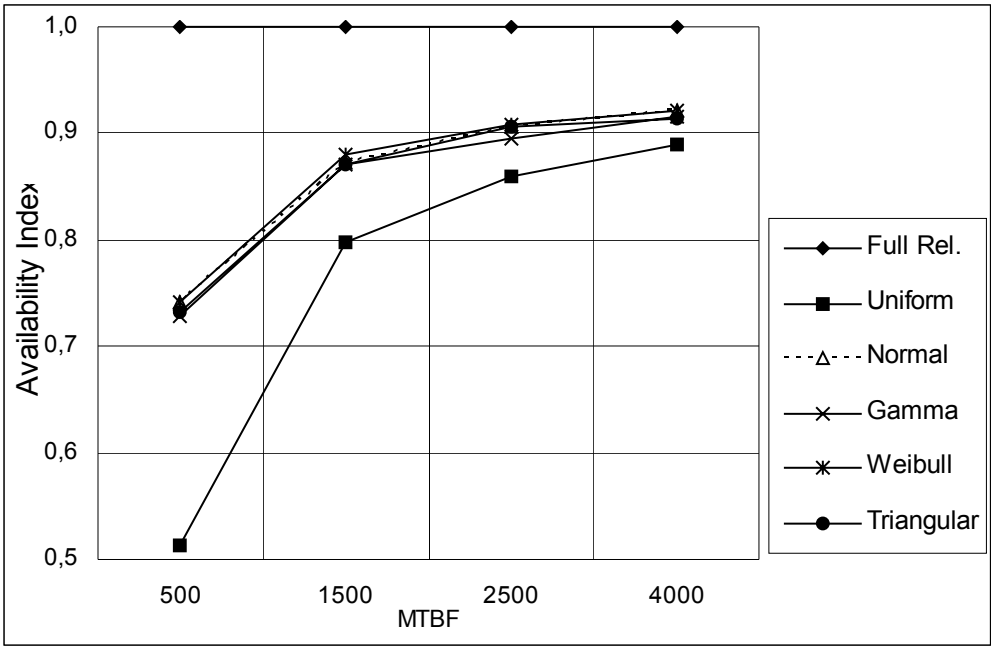


Figure 5. FMS OAI under various time to failure distributions and CM policy

The results of the analysis show that maintenance of any form has significant effect on the availability of the FMS as measured by its output rate. However, the type of maintenance applied is important and should be carefully studied before implementation. In the particular example studied, the best policy in all cases was the opportunity-triggered maintenance policy and the worst policy was the corrective maintenance policy. The amount of increase in system availability depends on the maintenance policy applied and the specific case studied. Implementation of any maintenance policy must also be justified by a detailed cost analysis.

The results presented in this chapter show a comparative analysis of specified maintenance policies with respect to operational availability measured by output rate. Future studies can be carried out on the cost aspects of various policies. The best cost saving policy can be determined depending on the specified parameters related to the repair costs and the preventive maintenance costs. In order to do cost related studies, realistic cost data must be collected from industry. The same models developed and procedures outlined in this paper can be used with cost data. Other possible maintenance policies must be studied and compared to those presented in this study. Combinations of several policies are also possible within the same FMS. For example, while a set of equipment is maintained by one policy, another set could be maintained by a different policy. These aspects of the problem may also be investigated by the models presented.

4. Analysis of the Effects of Maintenance Policies on Serial Lines

Multi-stage discrete part manufacturing systems are usually designed along a flow line with automated equipment and mechanized material flow between the stations to transfer work pieces from one station to the next automatically. CM and PM operations on serial lines can cause significant production losses, particularly if the production stages are rigidly linked. In-process inventories or buffer storages are introduced to decouple the rigidly-linked machinery and to localize the losses caused by equipment stoppages. Buffer storages help to smooth out the effect of variation in process times between successive stations and to reduce the effects of CM and PM in one station over the adjacent stations. While large buffer capacities between stages result in excessive inventories and costs, small buffer capacities result in production losses due to unexpected and planned stoppages and delays. One of the major problems associated with the design and operation of a serial production system is the determination of the effects of maintenance activities coupled with certain buffer capacities between the stations. Reliability and productivity calculations of multi-stage lines with maintenance operations and intermediate storage units can be quite complex. Particularly, closed form solutions are not possible when different types of maintenance operations are implemented on the machines.

Production line systems can take a variety of structures depending on the operational characteristics. Operation times can be stochastic or deterministic; stations can be reliable or unreliable; buffer capacities can be finite or infinite;

production line can be balanced or unbalanced; and material flow can be considered as discrete or continuous. Depending on the type of line considered and the assumptions made, complexity of the models vary. The objective in modeling these systems is to determine line throughput rate and machine utilizations as a function of equipment failures, maintenance policies, and buffer capacities,. Optimum buffer allocation results in maximum throughput rate. Algorithms and models are developed for buffer allocation on reliable and unreliable production lines for limited size problems. While closed form analytical models or approximations are restricted by several assumptions, models that can be coupled with numerical evaluation or computer simulation are more flexible and allow realistic modeling.

In this chapter we present a discrete mathematical model, which is incorporated into a generalized iterative simulation procedure to determine the production output rate of a multi-stage serial production line operating under different conditions, including random failures with corrective and preventive maintenance operations, and limited buffer capacities. The basic principal of the discrete mathematical model is to determine the total time a part *n* spends on a machine *i*, the time instant at which part *n* is completed on machine *i*, and the time instant at which part *n* leaves machine *i.* Figure 6 shows a multi-stage line with *m* machines and (*m+1*) intermediate buffer storages. Because each production machine is a highly complex combination of several instruments and working parts, it is assumed that more than one type of failure, which require different corrective actions, can occur on each machine and that each machine may receive more than one type of preventive maintenance actions. Effects of different maintenance policies on line production output rate are investigated.

Figure 6. A Serial Production Flow Line with n Stations and n+1 Buffers

Storages S_2,……,S_n are called intermediate storages, having finite capacity z_i, i=2,…,m. However, the initial input and the final output storages, namely storages 1 and m+1, are assumed to have unlimited capacities. The following notation is used in the formulation:

- R_{in}= Total duration of time that n^{th} part stays on the i^{th} machine not considering imposed stoppages due to maintenances or failures; i=1,2,…..,m.
- m = Number of machines on the line.
- P_{ijn} = Duration of preventive maintenance of j^{th} type on the i^{th} machine after machining of n^{th} part is completed; j=1,2,…….,np_i
- np_i = Number of preventive maintenance types performed on machine i
- t_{in} = Machining time for part n on machine i. This time can be assumed to be independent of n in the simulation program.
- r_{ijn} = Repair time required by i^{th} machine for correction of j^{th} type of failures which occur during the machining of n^{th} part; j=1,2,…..,nf_i
- nf_i = Number of failure types which occur on machine i.
- C_{in} = Instant of time at which machining of n^{th} part is completed on i^{th} machine.
- D_{in} = Instant of time at which n^{th} part departs from the i^{th} machine.
- D_{0n} = Instant of time at which n^{th} part enters the 1st machine.
- W_{in} = Instant of time at which i^{th} machine is ready to process n^{th} parts.

A part stays on a machine for two reasons: Either it is being machined or the machine is under corrective maintenance because a breakdown has occurred during machining of that part. R_{in}, which is the residence time of the n^{th} part on the i^{th} machine, without considering imposed stoppages for corrective maintenance, is given as follows:

$$R_{in} = t_{in} + \sum_{j=1}^{nf_i} r_{ijn} \qquad (14)$$

The duration of total preventive maintenance, P_{in}, performed on the i^{th} machine after completing the n^{th} part, is equal to the total duration of all types of preventive maintenances, P_{ijn}, that must be started after completion of n^{th} part as:

$$P_{in} = \sum_{j=1}^{np_i} P_{ijn} \qquad (15)$$

Each buffer B_i is assumed to have a finite capacity z_i, i=2,3,…….,m. The discrete mathematical model of serial line consists of calculating part completion times, C_{in}, and part departure times, D_{in}, in an iterative fashion.

4.1 Determination of Part Completion Times

Machining of part n cannot be started on machine i until the previous part, $n-1$, leaves machine i and until all the required maintenances, if necessary, are performed on machine i. Therefore the time instant at which i^{th} machine is ready to begin the n^{th} part, denoted by W_{in}, is given by the following relation:

$$W_{in} = max[D_{i,n-1}, C_{i,n-1}+P_{i,n-1}] \quad (16)$$

If $D_{i-1,n}<W_{in}$, then the n^{th} part must wait in storage buffer S_i, since it has left machine $i-1$ before machine i is ready to accept it. Therefore, machining of part n on machine i will start at instant W_{in}. If however, $D_{i-1,n}\geq W_{in}$, then machining of the n^{th} part on the i^{th} machine can start immediately after $D_{i-1,n}$. Considering both cases above, starting time of the n^{th} part to be machined on the i^{th} machine is:

$$max[D_{i-1,n}, D_{i,n-1}, C_{i,n-1}+P_{i,n-1}] \quad (17)$$

Since the n^{th} part will stay on machine i for a period of R_{in} time units, its machining will be completed by time instant C_{in} given by:

$$C_{in}= max[D_{i-1,n}, D_{i,n-1}, C_{i,n-1}+P_{i,n-1}] + R_{in} \quad (18)$$

Where

$$i=2,3,\ldots\ldots,m \text{ and } C_{1n}= max[D_{1,n-1}, C_{1,n-1}+P_{1,n-1}] + R_{1n} \quad (19)$$

Then,

$$D_{0n}< max[D_{1,n-1}, C_{1,n-1}+P_{1,n-1}], \quad (20)$$

assuming there are always parts available in front of machine 1.

4.2 Determination of Part Departure Times

The time instant at which n^{th} part leaves the i^{th} machine, D_{in}, is found by considering two cases.

Let $k = n - z_{i+1} - 1$. Then, in the first case:

$$C_{i,n} < \max[D_{i+1,k}, C_{i+1,k} + P_{i+1,k}] \tag{21}$$

which indicates that the n^{th} part has been completed on the i^{th} machine before machining of the $(n - z_{i+1})^{th}$ part has started on the $(i+1)^{th}$ machine. Since storage $i+1$, which is between machine i and $i+1$ and has capacity z_{i+1}, is full and machine i has completed the n^{th} part, the n^{th} part may leave the i^{th} machine only at the instant of time at which the $(n - z_{i+1})^{th}$ part of the $(i+1)^{th}$ machine has started machining. Therefore,

$$D_{i,n} = \max[D_{i+1,k}, C_{i+1,k} + P_{i+1,k}] \tag{22}$$

In the second case:

$$C_{i,n} > \max[D_{i+1,k}, C_{i+1,k} + P_{i+1,k}] \tag{23}$$

which indicates that, at the instant $C_{i,n}$ there are free spaces in buffer S_{i+1} and therefore part n can leave machine i immediately after it is completed; that is, $D_{i,n} = C_{i,,n}$ holds under this case. Considering both cases above, we have the following relations for $D_{i,,n}$:

$$D_{i,n} = C_{i,n} \qquad \text{if } n \le z_{i+1} + 1, \tag{24}$$

$$D_{i,n} = \max[C_{i,n}, D_{i+1,k}, C_{i+1,k} + P_{i+1,k}] \tag{25}$$

if $n > z_{i+1}+1$; $i = 1, 2, 3, \ldots\ldots, m-1$ and $k = n - z_{i+1} - 1$.

Since the last stage has infinite space to index its completed parts,

$$D_{m,,n} = C_{m,,n} \tag{26}$$

The simulation model, which is based on discrete mathematical model, can iteratively calculate $C_{i,n}$ and $D_{i,n}$ from which several line performance measures can be computed. Performance measures estimated by the above iterative computational procedures are: (i) Average number of parts completed by the line during a simulation period, T_{sim}; (ii) Average number of parts completed by each machine during the time, T_{sim}; (iii) Percentage of time for which each machine is up and down; (iv) Imposed, inherent and total loss factors for each machine; (v) Productivity improvement procedures.
In addition to the variables described for the discrete model in previous section, the simulation model can allow several distributions, including: exponential, uniform, Weibull, normal, log normal, Erlang, gamma, beta distributions and constant values to describe failure and repair times. After the production line related parameters and all data are entered into the simulator and necessary initializations are carried out by the program, iterative analysis of the production line is performed through the discrete mathematical model. The iterative analysis provides one simulation realization of the production line for a specified period of simulation time (T_{sim}). It basically calculates iteratively the time instant at which each part enters a machine, duration of its stay, and the time it leaves the machine. This is continued until, for example one shift is completed. The results of each iterative simulation are then utilized with statistical tests to determine if the specified conditions are met to stop the number of simulation iterations. If the conditions are not met, simulation iterations are continued with further runs. For each simulation realization, calculations of $R_{i,n}$, $C_{i,n}$, and $D_{i,n}$ are performed based on the need for repair or preventive maintenance operations as the parts flow through the system.

4.3 Statistical Analysis and Determination of Need for More Realizations

Reliable results cannot always be obtained from a single simulation realization. Therefore, additional runs have to be performed and the results tested statistically until the error in the line production rate is less than an epsilon value with a probability, both of which are predefined. This is accomplished as follows:

Let N_i = number of parts produced by the production line during the i^{th} simulation run. N_i is a random variable which approaches normal distribution as $t\rightarrow\infty$; that is, as the simulation time increases for ach realization, the sample random variable N_i approaches to an asymptotic normal distribution. The

mean value of N_i, $\overline{N}$, and its variance $V(N)$ are given by $\overline{N} = \sum_{i=1}^{n} (N_i)/n$ and $V(N) = \sum_{i=1}^{n} (N_i - \overline{N})^2/(n-1)$, where n represents the number of runs. The average line production rate, $\overline{Q}$, which is the average output rate per unit of time, is given by $\overline{Q} = \overline{N}/T_{sim}$ where T_{sim} is the simulation time for each realization, given in minutes. The quantity $\overline{Q}$ is the average line production rate in parts/minute. The variance of $\overline{Q}$, $V(\overline{Q})$, is determined by $V(\overline{Q}) = V(\overline{N}/T_{sim}) = V(\overline{N})/T_{sim}^2 = V(N)/n(T_{sim})^2$, since $V(\overline{N}) = V(N)/n$. The standard deviation of the average production rate, $\overline{Q}$, is $\sqrt{V(\overline{Q})}$ and the ratio of its standard deviation to its mean is expressed as by: $\sqrt{V(\overline{Q})}/\overline{Q} = [\sqrt{V(N)/nT_{sim}}]/[\overline{N}/T_{sim}] = [\sqrt{V(N)}]/[\overline{N}\sqrt{n}] = [\sqrt{V(N)/n}]/\overline{N}$. Since the number of parts N_i produced by the production line during a simulation period T_{sim} approaches the normal distribution, one can determine a confidence interval for $\overline{N}$ and $\overline{Q}$ using the normal distribution. $\Pr\left[\frac{N-\overline{N}}{\sqrt{V(N)}} < x\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-Z^2/2} dz$. The value of $\overline{N}$ is contained, with probability 1 – α, in the interval given by $\Pr\left[-Z_{\alpha/2} < \frac{N-\overline{N}}{\sqrt{V(N)}} < Z_{\alpha/2}\right] = 1-\alpha$. Since the mean output rate, $\overline{Q}$, is given by $\overline{Q} = \overline{N}/T_{\text{sim}}$, one can conclude that $\overline{Q}$ is normally distributed, as are N and $\overline{N}$. Therefore, a confidence interval for the actual mean output rate $\overline{\overline{Q}}$ would be given by $\Pr\left[-Z_{\alpha/2} < \frac{\overline{Q}-\overline{\overline{Q}}}{\sqrt{V(\overline{Q})}} < Z_{\alpha/2}\right] = 1-\alpha$ or by $\Pr\left[\overline{Q}(1 - Z_{\alpha/2}\sqrt{V(\overline{Q})}/\overline{Q}) < \overline{\overline{Q}} < \overline{Q}\left(1 + Z_{\alpha/2}\sqrt{V(\overline{Q})}/\overline{Q}\right)\right] = 1-\alpha$

$$\Pr\left[1 - \frac{Z_{\alpha/2}\sqrt{V(\overline{Q})}}{\overline{Q}} < \frac{\overline{\overline{Q}}}{\overline{Q}} < 1 + \frac{Z_{\alpha/2}\sqrt{V(\overline{Q})}}{\overline{Q}}\right] = 1-\alpha \qquad (27)$$

Our aim is to have an estimated output rate, $\overline{Q}$, as close to the actual mean output rate $\overline{\overline{Q}}$ as possible. To achieve this, $Z_{\alpha/2}\sqrt{V(\overline{Q})}/\overline{Q}$ is minimized by obtaining more runs. As this value gets closer to 0, $\overline{Q} \to \overline{\overline{Q}}$ with probability 1 – α. An ε value is entered by the user; the simulation program calculates $Z_{\alpha/2}\sqrt{V(\overline{Q})}/\overline{Q}$ after each iteration; compares this quantity with ε and terminates the program if it is less than ε. If it is not less than ε after a maximum number of iterations entered by the user, the program is still terminated to avoid excessive computation time. However, the results may not be as reliable.

4.4 Productivity Improvement Procedure

Operational characteristics, such as machining or operation times, equipment failures, corrective and preventive maintenance activities, and intermediate buffer capacities have significant effects on the production line efficiency. Assessment of the operational efficiency of an automated manufacturing line with storage units by computer simulation permits one to determine various possible parameters and dependent variables which have the most significant effects on productivity. Estimation indices are obtained for such variables as the total, inherent, and imposed relative time losses due to failures and stoppages for each machine. These variables are obtained by employing failure and repair times and nominal and relative production rates for each machine. These terms are defined as follows:

$Q_n(i) = 60/t(i)$ is the normal productivity of machine i, where $t(i)$ is the cycle time for machine i; $Q_r(i) = 60\overline{N}/T_{sim}$ is the relative productivity of machine i; $K_{\text{loss}}(i) = 1 - Q_r(i)/Q_n(i)$ is the total loss factor of machine i; $K_{\text{inh}}(i) = 1 - \bar{t}_r(i)/[\bar{t}_r(i) + \bar{t}_f(i)]$ is inherent loss factor of machine i; and $K_{\text{imp}}(i) = K_{\text{loss}}(i) - K_{\text{inh}}(i) = 1 - Q_r(i)/Q_n(i) - \bar{t}_r(i)/[\bar{t}_r(i) + \bar{t}_f(i)]$ is the imposed loss factor for machine i, i = 1, 2, ….., m. The terms $\bar{t}_f(i)$ and $\bar{t}_r(i)$ are mean times to failure and repairs, respectively of machine i. After determining these loss factors, the program can compare total loss factors for all machines. The machine or stage which has the highest total loss factor is then chosen for improvement. This machine's imposed and inherent losses are compared.

The following two suggestions are made.

(i) If $K_{\text{imp}}(i) > K_{\text{inh}}(i)$, it is suggested that the capacity of storages immediately preceding and succeeding machine i with highest total loss factor should be increased. Reliability and productivity of machine i–1 and machine i+1 should also be increased.

(ii) If $K_{\text{inh}}(i) > K_{\text{imp}}(i)$, stoppages are mainly caused by inherent failures, that is breakdowns. Therefore, the reliability of machine i should be increased by PM or its mean repair time should be decreased in order to gain improvement in total productivity. After the changes are made, simulation should be repeated to see the effects of the proposed changes in the line design.

4.5 Case Problem and Simulation Results

In order to illustrate the model developed above and to demonstrate the effects of maintenance operations on the productivity of serial production lines with intermediate buffers, two case problems are considered as follows.

Case problem 1: A balanced serial production line with 5 stations is considered. Operation times are 1.25 minutes on all stations; Number of failure types on each equipment is 2 failures; Distributions of time to failure and related parameters are Uniform (0, 120) and Uniform (0, 180) with means of 60 and 90 minutes respectively; Distributions of repair times and related parameters are Normal (5, 1) and Normal (7, 1.5); Buffer storage capacities between stations are varied from 1 to 10. When a preventive maintenance is implemented on the line, wear out failures are eliminated and only random failures, with constant failure rates, remain. Time between failures extend from uniform to exponential as explained in section 2.4.

Average line output rate (parts/min) = 0.674104
Standard deviation of line output rate = 0.0031194163

Machine Number	Total Parts Produced	Relative Production Rate	Nominal Production Rate	Imposed Loss Factor	Inherent Loss Factor
1	8102.250	40.515	48.000	0.081	0.075
2	8099.750	40.508	48.000	0.083	0.073
3	8096.250	40.484	48.000	0.080	0.077
4	8089.750	40.456	48.000	0.084	0.074
5	8089.250	40.448	48.000	0.084	0.073

Machine 5 has the maximum total loss factor.
Down time is mainly imposed: increase the capacity of storage adjacent to this machine, also increase reliability by PM; increase productivity of adjacent machines; reduce the CM times and try simulation again.
Maximum iteration is reached at 75.

Table 4. Iterative Simulation Output Results for Production Line Case 1

Two types of PM with intervals of 120 minutes and 180 minutes (corresponding to 96 parts and 144 parts) are assumed to be implemented to eliminate wear out failures; PM times are 2.5 and 3.5 time units. Time to failure distributions change to exponential with mean time to failures of 120 and 180 minutes. Distributions of repair times and related parameters are Normal (5, 1) and Normal (7, 1.5). Buffer storage capacities are again varied between 1 and 10 units. Parameters related to statistical tests were set as follows: α=0.05; $Z_{\alpha/2}$=1.96; ε=0.001, maximum number of iterations=200; and production simulation time was 12,000 minutes. Table 4 shows one output example for the balanced line case under CM&PM policy when maximum buffer capacity of 10 units are allowed between the stations. Average line output rate obtained is 0.674104 parts/minute. Station production rates, loss factors, and suggestion for improvements are also shown in the table. Related results obtained for all other cases of buffer sizes are summarized in Figure 7.

The effects of CM only policy and the policy of implementing PM with CM (CM and PM), under different buffer capacities are illustrated in the figure. An increase in production rate is achieved due to PM implementation for all cases of buffer capacities. The increase in production rate levels off as the buffer capacity is increased.

Case Problem 2: In addition to the balanced line case as shown in figure 7, three unbalanced cases, with a bottleneck station at the beginning, at the middle, and at the end of the line, are considered. Figure 8 shows the results for three unbalanced line cases under CM as well as CM with PM policies. It is assumed that a bottleneck station with operation time of 1.50 minutes is located at the beginning (designated as Bottleneck at Start=BS in figure 8), in the middle (Bottleneck in Middle=BM), or at the end of the line (Bottleneck at End=BE). As it is seen in the figure, all CM & PM cases result in higher production rate than CM only cases. Figure 8 also shows that when the bottleneck station is in the middle of the line, production rate is less than the cases of bottleneck being at the beginning or at the end of the line. These two last cases result in almost equal production rates as can be seen in the figure. The discrete model and the related program can also be used to perform an exhaustive search to find the optimum allocation of total fixed buffer capacity to the stages to maximize the line production rate.

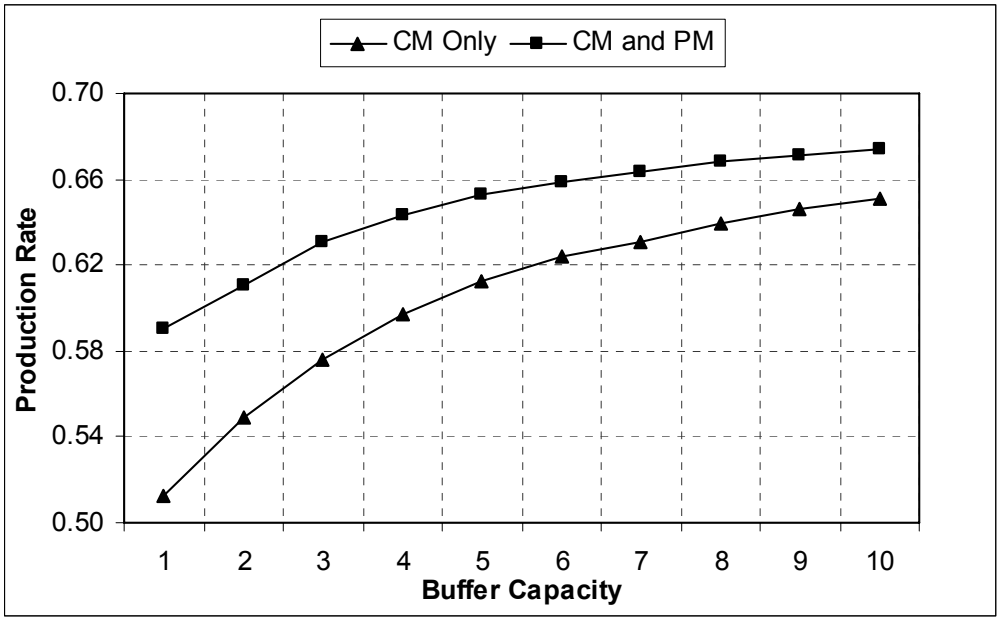


Figure 7. Comparison of CM policy with CM & PM policy for the balanced line

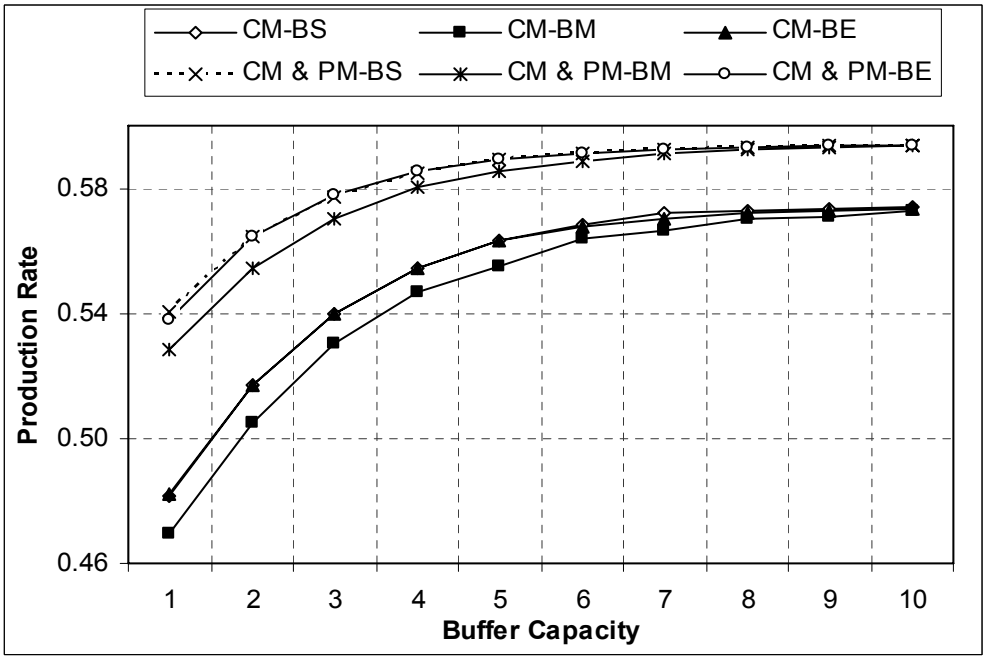


Figure 8. Comparison of CM policy with CM & PM policy for unbalanced line

5. Concluding Remarks

This chapter has presented some basic concepts in maintenance modeling for production systems. Mathematical models are developed for separation of different types of failure rates to evaluate effects of maintenance on equipment productivity. The models were first applied to a FMS through simulation and the results were discussed. It is found that PM of any type results in higher productivity than CM only. However, depending on the type of system considered, some PM policies perform better than others. The best policy must be determined by the analysis of the given system using the tools presented.

In order to analyze the effects of maintenance on the performance of serial production lines, a discrete mathematical model and an iterative computer simulation are developed for multi-stage production lines. The model allows several types of failures and maintenances to be incorporated into the analysis. Based on the discrete model, simulation approach incorporates a three-stage procedure which allows the user to enter a set of data describing the system under study, simulates the system until selected statistical criteria are satisfied and obtains output results. Specific recommendations for productivity increase can be applied until a satisfactory production output is achieved. The model is applied to the cases of balanced and unbalanced lines and the effects of PM are investigated. When PM was implemented in addition to CM, line productivity was significantly increased. The discrete model and the iterative simulation procedure proved to be very useful in estimating the production line productivity for complex realistic production systems. It allows the line designer or operation managers to evaluate the effects of storage-unit capacity and repair/maintenance policies on line productivity. As a future study, the suggested iterative model can be incorporated into an interactive visual computer software to be effectively utilized by design engineers and operation managers.

Acknowledgement

This chapter was prepared based on a research that was supported by Kuwait University Research Administration under the grant number EI02/03.

6. References

Albino, V., Carella, G., and Okogbaa, O. G. (1992). Maintenance Policies in Just-in-time Manufacturing Lines. *International Journal of Production Research*, Vol. 30, No. 2, 369-382.

Al-Najjar, B., Alsyouf, I. (2003). Selecting the Most Efficient Maintenance Approach Using Fuzzy Multiple Criteria Decision Making. *International Journal of Production Economics,* Vol. 84, No. 1, 85-100.

Ben-Daya, M. and Makhdoum, M. (1998). Integrated Production and Quality Model Under Various Preventive Maintenance Policies. *Journal of the Operational Research Society*, Vol. 49, 840-853.

Boukas, K. and Haurie, A. (1990). Manufacturing Flow Control and Preventive Maintenance: A Stochastic Control Approach. *IEEE Transactions on Automatic Control*, Vol. 35, No. 9, 1204-1031

Castro, H. F. D. and Cavalce, K. L.(2006). Maintenance Resources Optimization Applied to a Manufacturing System. *Reliability Engineering and System Safety*, Vol. 91, 413-420.

Chan, F. T. S., Lau, H. C. R., Ip, R. W. L., Chan, H. K., and Kong, S. (2005). Implementation of Total Productive Maintenance: A Case Study. *International Journal of Production Economics*, Vol. 95, No. 1, 71-94.

Cho, I. D. and Parlar, M. (1991). A Survey of Maintenance Models for Multi-unit Systems. *European Journal of Operational Research*, Vol. 51, 1-23.

Creehan, K. D. (2005). Establishing Optimal Maintenance Practices in a Traditional Manufacturing Environment. *Journal of the Chinese Institute of Industrial Engineers*, Vol. 22, No. 1, 11-18.

Dekker, R. and Smeitnik, E. (1994). Preventive Maintenance at Opportunities of Restricted Duration. *Naval Research Logistics,* Vol. 41, 335-353.

Dekker, R.(1996). Applications of Maintenance Optimization Models: A Review and Analysis. *Reliability Engineering and System Safety.* Vol. 51, 229-240.

Gupta, Y. P., Somers, T. M., and Grau, L. (1998). Modeling the Interrelationship Between Downtimes and Uptimes of CNC Machines. *European Journal of Operational Research*, Vol. 37, 254-271.

Kennedy, W. J. (1987). Issues in the Maintenance of Flexible Manufacturing Systems. *Maintenance Management International*, Vol. 7, 43-52

Kyriakidis, E. G. and Dimitrakos, T. D. (2006). Optimal Preventive Maintenance of a Production System with an Intermediate Buffer. *European Journal of Operational Research*, Vol. 168, 86-99.

Levitin, G. and Meizin, L. (2001). Structure optimization for continuous production systems with buffers under reliability constraints. *International Journal of Production Economics,* Vol. 70, 77-87.

Lin, C., Madu, N.C., and Kuei C. (1994). A Closed Queuing Maintenance Network for a Flexible Manufacturing System. *Microelectronics Reliability*, Vol. 34, No. 11, 1733-1744.

McKone, K. and Wiess, E. (1998). TPM: Planned and Autonomous Maintenance: Bridging the Gap Between Practice and Research. *Production an Operations Management,* Vol. 7, No. 4, 335-351.

Mobley, R. K. (1990). An Introduction to Predictive Maintenance. *Van Nostrand Reinhold*. New York.

Papadopoulos, H. T. and Heavey, C. (1996). Queuing Theory in Manufacturing Systems Analysis and Design: A Classification of Models for Production and Transfer Lines. *European J. of Operational Research,* Vol. 92, 1-27.

Pegden, C.D., Shannon, R.E., and Sadowski, R. P. (1995). *Introduction to Simulation Using SIMAN, 2nd edition*, McGraw Hill, New York.

Pintelon, L. M. and Gelders, L. F. (1992). Maintenance Management Decision Making. *European Journal of Operational Research,* Vol. 58, 301-317.

Savsar, M. and Biles, W. E. (1984). Two-Stage Production Lines with Single Repair Crew. *International J. of Production Res.*, Vol. 22, No.3, 499-514.

Savsar, M. (1997a). Modeling and Analysis of a Flexible Manufacturing Cell. *Proceedings of the 22nd International Conference on Computers and Industrial Engineering*, December 20-22, Cairo, Egypt, pp. 184-187.

Savsar, M. (1997b) Simulation Analysis of Maintenance Policies in Just-In-Time Production Systems. *International Journal of Operations & Production Management*, Vol. 17, No. 3, 256-266

Savsar, M. (2000). Reliability Analysis of a Flexible Manufacturing Cell. *Reliability Engineering and System Safety*, Vol. 67, 147-152

Savsar, M. and Youssef, A. S. (2004). An Integrated Simulation-Neural Network Meta Model Application in Designing Production Flow Lines. WSEAS Transactions on Electronics, Vol. 2, No. 1, 366-371.

Savsar, M. (2005a). Performance Analysis of FMS Operating Under Different Failure Rates and Maintenance Policies. *The International Journal of Flexible Manufacturing Systems*, Vol. 16, 229-249.

Savsar, M. (2005b) Buffer Allocation in Serial Production Lines with Preventiveand Corrective Maintenance Operations. *Proceedings of Tehran Interna-*

tional Congress on Manufacturing Engineering (TIMCE'2005), December 12-15, 2005, Tehran, Iran.

Savsar, M. (2006). Effects of Maintenance Policies on the Productivity of Flexible Manufacturing Cells. *Omega*, Vol. 34, 274-282.

Sheu, C. and Krajewski, L.J. (1994). A Decision Model for Corrective Maintenance Management. *International Journal of Production Research*, Vol. 32, No. 6, 1365-1382.

Sun, Y. (1994). Simulation for Maintenance of an FMS: An Integrated System of Maintenance and Decision-Making. *International Journal of Advance Manufacturing Technology*, Vol. 9, 35-39.

Valdez-Flores, C. and Feldman, R.M. (1989). A Survey of Preventive Maintenance Models for Stochastically Deteriorating Single-Unit Systems. *Naval Research. Logistics*, Vol. 36, 419-446.

Vatn, J., Hokstad, P., and Bodsberg, L. (1996). An Overall Model for Maintenance Optimization. *Reliability Engineering and System Safety*, Vol. 51, 241-257

Vineyard M. L. and Meredith J. R. (1992). Effect of Maintenance Policies on FMS Failures. *Int. J. of Prod. Research,* Vol. 30, No. 11, 2647-2657.

Vouros, G. A., Vidalis, M. I., Papadopoulos, H. T. (2000). A Heuristic Algorithm for Buffer Allocation in Unreliable Production Lines. *International Journal of Quantitative Methods*, Vol. 6, No. 1, 23-43.

Waeyenbergh, G. and Pintelon, L. (2004). Maintenance Concept Development: A Case Study. *International J. of Production Economics,* Vol. 89,395-405.

10

Zadehian Paradigms for Knowledge Extraction in Intelligent Manufacturing

A.M.M. Sharif Ullah and Khalifa H. Harib

1. Introduction

Manufacturing is a knowledge-intensive activity. The knowledge underlying a specific manufacturing process or system is often extracted from a small set of experimental observations. To automate the knowledge extraction process various machine learning methods have been used (Pham & Afifi, 2005; Monostori, 2003). Even though such methods are used, a great deal of human intelligence (knowledge extractor's judgment, preference) is required for getting good results (Ullah & Khalifa, 2006). As a result, a machine learning method that is able to utilize human cognition as straightforwardly as possible seems more realistic for extracting knowledge in manufacturing. In fact, human-assisted machine learning methods are in agreement with the modern concept of manufacturing automation—how to support humans with computers rather than how to replace humans by computers (Kals et al., 2004). Thus, for advanced manufacturing systems, the machine learning methods wherein humans and computers compliment each other and the course of knowleldge extraction is determine by the human cognition rather than by a fully automated algorethemic approach is desirable.
Artificial intelligence community has also started to recognize the need for human-assisted machining learning methods (i.e., human comprehensible machine learning methods):

"Humans need to trust that intelligent systems are behaving correctly, and one way to achieve such trust is to enable people to understand the inputs, outputs, and algorithms used as well as any new knowledge acquired through learning. As the use of machine learning increases in critical operations it is being applied increasingly in domains where the learning system's inputs and outputs must be understood, or even modified, by human operators…." (Dan Oblinger, AAAI Technical Report, WS-05-04, 2005.)

Now the question is how to develop human comprhensible machine learning methods for manufacturing? One of the possibilities is to integrate Inductive-Statistical Explanation introduced by Hempel (Hemple, 1965) and Fuzzy Set based computing introduced by Zadeh (Zadeh, 1965). Based on this contempletion this chapter is written. The strucutre of this chapter is as follows: Section 2 describles the Hempelian paradigm relevant to this chapter. Section 3 describles Zadehian paradigm relevant to this chapter. Section 4 describes the non-monotonic nature of real-life probability which is important for understanding the retionale behind the cross-fertilization between Hempelian and Zadehian paradigms. Section 5 describes the human comprehesible machine learning method for extraction knowledge knoweldge extraction process. Section 6 describles how the proposed knowledge extraction process is applied for modeling and simulation of nonlinear behaviors in manufacturing. Section 7 describles how the proposed knoweldge extraction process is applied to The presented knowledge extraction method can be used to establish the relationship between performance measures and control variables a machining operations using a small set of data. Finally, the concuding remarks are shown.

2. Hempelian Paradigm

Carl Gustav Hempel, a philosopher of natural science (Woodward, 2003), introduced two models of scientific explanation: 1) *Deductive-Nomological (D-N) Explanation* (Hempel & Oppenheim, 1948) and 2) *Inductive-Statistical (I-S) Explanation* (Hempel, 1965, 1968). In these models it is assumed that a scientific explanation deals with explanan (object that explains a problem) and explanandum (object that needs to be explained). D-N Explanation means that there is a logically provable universal law (logical positivism) and a "certain" conclusion can be made from the law if the underlying conditions are satisfied—the explanan must be true and the explanandum must be the logical consequence of it (explanan) for all cercusmtances. The logical setting of D-N Explanation is as follows:

$$\begin{array}{ll} (scientific\ law) & p \rightarrow q \\ (condition) & p\ is\ satisfied \\ \hline (certain\ conclusion) & q\ must\ be\ true \end{array} \qquad (1)$$

For example,

$$\begin{array}{ll} (scientific\ law) & bird \rightarrow fly \\ (condition) & It\ is\ a\ bird \\ \hline (certain\ conclusion) & It\ certainly\ can\ fly \end{array} \qquad (1.1)$$

In (1), p (bird) is called explanan and q (fly) is called explanandum. On the other hand, I-S Explanation means that there is no law-like relationship between explanan (p, bird) and explanandum (q, fly) and if one can prove that the probability of explanandum given explanan is very high then the explanandum is "most probably" the best explanation for the explanan. Note that in I-S Explanation the conclusion is not certain.

The logical setting of I-S Explanation is as follows:

$$\begin{array}{ll} (probability\ rule) & Pr(q \mid p) = r, \quad r \approx 1 \\ (condition) & p \qquad\qquad [r] \\ \hline (most\ probable\ conclusion) & q \end{array} \qquad (2)$$

For example,

$$\begin{array}{l} Pr(fly \mid bird) = r, \quad r \approx at\ least\ 0.95 \\ It\ is\ a\ bird \qquad\qquad [r] \\ \hline It\ most\ likely\ can\ fly \end{array} \qquad (2.1)$$

I-S Explanation is more suitable model of explanation (knowledge extraction) for manufacturing because in manufacturing the arguments are experimental data-driven rather than scientific law driven.

3. Zadehian Paradigm

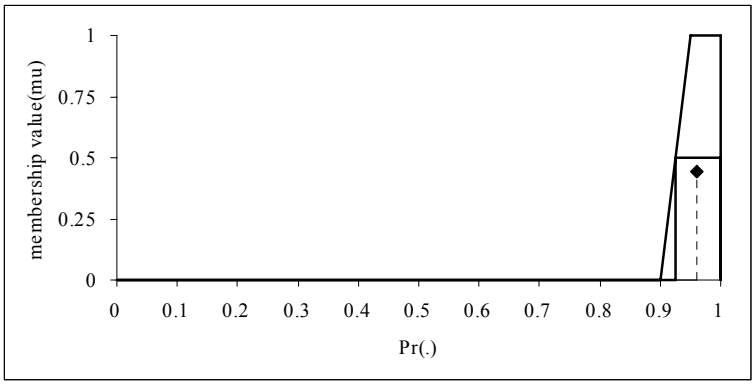
Lotfi Ashker Zadeh, a logician of human cognition, introduced some formal settings for dealing with uncertainty in man-machine systems. The main theme of his philosophy is that the human cognition has two main facets: *Partiality* and *Granularity*. Partiality means tolerance to partial truth—truth value of a proposition or an event is not only true and false, but also partially true or false. Granularity means formation of granules (words or phrases) assigned

un-sharply to a set of values or attributes. To known more about Zadeh's philosophy, refer to Zadeh, 1965; 1975; 1997; 2002; 2005a, 2005b. On of the mathematical entities underlying Granularity and Partiality is fuzzy number (Dubois & Prade, 1978). A fuzzy number A is defined by a membership function μ_A (user-defined) from real line to the interval 0 to 1 and it must satisfy the conditions of normality, convexity, continuity and compactness, as follows:

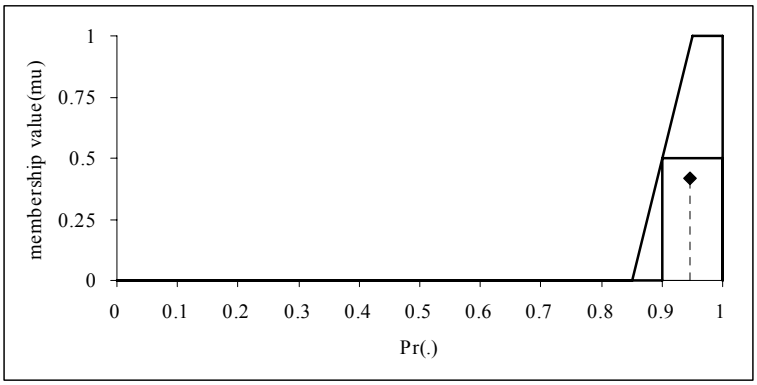
$$
\begin{aligned}
&\mu_A : R \to [0,1] \; such that \\
&(normality) \qquad \exists x \in R, \mu_A(x) = 1 \\
&(convexity) \qquad \mu_A(\lambda x + (1-\lambda) y) \geq min\{\mu_A(x), \mu_A(y)\} \quad for\ any\ x, y \in R, \lambda \in [0,1] \\
&(continuity) \qquad upper semi-continuous \\
&(compactness) \qquad Supp(A) = \{x \in R \mid \mu_A(x) > 0\}\ is\ a\ compact\ set
\end{aligned}
\tag{3}
$$

In semantics sense a fuzzy number A is a fuzzy subset of the real line whose highest membership values μ_A are clustered around a given real number (Dubois & Prade, 1978). Therefore, it is found useful in formally computing uncertain quantities like "probability is at least 0.95". Consider, for example, two fuzzy numbers illustrated in Fig. 1. Both fuzzy numbers define the same quantity A = "probability is at least 0.95" by using two different membership functions:

$$
\begin{aligned}
&\mu_A(Pr) : [0,1] \to [0,1] \\
&definition\ (a) \\
&Pr \mapsto max\left(min\left(\frac{Pr - 0.9}{0.95 - 0.9}, 1 \right), 0 \right) \\
&definition\ (b) \\
&Pr \mapsto max\left(min\left(\frac{Pr - 0.9}{0.95 - 0.9}, 1 \right), 0 \right)
\end{aligned}
\tag{3.1}
$$



definition (a)



definition (b)

Figure 1. Defining probability is at least 0.95 using fuzzy numbers.

The implications of these two definitions will be explained in the next section.

4. Linguistic Likelihood

In I-S Explanation, one is required to use *classical probability* (Kolmogorov, 1933; Billingsley, 1995; Hajeck, 2003) defined by a space called *probability space* (**Ω**, **F**, *Pr*) wherein **Ω** is a non-empty set called *universal set*; **F**, know as *filed*, is a set of all subsets of **Ω** that has **Ω** as a member and that is closed under complementation (with respect to **Ω**) and union; *Pr*, known as *probability function*, is a function from **F** to the real numbers. Pr is monotonic and follows three axioms as follows:

- (Axiom of Non-negativity) $Pr(X) \geq 0$, for all $X \in \mathbf{F}$,
- (Axiom of Normalization) $Pr(\mathbf{\Omega}) = 1$,
- (Axiom of Finite Additivity) $Pr(X \cup Y) = Pr(X) + Pr(Y)$ for all $X, Y \in \mathbf{F}$ such that $X \cap Y = \varnothing$.

To make sure that a value of Pr(X) follows the above mentioned axioms and other theorems and corollaries derived from the above axioms, the value of Pr(X) is determined by the following statistical procedure:

$$Pr(X) = \underset{N \to \infty}{Lim} \frac{N_X}{N} \qquad (4)$$

In (4), X is an event or proposition; N is the number of trails and N_X is the number of trials wherein X is found to be true. In practice, it is almost impossible to make infinite number of trials and determine the precise value of Pr(X). As a result, Pr(X) is estimated from a relatively large number of trails (from the relative frequency N_X/N, N being a relatively large number, i.e., from a relatively large set of data) or from a probability distribution assuming that X follows a distribution (e.g., normal distribution, binomial distribution, etc.). Sometimes the situation is much complex—the underlying probability distribution is unknown or the probability has to be estimated from a relatively small set of data. This is the case in most of the manufacturing situations.

Thus, the probability is perfect, but we can't elicit it perfectly" (O'Hagan & Oakley, 2004). In other words, all real-life probabilities are not monotonic (i.e., non-monotonic) or imprecise in nature. To deal with the non-monotonic nature of probability (i.e., imprecise probabilities) a model with upper and lower provision is used (Walley, 1991; Walley et al., 2004; de Cooman & Zaffalon, 2004; de Cooman et al., 2005; Lukasiewicz, 2005; Tonn, 2005) wherein a probability Pr(X) is expressed by lower provision, $\underline{Pr}(X)$, and upper provision, $\overline{Pr}(X)$, i.e., Pr(X) = $[\underline{Pr}(X), \overline{Pr}(X)]$. For example, consider Pr(fly | bird) = "at least 0.95". Here, the lower provision $\underline{Pr}(X)$ = 0.95 and upper provision is $\overline{Pr}(X)$= 1.

If someone calculates the value of Pr(X) from a limited number of observations, then (from the sense of imprecise probability) Pr(X) should be treated in such a way as if it is a range rather than a single value. One of the ways to achieve this is to use a set of linguistic likelihoods defined by appropriate fuzzy numbers and translate Pr(X) into the linguistic likelihood that contains Pr(X) most (Ullah & Harib, 2005). The translated linguistic likelihood can then be used to find upper and lower provisions or other quantities that are important from the view point of imprecise probability (Zadeh, 2002; 2005). For example, consider that one calculates Pr(X) = 0.85 from a limited number of observations.

This Pr(X) = 0.85 can be translated into linguistic likelihood labeled most-likely if the linguistic likelihoods illustrated in Figure 2 are used because $\mu_{\text{most-likely}}(0.85)$ is greater than $\mu_{\text{likely}}(0.85)$ and $\mu_{\text{most-unlikely}}(0.85)$. In fact, $\mu_{\text{most-likely}}(0.85)$ = 0.833, $\mu_{\text{likely}}(0.85)$ =0.167, and $\mu_{\text{most-unlikely}}(0.85)$ = 0 because

$$\mu_{most\text{-}likely}(Pr):[0,1]\to[0,1],\quad Pr\mapsto max\left(min\left(\frac{Pr-0.6}{0.9-0.6},1\right),0\right)$$

$$\mu_{likely}(Pr):[0,1]\to[0,1],\quad Pr\mapsto max\left(min\left(\frac{Pr-0.1}{0.4-0.1},1,\frac{0.9-Pr}{0.9-0.6}\right),0\right).$$

$$\mu_{most\text{-}unlikely}(Pr):[0,1]\to[0,1],\quad Pr\mapsto max\left(min\left(\frac{0.4-Pr}{0.4-0.1},1\right),0\right)$$

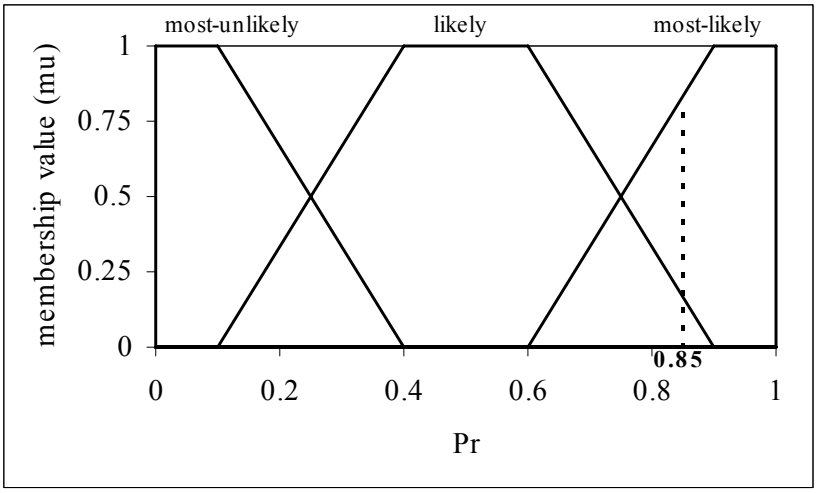


Figure 2. Three linguistic likelihoods.

Therefore, Pr(X) = 0.85 is no longer a crisp value "0.85". It is a fuzzy number labeled "most-likely". If needed, the lower and upper provisions can be calculated from "most-likely" using α-cuts:

$$most\text{-}likely_{\alpha}=\left\{Pr\mid\mu_{most\text{-}likey}(Pr)\geq\alpha,\alpha\in(0,1]\right\}.$$

For example, if α = 0.5, then most-likely$_{\alpha=0.5}$ = [0.75,1], i.e., $\underline{Pr}(X)$ = 0.75 and $\overline{Pr}(X)$ = 1. Moreover, if needed, the expected value of most-likely

$$E(most\text{-}likely)=\frac{\int_0^1\left[Pr\times\mu_{most\text{-}likely}(Pr)\right]dPr}{\int_0^1\mu_{most\text{-}likely}(Pr)dPr}=0.860$$

can be used to find out the average value of the imprecise probability Pr(X) = 0.85.

Compared to upper and lower provisions, expected value is more robust. The explanation is as follows: Recall the linguistic likelihoods shown in Fig. 1. E(probability is at least 0.95) = 0.96 for the first definition and = 0.95 for the other definition, respectively. This implies that even though the definition of a linguistic likelihood varies from person to person, the expected value does not vary much (for the above two cases the expected values are 0.96 and 0.95, a different of 1%, only). Therefore, if the expected value of a linguistic likelihood is used in an inference mechanism, the inferred output will not vary much from person to person. In other words, an inference mechanism will become robust if the expected value of the linguistic likelihood of an imprecise probability is used.
The above explanation can be summarized into the following procedure:

- Determine Pr from a small set of data
- Define a set of linguistic likelihoods, {L1,…,Ln} using appropriate fuzzy number in the universe of discourse [0,1]. Here, Li = {(Pr, $\mu_{Li}(Pr)$) |Pr, $\mu_{Li}(Pr)) \in [0,1]$}, $\forall i \in \{1,\ldots,n\}$. All $\mu_{Li}(Pr)$) follow the characteristics of fuzzy number.
- Translate Pr into LL $\in$ {L1,…,Ln} such that $\mu_{LL}(Pr) > \mu_{Lj}(Pr)$, Lj $\in$ {L1,…,Ln}-{LL}, $\forall j = 1,\ldots,n$. This means Pr |= LL.
- Determine the expected value, E(LL), of LL as follow:

$$E(LL) = \frac{\int (\mu_{LL}(Pr) \times Pr) dPr}{\int (\mu_{LL}(Pr)) dPr} \quad .$$

- Use E(LL) instead of Pr in further calculations. This means that Pr |= LL |= E(LL).

5. Knowledge Extraction Process

Based on the Hempel's I-S Explanation and Zadeh's fuzzy number based linguistic likelihoods the logical setting for extracting knowledge from a small set of numerical data is proposed, as follows:

$$\frac{Pr(q \mid p) \mid = most\text{-}likely, \qquad E(most\text{-}likely) \approx 1}{If \quad p \quad Then \quad most\text{-}likely \quad q} \tag{5}$$

In (5), "most-likely" is a metaphor that expresses the imprecise likeliness of occurring q given p. According to I-S Explanation the expected value of "most-likely" should be near to 1. Now, if p is found to be true, the extracted knowledge (i.e., the "If...Then..." rule "If p Then most-likely q") can be used to produce output. The output is a value in the range underlying most-likely q, i.e., a range little longer than q constrained by the expected value of "most-likely". For example, consider the arbitrary case shown in Fig. 3. As seen from Fig. 3, Pr(q|p) = 0.8 |= "most-likely", according to the definition linguistic likelihoods illustrated in Fig. 2.

Therefore, the rule in (5) holds for the arbitrary case shown in Fig. 3. If the input is a point in the range "p", then the output is in the range "q" for most of the cases and is out of the range "q" for a few cases. This means that most-likely q = q′ is a range little longer than q or $q \subseteq q'$. This leads to the following inference mechanism:

$$\frac{\begin{array}{ll}(Extracted) & If \quad p \quad Then \quad most\text{-}likely \quad q \\ (Given) & x \in p\end{array}}{y \in q' \qquad (q \subseteq q')} \tag{6}$$

To derive p and q from two fuzzy numbers A and B, α-cuts can be employed. As such, $p = A_{\alpha 1}$ and $q = B_{\alpha 2}$. In this case, q′ can be derived in the following manner:

$$\frac{\begin{array}{ll}(Extracted) & If \ A_{\alpha 1} \quad Then \quad most\text{-}likely \quad B_{\alpha 2} \\ (Given) & x \in A_{\alpha 1} \\ (Defined) & B'_{\alpha 2} = B_{\alpha 2} \\ & Supp(B') = (min(U_B), max(U_B)) \\ & \alpha' = \alpha 2 \times E(most\text{-}likely)\end{array}}{y \in B'_{\alpha'}} \tag{7}$$

In (7), B′ is a fuzzy number whose support spreads all over the universe of discourse of B and its α-cut at α2 is equal to that of B. Figure 4 depicts the inference mechanism. A seen from Fig. 4, the inferred output y is any value in the range $B'_{\alpha'}$, which is an α-cut of B′ at α′ = α2×E(most-likely). Since α′ < α2, $B'_{\alpha'} > B_{\alpha 2}$.

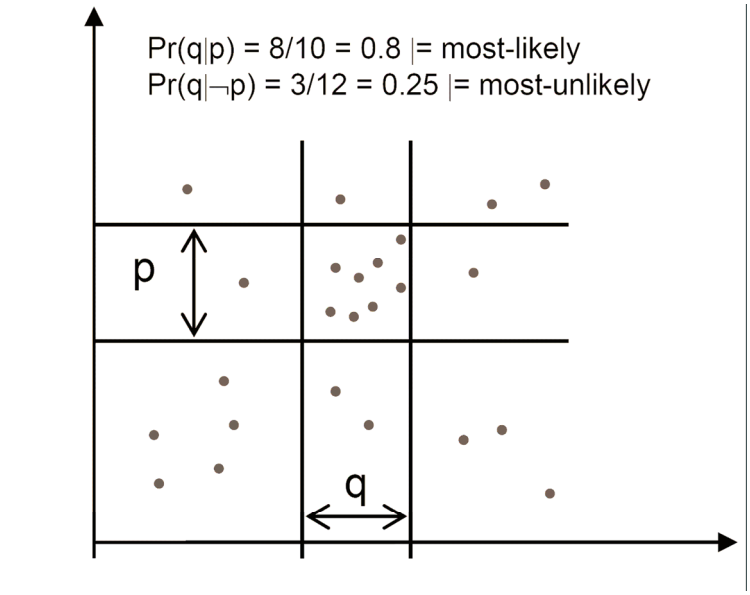


Figure 3. An Arbitrary relationship between p and q

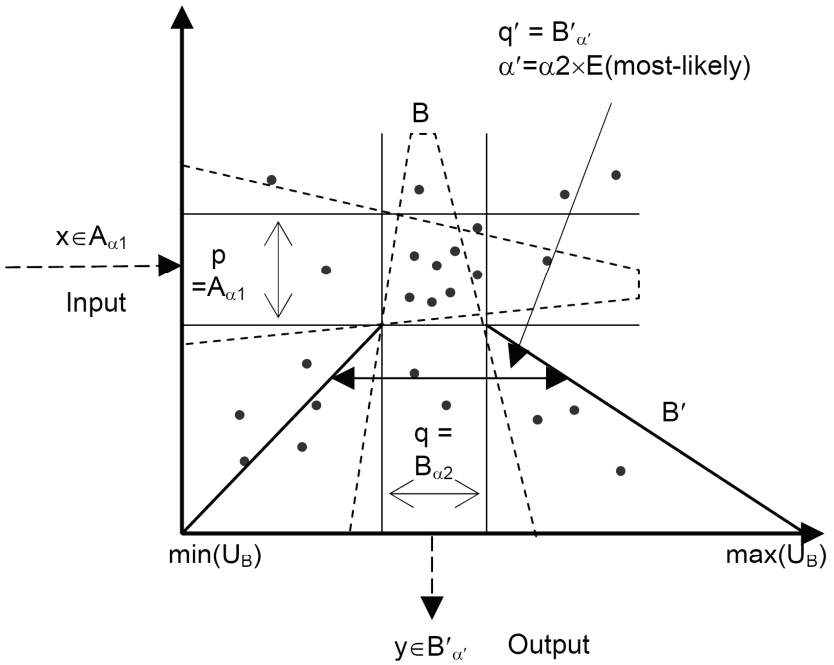


Figure 4. The Proposed Inference Mechanism

6. Nonlinear Signal Modeling and Simulation

This section shows how the proposed knowledge extraction method is applied to capture the dynamics of a highly nonlinear behavior Y(t) from its return map, i.e., a map of (Y(t),Y(t+1)). Particularly for capturing the dynamics of surface roughness the following form of rules are found useful (Ullah & Harib, 2004; 2006):

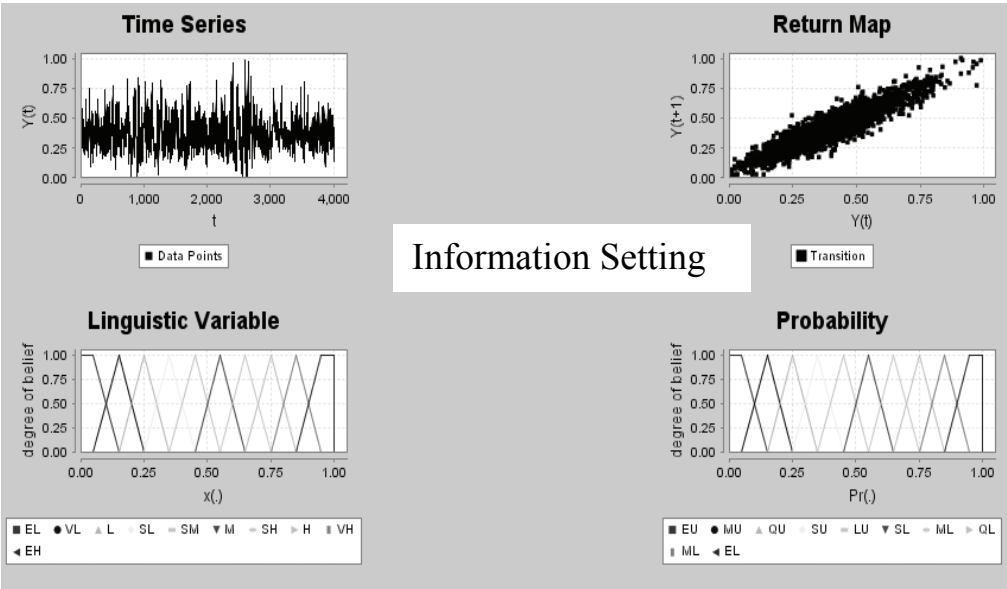
If Y(t) is extremely low Then Y(t+1) is most-likely very low

If Y(t) is very low Then Y(t+1) is most-likely very low

If Y(t) is extremely high Then Y(t+1) is most-likely extremely high

Here the phrase "most-likely" is metaphor for the linguistic likelihood associated with imprecise probability $Pr(Y(t+1)=State_i|Y(t)=State_i)$ of transition from a state $State_i$ to the same state after a unit interval. Figures 5 and 6 show some of the results and user interfaces of the software tool developed by the authors recently (Ullah & Harib, 2006).

Figure 5 shows how a user defines the states of Y(t) (under the message Linguistic Variable) and linguistic likelihoods (under the message Probability). In this stage the system automatically generates the return map from an input signal. The system then extracts the "if…then…" rules, as shown in Fig. 5.



(Extracted Rules)

(Human comprehensible rules)	(Machine comprehensible rules)
If Y(t) is Exteremely Low Then Y(t+1) is Moderately Likely Exteremely Low	If Y(t) is [0, 0.1) Then Y(t+1) is [0,0.415]
If Y(t) is Very Low Then Y(t+1) is Some Likely Very Low	If Y(t) is [0.1, 0.2) Then Y(t+1) is [0.055,0.56]
If Y(t) is Some Low Then Y(t+1) is Moderately Likely Some Low	If Y(t) is [0.2, 0.3) Then Y(t+1) is [0.13,0.545]
If Y(t) is Low Then Y(t+1) is Some Likely Low	If Y(t) is [0.3, 0.4) Then Y(t+1) is [0.165,0.67]
If Y(t) is Some Moderate Then Y(t+1) is Some Likely Some Moderate	If Y(t) is [0.4, 0.5) Then Y(t+1) is [0.22,0.725]
If Y(t) is Moderate Then Y(t+1) is Some Likely Moderate	If Y(t) is [0.5, 0.6) Then Y(t+1) is [0.275,0.78]
If Y(t) is Some High Then Y(t+1) is Less Unlikely Some High	If Y(t) is [0.6, 0.7) Then Y(t+1) is [0.27,0.865]
If Y(t) is High Then Y(t+1) is Some Likely High	If Y(t) is [0.7, 0.8) Then Y(t+1) is [0.385,0.89]
If Y(t) is Very High Then Y(t+1) is Less Unlikely High	If Y(t) is [0.8, 0.9) Then Y(t+1) is [0.315,0.91]
If Y(t) is Extremely High Then Y(t+1) is Moderately Likely Extremely High	If Y(t) is [0.9, 1] Then Y(t+1) is [0.585,1]

Figure 5. Knowledge extraction from the return map of nonlinear behavior (Ullah & Harib, 2006)

These rules can be used recurrently to simulate a surface roughness profile similar to the original one. For example consider the simulated surface roughness profiles shown in Fig. 6 wherein two consecutive simulations are shown. These signals can be connected piece-wise to produce a signal similar to the original one. See Ullah & Harib, 2006 for other computational issues associated with this simulation techniques and implications of such knowledge based technique from the context of exchanging information of nonlinear behaviors from one manufacturing system to another.

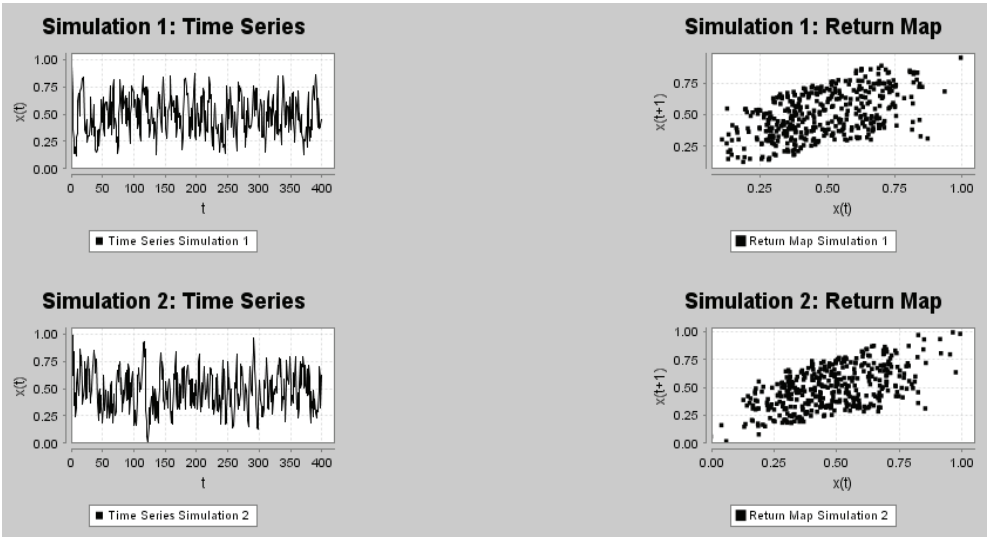


Figure 6. Simulated Surface Roughness (Ullah & Harib, 2006)

7. Knowledge Extraction for Machining Operations

In machining of materials, machining conditions (feed (f), depth of cut (a_p), width of cut (a_e) cutting velocity (V_c), tool nose radius (r_ε) and others process parameters) should be adjusted in such a way so that the required surface finish or productivity can be ensured. The presented knowledge extraction method can be used to establish the relationship between process Performance Measures (PMs) and process Control Variables (CVs). The relationship should provide a recommended range (RE) of a CV to ensure an acceptable range (AC) of a PM. For machining, surface finish (Ra, fractal dimension of roughness profile) and productivity (e.g., for turning productivity = $V_c \times f \times a_p$) are two common machining PMs. On the other hand, CV can be defined using a function $\phi(V_c, f, a_p, a_e, r_\varepsilon)$. See Ullah & Harib 2005a; 2005b for more details. The logical setting proposed for extracting rules for PMs and CVs is as follows:

$$
\begin{array}{l}
\left(Pr\left(PM = AC_{\alpha 1} \mid CV = RE_{\alpha 2}\right) \mid = most - likely\right) \wedge \\
\left(Pr\left(PM = AC_{\alpha 1} \mid CV = \neg RE_{\alpha 2}\right) \mid = most - unlikely\right) \\
E\left(most - likely\right) \approx 1 \\
E\left(most - unlikely\right) \approx 0 \\
\hline
\left(Extracted\right) \quad If\ CV\ is\ RE_{\alpha 2}\ Then\ PM\ is\ most - likely\ AC_{\alpha 1} \\
\left(Extracted\right) \quad If\ CV\ is\ not\ RE_{\alpha 2}\ Then\ PM\ is\ most - likely\ AC_{\alpha 1} \\
\left(Given\right) \quad CV \in PM_{\alpha 2} \\
\hline
\left(Output\right) \quad PM \in RE'_{\alpha'}
\end{array}
\qquad (8)
$$

Based on logical setting in (8), a JAVA™ based computing tool has been developed (Ullah & Harib, 2005b) to extract rules. For example, consider the case shown in Fig. 7. For this particular case, the user extracts "If…Then…" rules to establish the relationship between surface finish (Ra) and V_c, f, and r_ε. The interest is to set the values of V_c, f, and r_ε in such a way so that Ra remains relatively small. Therefore, the acceptable range of Ra is set by a trapezoidal fuzzy number AC = (0,1,1.5,2.5) as shown in Fig. 7. The user defines CV = $V_c \times f \times r_\varepsilon$ (in Fig. 7, V_c is shown by V and r_ε is shown by r) based on the judgement that surface roughness is affected mainly by f, r_ε, and V_c. The data of CV and PM plotted in Fig. 7 are corresponds to that of Ullah & Harib 2005b. As seen from Fig. 7, the desired rules according to the logical setting in (8) are found when recommended range of CV, i.e., RE, is set to be a fuzzy number RE = (90,150,140,90). The rules are:

If $V_c \times f \times r_\varepsilon$ is $RE_{0.5}$ Then Ra is Absolutely Likely $AC_{0.5}$

If $V_c \times f \times r_\varepsilon$ is not $RE_{0.5}$ Then Ra is Absolutely Unlikely $AC_{0.5}$

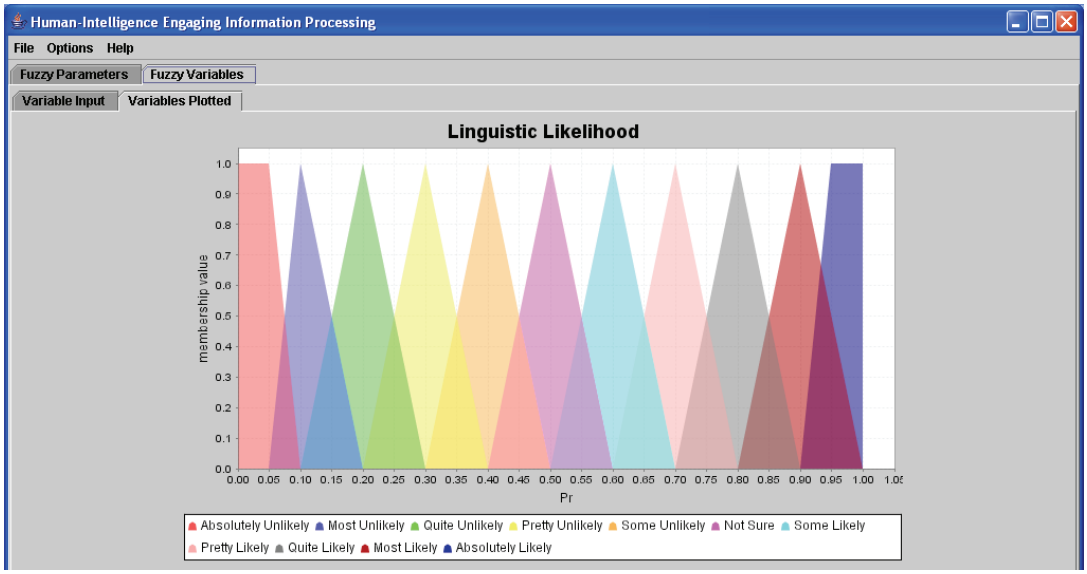
Here, $RE_{0.5}$ = [115,180]. According to the logical setting in (8), "Absolutely Likely $AC_{0.5}$" = "Absolutely Likely [0.5,2]" = [0.4806,2.8944].
This means that if someone keeps $V_c \times f \times r_\varepsilon$ in the range of [115,180], then Ra will remain in the range [0.480,2.8944]. Thus, the computer comprehensible rule is as follows:

If $V_c \times f \times r_\varepsilon$ is [115,180] Then Ra is [0.4806,2.8944].

The above rule can be further modified so that the modified rule remains valid for all most all recommended input points and predicts output not in the range much longer than the predicted range. One of the possible modifications is as follows:

If $V_c \times f \times r_\varepsilon$ is Moderate Then Ra is Fine.

In the modified rule, "Moderate" is a triangular fuzzy number (120,150,180). This implies that Supp(Moderate) = [120,180], i.e., a range that is narrowly included in [115,180]. On the other hand, "Fine" is a trapezoidal fuzzy number (0,0,1,3). This implies that Supp(Fine) = [0,3], i.e., a range that narrowly includes [0.4806,2.8944].



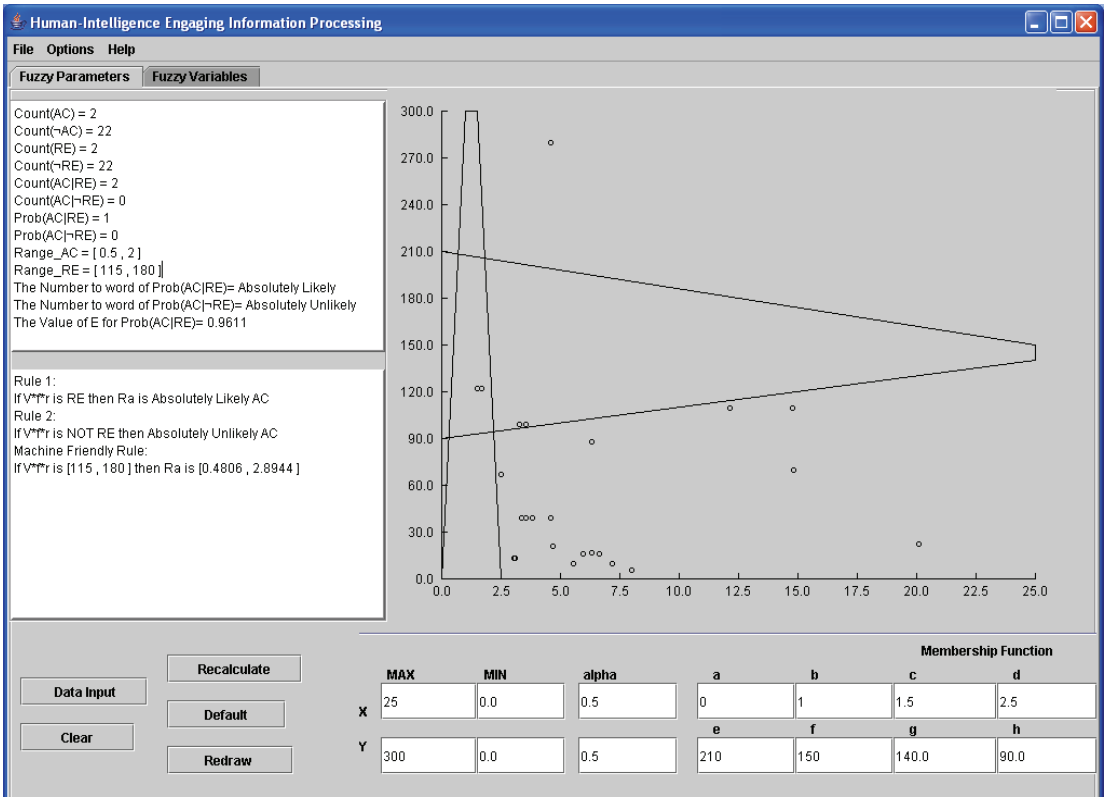


Figure 7. Knowledge Extraction for Machining of Materials

Note that the linguistic likelihoods (Absolutely Likely…Absolutely Unlikely) are user-defined as shown in the bottom of Fig. 7.

8. Concluding Remarks

Zadeh's partiality and granularity based computation has widely been applied in developing intelligent systems since its inception in 1965 whereas Hempel's empirical positivism (i.e., I-S Explanation) remains relatively untouched. There is a study that shows that Hempel's I-S Explanation underlies non-monotonic or default logic—a computational framework for making natural conclusion under evolving information (Tan, 1997). Hempel's I-S Explanation inspired inference from medical data has gained some attention (Gandjour & Lauterbach, 2003). This chapter has added a new dimension to the application potential of Hempel's I-S Explanation in intelligent systems development. Particularly it is

shown that cross-fertilization between Hempel's I-S Explanation and Zadeh's fuzzy number based imprecise probabilities provides the logical setting for human comprehensible machine learning methods that are able to extract human- and machine-friendly "If…Then…" rules from empirical facts (i.e., numerical data).

Two case studies are shown in this chapter to demonstrate the effectiveness of the method. In the first case study it is shown that the presented method is helpful in nonlinear signal modeling and simulation. Particularly, the case of knowledge based modeling and simulation of surface roughness (a very highly nonlinear behavior in manufacturing) is shown. This idea can be extended to model other nonlinear behaviors (productivity, cutting forces, cutting temperature, etc.) encountered in manufacturing. Since the nonlinear behaviors are stored by using a small set of "If…Then" rules, the method can be used to exchange the information of nonlinear behaviors from one manufacturing system to another.

The other case study deals with the extraction of "If…Then…" rules for machining. The goal is to get such rules that are able to predict the machining performance measures (e.g., surface roughness, productivity) for a given combination of cutting conditions (feed, depth of cut, cutting velocity, tool nose radius, etc.). Particularly, a rule is extracted to ensure fine surface finish by keeping the product of cutting velocity, feed rate and tool nose radius to a range. The issues of how to make the rules more general is also elaborated.
As the modern concept of manufacturing automation is "how to support humans with computers" rather than "how to replace humans by computers", the presented knowledge extraction method will provide valuable hints for the manufacturing systems developers to develop more human- and computer-friendly computing tools.

9. References

Billingsley, P. (1995). Probability and Measure (3rd ed.), John Wiley & Sons: New York.

de Cooman, G., and Zaffalon, M. (2004). Updating beliefs with incomplete observations, Artificial Intelligence, 159, (2004), 75–125.

de Cooman, G., Troffaes, M.C.M., and Mirand, E. (2005). n-Monotone lower previsions, Journal of Intelligent & Fuzzy Systems 16, 253–263.

Dubois, D. & Prade, H. (1978). Operations on fuzzy numbers, International Journal of Systems Science, 9(6), 613-626.

Gandjour, A. and Lauterbach, K.W. (2003). Inductive reasoning in medicine: lessons from Carl Gustav Hempel's 'inductive-statistical' model, Journal of Evaluation in Clinical Practice, 9, 2, 161–169.

Hajek, A. (2003), Interpretations of Probability, Stanford Encyclopedia of Philosophy, URL http:// plato.stanford.edu /archives/sum2003/ entries/ probability-interpret/.

Hempel, C.G. & Oppenheim, P. (1948). Studies in the Logic of Explanation, Philosophy of Science 15, 135-175.

Hempel, C. G. (1968). Maximal Specificity and Lawlikeness in Probabilistic Explanation, Philosophy of Science 35, 116–33.

Hempel, C.G. (1965). Aspects of Scientific Explanation, in Aspects of Scientific Explanation and Other Essays in the Philosophy of Science, (Ed.) C.G. Hempel, Free Press, New York, pp. 331-496.

Kals, H.J.J., Mentink, R.J., Wijnker, T.C. and Lutters D. (2004). Information Management and Process Integration in Manufacturing, CIRP Journal of Manufacturing Systems, 33, 1-10.

Kolmogorov, A.N. (1933). Grundbegriffe derWahrscheinlichkeitsrechnung. Springer, Berlin.

Lukasiewicz, T. (2005). Weak nonmonotonic probabilistic logics, Artificial Intelligence, 168, 119–161.

Monostori, L. (2003). AI and machine learning techniques for managing complexity, changes and uncertainties in manufacturing, Engineering Applications of Artificial Intelligence, 16, 4, 277-291.

O'Hagan, A. and Oakley, J.E. (2004). Probability is perfect, but we can't elicit it perfectly, Reliability Engineering and System Safety, 85, 239–248.

Pham, D.T. and Afify, A.A. (2005). Proceedings of the IMechE Part B, Journal of Engineering Manufacture, Vol. 219, B5, 395-412.

Salmon, W.C. (1999). The Spirit of Logical Empiricism: Carl G. Hempel's Role in Twentieth-Century Philosophy of Science, Philosophy of Science, 66, 333-350.

Tan, Y.-H. (1997). Is Default Logic a Reinvention of Inductive-Statistical Reasoning?, Synthese, 110, 357-379.

Tonn, B. (2005). Imprecise probabilities and scenarios. Futures, 37, 767–775.

Ullah, A.M.M.S. and Harib, K.H. (2004). Novel Techniques for Knowledge Formulation for Intelligent Manufacturing Engineering, Proceedings of

the 4th CIRP International Seminar on Intelligent Computation in Manufacturing Engineering (CIRP ICME 04), 30 June-2 July 2004, Sorrento (Naples), Italy.

Ullah, A.M.M.S. and Harib, K.H. (2005a). Manufacturing Process Performance Prediction by Integrating Crisp and Granular Information, Journal of Journal of Intelligent Manufacturing 16, 3, 319-332.

Ullah, A.M.M.S. and Harib, K.H. (2005b). A Human-Assisted Knowledge Extraction Method for Machining Operations, under review in Advanced Engineering Informatics.

Ullah, A.M.M.S. and Harib, K.H. (2006). Knowledge extraction from time series and its application to surface roughness simulation, to appear in Information Knowledge and Systems Management.

Walley, P. (1991). Statistical Reasoning with Imprecise Probabilities. Chapman and Hall: London.

Walley, P., Pelessoni, R. and Vicig, Paolo. (2004). Direct algorithms for checking consistency and making inferences from conditional probability assessments, Journal of Statistical Planning and Inference, 126, 119 – 151.

Woodward, J.F. (2003). Scientific Explanation: Stanford Encyclopedia of Philosophy, http://plato.stanford.edu/entries/scientific-explanation.

Zadeh L.A. (2005). From imprecise to granular probabilities, Fuzzy Sets and Systems, 154(3), 370-374.

Zadeh, L.A. (1965). Fuzzy sets, Information and Control, 8, 338–353.

Zadeh, L.A. (1975). Fuzzy logic and approximate reasoning, Synthese, 30(3–4), 407–428.

Zadeh, L.A. (1997). Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, Fuzzy Sets and Systems, 90, 2, 111-127.

Zadeh, L.A. (2002). Toward a perception-based theory of probabilistic reasoning with imprecise probabilities, Journal of Statistical Planning and Inference, 105, 233–264

Zadeh, L.A. (2005). Toward a generalized theory of uncertainty (GTU)—an outline, Information Sciences, 172:1-2, 1-40.

11

PURE: A Fuzzy Model for Product Upgradability and Reusability Evaluation for Remanufacture

Ke Xing, Kazem Abhary and Lee Luong

1. Introduction

Remanufacture strategy has been broadly applied to curb the environmental impacts of waste from discarded products. The strategy strives for waste reduction by promoting reutilisation and service life extension of end-of-life products. However, challenged by fast changing technologies and fashions, to rebuild a product just "as it is (or was)" often falls short in making it favourable in the market environment. It is important that the upgrade strategy can be incorporated with remanufacture to achieve a real effective reutilisation of products. In this paper, a mathematical model of product upgradability is proposed to measure a product's potential to accommodate incremental changes/improvements in its functionality during remanufacture. By using fuzzy set theory, the evaluation model represents and examines the product's fitness in terms of its technological, functional, physical, as well as structural characteristics, providing a number of indications for further improvement to base on.

1.1 Remanufacture and its Challenges

As a life cycle strategy supporting product service life extension, remanufacture has become a "rising giant" worldwide in recent years, providing very promising solutions to reduce the demands on landfill space and virgin materials through salvaging the reusable parts of retired products. It is regarded as the ultimate form of recycling as it recaptures the value-added to a product when it was first produced without reducing the product to its elemental form, e.g. materials or chemicals.

Generally, the physical as well as functional conditions of those end-of-life products are restored by rebuilding or replacing component parts wherever necessary and reassembling them to form "remanufactured products" (Rose

2000; Ijomah et al. 1999). However, current trends of development in technologies and marketing present big challenges to the effectiveness of remanufacturing practice. The evolutionary changes to products often lead to higher market expectations on their functionality and make it more difficult for remanufactured products to be in line with the "like-new" criteria and customer demands. Also, the product variety generated by such changes and associated with customisation creates a new and more competitive arena for the remanufactured products to strive for their market status. In both circumstances, a conventional restoration of a product just "as it was" is becoming insufficient, sometimes even unworthy, to make it acceptable to the consumers. Therefore, a more ambitious upgrade strategy should be implemented in conjunction with remanufacture to champion its effectiveness and attractiveness.

1.2 Product Upgradability in the Context of Remanufacture

Different from the upgrade "on the drawing board" in the design phase, upgrading products through remanufacture is more based on the physical reuse of their current configurations. The improvement is added-in to the existing frames. During remanufacture, the intended changes to the existing physical configurations should be minimised, or even prohibited. Compared with the upgrading work in the design stage, there are more constraints and less freedom for incorporating new functions or improvements "off the drawing board" in the context of remanufacture. The most convenient as well as typical example of such upgrade is the plug-in feature of personal computers.

When a product life cycle process is analysed, the performance or potential of this practice is addressed and measured in terms of its "*-ability*", which stands for the characteristics, or the virtue, of the process (Huang 1996). Upgrade is one of the elements of product life cycle processes. Therefore, following the pattern of "*x-ability*", in its simplified form product *upgradability* can be defined as the level of potential of a product to be effectively as well as efficiently upgraded.

Given the commonalities and, more importantly, the differences of product upgrade in the two circumstances, the upgradability of a product in remanufacture can be regarded as a reflection of its characteristics supporting new functions or improvement and reuse at the engineering metrics level, the component level, and the structure level, featuring the projections from customer demands to the different hierarchies of domains for the configuration of products.

Generically, the fitness of engineering metrics settings to upgrade is represented by the compatibility of the current parameters to the new functions or improvements. At the component level, whether the product is upgradable in the context remanufacture is influenced by the reusability of the components. Understandably, unless the key components are reusable for an extended service life, the product is not worth to be remanufactured and thus unsupportive for upgrading. A modular structure is ideal for ease of swap and upgrade during remanufacturing processes. Therefore, at the structure level, the upgradability of a product can be addressed in terms of its modularity feature. From this perspective, the upgradability of a product in the context of remanufacture is represented in the form of a system of characteristics from different levels. It is a joint effect, or integration, of parametrical compatibility, component reusability and structural modularity.

1.3 Current Works on Product Upgradability and Limitations

The incorporation of refurbishments and technical upgrades is regarded as a resources and energy efficient way to achieve product life cycle extension in remanufacturing environment (Guide & Srivastava 1997). As a form of retrofitting, upgrading products in association with remanufacture in various industry contexts has been quite extensively researched, ranging from the improvement of productive life of injection moulding machines (Overfield 1979), the rebuild of used machine tools for performance (Drozda 1984) and the renewal of power generation equipment (Beck & Gros 1997), to the rotorcraft modernisation for US Marine Corps (Wall 2004). In these works, upgrade opportunity of the products was studied as the complements to where remanufacture alone fell short to provide holistic solutions. However, the investigations in those works were mainly focused on operational issues rather than studying the inherent ease-of-upgrade characteristics of the products.

The major aims of assessing product upgradability are to make a long-term upgrade plan for multi-generations of a product during its use or remanufacturing stage and assist designers to derive a suitable design solution to for the whole product (Umemori et al. 2001). The effects of structural configuration on system upgradability was studied by Pridmore et al. (1997) when they investigated the favourable configuration forms for rapid prototyping of application-specific signal processor, which enable both hardware and software reuse with open interface standards. The techno-financial dominance of upgrade option

was highlighted in work of Wilhelm et al. (2003) through emphasizing the content and timing of upgrades in maximising the life cycle revenue of a family of high-tech products. Shimomura and the colleagues (1999) stressed the significance of a product's upgradability to the extension of its service life and the reduction of resource and energy consumption. By comparing upgradable artefacts and design with traditional modes, they proposed two basic guiding principles for improving product upgradability, namely "function independent" and "function insensitive". Umeda and his team furthered the research and implemented the two principles as the basis to analyse the linkage between function parameters and design parameters, represented as a casual relation graph, and evaluate the upgrade potential of products (Umemori et al. 2001; Ishigami et al. 2003)

Nevertheless, the reviews on the features of contemporary works show that the synergy of technology improvement, module introduction, system reliability and component reusability is not sufficiently reflected by the current representations and measures of products' potential for upgrade. The works focusing on the modelling of product upgradability are scars and there is lack of a systematic way to incorporate the key technical factors from the three major product domains in the formulation of upgradability evaluation approaches.
In order to overcome the limitations of the current research and methods, this paper is to propose an integrated evaluation approach for a product's upgrade potential, Product Upgradability and Reusability Evaluator (PURE), with an integral consideration of the fitness of its functional, physical and structural characteristics, and the ability to identify the weaknesses in its configurations for improvement. The implementation of this approach is intended to incorporate with the design of products to represent and measure their upgradability, identifying the weaknesses in their configurations for design improvement.
In the subsequent arrangement of the paper, Section 2 is dedicated to propose three key technical indicators to represent and measure the upgradability features of products. Following that, the approach of PURE is elaborated in Section 3 with the formulation of mathematical models for product upgradability and its indicators by using Fuzzy Set Theory. Then, in Section 4, a case study on a Solar Roof Home Heating System is provided to demonstrate its effectiveness in evaluating product upgradability for different upgrade scenarios. Finally, the features of this upgradability evaluation approach and the future work will be discussed in the summary.

2. Upgradability Indicators of PURE

As discussed in the previous section, in the context of remanufacture product upgradability is regarded as an integrated reflection of the characteristics of being compatible in functional parameters, containing reusable "core" (key components), and having a modular structure. In the approach of PURE, three technical factors are defined as the indicators for the corresponding characteristics, and they are adopted for the measure of a product's suitability to upgrade.

2.1 Indicator of Compatibility to Generational Variety (CGV)

Generally, generational variety is created through introducing a variety of evolutionary changes or incremental improvements to the functionality of a product on the base of the existing settings. As far as product upgradability be concerned, when the new functions or improvements are mapped as customer requirements to corresponding engineering metrics (EM) of product functions in the functional domain, it is important that the current parameters of those EMs have certain levels of fitness to meet the new requirements and ability to accommodate the changes. For any individual metric, its susceptibility to the direct or indirect impacts of upgrade rests with the conformance of its parametrical setting with the possible new performance requirement imposed on it by the functionality improvement. Practically, the smaller the gap between the current performance and the new ideal performance of an engineering metric, the higher the contribution it has to the ease-of-upgrade potential of the product. By forecasting the possible functional changes and their effects on the EMs of the product, such fitness can be represented and measured as the reflection of the degrees of disagreement, or distance, between the current and the expected EM values.
An indicator of Compatibility to Generational Variety (CGV) is proposed and used in this approach to signify and assess the fitness of EMs and their current valuation to the ease-of-upgrade potential of products in the functional domain. It intends to reveal whether and how much the present metrical settings of interest are parametrically in tune with the performance requirements imposed on them in the upgraded model. By using CGV to represent and measure the fitness of a given engineering metric, it is capable of reflecting:

1. the suitability of this engineering metric for being a core metric,

2. the fitness of the current parametrical setting to serve the product platform for generational variety, and
3. the level of difficulty to configure this engineering metric as a core metric of product platform and its impact to the product's upgradability.

For an EM, the higher the value of its CGV, the less sensitive it might become in facing the changes of functions.

2.2 Indicator of Fitness to Extended Utilisation (FEU)

For remanufacture and upgrade, the "core" of a product consists of the valuable components that perform the enabling functions. In the given context, a crucial trait of those core components to support the ease-of-upgrade potential is their fitness to serve an extended use life. For components in the physical domain, this fitness represents the level of reusability, which is inherently constrained by the functional and the physical states. By studying the correlation between product upgradability and product reusability, it is logical to assert that being able to reuse is a necessary condition for a product's upgradability. After all, any upgrade option applied to a product is to enable it to serve longer with better and/or more functions. A product's reuse potential is defined by the states of its functional components.

In the PURE, an indicator of Fitness to Extended Utilisation (FEU) is designed to feature and measure the reusability of components and the product that they reside in. The factor of FEU can be defined as the integration of the effects of functional reusability (FRe) and physical reusability (PRe). For a component, its functional reusability (FRe) represents the technological fitness to service life extension, or in other words the potential of remaining functionally and technologically conform with the expectation of users after a period of use time.

By focusing on essentiality, the key factors contributing to components' FRe state are their technological maturity features, functional performance features, and design cycles. The PRe value of a component is featured as the reflection of its physical condition to serve an extended use life. After serving a certain period of time, this fitness represents the chance of performing the intended function satisfactorily for another unit of time under the specified conditions. Hence, PRe is inherently time-dependent and associated with the level of reliability (R) and the failure rate of components.

2.3 Indicator of Life-cycle Oriented Modularity (LOM)

Intrinsically, the architectural complexity of a product is influenced by states of the links or relationships among its components. Modularisation is an effective way of integrating manifold component relationships to simplify the structural configuration of products, facilitating multiple life-cycle processes and their virtues. A modular product serves a common interest of upgrade and remanufacture by facilitating the separation, swap and insertion of interested components. The modularity of a product is an indication of the degrees of intra-module correlations and inter-module dependence in its structural configuration. These two features are represented in the forms of Correspondence Ratio (CR) and Cluster Independence Index (CI) (Newcomb et al. 1998; Hata et al. 2001).

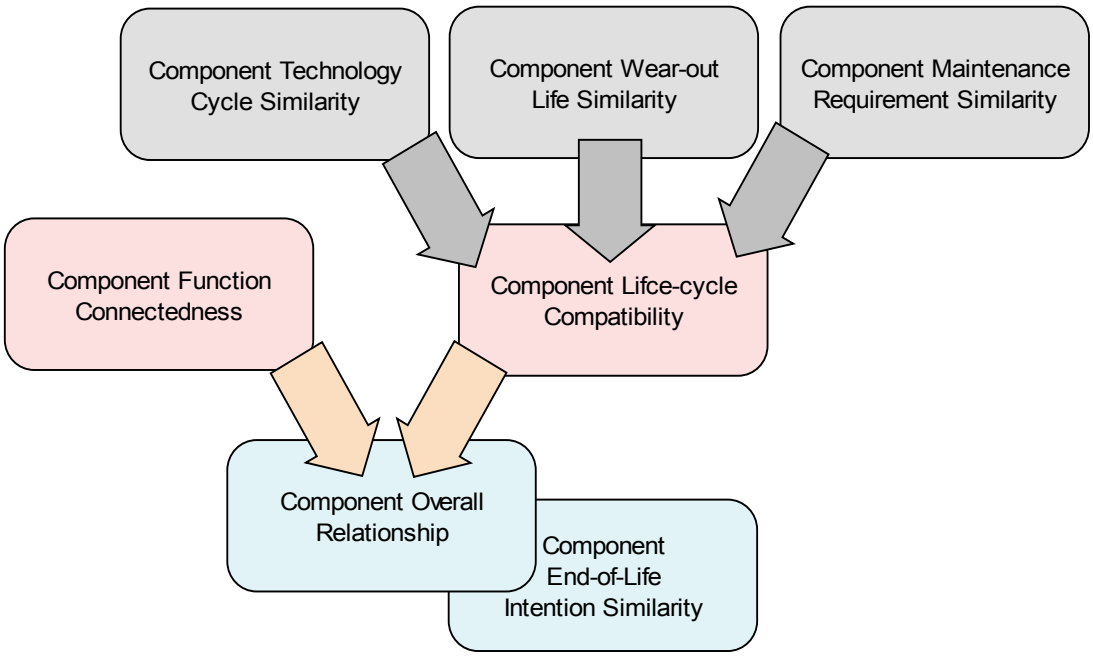


Figure 1. A Hierarchy of Component Relationships

Accordingly, the indicator of Life-Cycle Oriented Modularity (LOM) is introduced to provide a comprehensive vision and address the life cycle concerns that are influential to both upgrade and remanufacture in the measure of the two indexes and product modularity. CR measures the strength of connections

among the components within a module. Components having similar life cycle features (e.g. technological life, physical life, service requirements, etc.) and strong functional connectivity (in terms of material, signal, and energy flows) often tend to have similar end-of-life paths. Therefore, a set of component relationships (Figure 1) proposed by the authors (Xing 2003) is used to evaluate the CR value of each module formed in the product. For CI, the focal point of evaluation is at the physical links among the components of a module in comparison with the total links to the module, which suggests the potential complexity and amount of required effort involved in upgrade and remanufacturing operations.

3. Models and Evaluation Mechanism of PURE

Implemented in the remanufacture context, the approach of PURE is a formulated mathematical model that aims at assessing the potential of a product to be upgraded and reused contributing to its service life extension through remanufacturing operations.

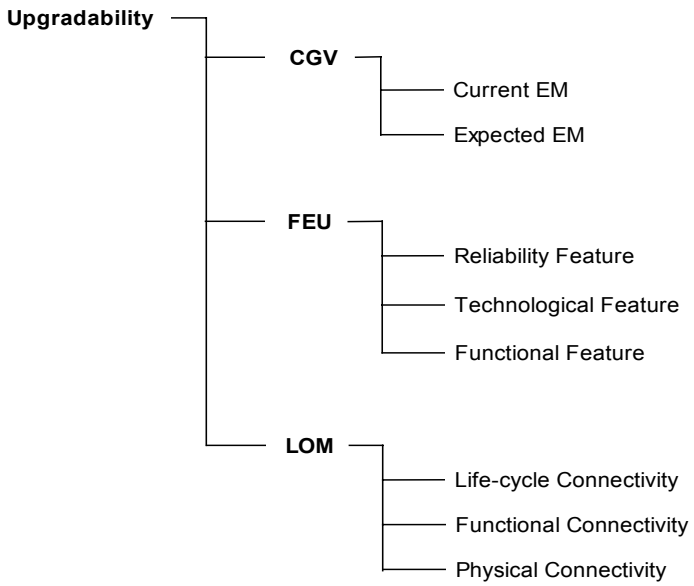


Figure 2. Scope of Product Upgradability and Its Indicators

Suggested by the name of the approach, the introduction of the featuring indicators of ease-of-upgrade characteristics and the modelling of product up-

gradability are conducted on the basis of the inherent connections and commonalities of product upgradability and product reusability.
In the last section, three featuring indicators for product upgradability are defined and discussed. Each of them is corresponding to a particular style of product representation and addresses one major characteristic pertinent to products' upgrade and reutilisation potential. Together, CGV, FEU and LOM contribute to the overall upgradability level of a product. The figure below (Figure 2) provides an overview of product upgradability, upgradability indicators and the associated factors.

3.1 Fuzzy Nature of Product Upgradability Representation and Evaluation

In this work, the product upgradability issue is studied in the context of remanufacture, which is featured by product life-cycle characteristics, and the PURE approach is to incorporate with product design for upgradability improvement. Therefore, the initiation of product upgradability evaluation is inevitably coupled with the inherent uncertainties residing in the collection and interpretation of product life-cycle information in the early phase of product development. Essentially, the sources of those uncertainties are the qualitative descriptions of pertinent product characteristics, the subjectivity of decision-making based on expertise and empirical experience, the lack of accurate design data and product life cycle information, and the design factors that are usually not within the control of designers (Wang *et al*., 1999; Xing *et al*., 2002).
The same situation exists for the identifications of upgrade opportunities and scenarios. The technological evolution characteristics of the product, the possible changes in customer expectations, and the compatibility of components to the changes are examined for the prediction of progressive changes in a product's functionality. Given the fact that in general those design information are expertise-based, qualitatively expressed and subjectively assessed, the phenomenon of fuzziness exists and it is largely a matter of degrees of belief regarding to what, when and how the functional changes or improvements will happen and their impacts to the product's function system. Furthermore, in the design of a product, the configuration of the product's structure, where modularity is usually a preferred feature to achieve, is based on examining the interactions among the components in abstract forms of semantic links, such as containment, alignment, affiliation, etc. (Ishii 1994; de Souza *et al*., 1996). Under the circumstances of remanufacture and upgrade, the implications of life cycle

characteristics in product upgradability considerations introduce extra dimensions into the relationships of components in addition to their physical links, geometrical interactions and functional connections. Linguistic values, such as very strong/high, strong/high, medium, weak/low and very weak/low, are usually applied to represent the levels of similarity, connectivity or compatibility among its components, which are very difficult to be depicted in a binary way. Intrinsically, these qualitatively classified (i.e. highly related or moderately related) and often subjectively valued (i.e. five-category score assignment) attributes exhibit the very characteristic of fuzziness and are of fuzzy concept (Li *et al*., 2000).

Fuzzy set theory is firstly introduced by Zadeh (1965) and fuzzy approaches have been effectively used to solve the vagueness or fuzziness in uncertain information, subjective decision-making and multi-attribute related problems. As a powerful mathematic tool, the advantage of applying fuzzy set theory in tackling product upgradability representation and evaluation problems, which have fuzzy features implicated, is significant and promising. To the design practices for upgradability associated with reuse or remanufacture, fuzzy set theory can have its major contributions in the aspects:

1. the assessment of the fitness of a product to remanufacture and functionality improvement,
2. component categorisation and clustering for efficient maintenance or replacement based on the similarity or compatibility of their characteristics, and
3. the evaluation of component life cycle interconnections in the forms of semantic linkage or design alternatives with regard to the given linguistic measures.

Therefore, two of the major elements of fuzzy set theory hereby present themselves as very useful tools for the above-mentioned tasks – *fuzzy membership values and functions* and *fuzzy relationships*. In PURE, they are used to formulate the mathematical models for the key characteristics of product upgradability.

3.2 Basic Notions for Product Upgradability Evaluation

Although the implication of cost concerns and the impacts of economic factors are critical in determining the viability of end-of-life options of products, the

emphasis on their roles could be easily taken with bias over the exploration on other essential life cycle factors and the technical solutions for their improvement. As upgradability is regarded as an intrinsic virtue of a product, the formulation of the PURE model in this work is based on the key technical factors that contribute to the ease-of-upgrade characteristics of a product.

Essentially, the suitability of a product to upgrade in the event of remanufacture is dependent of its type, mode of upgrade, and structural as well as physical features. In this work, the focus of interest is set on electr(on)ic products and electro-mechanical equipment which are most frequently treated in remanufacturing practice and generally more amiable to upgrade. Furthermore, the context of remanufacture in which product upgradability is concerned dictates that the functionality upgrade considered in this work are achieved through improving the hardware (e.g. adding or replacing parts, modules or subassemblies) rather than the software of a product. Considering the features and requirements of remanufacture and hardware upgrade, the product undergone the processes has to be structurally separable and physically serviceable. Such characteristics suggest that the product of interest should be a durable multi-component system and can be repaired once failures have occurred. Non-destructive separation of components is imperative to upgrade and remanufacture purposes. In accordance with the discussions above, following basic conditions are further assumed as the basis for the modelling and measure of product upgradability:

- *Assumption 1*: Products or systems considered in this work are reparable electro-mechanical products or systems working in normal conditions with standard usage, which exclude the impacts exerted by any external abnormalities, such as abuse and unforeseeable changes of the ambient environment,
- *Assumption 2*: Components examined in this work are the functional parts of a product or system. Connective parts, such as fasteners, cables, wires, harness, etc., are left out of the consideration to minimise distraction in that they normally have very long physical as well as technological life, but little functional importance to upgrade,
- *Assumption 3*: The fitness of an engineering metrics or a component in respect of any given upgrade option and reutilisation strategy is considered as independent of the conditions of other engineering metrics or components in the product, which is to simplify the complexity of the study, concentrate the examination on the essential factors of components, and model

upgradability truly on the basis of the inherent characteristics of each individual element of product representations.

- *Assumption 4*: A product is regarded as upgradable if its level of upgradability is higher than a threshold value that can be set by designers or companies. In this case, the qualitative concept of upgradability can be translated as a measurable indication based on expertise, facilitating the interpretation of the status of "being upgradable".

3.3 Formulation of PURE

Although a virtue of intrinsic characteristic, the upgradability of a product is in fact quite conditional and closely associated with the time constraint and the functional changes to apply. To ensure the success of the evaluation of a product's ease-of-upgrade potential, a prerequisite is to analyse the targeted product regarding to its functional and life cycle features. Considering technological, functional, physical, and time implications in upgrade and remanufacture, it is essential to facilitate the upgradability evaluation by identifying the following basic information:

1. The product function system, its key components and interconnections among the components,
2. The technical characteristics of the product and its components,
3. The time frame, or the planning horizon, for the upgrade and remanufacture considerations, and
4. The number and types of possible incremental changes or improvements to be applied to the functionality of the product within the defined time frame.

Based on the information above, the formulation of the PURE to assess product upgrade potential takes the steps to construct the mathematical models for the key indicators and the overall upgradability.

3.3.1 Modelling and Measure of CGV

The generic mappings among the domains of customer requirements, engineering metrics, and components are depicted in the form as Figure 3 and facilitated by implementing QFD. Therefore, a product system can be represented a set $\Upsilon(r_1, r_2, \ldots, r_m)$, where r_i is an engineering metric of the functions of

the product and m (≥1) stands for the number of engineering metrics.

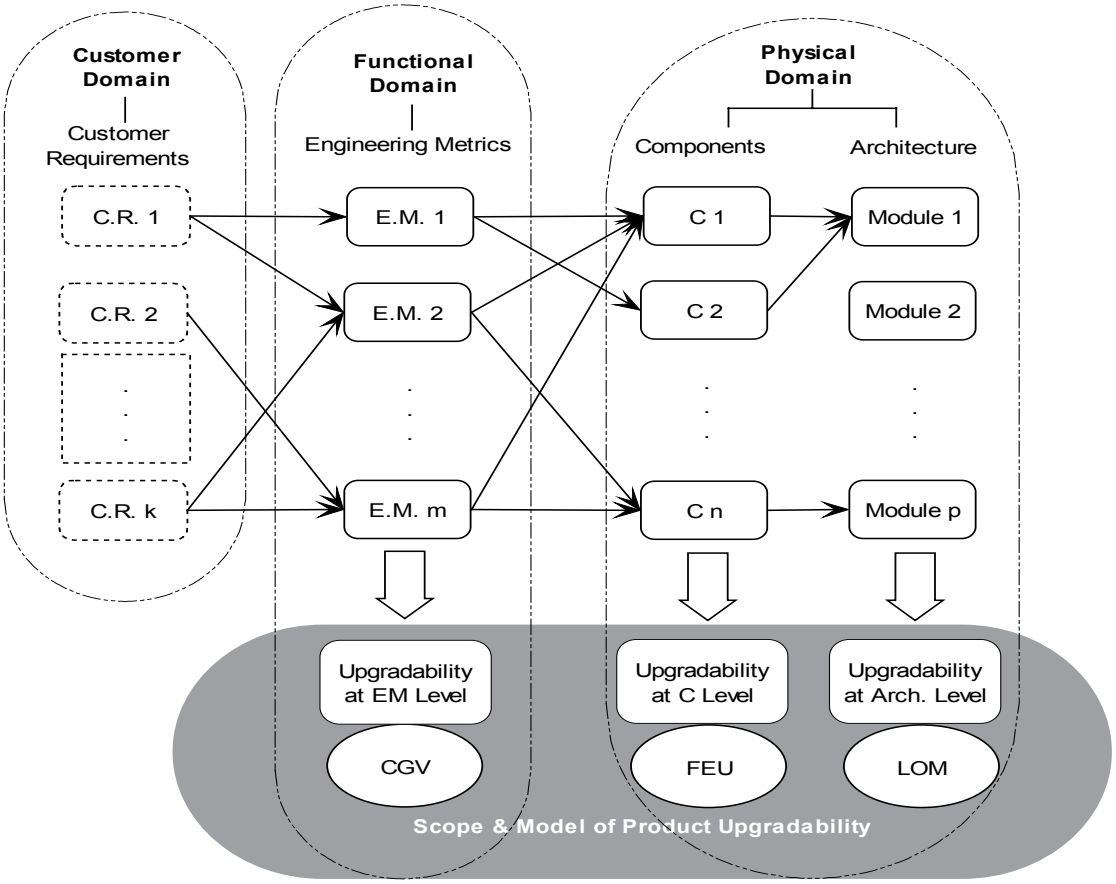


Figure 3. Mappings among the Design Domains for Product Upgradability Representation

For each engineering metric r_i (i = 1,2, …, m), it is regarded a function of a number of corresponding components composing a subset of $\boldsymbol{C}(c_1, c_2, \ldots, c_n)$ (n≥1), which is expressed as

$$r_i = g(\{c\}), \quad \{c\} \subseteq C \tag{1}$$

With the introduction of generational changes, often a new performance requirement on any an existing engineering metric r_i ($r_i \in \Upsilon$) is expected. Exhibiting the same feature, this engineering metric of the new generation product is denoted as r_i^e. Assuming that r_i^e is always not worse than r_i in terms of functionality, the state of the current r_i to the change, CGV_i, is measured as a mem-

bership value of fitness by the following fuzzy membership function and illustrated as Figure 3. The coefficient τ is a vector of the standard gradient of improvement for a given EM. A positive τ designates the improvement of "increasing" the current value, while a negative τ stands for the opposite direction. The coefficient κ valued as 0.1, 0.3, 0.5, 0.7, or 0.9 represents the ascending level of difficulty or significance of the improvement.

$$CGV_i = \exp\left[-\kappa\left(\frac{\gamma_i^e - \gamma_i}{\tau}\right)\right] \qquad (2)$$

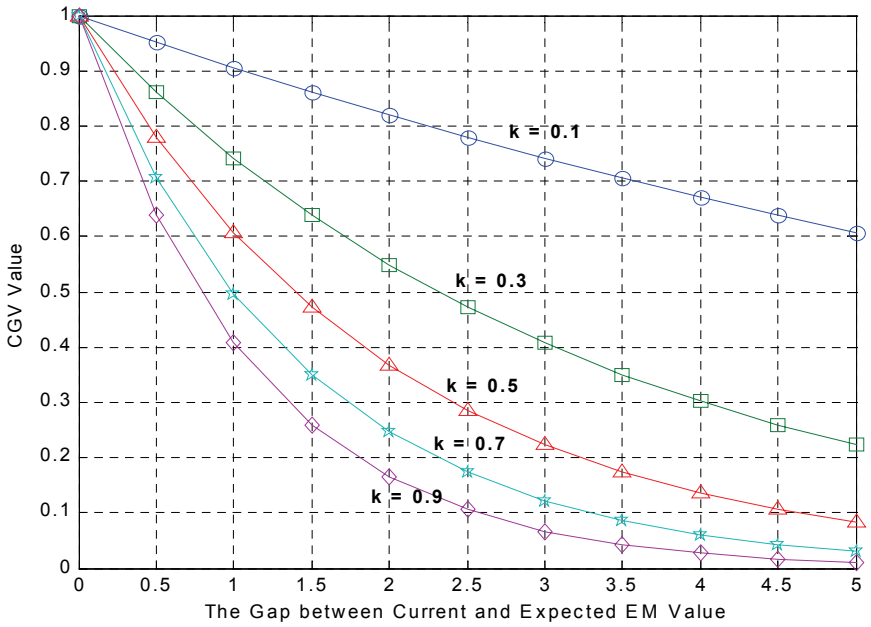


Figure 4. Function of CGV (τ= 2)

Given the different importance of engineering metrics to the functionality of a product, the overall CGV at the system level, CGV_{sys}, is modelled as the aggregation of the weighted CGV of all the *r*s of $\boldsymbol{\Upsilon}$. The assignment of weight to engineering metrics is conducted on the basis of function analysis and their correspondence to the fulfilment of the identified customer requirements. Metrics corresponding to functions that address important customer requirements should be given higher values of weight for their impacts on the shape-up of CGV_{sys}.

$$CGV_{sys} = \sum_{i=1}^{m} w_i CGV_i \ \ where \ \ w_i \ \ is \ the \ weight \ \ of \ \ r_i \ \ and \ \ \sum w_i = 1 \qquad (3)$$

According to the ways that CGV_i and CGV_{sys} are measured, the more engineering metrics of a product affected by functional changes during upgrade, the lower the level of CGV_{sys} will become and the more difficult for the product to be upgraded. If a product is highly specified or compactly designed, any changes introduced to the functionality of the product by an upgrade plan could affect the whole system of engineering metrics. In such circumstances, many of the engineering metrics would have the "expected values" identified as being significantly different from their current parametrical settings. It suggests that a great deal of effort might be needed to bridge such "gaps" between the two sets of values. Consequently, the CGV value of the product and its potential to upgrade will become much lower than the designs that are less specified or compact.

3.3.2 Modelling and Measure of FEU

In the event of remanufacture and upgrade, the suitability of a component to reuse is measured as functional reusability (FRe) and physical reusability (PRe). For any component c_i ($c_i \in \boldsymbol{C}$, i = 1, 2, …, n), FRe_i and PRe_i are described by the grades of their membership values to the states of being "desirable" and "reliable" after serving a certain period of time t under the specified conditions. Regarding to the status of "being reusable", the membership degree quantifying the PRe level of a component is computed against the specified boundary values of reliability, usually standing for the minimum expected and the ideal reliability state. If an exponential feature and a constant instantaneous failure rate are assumed, for any a component c_i its $PRe_i(t)$ can be expressed as Eq.4 where λ_i is the instantaneous failure rate and R_{min} is the minimum expected reliability.

$$PRe_i(t) = \frac{\max\left[0, (e^{-\lambda_i t} - R_{\min})\right]}{1 - R_{\min}} \qquad (4)$$

A practical way to define the R^{min} value for each component is through reliability allocation at the component level of product representation. Usually, the minimum reliability is much more obvious and easier to specify at the system

level. After setting the $R^{\min}$ of the system, the corresponding minimum acceptable reliability of each subsystem or component in the system can be identified through the analysis of the hierarchy of the system representation and the reliability block diagram which demonstrates the configuration of the product function system. A methodical process of top-down allocation of reliability through the systems hierarchy is detailed in the ref. (Blanchard et al. 1995).

As component desirability is a very subjective and fuzzy concept, it is difficult to have a direct quantitative measure of desirability level and FRe. Nevertheless, given the associations between the length of technology cycle (TC), the maturity of technology, and the suitability for reutilisation, to use the membership degree of TC to the concept of "being long" to represent the status of "being desirable" becomes a quite reasonable option. It is suggested that the maximum TC of "New Economy/I.T. " products is 5.5 years (Rose 2000). Based on this, 5.5-year life is used as the empirical rough pre-evaluation standard for the fast classification such as "rather short" or "rather long". A Sigmoid-Shape membership function is adopted for the evaluation. Expressed as the probability of not being obsolescent by time t and assumed to exhibit an exponential feature, the $FRe_i(t)$ of component c_i is defined as Eq.5

$$FRe_i(t) = \frac{1}{1 + \exp\left[-0.5 * \left(ETL_i(t) - 5.5\right)\right]}, \quad i = 1, 2, \ldots, n \qquad (5)$$

ETL_i is the effective technology life of component c_i. For a component, its ETL is equivalent to its theoretical technology life – TC when the component is unrivalled or at the start of its service ($t \leq$ DC). ETL degrades along with time when more new designs are introduced ($t >$ DC).

Linguistic Values	Scores
Very High	0.90 ~ 1.00
High	0.70 ~ 0.89
Medium	0.50 ~ 0.69
Low	0.30 ~ 0.49
Very Low	0.00 ~ 0.29

Table 1. Measure and value range of FL

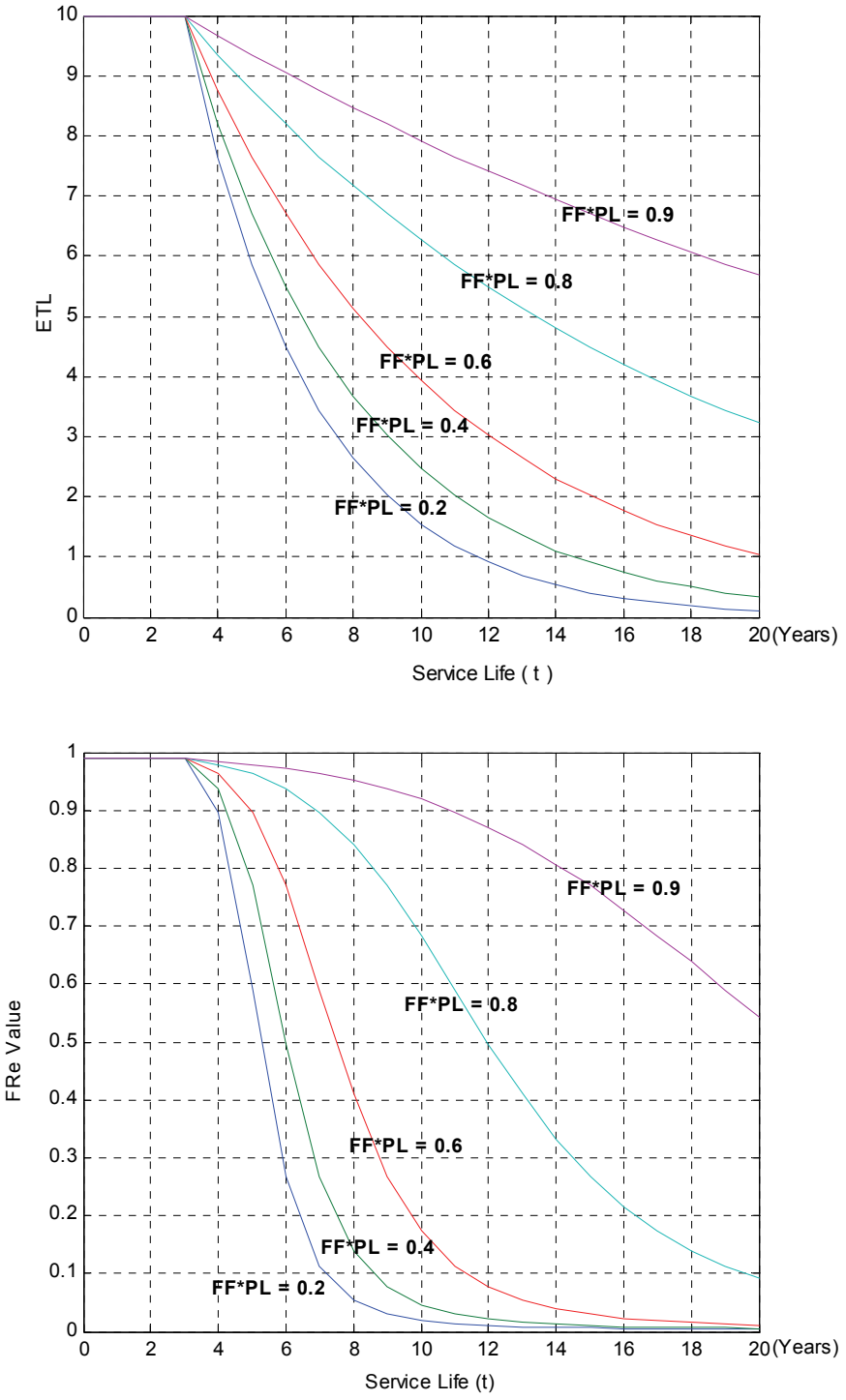
Given the fact that components using the same technology have different performance levels and usually the one with better performance serves longer, the pace of degradation of a component's ETL is dictated by its functionality level (FL) which denotes the quality to function in comparison with the industry benchmarks or the best design in the market (Table 1).
Consequently, ETL can be considered as a function of TC, DC, FL, and the time factor t (Figure 5). By assuming an exponential mode for the degradation of ETL, for any component c_i, the function of its ETL is expressed as Eq.6.

$$ETL_i(t) = TC_i \exp\left[-(1 - FL_i) * \max\left[0, \left(\frac{t}{DC_i} - 1\right)\right]\right], \; i = 1, 2, , n \qquad (6)$$

The formation of FEU model is based on the integration of the model of PRe and the model of FRe. In the circumstances of remanufacture, the "shorter plank" of FRe and PRe plays a crucial role in determining the fitness of components to serve. While, the other one provides a strengthening effect which could help to enhance the chance of reutilisation, but will not dramatically improve component reusability. Therefore, at the moment of t the FEU of component c_i (i = 1, 2, …, n) is measured by Eq.7 and the reusability of the entire system is represented in the form of Eq.8. Apparently, the more important a component is, the greater the impact that it exerts on the FEU value of the product, which is in accordance with the notion of the role of key components. As assumed, only key and auxiliary components are considered in the product representation and the modelling. The contributions of the components to the ultimate reusability at the system level, FEU_{sys}, are directly related with their importance to the functionality of the product, which can be identified through the mappings among customer requirements, engineering metrics and components.

$$FEU_i = \min\left[FRe, PRe_i\right]^{\left(1 - \frac{\max\left[FRe_i\, PRe_i\right]}{2}\right)}, \; i = 1, 2, , n \qquad (7)$$

$$FEU_{sys} = \sum_{i=1}^{n} w_i FEU_i, \quad \sum_{i=1}^{n} w_i = 1 \qquad (8)$$



ETL-function FRe-function

Figure 5. Functions of ETL and FRe (TC=10; DC= 3)

3.3.3 Modelling and Measure of LOM

As described in previous sections, the modularity of a product is assessed Correspondence Ration (CR) and Cluster Index (CI). If the product consists of M modules, for each individual module its CR level is determined by the intensity of the interconnections among the constituent components, and its CI level is an indication of physical independence from the other modules. The states of components' interconnections are determined by their functionality connectedness and similarities in technological life, physical life, and service requirements, which are intrinsically of fuzzy nature. The overall relationship among the components in the same module, denoted as $\tilde{R}$, is measured as the basis for the evaluation of CR level of the module. The evaluations of those fuzzy component relationships in the context of remanufacture are elaborated in detail in the ref. (Xing et al. 2003). For any module m_i (i = 1, 2, …, M) with p components, the indicator of CR_i is calculated as:

$$CR_i = \begin{cases} 1 & , \quad if \quad p = 1 \\ \dfrac{\sum_{k=1}\sum_{l=1} \tilde{R}(k,l)}{p(p-1)/2} & , \quad otherwise \end{cases} \tag{9}$$

The level of CI level is very critical to the level of difficulty and the operational effort of product reprocessing in the context of remanufacture. Remanufacture is featured by a complete disassembly of a large quantity of products. The investigation on the empirical data from disassembled products with more than one module shows that the interaction metric of a module becomes high or very high when its inter-module physical links are approximately 3 or greater (Allen & Calson-Skalak, 1998). Therefore, the value of 3 is used as a reference value to measure the membership of the inter-module links of a module to the concept of "having high interaction metric", featured by Eq.10, where the value of -2 is adopted to ensure that the interaction metric is close to zero when inter-module links approach zero. By incorporating the consideration of module interaction metric into the measure of module independence, the CI value of module m_i is calculated by using Eq.11.

$$\alpha_i = \frac{1}{1 + \exp\left[-2 * \left(LINKS_i^{Inter-module} - 3\right)\right]}, \quad i = 1, 2, \ldots, M \tag{10}$$

$$CI_i = 1 - \left(\frac{LINKS_i^{Inter-module}}{LINKS_i^{Total}} \right) * \alpha_i \qquad (11)$$

According the equations above, by reducing the constituent components in each module, the number of modules in the product is increased and in turn the CR values of individual modules could be improved. However, a most probable adverse effect resulted from such change would be the decrease of the number of intra-module physical links in the total number of links and thus a lower CI value for each module. Upgrade through remanufacture is often conducted in a mass production scale. All the modules are disconnected from each other and treated before being recombined with new components. The CI status of the product directly reflects the fitness of its structural configuration to meet the requirements of remanufacture. Having less modules and weak inter-module links is highly important to the interest of product upgradability in this particular context. Relatively, CR is less important in such circumstances. The increase of module quantity associated with a larger CR is not in favour of remanufacturing operations. Consequently, LOM is formulated as Eq.12, where 0.5 is the theoretical maximum weight of importance.

$$LOM = \left(\frac{\sum_{i=1}^{M} CI_i}{M} \right)^{\left(1 - 0.5 * \frac{\sum_{i=1}^{M} CR_i}{M} \right)} \qquad (12)$$

3.3.4 Modelling and Measure of Product Upgradability

After modelling the three characteristics, the upgradability index of a product (PUI) is measured as a function of CGV, FEU and LOM. As a tripod stool that is usually crippled by its shortest leg, the overall upgradability of a product supported by three ease-of-upgrade characteristics is to a large extent determined by the "weakest link" of them. Eq.12 below presents the formulation of the mathematical expression of PUI.

$$PUI = \min\left[CGV_{sys},\ FEU_{sys},\ LOM_{sys}\right]^{\left[1-\frac{h}{3}\right]} \qquad (12)$$
$$\left(CGV_{sys} + FEU_{sys} + LOM_{sys}\right) - \min\left[CGV_{sys},\ FEU_{sys},\ LOM_{sys}\right]$$

The success of the PURE relies on its ability to identify the inherent fitness of a product to the given upgrade scenario, consisting of a plan for functional improvement and the timeframe (planning horizon) for the implementation of intended upgrade. The upgradability evaluation results from the PURE model should be able to provide users informative indications for any redesign or reconfiguration of the product for ease-of-upgrade features of to base on.

4. Case Study

In this case example, the application of PURE on a Roof-Integrated Solar Air Heating System (RISAHS) is analysed. This technology is being developed by the Sustainable Energy Centre at the University of South Australia. The RISAHS utilises solar energy to provide space heating for a building. As with traditional solar air heating systems a *collector* is required to absorb the solar energy and is used to heat air. This air is distributed throughout the building via a *fan* and *ducting*.

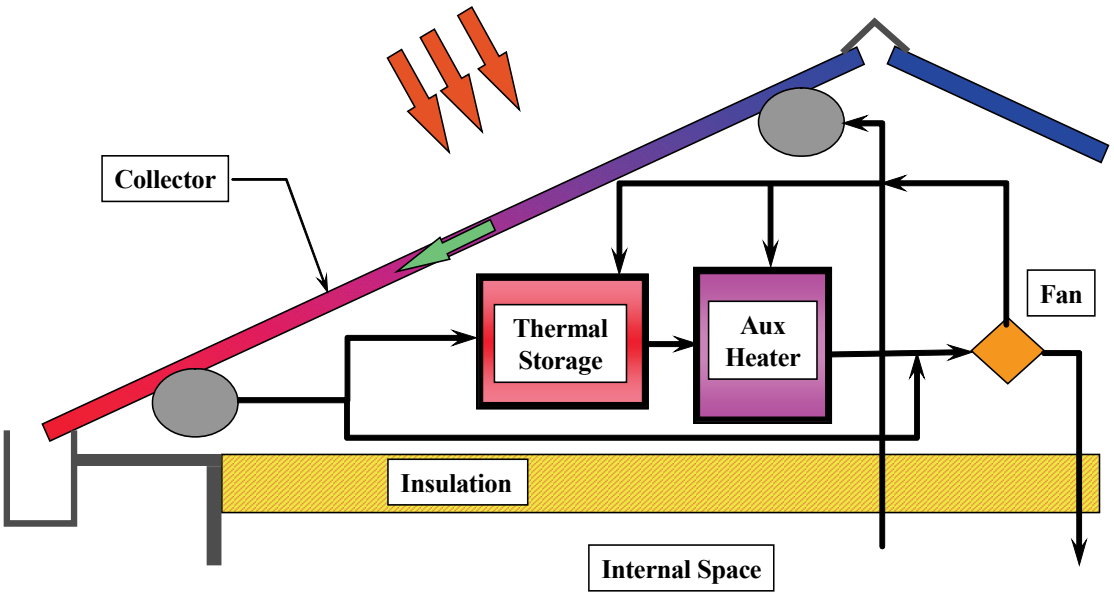


Figure 6. A Schematic View of Roof Integrated Solar Air Heating System

The collector can provide heating when there is a high level of sunshine. A *thermal storage unit* (TSU), which is charged by the collector, is used to store heat for times when there is inadequate levels of sunshine, and is charged by the collector. For times when the storage facility is empty, an *auxiliary backup heater* is used (Belusko et al. 2004). A *control panel* with a control system is integrated into the system to control the different heating operations of the system. The schematic representation of the RISAHS is shown as Figure 6.

The upgrade plan applicable to the system is assumed in the context of remanufacture. Table 2 presents the basic information of the engineering metrics and their current as well as expected values.

Code	Engineering Metrics	Unit	Weight	Current Value	Expected Value	τ	κ
EM1	Average temperature	deg.C	0.31	2	0	-1.75	0.3
EM2	Volume of outside air	air changes/day	0.08	3.75	15	7.5	0.1
EM3	System min. heating capacity/heating load	kw/kw	0.17	0.75	0.9	0.2	0.3
EM4	Max. room temp. - Min. room temp.	deg.C	0.06	2	1	-1	0.1
EM5	Max air velocity at head height	m/s	0.08	0.5	0.25	-0.325	0.1
EM6	Solar heating/total heating	MJ/MJ	0.10	0.45	0.7	0.2	0.5
EM7	The conventional energy use	MJ/m2	0.10	20.6	13.7	-4.3	0.5
EM8	Amount of CO2 produced	kg/m2	0.10	9.8	0.1	-2.3	0.5

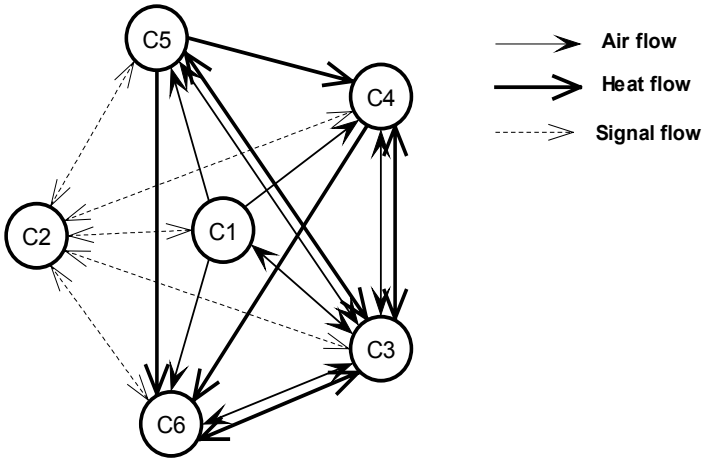
Table 2. Engineering metrics information of the RISAHS

As described above, there are six major functional components in this RISAHS, having different operation time per annum. In this case study, we assume that 1) the planning horizon for the upgrade consideration is 5 years, and the minimum acceptable level of reliability for the components in the system at any time is 0.3. The information about the importance rankings, technological life, physical life and reliability feature of the components are listed in Table 3.

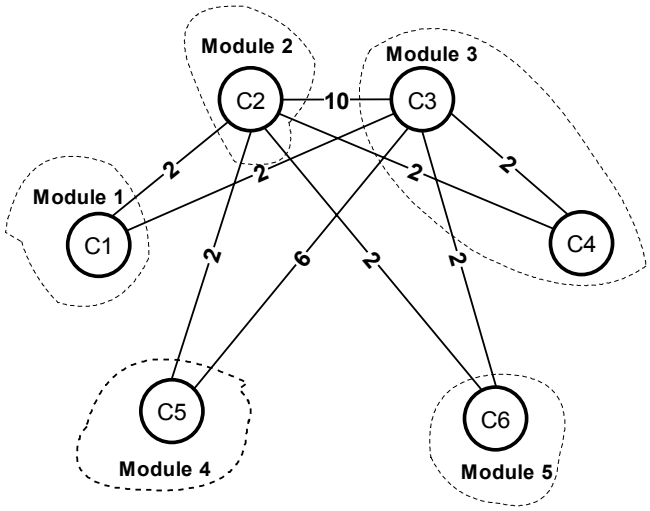
Code	Component	Weight	Technology Life (yr)	Design Cycle (yr)	Functionality Level	Physical Life	Failure Rate	Working Hours/yr
C1	Fan	0.1	13	5	0.65	15	0.00001	2361
C2	Controls	0.1	3	2	0.55	15	0.000001	2361
C3	Ducting	0.05	15	7	0.75	15	0.00002	2361
C4	TSU	0.25	8	5	0.35	15	0.00001	1458
C5	Collector	0.35	8	5	0.4	15	0.00001	1180
C6	Aux. Heater	0.15	10	5	0.8	15	0.00001	694

Table 3. Component information of the RISAHS

The structural configuration of RISAHS adopted in the current design is a dispersed arrangement of the components. The system consists of six components which are regarded as one-component modules. The components are connected with each other by their functional as well as physical interactions. The function flows between the components in RISAHS are demonstrated as Figure 7-(a), while the physical links among them are depicted by a graph in Figure 7-(b). The value assigned to each edge of the graph stands for the number of links existing between two components.



(a) Function Flows among the Components



(b) Physical Links among the Components

Figure 7. Interactions among RISAHS Components

To assess the modularity of the system, Figure 8 illustrates the number of physical links, the level of similarity in service requirements, and the level of functional connectedness among the six modules. Again, the equal significance is considered for the effects of the ratio of intra-module physical links and the ratio of intra-module component life cycle correspondence.

On the basis of the information provided above, the CGV, FEU and LOM values are calculated for the engineering metrics, the components and the entire system. The results of the evaluation are presented in Table 4. Based on these technical characteristics, the result shows that the system has a *Medium*

LINKS	C1	C2	C3	C4	C5	C6
C1	0	2	2	0	0	0
C2	2	0	10	2	2	2
C3	2	10	0	2	6	2
C4	0	2	2	0	0	0
C5	0	2	6	0	0	0
C6	0	2	2	0	0	0

SRS	C1	C2	C3	C4	C5	C6
C1	1	0.3	0.3	0.8	0.8	0.9
C2	0.3	1	0.1	0.3	0.3	0.4
C3	0.3	0.1	1	0.3	0.3	0.2
C4	0.8	0.3	0.3	1	0.9	0.7
C5	0.8	0.3	0.3	0.9	1	0.7
C6	0.9	0.4	0.2	0.7	0.7	1

FC	C1	C2	C3	C4	C5	C6
C1	1	0.9	0.9	0.8	0.8	0.8
C2	0.9	1	0.5	0.4	0.5	0.5
C3	0.9	0.5	1	0.3	0.2	0.3
C4	0.8	0.4	0.3	1	0.8	0
C5	0.8	0.5	0.2	0.8	1	0
C6	0.8	0.5	0.3	0	0	1

Figure 8. Physical Links, Service Requirement Similarity and Functional Connectedness among RISAHS Components

Code	Weight	CGVi	CGVsys	PUI				
EM1	0.31	0.77	0.71	0.51				
EM2	0.08	0.86						
EM3	0.17	0.80						
EM4	0.06	0.91						
EM5	0.08	0.93						
EM6	0.10	0.54						
EM7	0.10	0.45						
EM8	0.10	0.45						
Code	**Weight**	**FRei**	**PRei**	**FEUi**	**FEUsys**	**CRi**	**CIi**	**LOM**
C1	0.10	0.91	0.84	0.91	0.83	1.00	0.12	0.25
C2	0.10	0.12	0.98	0.34		1.00	0	
C3	0.05	0.99	0.69	0.84		1.00	0	
C4	0.25	0.78	0.90	0.87		1.00	0.12	
C5	0.35	0.78	0.92	0.87		1.00	0	
C6	0.15	0.91	0.95	0.95		1.00	0.12	

Table 4. Upgradability evaluation results for the RISAHS

level of upgradability in the context of remanufacture after 5 years into its service life. It has a very good reusability due to the high reliability and technological maturity of its constituents. However, the low modularity of the system's structure coupled with a large number of inter-module physical links exhibits the downside of its configurations, suggesting that it might need a great amount of effort and/or time to disconnect those components during upgrade or reprocessing.

As the "weakest link", the structural features of the system need to be changed, if technically possible, in favour of upgradability improvement. On the other hand, in this evaluation approach only the number of links is considered to represent the complexity of the structural configuration of product or system. All types of physical links among components are treated equally regardless of their nature. Incorporating the severity ranking of each type of links with the number of links could be more informative in reflecting their impact on product upgradability. This issue will be considered in the refinement of the PURE approach.

5. Conclusion

This paper highlights the concept of product upgradability and its importance to the success of remanufacture. A new approach, the PURE (Product Upgradability and Reusability Evaluator), is proposed to model and assess the upgradability of a product in the context of remanufacture. By focusing on the essential technical characteristics, the upgrade potential of a product is measured at three domains of product representation, namely the engineering metrics domain, the component domain and the structural domain. Correspondingly, the indicators of compatibility to generational variety (CGV), fitness to extended utilisation (FEU), and life-cycle-oriented modularity (LOM) are proposed for the upgradability evaluation purpose. A simple example and a case study on a solar air heating system presented in this paper demonstrate that the results provided by the PURE are quite in line with common engineering knowledge about the technical features of product upgradability. Furthermore, the three indicators and their coefficients are able to provide companies good information about the readiness of the product to the given scenarios of upgrade at various levels (engineering metrics, component, and structure). Those values can be used for decision making in the redesign of the product, pointing out the aspects where improvements are needed. Nevertheless, a major

drawback of the PURE model at this stage is that cost factors are not directly included in the modelling and evaluation of upgradability. Actually, one of the basic ideas for the development of this approach at this stage is to deliberately ignore the externality of cost and market issues and only focus on the essential roles of technical characteristics. The lack of economic perspective is acknowledged as one of the limitations of the current version of the approach. But, the development of this upgradability evaluation approach is just the first step of an ongoing research. In the next step, an approach for design optimisation of upgradability will be developed on the basis of PURE. Cost issues will be well considered by then and incorporated, as a major constituent, into the objective function. The other issues that are outside the scope of this paper include how to develop upgrade plans, how to identify planning horizon, and how to assess the possibility-to-change of components. These problems will be researched by our future work, together with the prediction of function/technology changes when a more comprehensive design framework is developed.

At this stage, the PURE is just a general framework for the representation and measure of product upgradability in the technical sense. Further refinement of the formulations of the three evaluation indicators or the inclusion of new indicators can be accommodated by the current structure of the PURE model to adapt to any particular products or upgrade scenarios.

6. Reference

Allen, K. R. and S. Carlson-Skalak (1998). Defining Product Architecture During Conceptual Design. Proceedings of DETC98: 1998 ASME Design Engineering Technical Conference, Atlanta, GA.

Beck, R. C. and S. E. Gros (1997). 'Remanufactured' Power Generation Equipement - An Economical Solution for Self-generation, Process or Peaking Applications. International Exhibition & Conference for the Power Generation Industries - Power-Gen, 1997.

Belusko, M., W. Saman, et al. (2004). "Roof Integrated Solar Heating System with Glazed Collector." Solar Energy 76(1-3): 61-69.

Blanchard, B. S., D. Verma, et al. (1995). Maintainability: A Key to Effective Serviceability and Maintenance Management. New York, Wiley & Sons.

De Souza, R. B., K. Ab, et al. (1996). "An Integrated Decision-Support Environment for DFX Audit." Journal of Electronics Manufacturing Vol. 6 , No. 3: 163-171.

Droda, T. J. (1984). "Machine Tool Remanufacturing." Manufacturing Engineering 92(2): 88-98.

Hata, T., S. Kato, et al. (2001). Design of Product Modularity for Life Cycle Management. Proceedings of 2nd International Symposium On Environmentally Conscious Design and Inverse Manufacturing, 2001, Eco'01.

Huang, G. Q. (1996). Design for X: Concurrent Engineering Imperatives. London, Chapman & Hall.

Ijomah, W., J. P. Bennett, et al. (1999). Remanufacturing Evidence of Environmental Conscious Business Practice in UK. EcoDesign 99': First International Symposium on Environmentally Conscious Design and Inverse Manufacturing, 1999., Tokyo.

Ishii, K., C. F. Eubanks, et al. (1994). "Design for Product Retirement and Material Life-Cycle." Materials & Design **15(4)**: 225-233.

Ishigami, Y., H. Yagi, et al. (2003). Development of A Design Methodology for Upgradability Involving Changes of Functions. Proceedings of the 3rd International Symposium on Environmentally Conscious Design and Inverse Manufacturing, EcoDesign '03.

Li, F.-Y., G. Liu, et al. (2000). "A Study on Fuzzy AHP Method in Green Modularity Design." China Mechanical Engineering (In Chinese) **11(9)**.

Li, J.-z., P. Shrivastava, et al. (2004). A Distributed Design Methodology for Extensible Product Life Cycle Strategy. Proceedings of IEEE International Symposium on Electrionics and the Environment 2004.

Newcomb, N. J., B. Bras, et al. (1998). "Implications of Modularity on Product Design for the Life Cycle." Journal of Mechanical Design vol. 120: pp 483-490.

Overfield, J. A. (1979). Maintenance and Remanufacturing of Injection Molding Machines. Chicago, National Bureau of Standards, Inject Molding Division and Chicago Section: X.1-X.7.

Pridmore, J., G. Bunchanan, et al. (1997). "Model-Year Architectures for Rapid Prototyping." Journal of VLSI Signal Processing 15: 83-96.

RMIT (2001). EcoRedesign Guidlines,
http://www.cfd.rmit.edu.au/outcomes/erdnews/ERD1/ERDguide.html (29/06/2001 & 04/07/2001).

Rose, C. M. (2000). Design for Environment: A Method for Formulating Product End-of-Life Strategies. Dept. of Mechanical Engineering. Stanford, Stanford University: pp 185.

Shimomura, Y., Y. Umeda, et al. (1999). A Proposal of Upgradable Design. Proceedings of First International Symposium On Environmentally Conscious Design and Inverse Manufacturing, EcoDesign '99.

Umemori, Y., S. Kondoh, et al. (2001). Design for Upgradable Products Considering Future Uncertainty. Proceedings of EcoDesign 2001: Second International Symposium onEnvironmentally Conscious Design and Inverse Manufacturing, 2001.

Wang, J., G. Duan, et al. (1999). "Current Situation and Future Developing Trends of Green Manufacturing Technology Based on Products'Life Cycle." Computer Integrated Manufacturing System, CIMS (In Chinese **5(4)**: 1-8.

Well, R. (2004). Make it New. Aviation Week and Space Technology. 161: 56-57.

Wilhelm, W. E., P. Damodaran, et al. (2003). "Prescribing the Content and Timing of Product Upgrades." IIE Transactions 35(7): 647-663.

Xing, K., L. Luong, et al. (2002). The Development of A Quantitative Method for Product End-of-Life Strategy (EOLS) Planning. Proceedings of 4th International Symposium on Tools and Methods of Competitive Engineering (TMCE) 2002.

Xing, K. (2003). Development of An Integrated Framework for Design for Recyclability. School of Adv. Manufacturing & Mechanical Engineering. Adelaide, SA, University of South Australia: 239.

Xing, K., B. Motevallian, et al. (2003). A Fuzzy-graph based Product Structure Modularisation Approach for Ease-of-recycling Configuration. Proceedings of 9th International Conference on Manufacturing Excellence, ICME 2003.

Zadeh, L. A. (1965). "Fuzzy Sets." Information and Control 8: 338-353.

12

Distributed Architecture for Intelligent Robotic Assembly Part I: Design and Multimodal Learning

Ismael Lopez-Juarez and Reyes Rios-Cabrera

1. Introduction

1.1 General Description

In this chapter we describe the general framework design for a distributed architecture to integrate multiple sensorial capabilities within an intelligent manufacturing system for assembly. We will formally define the model of the M2ARTMAP multimodal architecture. We present some simulated results of the model using a public domain multimodal data base. Initial findings have indicated the suitability to employ the M2ARTMAP model in intelligent systems when three or less modalities are involved. Taking into account these results and the M2ARTMAP's modularity it was decided to integrate this model using the CORBA (Component Object Request Broker Architecture) middleware to develop robotic assembly tasks that includes contact force sensing, an invariant object recognition system and natural language processing. This chapter introduces the overall system in the intelligent cell and the task planner (SIEM) and the object recognition system (SIRIO) are described in detail in part II and Part III.

Design and experiments of the distributed architecture using a Local Area network (LAN) and an industrial robot KUKA KR-15 to fusion different modalities in a common task are first described. The modalities are: vision, contact force sensing (tactile), and natural language processing using context free grammars. The vision modality and force sensing are implemented based on a FuzzyARTMAP neural network and a main coordinator. The testbed for the application and a general distributed environment using CORBA as a middleware is described. Later several learning simulations using M2ARTMAP, are presented (Lopez-Juarez & Ordaz-Hernandez, 2005). The main distributed objects are described in detail, the experimentation is presented and the results analyzed. Finally, future work is proposed.

1.2 The architecture

Robots in unstructured environments have to be adaptable to carry out operation within manufacturing systems. Robots have to deal with its environment, using available sensors and their adaptability will depend on how flexible they are (Wu et al., 1999). To create that system it is necessary to integrate different techniques of artificial intelligence, on-line learning, sensorial capabilities and distributed systems.

This work stands on and improves the design of intelligent agents for assembly (Lopez-Juarez et al., 2005a) by integrating the fusion of different modalities using a distributed system based on CORBA. In the chapter it is designed a distributed architecture where different sensorial modalities, operating systems, middleware and programming languages are integrated to perform mechanical assembly by robots.

A task coordinator referred as the SICT (*Sistema Inteligente Coordinador de Tareas, in Spanish*), which plans the assembly and general coordination with the vision and system and the assembly system was designed. The SICT is described, it was built using several operating systems (Linux and windows), two ORB's (Object Request Broker) that is ORBit and omniORB and the programming languages are C and C++. The communication model includes the schema client-server in each module of the system.

1.3 Description of the Manufacturing System

The manufacturing system used for experimentation is integrated by a KUKA KR15/2 industrial robot. It also comprises a visual servo system with a ceiling mounted camera as shown in figure 1.

The main operation of the manufacturing system is about the peg-in-hole insertion where, there robot grasps a male component from a conveyor belt and performs the assembly task on a working table where the fixed female component is located. A main coordinator starts the assembly cycle using the vision system that obtains an image from the male component and calculates the object's pose estimation, later it sends information to the coordinator from two defined zones:

Zone 1 which is located on the conveyor belt. The vision system searches for the male component and determines the pose information needed by the robot.

Zone 2 is located on the working table. Once the vision system locates the female component, it sends the information to the coordinator which executes the assembly with the available information.

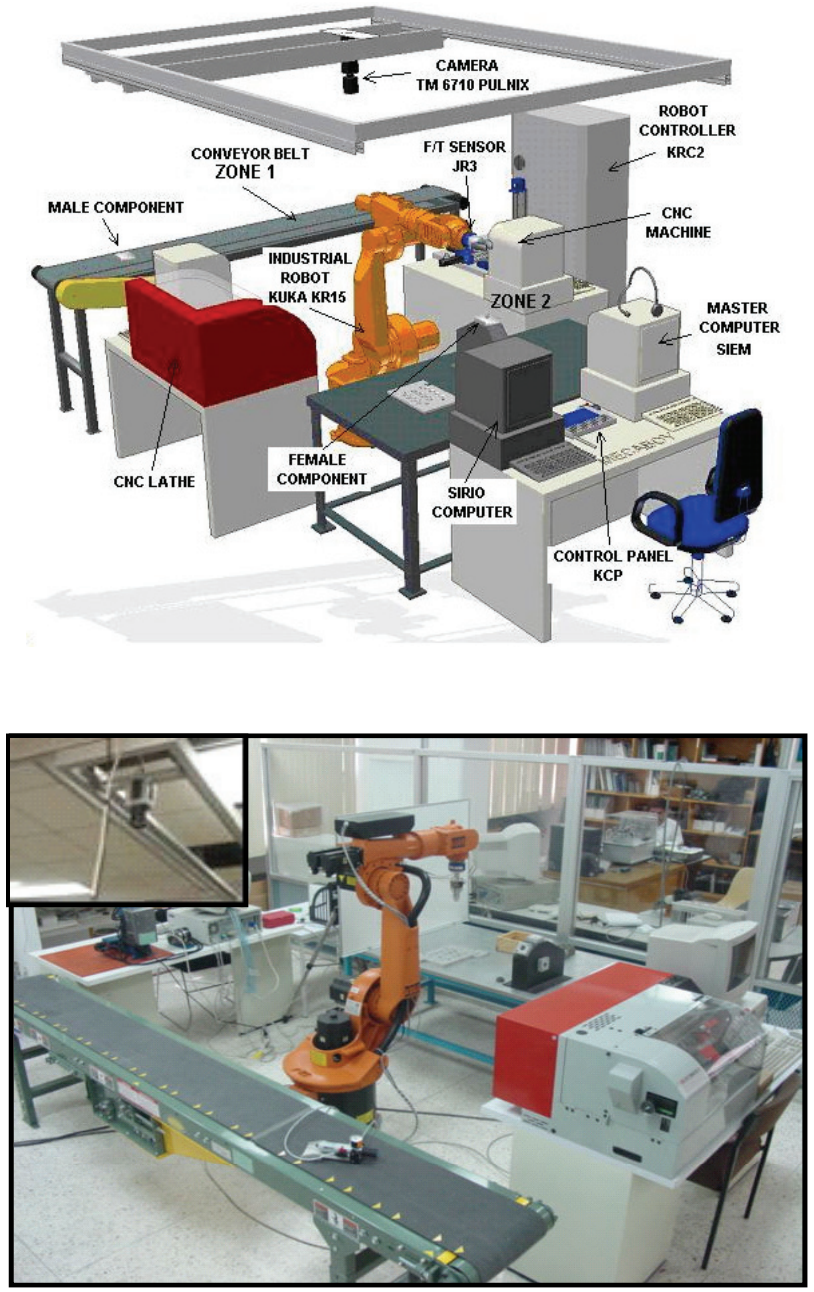


Figure 1. Manufacturing System

The NNC (Neural Network Controller) for assembly is called SIEM (*Sistema Inteligente de Ensamble Mecánico, in Spanish*) and is based on a FuzzyARTMAP neural network working in fast learning mode (Carpenter et al., 1992). The vision system, called SIRIO (*Sistema Inteligente de Reconocimiento Invariante de Objetos*), also uses the same neural network to learn and classify the assembly components. The SIRIO was implemented with a high speed camera CCD/B&W, PULNIX 6710, with 640x480 resolution. The camera movement on the X-Y plane was implemented using a 2D positioning system.

2. State of the Art

2.1 Distributed Robotic Systems

The concept of distributed systems and other technologies recently have made possible the creation of new application called "Networked Robot Systems". The reduction in cost is one of the several advantages of creating distributed systems. This is mainly for the use and creation of standard components and infrastructures (Amoretti, 2004); in addition the CORBA standard solves the heterogeneity problem which is found in the robotic systems in a network. It permits the interaction and interoperability of different systems developed with different technologies, programming languages, operation systems or hardware.

Currently, the development of robot systems base on distributed components is being developed by different researchers. In (Amoretti et al., 2003), Michael Amoretti et al., present an analysis of three techniques for data distributing of sensors through the network. The first technique is called *Callback*, where the clients call a method of a server and at the same time send an object reference to inform the server to which client has to send the answer and information. When the server finishes the asked task, it checks the object number to which it has to send the results. The second technique is based on the services or events of CORBA. Where the servers produce events and the clients receive them using an *event channel*. The event channel conducts the events of the servers to the clients, without having information about the clients and vice versa. In spite of the advantages of the event channel, when it is used, it generates negative aspects, such as the data type, since it has to be of type "any" or IDL, and it makes the communication not very secure. Clients have to convert the data

to their respective type and another problem of the event channel is that it could saturate the network, since it does not have an event filter and sends all messages to all clients. The event services does not contemplate the use of QoS (Quality of Service), related with the priority, liability and order.
The third technique is based on the Notification Service of CORBA. It is an improvement of the Service of Events. The most important improvement includes the use of QoS. In the notification service each client uses the events in which it is interested.

The implementation of the Callback technique offers a better performance than the others; however the ones based on the event channel are easily scalable. The technique used in our research is Callback since the number of clients, is not bigger of 50.
In (Amoretti, 2004) it is proposed a robotic system using CORBA as communication architecture and it is determined several new classes of telerobotic applications, such as virtual laboratories, remote maintenance, etc. which leads to the distributed computation and the increase of new developments like teleoperation of robots. They used a distributed architecture supporting a large number of clients, written in C++ and using CORBA TAO as middleware, but it is an open architecture, and it does not have intelligence, just remote execution of simple tasks.
In (Bottazzi et al., 2002), it is described a software development of a distributed robotic system, using CORBA as middleware. The system permits the development of Client-Server application with multi thread supporting concurrent actions. The system is implemented in a laboratory using a manipulator robot and two cameras, commanded by several users. It was developed in C++ and using TAO.
In (Dalton et al., 2002), several middleware are analyzed, CORMA, RMI (Remote Method Invocation) and MOM (Message Oriented Middleware). But they created their own protocol based on MOM for controlling a robot using Internet.

In (Jia et al., 2002), (Jia et al., 2003) it is proposed a distributed robotic system for telecare using CORBA as communication architecture. They implemented three servers written in C++, the first one controls a mobile robot, the second one is used to control an industrial robot and the last one to send real time video to the clients. On the other side of the communication, it is used a client based on Web technology using Java Applets to make easier the use of the system in Internet. In (Jia et al., 2003), the authors increased the number of servers

available in the system, with: a user administrator and a server for global positioning on the working area.
In (Corona-Castuera & Lopez-Juarez, 2004) it is discussed how industrial robots are limited in terms of a general language programming that allows learning and knowledge acquisition, which is probably, one of the reasons for their reduced use in the industry. The inclusion of sensorial capabilities for autonomous operation, learning and skill acquisition is recognized. The authors present an analysis of different models of Artificial Neuronal Networks (ANN) to determine their suitability for robotic assembly operations. The FuzzyART-MAP ANN presented a very fast response and incremental learning to be implemented in the robotic assembly system. The vision system requires robustness and higher speed in the image processing since it has to perceive and detect images as fast as or even faster than the human vision system. This requirement has prompted some research to develop systems similar to the morphology of the biological system of the human being, and some examples of those systems, can be found in (Peña-Cabrera & Lopez-Juarez, 2006), (Peña-Cabrera et al., 2005), where they describe a methodology for recognising objects based on the Fuzzy ARTMAP neural network.

2.2 Multimodal Neural Network

A common problem in working in multimodality for robots systems is the employment of data fusion or sensor fusion techniques (Martens, S. & Gaudiano, P., 1998 and Thorpe, J. & Mc Eliece, R., 2002). Multimodal pattern recognition is presented in (Yang, S. & Chang, K.C., 1998) using Multi-Layer Perceptron (MLP). The ART family is considered to be an adequate option, due to its superior performance found over other neural network architectures (Carpenter, G.A. et al., 1992). The adaptive resonance theory has provided ARTMAP-FTR (Carpenter, G.A. & Streilein, W.W, 1998), MART (Fernandez-Delgado, M & Barro Amereiro, S, 1998), and Fusion ARTMAP (Asfour, et al., 1993) —among others— to solve problems involving inputs from multiple channels. Nowadays, G.A. Carpenter has continued extending ART family to be employed in information fusion and data mining among other applications (Parsons, O. & Carpenter, G.A, 2003).

The Mechatronics and Intelligent Manufacturing Systems Research Group (MIMSRG) at CIATEQ performs applied research in intelligent robotics, concretely in the implementation of machine learning algorithms applied to as-

sembly tasks —using distributed systems contact forces and invariant object recognition. The group has obtained adequate results in both sensorial modalities (tactile and visual) in conjunction with voice recognition, and continues working in their integration within an intelligent manufacturing cell. In order to integrate other sensorial modalities into the assembly robotic system, an ART-Based multimodal neural architecture was created.

3. Design of the Distributed System

3.1 CORBA specification and terminology

The CORBA specification (Henning, 2002), (OMG, 2000) is developed by the OMG (Object Management Group), where it is specified a set of flexible abstractions and specific necessary services to give a solution to a problem associated to a distributed environment. The independence of CORBA for the programming language, the operating system and the network protocols, makes it suitable for the development of new application and for its integration into distributed systems already developed.

It is necessary to understand the CORBA terminology, which is listed below:

A **CORBA object**	is a virtual entity, found by an ORB (Object Request Bro ker, which is an ID string for each server) and it accepts petitions from the clients.
A **destine object**	in the context of a CORBA petition, it is the CORBA ob ject to which the petition is made.
A **client**	is an entity which makes a petition to a CORBA object.
A **server**	is an application in which one or more CORBA objects run.
A **petition**	is an operation invocation to a CORBA object, made by a client.
An **object reference**	is a program used for identification, localization and di rection assignment of a CORBA object.
A **server**	is an identity of the programming language that imple ments one or more CORBA objects.

The petitions are showed in the figure 2: it is created by the client, goes through the ORB and arrives to the server application.

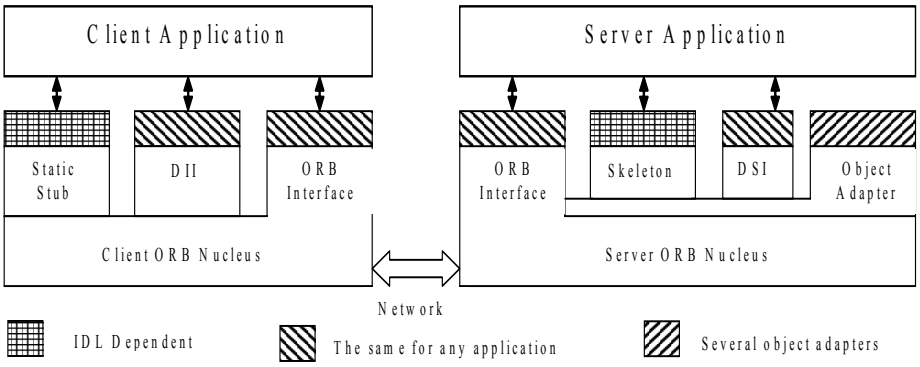


Figure 2. Common Object Request Broker Architecture (COBRA)

- The client makes the petitions using static stub or using DII (Dynamic Invocation Interface). In any case the client sends its petitions to the ORB nucleus linked with its processes.
- The ORB of the client transmits its petitions to the ORB linked with a server application.
- The ORB of the server redirect the petition to the object adapter just created, to the final object.
- The object adapter directs its petition to the server which is implemented in the final object. Both the client and the sever, can use static skeletons or the DSI (Dynamic Skeleton Interface)
- The server sends the answer to the client application.

In order to make a petition and to get an answer, it is necessary to have the next CORBA components:
Interface Definition Language (**IDL**): It defines the interfaces among the programs and is independent of the programming language.

Language Mapping: it specifies how to translate the IDL to the different pro gramming languages.

Object Adapter: it is an object that makes transparent calling to other ob jects.

Protocol Inter-ORB: it is an architecture used for the interoperability among different ORBs.

The characteristics of the petitions invocation are: transparency in localization, transparency of the server, language independence, implementation, architecture, operating system, protocol and transport protocol. (Henning, 2002).

3.1 Architecture and Tools

The aim of having a coordinator, is to generate a high level central task controller which uses its available senses (vision and tactile) to make decisions, acquiring the data on real time and distributing the tasks for the assembly task operation.

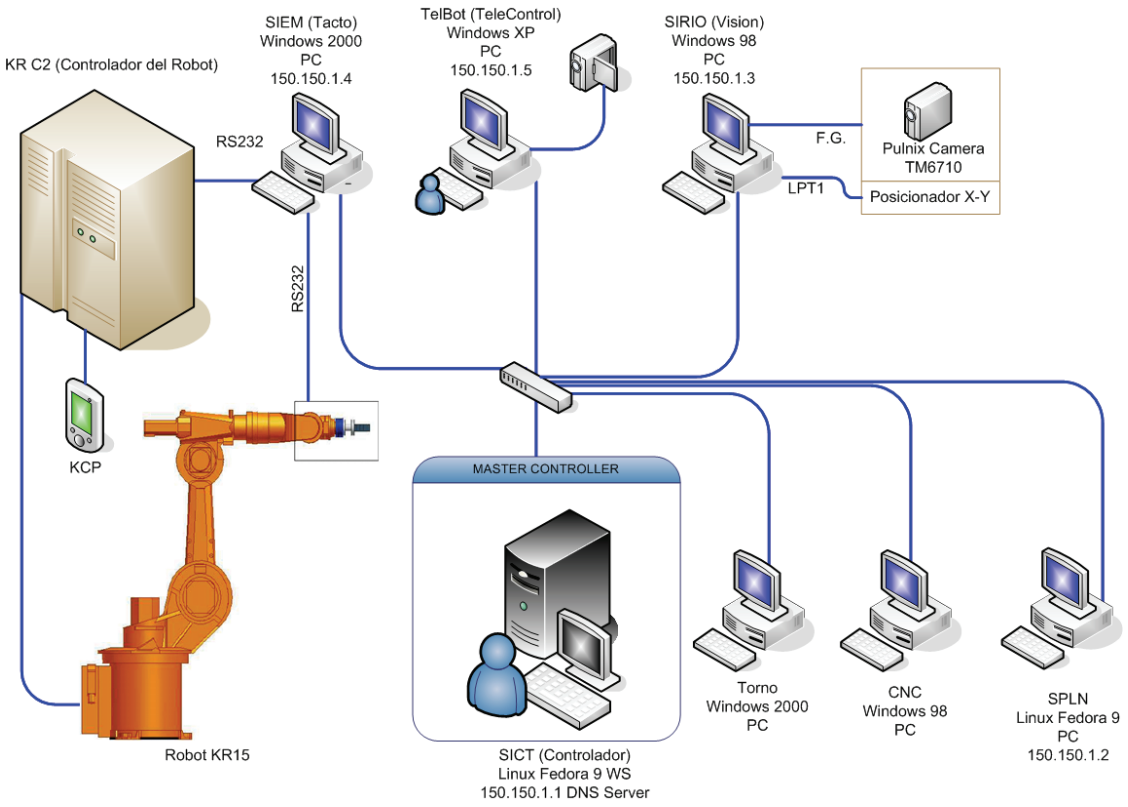


Figure 3. Distributed Manufacturing Cell

Figure 3 shows the configuration of the network and the main components of the distributed cell, however, the active ones are: SIRIO, SIEM, SICT and SPLN. The system works using a multiple technology architecture where different operating systems, middleware, programming language and graphics tools were used, as it can be seen in figure 4. It describes the main modules of the manufacturing cell SIEM, SIRIO, SICT and SPLN.

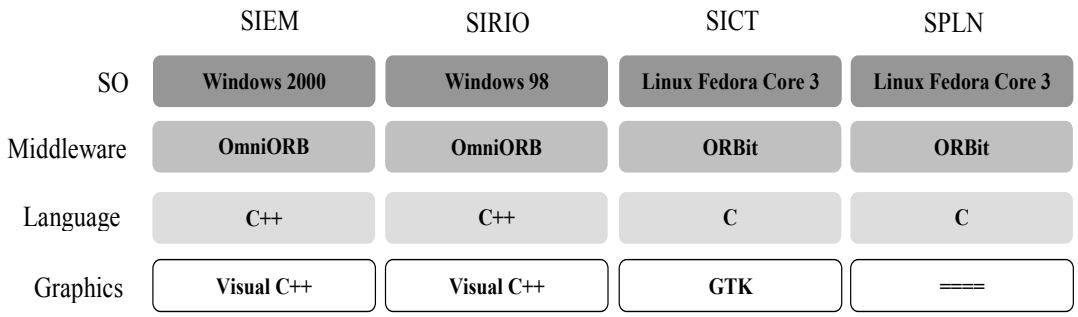


Figure 4. Different operating systems, middleware, programming languages and graphic tools.

The architecture of the distributed system uses a Client/Server in each module. Figure 5 shows the relationship client-server in SICT for each module. But with the current configuration, it is possible a relationship from any server to any client, since they share the same network. It is only necessary to know the name of the server and obtain the IOR (Interoperable Object Reference).

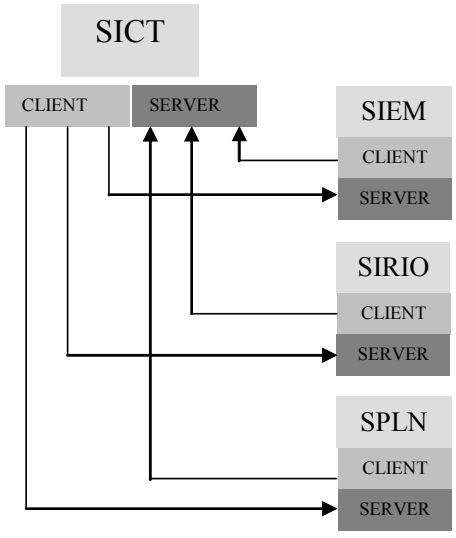


Figure 5. Client/server architecture of the distributed cell

The interfaces or IDL components needed to establish the relations among the modules SICT, SIRIO, SIEM and SPLN are described in the following section.

4. Servers Description

4.1 SICT Interface

This module coordinates the execution of task in the servers (this is the main coordinator). It is base in Linux Fedora Core 3, in a Dell Workstation and written in C language using gcc and ORBit 2.0. For the user interaction of these modules it was made a graphic interface using GTK libraries.

The figure 6 shows the most important functions of the IDL.

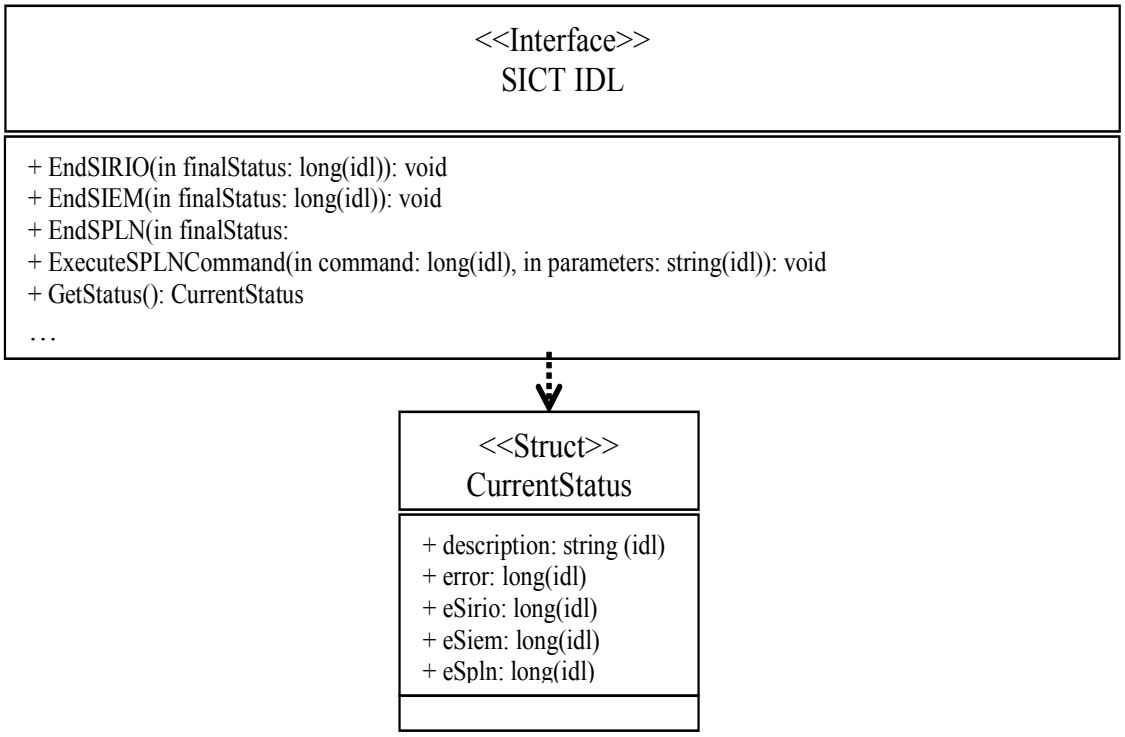


Figure 6. SICT Interface

iSICT: the functions of this interface are used for SIRIO and SIEM to indicate that they have finished a process. Each system sends to SICT a finished process acknowledgement of and the data that they obtain. SICT makes the decisions about the general process. The module SPLN uses one of the functions of SICT to ask it to do a task, sending the execution command with parameters. The figure 7 shows the main screens of the coordinator.

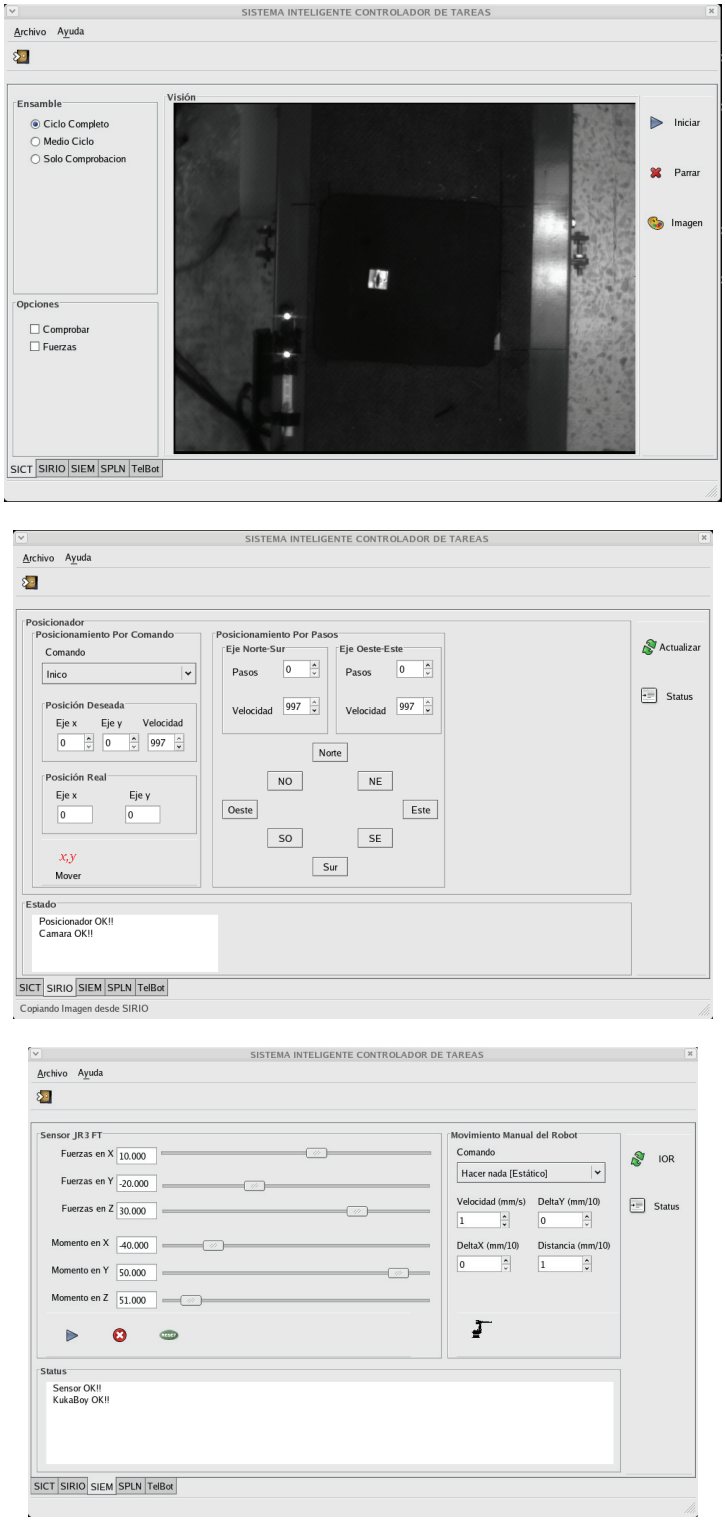


Figure 7. Controls of the interface SICT

4.2 SIRIO Interface

This system is the vision sense of the robot, using a camera Pulnix TM6710, which can move around the cell processing the images in real time. SIRIO carries out a process based on different marks. It calculates different parameters of the working pieces, such as orientation, shape of the piece, etc. This system uses Windows 98 and is written in Visual C++ 6.0 with OmniORB as middleware.

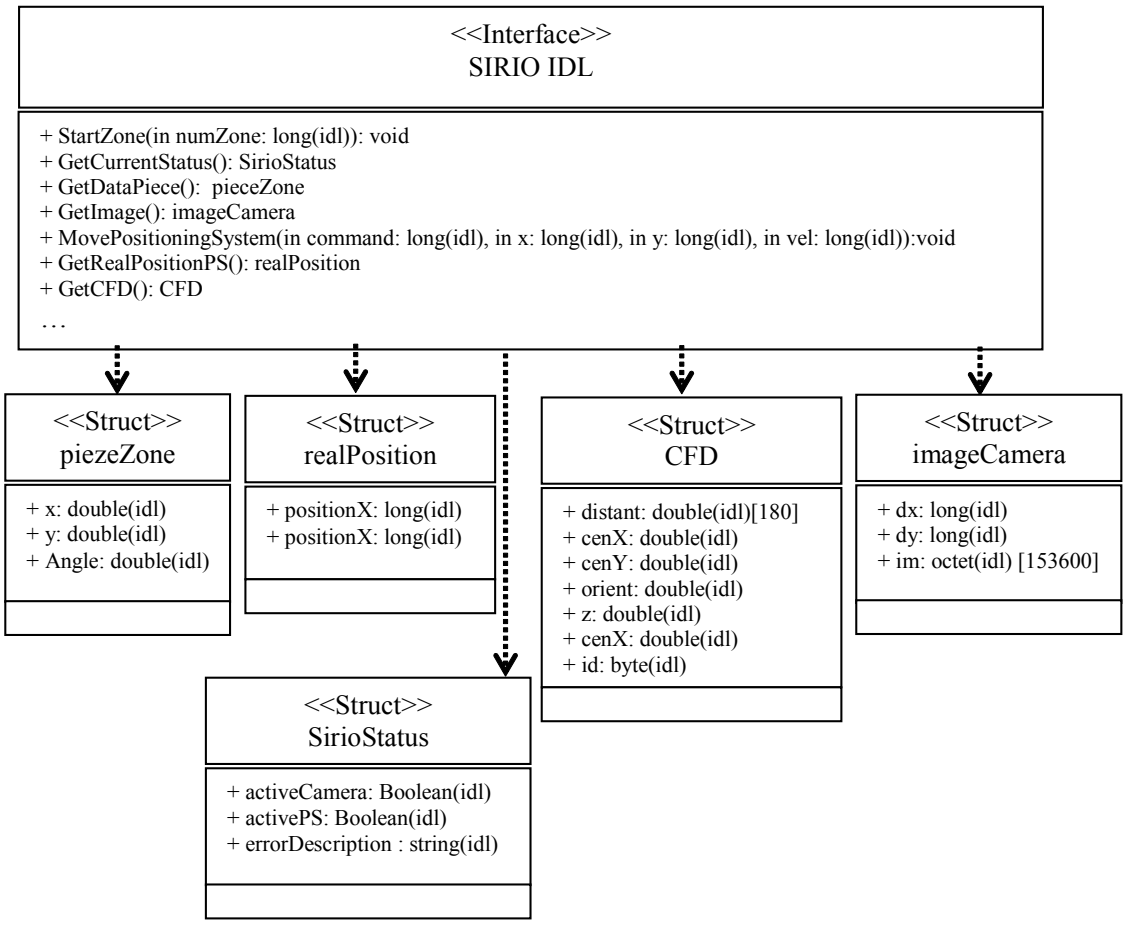


Figure 8. SIRIO Interface

iSIRIO interface contains functions used by the SICT to initialize the assembly cycle, to obtain the status of SIRIO, an image in real time or to move the camera over the manufacturing cell. The function StartZone, calls a process located in SIRIO to make the positioning system move to different zones of the cell. The function GetCurrentStatus is used to get the current status of the SIRIO

module, and it sends information about the hardware. When SIRIO finishes processing an image it sends an acknowledgement to SICT and this ask for the data using the function GetDataPiece which gives the position and orientation of the piece that the robot has to assembly.

The function GetImage gives a vector containing the current frame of the camera and its size. The function MovePositioningSystem is used by SICT to indicate to SIRIO where it has to move the camera. The movements are showed in table 1, where it executes movements using the variables given by the client that called the function.

Tabla 1.	**Command**	**Tabla 2.**	**X**	**Tabla 3.**	**Y**	**Tabla 4**	**Speed**
Tabla 5.	Start	Tabla 6.	No	Tabla 7.	No	Tabla 8.	Yes
Tabla 9.	Zone 1	Tabla 10.	No	Tabla 11.		Tabla 12.	Yes
Tabla 13	Zone 2	Tabla 14.	No	Tabla 15.	No	Tabla 16.	Yes
Tabla 17	Moves to (x,y)	Tabla 18.	Yes	Tabla 19.	Yes	Tabla 20.	Yes

Table 1. Commands for moving the positioning system.

The function GetRealPositonPS obtains the position (x, y) where the positioning system is located.

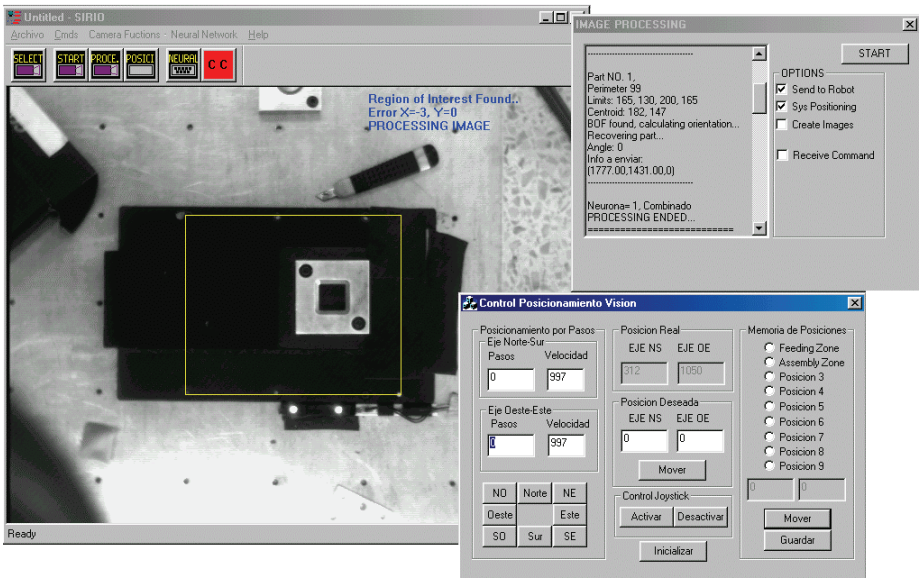


Figure 9. SIRIO main scream

The last function GetCFD(), gets the CFD (Current Frame Descriptor) of a piece. The piece is always the last the system used, or the one being used. The CFD contains the description of a piece. For more details the reader is referred to part III of this work (Peña-Cabrera, M. & Lopez-Juarez, I, 2006).

4.3 SIEM Interface

This contact force sensing system resembles the tactile sense, and uses a JR3 Force/Torque (F/M) sensor interacting with the robot and obtaining contact information from the environment. SIEM is used when the robot takes a piece from the conveyor belt or when or when an assembly is made. The robot makes the assemblies with incremental movements and in each movement, SIEM processes and classifies the contact forces around the sensor, using the neural network to obtain the next direction movement towards the assembly. SIEM is implemented in an industrial parallel computer using Windows 2000 and written in Visual C++ 6.0 and OmniORB.

Figure 10 shows the main functions of the IDL SIEM.

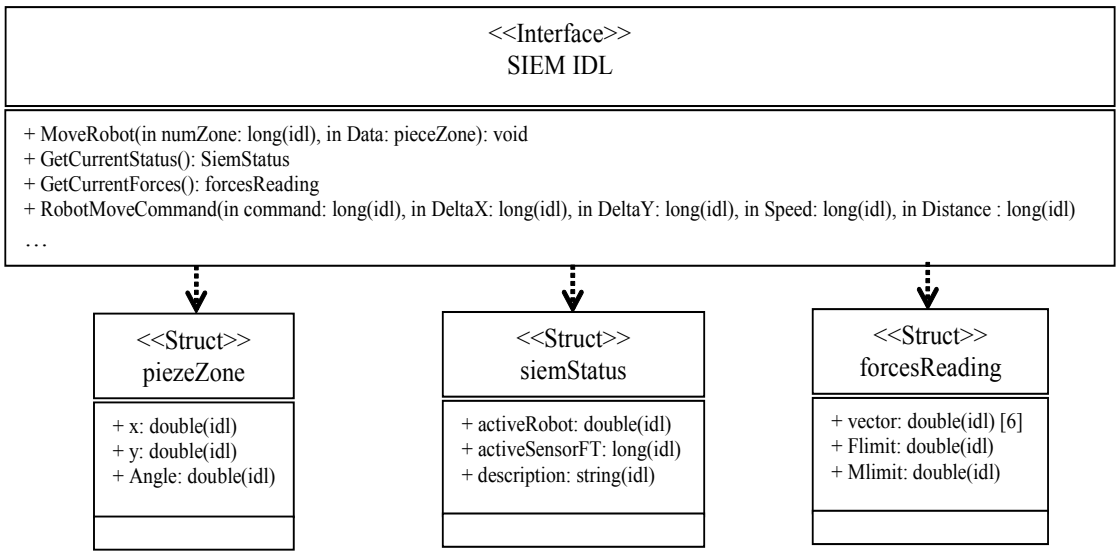


Figure 10. SIEM Interface

iSIEM: SICT moves the robot thought SIEM, obtains the components state and the reading of the current forces in the different zones of the manufacturing cell. The function GetCurrentStatus, is used to obtain the status of the hard-

ware (sensor F/T and communication) and software of the SIEM. The function MoveRobot is used when SIRIO finishes an image processing and sends information about the piece to the task coordinator.

The GetCurrentForces function helps the SICT to acquire force data from the JR3 Force/Torque (F/T) sensor at a selected sampling rate. This function returns a data vector with information about the force and torque around X, Y and Z axis.

Finally, the function RobotMoveCommand is used by the SICT to indicate appropriate motion commands to SIEM. These types of motions are shown in Table 2. Here is also shown the required information for each command (distance, speed). The windows dialog is shown in Figure 11.

Command	Distance	Speed
Do nothing [static]	No	No
Home position	No	No
Coordinates world	No	No
Tool Coordinates	No	No
Axe by Axe Coordinates	No	No
Base Coordinates	No	No
Movement X+	Yes	Yes
Movement X-	Yes	Yes
Movement Y+	Yes	Yes
Movement Y-	Yes	Yes
Movement Z+	Yes	Yes
Movement Z-	Yes	Yes
Rotation X+	Yes	Yes
Rotation X-	Yes	Yes
Rotation Y+	Yes	Yes
Rotation Y-	Yes	Yes
Rotation Z+	Yes	Yes
Rotation Z-	Yes	Yes
Diagonal X+Y+	Yes	Yes

Command	Distance	Speed
Diagonal X+Y-	Yes	Yes
Diagonal X-Y+	Yes	Yes
Diagonal X-Y-	Yes	Yes
Finish Communication	No	No
Open griper	No	No
Close griper	No	No
Rotation A1+	Yes	Yes
Rotation A1-	Yes	Yes
Rotation A2+	Yes	Yes
Rotation A2-	Yes	Yes
Rotation A3+	Yes	Yes
Rotation A3-	Yes	Yes
Rotation A4+	Yes	Yes
Rotation A4-	Yes	Yes
Rotation A5+	Yes	Yes
Rotation A5-	Yes	Yes
Rotation A6+	Yes	Yes
Rotation A6-	Yes	Yes

Table 2. Commands to move the robot

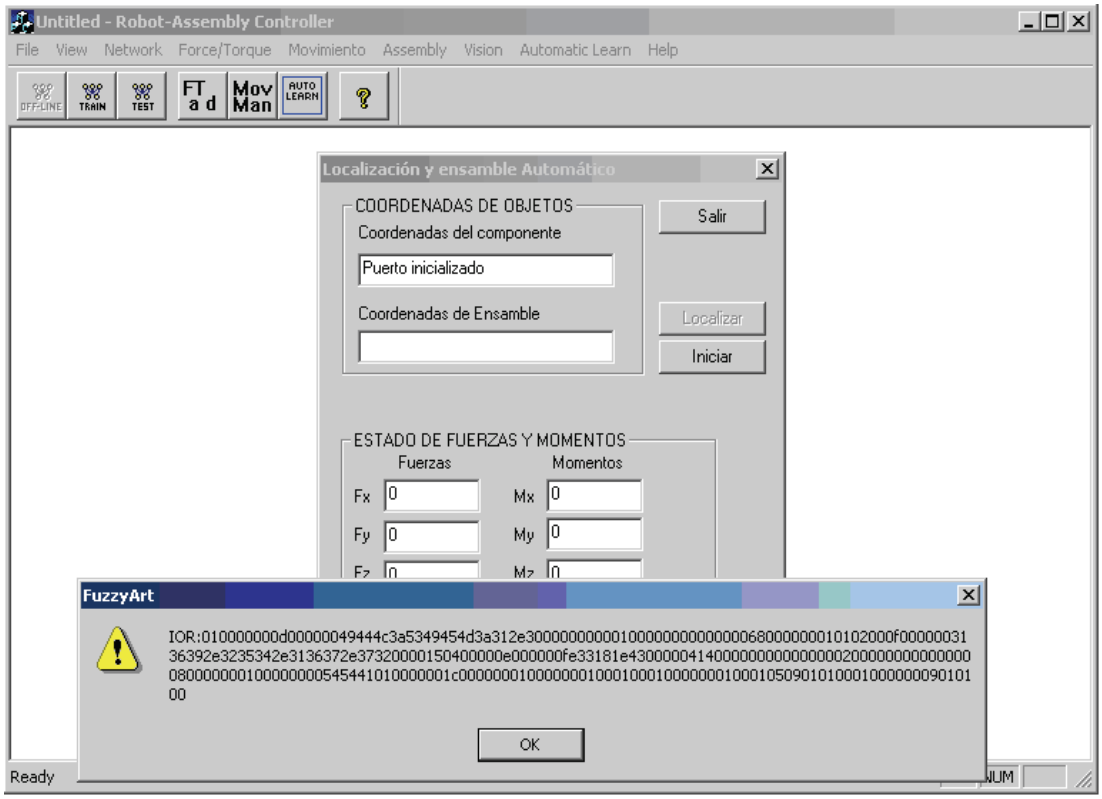


Figure 11. SIEM screen

4.4 SPLN Interface

The system provides a user interface to receive directions in natural language using natural language processing and context free grammars. After the instruction is given, a code is generated to execute the ordered sentences to the assembly system. The SPLN is based on Linux Fedora Core 3 operating system using a PC and programmed in C language and a g++, Flex, Yacc and ORBit 2.0. compiler.

iSPLN: This interface receives the command status from the SPLN, and gets the system's state as it is illustrated in Figure 12.
EndedTask is used by the SICT to indicate the end of a command to the SPLN like the assembly task. As a parameter, SICT sends to SPLN the ending of the task. GetStatus function serves to obtain the general state of the SPLN.

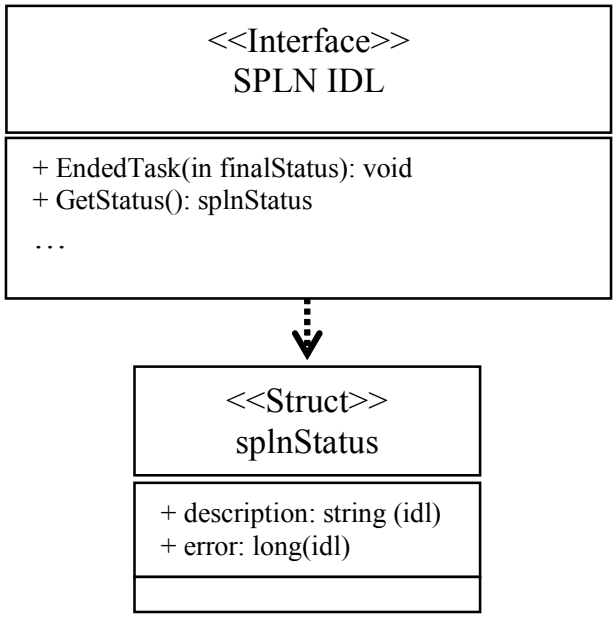


Figure 12. Interface SPLIN

5. Multimodal Architecture.

We have described so far the architecture of the distributed system, where each module has its own FuzzyARTMAP neural network, its own KB (Knowledge Base) and configuration. The architecture showed in figure 13 shows the current architecture module M2ARTMAP. Currently, there is a coordinator substituting the Integrator and Predictor in the upper level. M2ARTMAP has demonstrated to be faster in training and learning than a single Fuzzy ARTMAP making a fusion of all senses at the same time, for more details see (Lopez-Juarez et al., 2005).

– **Predictor** is the final prediction component that uses modalities' predictions.

– **Modality** is the primary prediction component that is composed by an artificial neural network (ANN), an input element (Sensor), a configuration element (CF), and a knowledge base (KB).

– **Integrator** is the component that merges the modalities' predictions by inhibiting those that are not relevant to the global prediction activity, or stimulating those who are consid ered of higher reliability —in order to facilitate the Pre dictor's process.

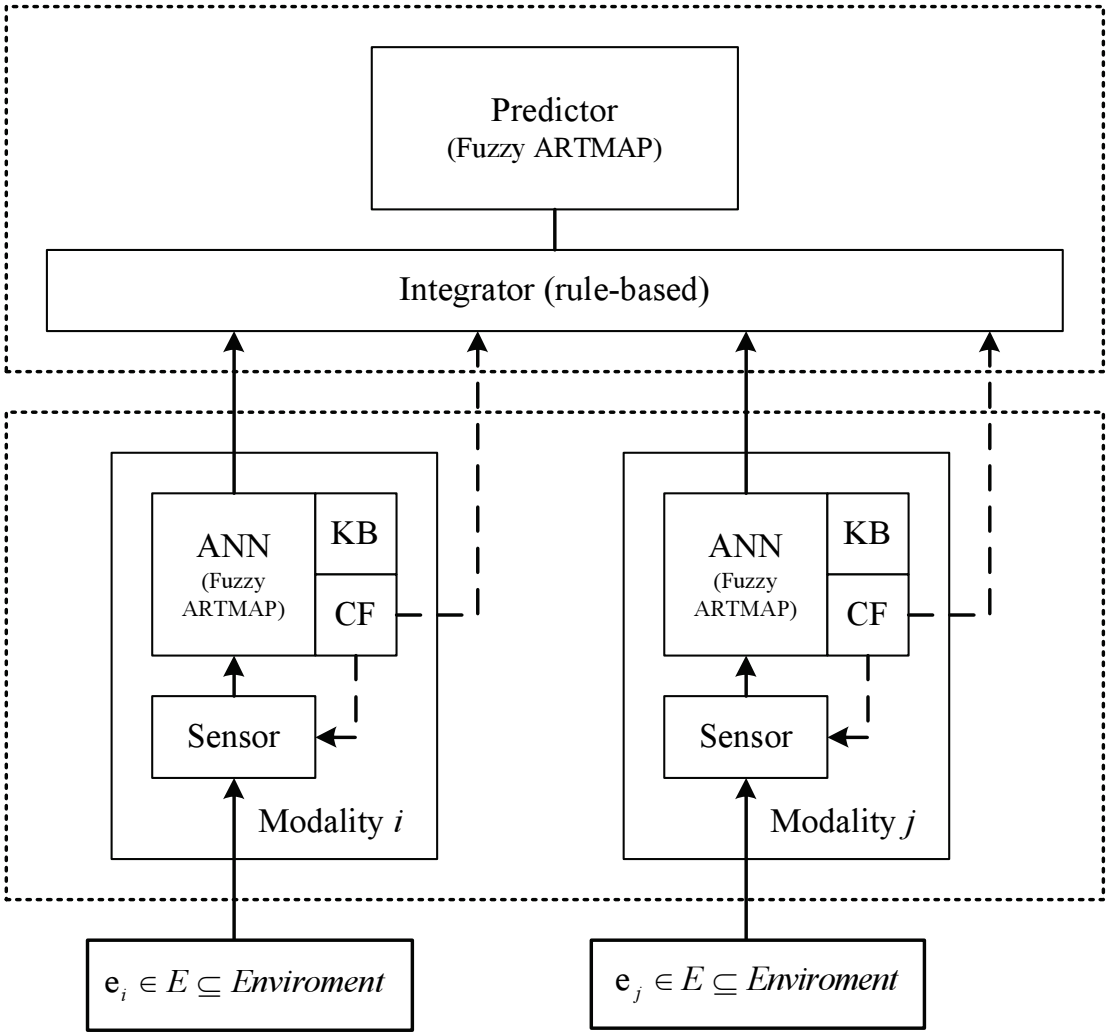


Figure 13. Multimodal neural architecture, M2ARTMAP, integrated by three main components organized in two layers: Modality (several found at the lower layer), Predictor and Integrator (at the upper layer)

5.1 Multimodal simulations

Fuzzy ARTMAP and M2ARTMAP systems were simulated using the *Quadruped Mammal* database (Ginnari, J.H.; et al., 1992) which represents four mammals (dog, cat, giraffe, and horse) in terms of eight components (head, tail, four legs, torso, and neck). Each component is described by nine attributes (three location variables, three orientation variables, height, radius, and texture), for a total of 72 attributes. Each attribute is modelled as a Gaussian process with mean and variance dependent on the mammal and component (e.g. the radius of a horse's neck is modelled by a different Gaussian from that of a dog's neck or a horse's tail). At this point, it is important to mention that Quadruped Mammal database is indeed a structured quadruped mammal instances generator that requires the following information to work: animals <seed> <# of objects>.

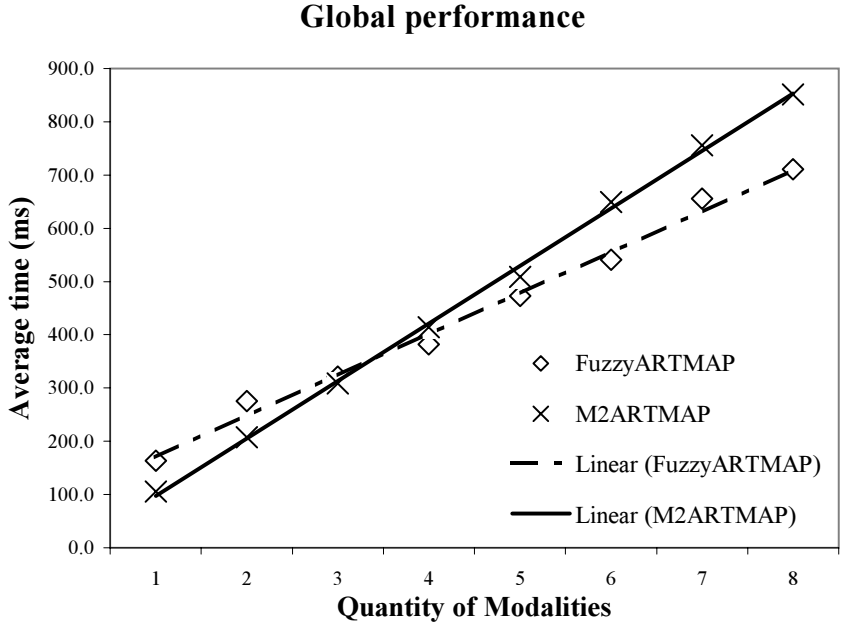


Figure 14. Performance comparison of Fuzzy ARTMAP vs. M2ARTMAP. (a) Training phase. (b) Testing phase. (C) Global performance

In the first set of simulations, both Fuzzy ARTMAP and M2ARTMAP where trained (in one epoch) and tested with the same set of 1000 exemplars produced with seed = 1278. Both architectures achieved 100% prediction rates.

In the next set of simulations, Fuzzy ARTMAP and M2ARTMAP where applied to a group of 384 subjects (91 variations of the choice parameter and 4 variations of the base vigilance), both architectures where trained (again in one epoch) using the set of 1000 exemplars produced with seed = 1278 and tested using the set of 1000 exemplars produced with seed = 23941. Once again, both achieved 100% prediction rates. Nevertheless, M2ARTMAP's recognition rates where slower than expected. Thus, a t-Student paired test was conducted to constraint the difference between both architectures. It was confirmed that M2ARTMAP's recognition rate was at most 5% slower than Fuzzy ARTMAP's recognition rate, by rejecting the null hypothesis with a 1-tail *p*-value less than 0.0001.

The global performance of the M2ARTMAP indicated that its performance it is superior when three or less modalities are used, which was considered acceptable since in a manufacturing environment is likely to encounter two or three modalities at most. The global comparison between M2ARTMAP and the FuzzyARTMAP architecture is illustrated in Figure 14. (The reader is referred to (Lopez-Juarez, I. & Ordaz-Hernandez, 2005) for complete details).

6. Results from the implementation of the Distributed System.

6.1 General Description

36 robotic assembly cycles were performed. Three modules were involved in the assessment SICT, SIRIO and SIEM. At the start of the operation SICT indicates to SIRIO to initialise the image processing in Zone 1, which corresponds to the area where the male component is grasped from the belt conveyor. Later this information is being sent to the SIEM which in turns moves the robot manipulator to pick up the component. At the same time the camera moves on the working space detecting the Zone 2, where the fixed, female component is located. This information is also sent to the SIRIO, to direct the robot towards the assembly point. Once the part is grasped and in contact with the female component the assembly operation is solely directed by the SIEM.

Table 3 contains the results from the 36 assembly cycles. The table provides information about the geometry of the component, chamfer, operation time, position error, based on the centroid location and the component rotation and finally the predicted type of assembly by the SIRIO module.

Test for grasping the parts was made from zone 1 for each geometry. Each type was placed three times in the zone with 10º orientation difference and four different locations.

In the assembly zone (zone 2) the location and orientation of the female component was constant. However, this information was never available to the SIEM or robot controller. So, every time this distance had to be calculated.

The first 18 assembly cycles were performed with chamfered female components and the remaining 18, without a chamfer. This can be observed in Table 3.

The total time corresponds to the assembly cycle, including insertion time, camera positioning, robot motion, image processing and home positioning. It is important to mention that all speeds were carefully chosen since it was a testing session. The system was later improved to work faster as it can be seen in (Corona-Castuera & Lopez-Juarez, 2006). In this case for the assembly zone, there was always an initial error, to test the error recovery.

#	Piece	Ch	Time (Min)	ZONE 1			Error Zone 1			ZONE 2		Error Zone 2		Clasific.
				Xmm	Ymm	RZº	Xmm	Ymm	RZº	Xmm	Ymm	Xmm	Ymm	
1	square	Yes	1,14	57,6	143,1	0°	2,4	1,9	0	82,8	102,0	0,3	1,8	square
2	square	Yes	1,19	56,6	44,8	12°	15,2	0,2	2	82,8	101,1	0,2	1,2	square
3	square	Yes	1,13	172,8	46,7	23°	2,20	-1,7	3	83,8	162,0	-0,9	2,1	square
4	rad	Yes	1,72	176,7	145,1	29°	-1,70	-0,1	-1	79,6	103,0	3,9	2,9	rad
5	rad	Yes	1,86	56,6	143,1	36°	3,4	1,9	-4	82,8	103,0	0,6	2,9	rad
6	rad	Yes	1,29	58,5	44,8	55°	15,2	0,2	5	80,7	102,0	1,5	2,3	rad
7	circle	Yes	1,12	172,8	46,7	57°	2,20	-1,7	-3	82,8	101,1	0,4	1,2	circle
8	circle	Yes	1,13	176,7	145,1	104°	-1,70	-0,1	34	83,8	103,0	0	3	circle
9	circle	Yes	1,24	56,6	143,1	79°	3,4	1,9	-1	79,6	102,0	3,2	2,2	circle
10	square	Yes	1,7	56,6	42,8	66°	17,2	2,2	-24	79,6	102,0	3,5	1,9	square
11	square	Yes	1,22	172,8	45,7	123°	2,20	-0,7	23	83,8	102,0	-2	2,4	square
12	square	Yes	1,93	178,7	144,1	110°	-3,70	0,9	0	80,7	102,0	0	0	square
13	rad	Yes	1,79	55,6	143,1	116°	4,4	1,9	-4	82,8	102,0	1	2,4	rad
14	rad	Yes	1,83	59,5	43,8	124°	16,2	1,2	-6	80,7	103,0	-0,5	2,3	rad
15	rad	Yes	1,85	174,7	44,8	145°	0,30	0,2	5	82,8	102,0	1,8	3,1	square
16	circle	Yes	1,76	176,7	147	143°	-1,70	-2	-7	80,7	102,0	-0,4	2,2	circle
17	circle	Yes	1,21	57,6	144,1	164°	2,4	0,9	4	82,8	102,0	2,2	2,4	circle
18	circle	Yes	1,23	60,5	45,7	175°	14,3	-0,7	5	83,8	103,0	-0,3	2,4	circle
19	square	No	1,21	174,7	48,6	0°	0,30	-3,6	0	81,7	103,0	-0,7	3,3	square
20	square	No	1,13	177,7	147	12°	-2,70	-2	2	82,8	103,0	0,5	1	square
21	square	No	1,8	57,6	142,1	15°	2,4	2,9	-5	83,8	102,0	-0,5	0,8	square
22	rad	No	1,84	61,5	45,7	26°	14,3	-0,7	-4	83,8	148,0	-1,6	-0,2	rad
23	rad	No	1,26	175,7	49,6	36°	-0,70	-4,6	-4	82,8	103,0	-2,7	0,2	rad
24	rad	No	1,21	179,6	147	49°	-4,60	-2	-1	82,8	103,0	-1,7	1,3	rad
25	circle	No	1,13	59,5	147	63°	0,5	-2	3	82,8	102,0	-1,1	0,9	circle
26	circle	No	1,2	61,5	46,7	84°	13,3	-1,7	14	79,6	102,0	-1,7	0,2	circle
27	circle	No	1,11	176,7	49,6	77°	-1,70	-4,6	-3	82,8	102,0	1,2	0,5	circle
28	square	No	1,71	178,7	148	71°	-3,70	-3	-19	81,7	103,0	-1,7	0,1	square
29	square	No	1,71	56,6	143,1	108°	3,4	1,9	8	79,6	103,0	0,6	0,9	square
30	square	No	1,25	59,5	42,8	105°	17,2	2,2	-5	83,8	102,0	13,2	-9,3	square
31	rad	No	1,71	174,7	46,7	116°	0,30	-1,7	-4	82,8	102,0	2,6	1,1	rad
32	rad	No	2,88	176,7	146	131°	-1,70	-1	1	82,8	102,0	-2,1	0,1	rad
33	rad	No	1,82	57,6	143,1	131°	2,4	1,9	-9	82,8	102,0	-1,8	0,3	rad
34	circle	No	1,28	58,5	43,8	145°	16,2	1,2	-5	78,5	103,0	-1,8	0,3	circle

#	Piece	Ch	Time (Min)	ZONE 1			Error Zone 1			ZONE 2		Error Zone 2		Clasific.
				Xmm	Ymm	RZº	Xmm	Ymm	RZº	Xmm	Ymm	Xmm	Ymm	
35	circle	No	1,14	174,7	46,7	164°	0,30	-1,7	4	82,8	102,0	-2,8	0,3	circle
36	circle	No	1,16	176,7	146	170°	-1,70	-1	0	82,8	102,0	2,7	1,2	circle

Table 3. Information of 36 assembly cycles for testing, with controlled speed

6.2 Time in Information transfers

During the testing session, different representative time transference was measured. This was accomplished using time counters located at the beginning and at the end of each process.

In the following graphs the 36 results are described . In Figure 15 a graph assembly vs time is shown. In the graph the timing between the data transfer between the imageCamera data to the iSIRIO it is shown. From bottom up, the first graph shows the timing SIRIO took to transfer the information in a matrix form; the following graph represents timing between the transfers of image information. The following graph shows the time the client SICT used to locate the image information to the visual component. Finally, the upper graph shows the total time for image transfer considering all the above aspects.

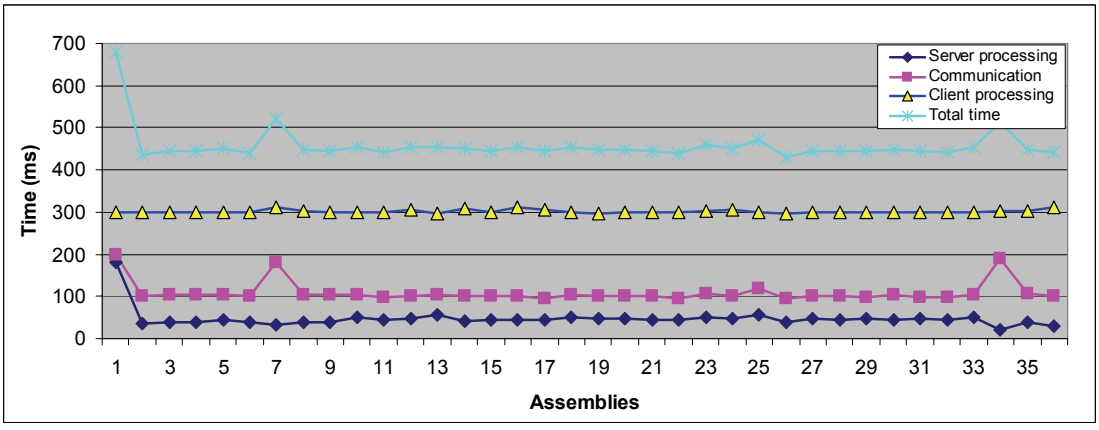


Figure 15. Transference of the information in the structure imageCamera of the SIRIO interface

Figure 16 shows the graphs corresponding to timing of the 36 operations for the transmission of pieceZone data type from the iSirio interface. The first graph from bottom up show the time that the Server took to locate the infor-

mation in the structure pieceZone, the second graph shows the communication time and finally the total operation time.

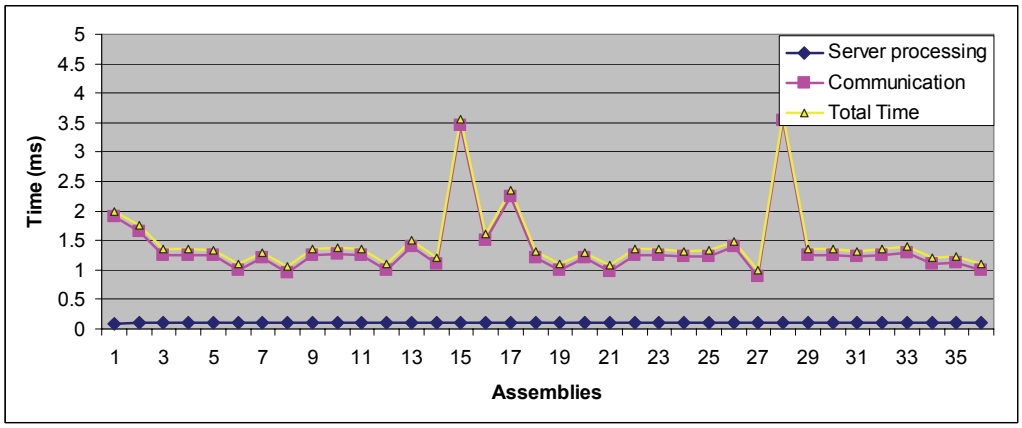


Figure 16. Transference of the information in the structure pieceZone of the SIRIO interface

Figure 17 shows the transference time of the sirioStatus data in the iSIRIO interface, where the first graph from bottom up represents the information transference time, the second graph represents the SIRIO processing time verify the camera status and the camera positioning system and finally the graph that represents the sum of both. It is important to mention that the timing is affected by the Server processing due to the process to verify the location of the positioning system.

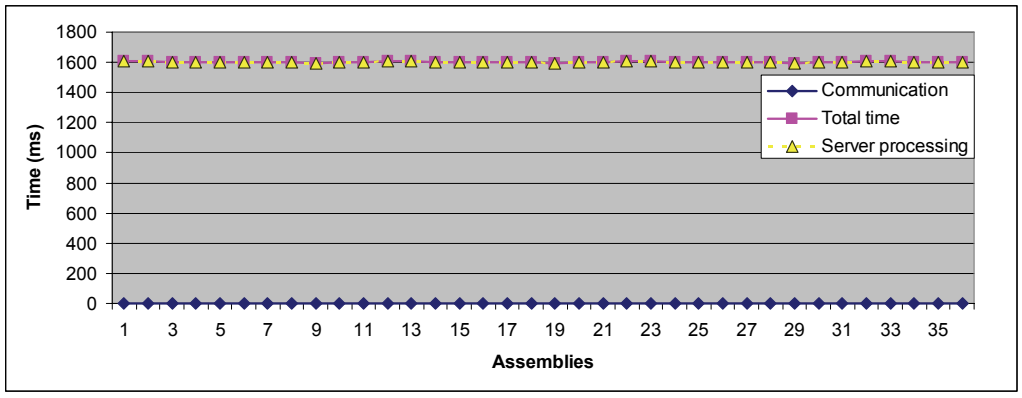


Figure 17. Transference of information in the structure sirioStatus of the SIRIO interface

Finally, figure 18 shows the transference time from the F/T vector through the forcesReading data type from the interface iSIEM. The first graph from bottom up represents the communication times, the second, the time the SICT client took to show the information in visual components. The upper graph represents the time SIEM Server took to read the sensor information and finally a graph that represents the total time for each transference.

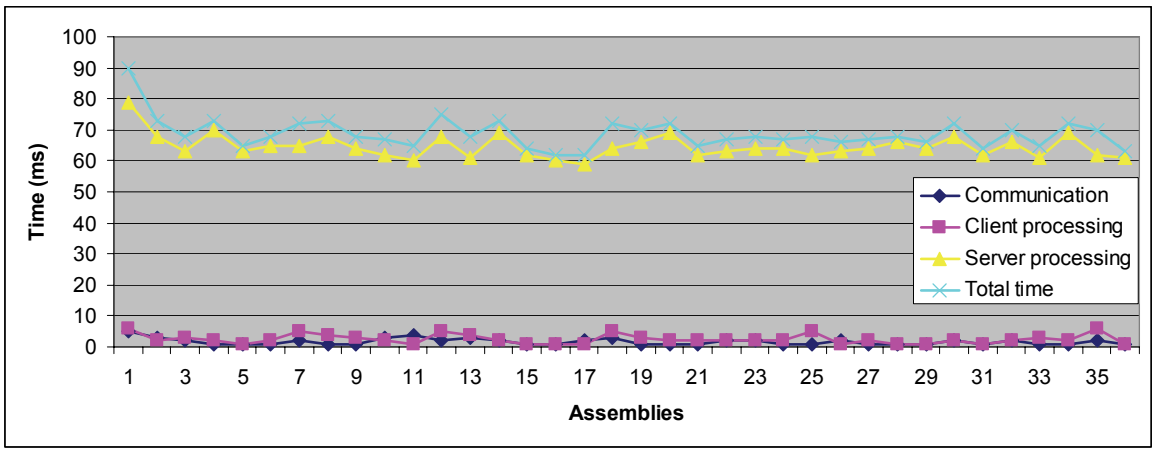


Figure 18. Transference of information in the structure forcesReading of the SIEM interface

6.3 Failure Measurement

We observed this point to statistically obtain the reliability of our system according to the performance of our system. In the 36 assemblies carried out, (18 chamfered and 18 chamferless), we obtained a 100% success, in spite of the variety of the intercommunications and operations of the modules. During the 36 assembly cycles we did not register any event which caused the cycle to abort.

6.4 Robot Trajectories in the Assembly Zone

A robot trajectory describes the movements that the robot developed in the X, Y and Rz directions starting from an offset error to the insertion point. In Figure 19 the followed trajectories for the first 18 insertions during circular chamfered insertion are shown whereas in Figure 20 the corresponding trajectories for circular chamfered insertion are illustrated. In both cases a random offset was initially given.

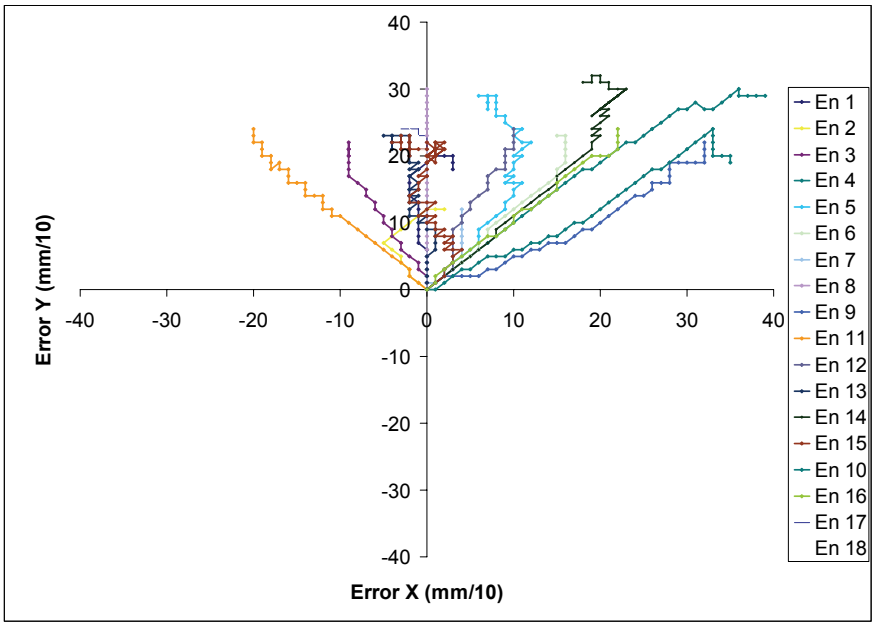


Figure 19. Trajectories for circular shamfered insertion

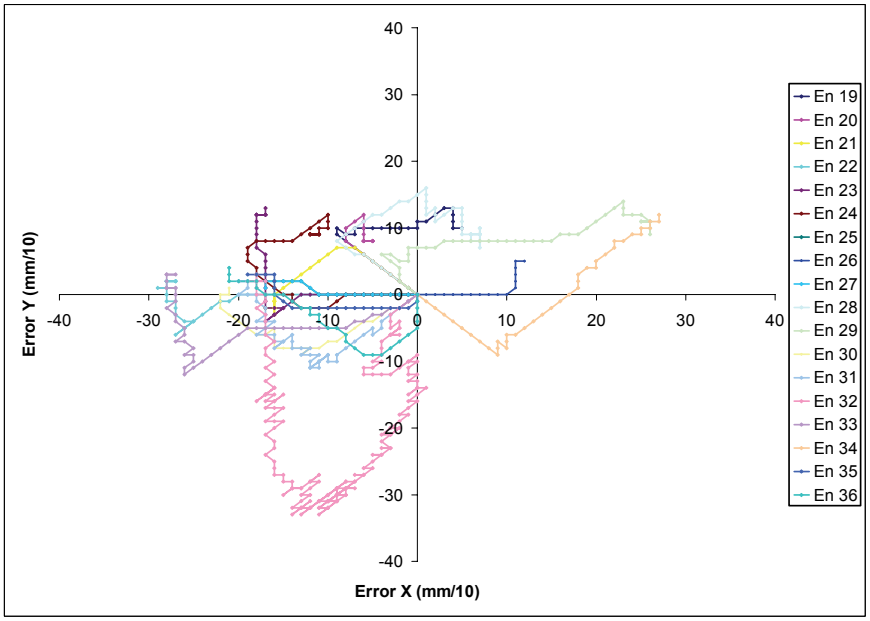


Figure 20. Trajectories for circular chamferless insertion

8. Conclusion and Future Work

We have explained how the distributed system has been structured to perform robotic assembly operations aided by visual and contact force sensing information. The multimodal architecture M2ARTMAP was simulated in previous work, where global results motivated the implementation of the system by including visual and tactile information in two modules.
The current system has been tested in an intelligent manufacturing system. SIEM and SIRIO modules were incorporated successfully. Still further work is envisaged to fuse both visual and contact force sensing information as well as to include redundant and complementary information sensors.

Acknowledgements

The authors wish to thank the following organizations who made possible this research through different funding schemes: Deutscher Akademischer Austausch Dienst (DAAD), Consejo Nacional de Ciencia y Tecnologia (CONACyT) and the Consejo de Ciencia y Tecnologia del Estado de Queretaro. (CONCyTEQ).

10. References

Amoretti, Michele, Stefano Bottazzi, Monica Reggiani, Stefano Caselli., (2003). "Evaluation of Data Distribution Techniques in a CORBA-based Telerobotic System" *Proc. of the 2003 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS 2003),* October, Las Vegas, NV.

Amoretti, Michele, Stefano Bottazzi, Stefano Caselli, Monica Reggiani, (2004), "Telerobotic Systems Design based on Real-Time CORBA", *Journal of Robotic Systems* Volume 22, Issue 4 , PP. 183 – 201.

Asfour, Y.R., Carpenter, G.A., Grossberg, S., Lesher, G.W. (1993). Fusion ARTMAP: An adaptive fuzzy network for multi-channel classification. *In: Third International Conference on Industrial Fuzzy Control and Intelligent Systems* [IFIS-93], IEEE Press 155–160

Barney Dalton, Ken Taylor, (2000). "Distributed Robotics over the Internet", *IEEE Robotics and Automation.* 7(2): 22-27.

Bottazzi, S., S. Caselli, M. Reggiani, M. Amoretti, (2002). "A Software Framework based on Real-Time CORBA for Telerobotic Systems", *Proceedings of*

the 2002 IEEE/RSJ Int. Conference on Intelligent Robots and Systems, EPFL, Lausanne, Switzerland, October.

Birney, Ewan, Michael Lausch, Todd Lewis, Stéphane Genaud, and Frank Rehberger (2003). *ORBit Beginners Documentation V1.6*

Carpenter, G.A., Grossberg, S., Iizuka, K. (1992a). Comparative performance measures of fuzzy ARTMAP, learned vector quantization, and back propagation for handwritten character recognition. In: *International Joint Conference on Neural Networks.* Volume 1., IEEE (1992) 794–799.

Carpenter, G.A., Grossberg, J., Markunzon, N., Reynolds, J.H., Rosen, D.B. (1992b). Fuzzy ARTMAP: a neural network architecture for incremental learning of analog multidimensional maps. *IEEE Trans. Neural Networks* Vol. 3 No. 5 678-713.

Carpenter, G.A., Streilein, W.W. (1998). ARTMAP-FTR: a neural network for fusion target recognition with application to sonar classification. In: AeroSense: *Proceedings of SPIE's 12th Annual Symposium on* Aerospace/*Defense Sensing, Simulation, and Control. SPIE Proceedings*, Society of Photo-Optical Instrumentation Engineers.

Corona-Castuera, J., I Lopez-Juarez, (2004). "Intelligent Task Level Planning for Robotic Assembly: Issues and Experiments" Mexican International Conference on Artificial Inteligence (MICAI'2004) *Lecture Notes on Computer Science,* Springer Verlag, ISBN 3-540-21459-3.

Corona-Castuera, J. & Lopez-Juarez, I. (2006). Distributed Architecture for Intelligent Robotic Assembly, Part II: Design of the Task Planner. ADVANCED TECHNOLOGIES: Research-Development-Application. Submitted for publication.

Distributed Systems Research Group. "CORBA comparison Project", final report Charles University, Prague, Czech Republic. August 16, 1999.

Fernandez-Delgado, M., Barro Amereiro, S. (1998). MART: A multichannel art-based neural network. *IEEE Transactions on Neural Networks* 9 139–150

Ginnari, J.H., Langley, P., Fisher, D. (1992). : Quadruped mammals. Found as Quadruped Animals. *Data Generator at UCI Machine Learning Repository* http://www.ics.uci. edu/~mlearn/MLRepository.html.

Henning, Michi, Steve Vinoski, (2002). "Programación Avanzada en CORBA con C++", Addison Wesley, ISBN 84-7829-048-6.

Jia, Songmin, Yoshiro Hada, Gang Ye, Kunikatsu Takase, (2002). "Distributed Telecare Robotic Systems Using CORBA as a Communication Architec-

ture" *IEEE International Conference on Robotics & Automation,* Washington, DC.

Jia, Yoshiro Hada, Kunikatsu Takase, (2003). "Development of a Network Distributed Telecare Robotic System Using CORBA," *Proceedings of the 2003 IEEE Int. Conference on Robotics, Intelligent Systems and Signal Processing,* Changsha, China, October.

Lopez-Juarez, I; J. Corona-Castuera, M. Peña-Cabrera, K. Ordaz-Hernandez, (2005a), "On The Design of Intelligent Robotic Agents for Assembly", In special issue on Intelligent Embedded Agents", *Journal of Information Sciences.* Elsevier 171(2005) 377-402.

Lopez-Juarez, I.; K. Ordaz-Hernandez, M. Peña-Cabrera, J. Corona-Castuera and R. Rios-Cabrera, (2005b). "On The Design Of A Multimodal Cognitive Architecture for Perceptual Learning in Industrial Robots," *Mexican Int. Conf. on Artificial Intelligence, (MICAI 2005), LNAI 3789*, PP.1052-1061 Springer-Verlag Berlin Heidelberg.

Martens, S., Gaudiano, P., Carpenter, G.A. (1998). Mobile robot sensor integration with fuzzy ARTMAP. *In: IEEE ISIC/CIRA/ISAS Joint Conference, IEEE*

Object Management Group, (2000). *The Common Object Request Broker: Architecture and Specification*, Revision 2.4, October 2000.

Parsons, O., Carpenter, G.A. (2003). Artmap neural networks for information fusion and data mining: Map production and target recognition methodologies. Neural Networks 16.

Peña-Cabrera, Mario, Ismael Lopez Juarez, Reyes Rios Cabrera, Roman Osorio, (2004). "Un Proceso de Aprendizaje para Reconocimiento de Objetos en Línea en Tareas Robotizadas", *3ª Conferencia Iberoamericana en Sistemas, Cibernética e Informática (CISCI 2004),* Orlando, Florida, EE.UU., ISBN: 980-6560-15-9.

Peña-Cabrera, M. & Lopez-Juarez, I. (2006). Distributed Architecture for Intelligent Robotic Assembly, Part III: Design of the Invariant Object Recognition System. ADVANCED TECHNOLOGIES: Research-Development-Application. Submitted for publication.

Ríos-Cabrera R., Peña-Cabrera M., Goñi-Hernández F., Lopez-Juarez I., (2004a)., "Object Recognition Methodology for Part Grasping in a Manufacturing Cell", *International Symposium on Robotics and Automation (ISRA'2004),* Querétaro Qro., ISBN: 970-9702-00-9.

Ríos-Cabrera, R., (2004b). "Distribución de datos en una celda de manufactura flexible", *Reporte interno CIATEQ, A.C.* 2do. Sem. 2004, proy. 620088.

Rios-Cabrera, R., Lopez-Juarez I., Corona-Castuera J, Chaparro-Sanchez R, Peña-Cabrera M, (2005). "Integración de Lenguaje Natural, Visión y Sensado de Fuerzas en una Celda de Manufactura Flexible", *4ª Conferencia Iberoamericana en Sistemas, Cibernética e Informática (CISCI 2005),* Orlando, Florida, EE.UU., ISBN: 980-6560-26-4

Thorpe, J., McEliece, R. (2002). Data fusion algorithms for collaborative robotic exploration. *Progress Report* 42-149, California Institute of Technology.

Yang, S., Chang, K.C. (1998). Multimodal pattern recognition by modular neural network. Optical Engineering 37. 650–659.

Wu, L., S. L. Oviatt, P. R. Cohen, (1999)., "Multimodal Integration – A Statical View", *IEEE Transactions on Multimedia*, vol 1 , Num. 4, pp 334-341.

13

Distributed Architecture for Intelligent Robotic Assembly Part II: Design of the Task Planner

Jorge Corona-Castuera and Ismael Lopez-Juarez

1. Introduction

In previous chapter it has been described the overall architecture for multimodal learning in the robotic assembly domain (Lopez-Juarez & Rios-Cabrera, 2006). The acquisition of assembly skills by robots is greatly supported by the effective use of contact force sensing and objects recognition. In this chapter, we will describe the robot's ability to acquire and refine its knowledge through operations (i.e. using contact force sensing during fine motions) and how a manipulator can effectively learn the assembly skill starting from scratch.

The use of sensing to reduce uncertainty significantly extends the range of possible tasks. One source of uncertainty is that the programmer's model of the environment is incomplete. Shape, location, orientation and contact states have to be associated to movements within the robot's motion space while it is in constraint motion. Compliant motion meets external constraints by specifying how the robot's motion should be modified in response generated forces when constraints are violated. Generalizations of this principle can be used to accomplish a wide variety of tasks involving constrained motion, e.g., inserting a peg into a hole or following a weld seam under uncertainty.

The success of robotic assembly operations therefore, is based on the effective use of compliant motion, the accuracy of the robot itself and the precise knowledge of the environment, i.e. information about the geometry of the assembly parts and their localisation within the workspace. However, in reality uncertainties due to manufacturing tolerances, positioning, sensing and control make it difficult to perform the assembly. Compliant motion can be achieved by using passive devices such as the Remote Centre Compliance (RCC) introduced by Whitney (Whitney & Nevis, 1979) or other improved versions of the device (Joo & Miyasaki, 1998). Other alternative is to use Active Compliance, which actually modifies either the position of the manipulated component as a response to constraint forces or the desired force. Some com-

mercial devices have emerged in recent years to aid industrial applications (Erlbacher, 2004).

Active compliance can be roughly divided into fine motion planning and reactive control. Fine motion planning relies on geometrical path planning whereas reactive control on the synthesis of an accommodation matrix or mapping that transform the corresponding contact states to corrective motions. A detailed analysis of active compliance can be found in (Mason, 1983) and (De Schutter & Brussel, 1988). Perhaps, one of the most significant works in fine motion planning is the work developed by Lozano-Perez, Mason and Taylor known as the LMT approach (Lozano-Perez, et al, 1984). The LMT approach automatically synthesizes compliant motion strategies from geometric descriptions of assembly operations and explicit estimates of the errors in sensing and control. Approaches within fine motion planning can also be further divided into model-based approaches and connectionist-based approaches though, some reactive control strategies can be well accommodated within the model-based approach. In either case, a distinctive characteristic in model-based approaches is that these take as much information of the system and environment as possible. This information includes localisation of the parts, part geometry, material types, friction, errors in sensing, planning, and control, etc. On the other hand, the robustness of the connectionist-based approaches relies on the information given during the training stage that implicitly considers all the above parameters.

In this chapter we present a "Task Planner", connectionist-based approach that uses vision and force sensing for robotic assembly when assembly components geometry, location and orientation is unknown at all times. The assembly operation resembles the same operation as carried out by a blindfold human operator. The task planner is divided in four stages as suggested in (Doersam & Munoz, 1995) and (Lopez-Juarez, 2000):

Pre-configuration: From an initial configuration of the hand/arm system, the expected solutions are the required hand/arm collision-free paths in which the object can be reached. To achieve this configuration, it is necessary to recognize invariantly the components and determining their location and orientation.

Grasp: Once the hand is in the Pre-configuration stage, switching strategies between position/force controls need to be considered at the moment of contact and grasping the object. Delicate objects can be broken without a sophisticated contact strategy even the Force/Torque (F/T) sensor can be damaged.

Translation: After the object is firmly grasped, it can be translated to the assembly point. The possibility of colliding with obstacles has to be taken into account.

Assembly Operation: The assembly task requires robust and reactive positions/force control strategies. Mechanical and geometrical uncertainties make high demands on the controller.

The pre-configuration for recognition and location of components as well as the assembly operation are based on FuzzyARTMAP neural network architecture, situated under the connectionist-based approach employing reactive contact forces.
In this approach, the mapping between contact states and arm motion commands is achieved by using fuzzy rules that create autonomously an Acquired-Primitive Knowledge Base (ACQ-PKB) without human intervention. This ACQ-PKB is then further used by the Neural Network Controller (NNC) for compliance learning.

2. Related Work

The use of connectionist models in robot control to solve the problem under uncertainty has been demonstrated in a number of publications, either in simulations (Lopez-Juarez & Howarth, 1996), (Asada, 1990), (Cervera & del Pobil, 1996), or being implemented on real robots (Cervera & del Pobil, 1997), (Gullapalli, et al, 1994), (Howarth, 1998), (Cervera & del Pobil, 2002). In these methods, Reinforcement Learning (RL), unsupervised and supervised type networks have been used.

The reinforcement algorithm implemented by V. Gullapalli demonstrated to be able to learn circular and square peg insertions. The controller was a back-propagation network with 11 inputs. These are the sensed positions and forces: (X, Y, Z, θ_1, θ_2) and (F_x, F_y, F_z, m_x, m_y, m_z). The output of the network was the position commands. The performance of the operation was evaluated by a parameter r, which measured the performance of the controller. r varied between 0 to 1 and was a function of the sensed peg position and the nominal hole location. The network showed a good performance after 150 trials with insertion times lower than 100 time steps (Gullapalli, 1995). Although the learning capability demonstrated during experiments improved over time, the network was unable to generalise over different geometries. Insertions are reported with

both circular and square geometries; however, when inserting the square peg, its rotation around the vertical axis was not allowed, which facilitated the insertion. M. Howarth followed a similar approach, using also backpropagation in combination with reinforcement learning. In comparison with Gullapalli's work, where the reinforcement learning values were stochastic, Howarth's reinforcement value was based on two principles: minimization of force and moment values and continuation of movement in the assembly direction. This implied that whenever a force or moment value was above a threshold, an action (i.e., reorientation), should occur to minimize the force. Additionally, movements in the target assembly direction were favoured. During simulation it was demonstrated that 300 learning cycles were needed to achieve a minimum error level with his best network topology during circular insertions (Howarth, 1998). A cycle meant to be an actual motion that diminished the forces acting on the peg. For the square peg, the number of cycles increased dramatically to 3750 cycles. These figures are important, especially when fast learning is desired during assembly.

On the other hand, E. Cervera using SOM networks and a Zebra robot (same used by Gullapalli) developed similar insertions as the experiments developed by Gullapalli. Cervera in comparison with Gullapalli improved the autonomy of the system by obviating the knowledge of the part location and used only relative motions. However, the trade-off with this approach was the increment of the number of trials to achieve the insertion (Cervera & del Pobil, 1997); the best insertions were achieved after 1000 trials. During Cervera's experiments the network considered 75 contact states and only 8 out of 12 possible motion directions were allowed. For square peg insertions, there were needed 4000 trials to reach 66% success of insertion with any further improvement. According to Cervera's statement, "We suspect that the architecture is suitable, but the system lacks the necessary information for solving the task", the situation clearly recognises the necessity to embed new information in the control system as it is needed.

Other interesting approaches have also been used for skill acquisition within the framework of Robot Programming by Demonstration that considers the characteristics of human generated data. Work carried out by (Kaiser & Dillman, 1996) shows that skills for assembly can be acquired through human demonstration. The training data is first pre-processed, inconsistent data pairs are removed and a smoothing algorithm is applied. Incremental learning is achieved through Radial Basis Function Networks and for the skill refinement;

the Gullapalli's Stochastic Reinforcement Value was also used. The methodology is demonstrated by the peg-in-hole operation using the circular geometry. On the other hand (Skubic & Volz, 2000 b), use a hybrid control model which provides continuous low-level force control with higher-level discrete event control. The learning of an assembly skill involves the learning the mapping of force sensor signals to Single-Ended Contact Formations (SECF), the sequences of SECFs and the transition velocity commands which move the robot from the current SECF to the next desired SECF. The first function is acquired using supervised learning. The operator demonstrates each SECF while force data is collected, and de data is used to train a state classifier. The operator then demonstrates a skill, and the classifier is used to extract the sequence of SECFs and transitions velocities which comprise the rest of the skill.

The above approaches can be divided in two groups, those providing autonomous assembly skill and those which teach the skill by demonstration. These approaches have given some inputs to our research and the work presented here is looking to improve some of their limitations. In Gullapalli's work the hole location has to be known. Howarth improved the autonomy by obviating the hole's location; however, the lengthy training process made this approach impractical. Cervera considered many contact states, which worked well also during the assembly of different type of components. In the case of teaching the skill by demonstration, the method showed by Kaiser and Dillman was lengthy for real-world problems and the work by Skubic and Volz assumes that during supervised training the operator must know which SECF classes to include in the set.
The integration of vision systems to facilitate the assembly operations in uncalibrated workspaces is well illustrated in (Jörg, et al, 2000) and (Baeten, et al, 2003) using eye-in-hand vision for different robotic tasks.

3. Workplace Description

The manufacturing cell used for experimentation is integrated by a KUKA KR15/2 industrial robot. It also comprises a visual servo system with a ceiling mounted camera as shown in figure 1. The robot grasps the male component from a conveyor belt and performs the assembly task in a working table where the female component is located. The vision system gets an image to calculate the object's pose estimation and sends the information to the robot from two predefined zones:

Zone 1 which is located on the conveyor belt. The vision system searches for the male component and determines the pose information needed by the robot.
Zone 2 is located on the working table. Once the vision system locates the female component, it sends the information to the NNC.

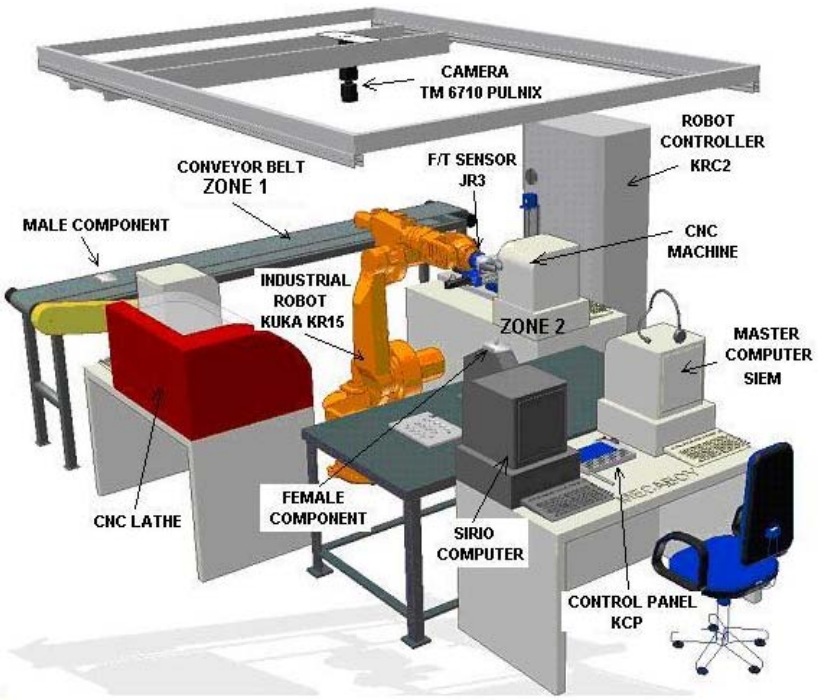


Figure 1. Manufacturing cell

The NNC for assembly is called SIEM (Sistema Inteligente de Ensamble Mecánico) and is based on a FuzzyARTMAP neural network working in fast learning mode (Carpenter, et al, 1992). The vision system, called SIRIO (Sistema Inteligente de Reconocimiento Invariante de Objetos), also uses the same neural network to learn and classify the assembly components (Pena-Cabrera & Lopez-Juarez, 2006). The SIRIO was implemented with a high speed camera CCD/B&W, PULNIX 6710, with 640x480 resolution; camera movements on the X and Y axis were implemented using a 2D positioning system.

For experimental purposes three canonical peg shapes were used: circular, square and radiused-square as it is shown in figure 2. Both, chamfered and chamferless female components were employed during experimentation.

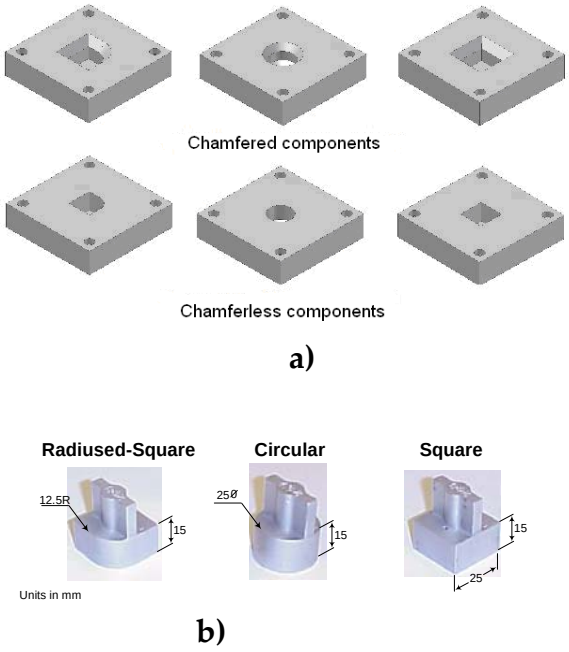


Figure 2. a) Female assembly components, b) Male assembly components

4. Assembly methodology

4.1 Pre-Configuration

4.1.1 Starting from scratch

Initially, the robot system does not have any knowledge. To accomplish the very first assembly the robot has to acquire a Primitive Knowledge Base (PKB) using an interactive method.

a) Given Primitive Knowledge Base (GVN-PKB)
The formation of the PKB basically consists of showing the robot how to react to individual components of the F/T vector. This procedure results in creating the required mapping between contact states and robot motions within the motion space– linear, angular and diagonal movements- , this is illustrated in figure 3. The Given PKB (GVN-PKB) used for the experiments reported in this chapter considered rotation around Z axis and diagonal motions as it is illustrated in figure 4.

Using the above mentioned GVN-PKB to start the learning of the assembly skill, it showed to be effective, however the robot still lacked for autonomy and it was realized that sometimes the robot did not used all the information given in the GVN-PKB and also it was noticed a difference between the taught contact forces the actual forces occurring during assembly so that an autonomously created PKB was needed in order to provide complete self-adaptive behaviour to the robot.

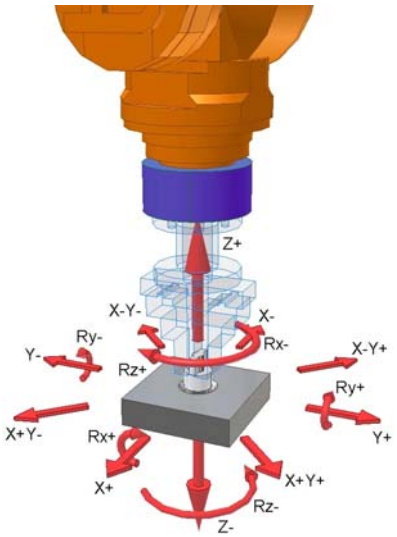


Figure 3. Motion space

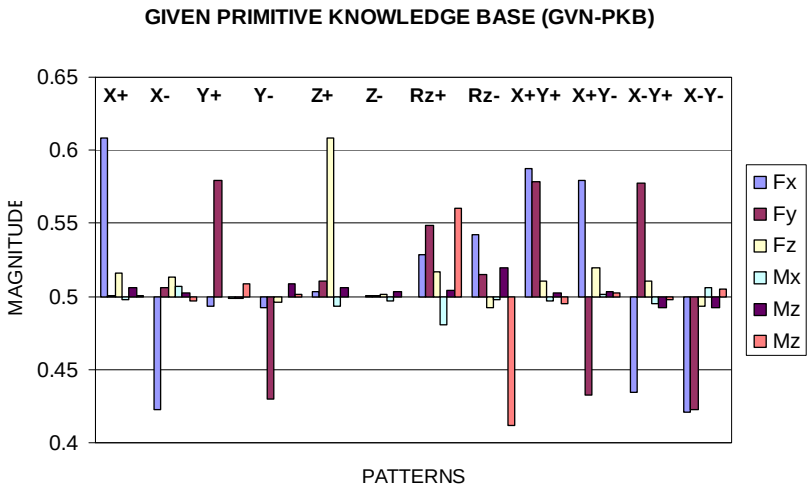


Figure 4. Given PKB (GVN-PKB)

b) Acquired Primitive Knowledge Base (ACQ-PKB)

It was decided to embed a fuzzy logic mechanism to autonomously acquire an initial knowledge from the contact states. That is, learning the mapping from scratch without knowledge about the environment. The only instruction given to the robot was the task – assembly – in order to start moving downwards. When the contact is made the robot starts acquiring information about the contact states following fuzzy rules and autonomously generating the corresponding motion commands and forming the Acquired PKB (ACQ-PKB). During the first contact, the fuzzy algorithm determines the type of operation: chamfered or chamferless assembly and chooses the rules to apply depending of moments and forces magnitude presents in X and Y directions.

Fuzzy logic have proved to be useful to model many decision taking processes in presence of uncertainty or where no precise knowledge of the process exist in an attempt to formalize experience and empiric knowledge of the experts in a specific process. The initial knowledge from our proposal comes from a static and dynamic force analysis when the components are in contact assuming that there is an error in the position with respect to the centre of insertion. With the aid of dynamic simulation software (ADAMS), the behaviour of the contact impact is obtained for different situations which are to be solved by the movements of the manipulator.

There are 12 defined motion directions (X+, X-, Y+, Y-, Z+, Z-, Rz+, Rz-, X+Y+, X+Y-, X-Y+ and X-Y-) and for each one there is a corresponding contact state. An example of these contact states for a chamfered female squared component is shown in figure 5. The contact states for linear motion X+, X-, Y+, Y-, and linear combined motions X+Y+, X+Y-, X-Y+, X-Y- are shown in figure 5(a). In figure 5(b), it is shown a squared component having four contact points. Figures 5(c) and 5(d) provide additional patterns for rotation Rz- and Rz+ respectively when the component has only one point of contact. The contact state for mapping Z+ is acquired making vertical contact between component and a horizontal surface, Z- direction is acquired with the component is in free space. This approach applies also for chamfered circular and radius-squared components as well as the chamferless components.

It is stated to use the following considerations for the generation of the fuzzy rules: a) Number of linguistic values: 2 (minimum, maximum), b) Number of input variables: 12 (*Fxp, Fxn, Fyp, Fyn, Fzp, Fzn, Mxp, Mxn, Myp, Myn, Mzp, Mzn*) and c) Maximum number of rules: $12^2 = 144$ (only 24 were used).

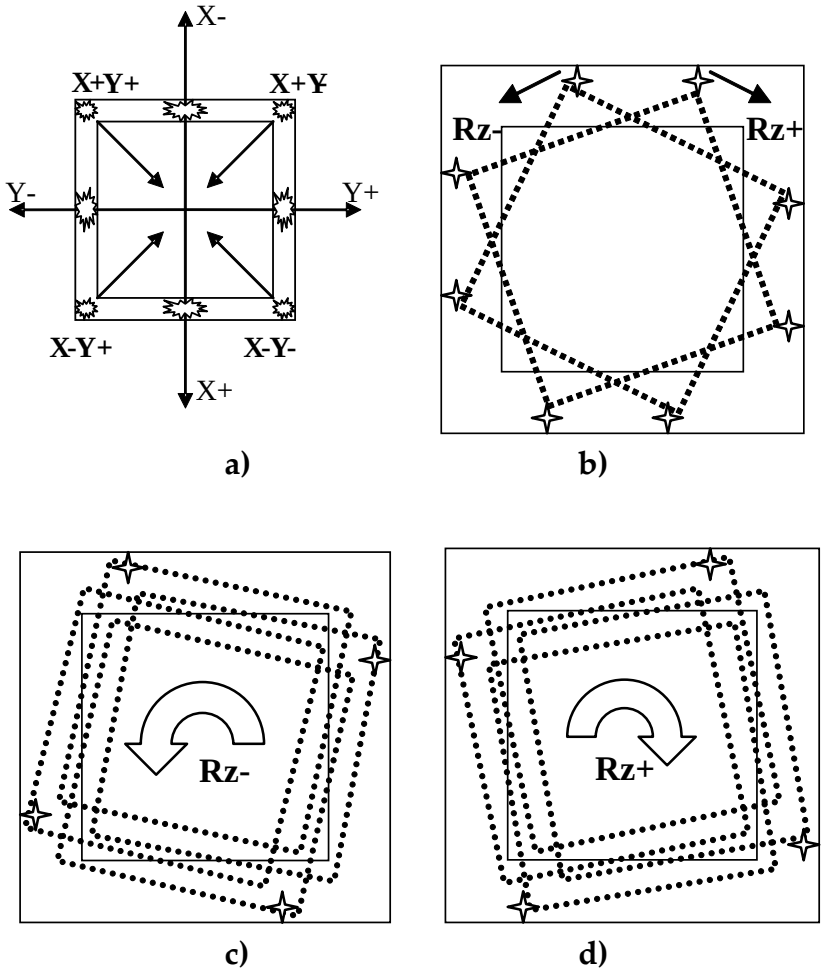


Figure 5. Contacts between chamfered components while acquiring the primitive knowledge base,

a) Linear movements,
b) Pure rotation Rz+ and Rz-,
c) Rotation Rz-,
d) Rotation Rz+.

The membership functions are stated as showed in figure 6. Forces and moments have normalised values between 0 and 1. The normalization was *ad-hoc* and considered the maximum experimental value for both, force and moment values. No belong functions were defined for the output, because our process does not includes defuzzification in the output. The function limit values are chosen heuristically and according to previous experience in the assembly operation.

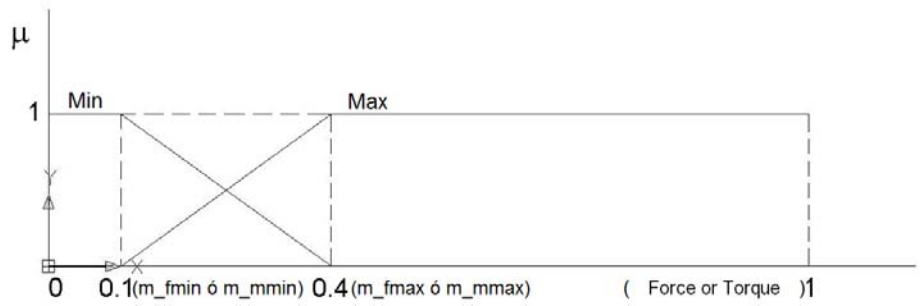


Figure 6. Membership functions

Having those membership values, antecedents and consequents defined, then the Rule Statement can be generated and the ACQ-PKB created. An example of these rules for chamfered assembly is given in table 1.

IF	Fxp	Fxn	Fyp	Fyn	Fzp	Fzn	Mxp	Mxn	Myp	Myn	Mzp	Mzn	THEN	DIR
IF	Max	Min	Min	Min	Max	Min	Min	Min	Max	Min	Min	Min	*THEN*	X+
IF	Max	Min	Min	Min	Max	Min	Max	Min	Max	Min	Min	Min	*THEN*	X+
IF	Min	Max	Min	Min	Max	Min	Min	Min	Min	Max	Min	Min	*THEN*	X-
IF	Min	Max	Min	Min	Max	Min	Min	Max	Min	Max	Min	Min	*THEN*	X-
IF	Min	Min	Max	Min	Max	Min	Min	Max	Min	Min	Min	Min	*THEN*	Y+
IF	Min	Min	Max	Min	Min	Min	Min	Min	Min	Max	Min	Min	*THEN*	Y+
IF	Min	Min	Min	Max	Max	Min	Max	Min	Min	Max	Min	Min	*THEN*	Y-
IF	Min	Min	Min	Max	Max	Min	Min	Min	Min	Min	Min	Min	*THEN*	Y-
IF	Min	Min	Min	Min	Max	Min	Min	Min	Min	Min	Min	Min	*THEN*	Z+
IF	Min	Min	Min	Min	Min	Min	Min	Min	Min	Min	Min	Min	*THEN*	Z-
IF	Min	Min	Min	Min	Max	Min	Min	Min	Min	Min	Max	Min	*THEN*	Rz+
IF	Max	Min	Min	Min	Max	Min	Min	Min	Max	Min	Max	Min	*THEN*	Rz+
IF	Min	Max	Min	Min	Max	Min	Min	Min	Min	Max	Max	Min	*THEN*	Rz+
IF	Min	Min	Max	Min	Max	Min	Min	Max	Min	Min	Max	Min	*THEN*	Rz+
IF	Min	Min	Min	Max	Max	Min	Max	Min	Min	Min	Max	Min	*THEN*	Rz+
IF	Min	Min	Min	Min	Max	Min	Min	Min	Min	Min	Min	Max	*THEN*	Rz-
IF	Max	Min	Min	Min	Max	Min	Min	Min	Max	Min	Min	Max	*THEN*	Rz-
IF	Min	Max	Min	Min	Max	Min	Min	Min	Min	Max	Min	Max	*THEN*	Rz-
IF	Min	Min	Max	Min	Max	Min	Min	Max	Min	Min	Min	Max	*THEN*	Rz-
IF	Min	Min	Min	Max	Max	Min	Max	Min	Min	Min	Min	Max	*THEN*	Rz-
IF	Max	Min	Max	Min	Max	Min	Min	Max	Max	Min	Min	Min	*THEN*	X+Y+
IF	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Min	Min	*THEN*	X+Y-
IF	Min	Max	Min	Max	Max	Min	Min	Max	Min	Max	Min	Min	*THEN*	X-Y+
IF	Min	Max	Min	Max	Max	Min	Max	Min	Min	Max	Min	Min	*THEN*	X-Y-

Table 1. Fuzzy rules for chamfered assembly

For chamferless assembly another knowledge base would have to be generated using similar rules as shown above, but without considering force in axis X and Y. The reason is that these forces in comparison with the moments generated around those axes are very small. The inference machine determines the rules to apply in a given case.

To quantify the fuzzy output response a fuzzy logic membership value is used. For the "AND" connector we used the product criteria (Driankov, et al, 1996), and to obtain a conclusion, the maximum value for the fuzzy outputs in the expression (1) response was used.

$$
\begin{aligned}
&X+ = Fxp_{max} * Fxn_{min} * Fyp_{min} * Fyn_{min} * Fzp_{max} * Mxp_{min} * Mxn_{min} * Myp_{max} * Myn_{min} * Mzp_{min} * Mzn_{min}\\
&X+ = Fxp_{max} * Fxn_{min} * Fyp_{min} * Fyn_{min} * Fzp_{max} * Mxp_{max} * Mxn_{min} * Myp_{max} * Myn_{min} * Mzp_{min} * Mzn_{min}\\
&X- = Fxp_{min} * Fxn_{max} * Fyp_{min} * Fyn_{min} * Fzp_{max} * Mxp_{min} * Mxn_{min} * Myp_{min} * Myn_{max} * Mzp_{min} * Mzn_{min}\\
&X- = Fxp_{min} * Fxn_{max} * Fyp_{min} * Fyn_{min} * Fzp_{max} * Mxp_{min} * Mxn_{max} * Myp_{min} * Myn_{max} * Mzp_{min} * Mzn_{min}\\
&Y+ = Fxp_{min} * Fxn_{min} * Fyp_{max} * Fyn_{min} * Fzp_{max} * Mxp_{min} * Mxn_{max} * Myp_{min} * Myn_{min} * Mzp_{min} * Mzn_{min}\\
&Y+ = Fxp_{min} * Fxn_{min} * Fyp_{max} * Fyn_{min} * Fzp_{min} * Mxp_{min} * Mxn_{min} * Myp_{min} * Myn_{max} * Mzp_{min} * Mzn_{min}\\
&Y- = Fxp_{min} * Fxn_{min} * Fyp_{min} * Fyn_{max} * Fzp_{max} * Mxp_{max} * Mxn_{min} * Myp_{min} * Myn_{max} * Mzp_{min} * Mzn_{min}\\
&Y- = Fxp_{min} * Fxn_{min} * Fyp_{min} * Fyn_{max} * Fzp_{max} * Mxp_{min} * Mxn_{min} * Myp_{min} * Myn_{min} * Mzp_{min} * Mzn_{min}\\
&Z+ = Fxp_{min} * Fxn_{min} * Fyp_{min} * Fyn_{min} * Fzp_{max} * Mxp_{min} * Mxn_{min} * Myp_{min} * Myn_{min} * Mzp_{min} * Mzn_{min}\\
&Z- = Fxp_{min} * Fxn_{min} * Fyp_{min} * Fyn_{min} * Fzp_{min} * Mxp_{min} * Mxn_{min} * Myp_{min} * Myn_{min} * Mzp_{min} * Mzn_{min}\\
&Rz+ = Fxp_{min} * Fxn_{min} * Fyp_{min} * Fyn_{min} * Fzp_{max} * Mxp_{min} * Mxn_{min} * Myp_{max} * Myn_{min} * Mzp_{max} * Mzn_{min}\\
&Rz+ = Fxp_{max} * Fxn_{min} * Fyp_{min} * Fyn_{min} * Fzp_{max} * Mxp_{min} * Mxn_{min} * Myp_{max} * Myn_{min} * Mzp_{max} * Mzn_{min}\\
&Rz+ = Fxp_{min} * Fxn_{max} * Fyp_{min} * Fyn_{min} * Fzp_{max} * Mxp_{min} * Mxn_{min} * Myp_{min} * Myn_{max} * Mzp_{max} * Mzn_{min}\\
&Rz+ = Fxp_{min} * Fxn_{min} * Fyp_{max} * Fyn_{min} * Fzp_{max} * Mxp_{min} * Mxn_{max} * Myp_{min} * Myn_{min} * Mzp_{max} * Mzn_{min}\\
&Rz+ = Fxp_{min} * Fxn_{min} * Fyp_{min} * Fyn_{max} * Fzp_{max} * Mxp_{max} * Mxn_{min} * Myp_{min} * Myn_{min} * Mzp_{max} * Mzn_{min}\\
&Rz- = Fxp_{min} * Fxn_{min} * Fyp_{min} * Fyn_{min} * Fzp_{max} * Mxp_{min} * Mxn_{min} * Myp_{min} * Myn_{min} * Mzp_{min} * Mzn_{max}\\
&Rz- = Fxp_{max} * Fxn_{min} * Fyp_{min} * Fyn_{min} * Fzp_{max} * Mxp_{min} * Mxn_{min} * Myp_{max} * Myn_{min} * Mzp_{min} * Mzn_{max}\\
&Rz- = Fxp_{min} * Fxn_{max} * Fyp_{min} * Fyn_{min} * Fzp_{max} * Mxp_{min} * Mxn_{min} * Myp_{min} * Myn_{max} * Mzp_{min} * Mzn_{max}\\
&Rz- = Fxp_{min} * Fxn_{min} * Fyp_{max} * Fyn_{min} * Fzp_{max} * Mxp_{min} * Mxn_{max} * Myp_{min} * Myn_{min} * Mzp_{min} * Mzn_{max}\\
&Rz- = Fxp_{min} * Fxn_{min} * Fyp_{min} * Fyn_{max} * Fzp_{max} * Mxp_{max} * Mxn_{min} * Myp_{min} * Myn_{min} * Mzp_{min} * Mzn_{max}\\
&X+Y+ = Fxp_{max} * Fxn_{min} * Fyp_{max} * Fyn_{min} * Fzp_{max} * Mxp_{min} * Mxn_{max} * Myp_{max} * Myn_{min} * Mzp_{min} * Mzn_{min}\\
&X+Y- = Fxp_{max} * Fxn_{min} * Fyp_{min} * Fyn_{max} * Fzp_{max} * Mxp_{max} * Mxn_{min} * Myp_{max} * Myn_{min} * Mzp_{min} * Mzn_{min}\\
&X-Y+ = Fxp_{min} * Fxn_{max} * Fyp_{max} * Fyn_{min} * Fzp_{max} * Mxp_{min} * Mxn_{max} * Myp_{min} * Myn_{max} * Mzp_{min} * Mzn_{min}\\
&X-Y- = Fxp_{min} * Fxn_{max} * Fyp_{min} * Fyn_{min} * Fzp_{max} * Mxp_{max} * Mxn_{min} * Myp_{min} * Myn_{max} * Mzp_{min} * Mzn_{min}
\end{aligned}
\tag{1}
$$

Once the algorithm values have been generated, a routine which allows the manipulator for autonomous database generation is created. The mapping acquisition between generated contact states-arm motion commands starts from the insertion centre. This information is determined by calculating the centroid of the component by the vision system. Positional errors due to the image

processing are about 1 mm to 2 mm which were acceptable for the experimental work since the assembly was always successful. The manipulator starts moving in every possible direction generating a knowledge database. The results given in this research considered only 24 patterns as indicated in the fuzzy rules shown in table 1, omitting the rotations around the X and Y axis since only straight insertions were considered. Some patterns generated with this procedure for the chamfered and chamferless square peg insertion can be observed in figure 7.

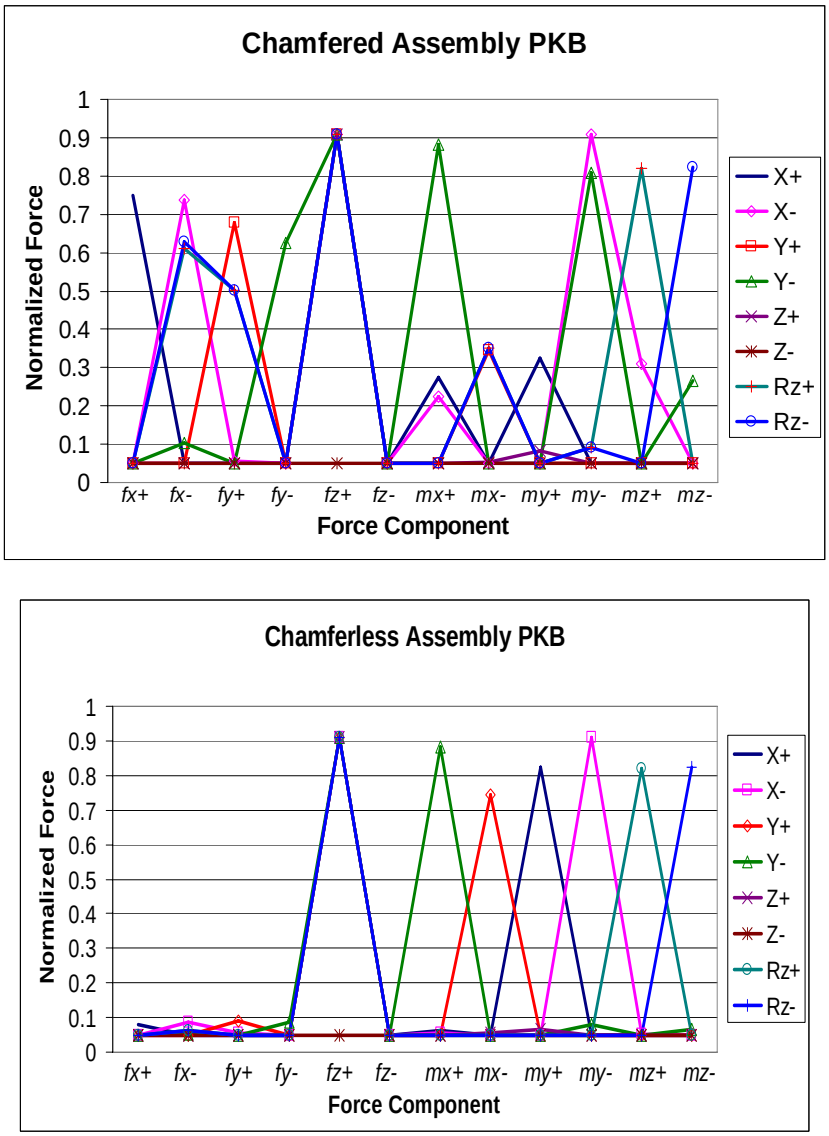


Figure 7. ACQ-PKB, left chamfered assembly, right chamferless assembly

In order to get the next motion direction the forces are read, normalized and classified using the NNC on-line. The F/T pattern obtained from the sensor provides a unique identification. The F/T vector (2) comprises 12 components given by the 6 data values (positive and negative).

$$[Current\,F/T] = \left[fx,\ fx-,\ fy,\ fy-,\ fz,\ fz-, mx, mx-, my, my-, mz, mz-\right]^T \quad (2)$$

4.1.2 Acquiring location and component type

The SIRIO system employs the following methodology: a) Finding the region of interest (ROI), b) Calculate the histogram of the image, d) Search for components, e) Centroid calculation, f) Component orientation, g) Calculate Boundary Object Function (BOF), distances between the centroid and the perimeter points, h) Descriptor vector generation and normalization (CFD&POSE) and i) Information processing in the neural network.
The descriptive vector is called CFD&POSE (Current Frame Descriptor and Pose) and it is conformed by (3):

$$[CDF\ \&\ POSE] = [D_1, D_2, D_3, ..., D_n, X_c, Y_c, \theta, Z, ID]^T \quad (3)$$

Where: $\mathbf{D}_i$ are the distances from the centroid to the perimeter of the object. (180 data values)

- $\mathbf{X}_C$, $\mathbf{Y}_C$, are the centroid coordinates.
- ϕ, is the orientation angle.
- **Z** is the height of the object.
- **ID** is a code number related to the geometry of the components.´

With this vector and following the above methodology, the system has been able to classify invariantly 100% of the components presented on-line even if they are not of the same size, orientation or location and for different light conditions, see (Pena-Cabrera & Lopez-Juarez, 2006) for details.

The CFD&POSE vector is invariant for each component and it is used for classification. The vector is normalized to 185 data dimension and normalized in

the range [0.0 – 1.0]. The normalization of the BOF is accomplished using the maximum divisor value of the vector distance. This method allows having very similar patterns as input vectors to the neural network, getting a significant improvement in the operation system. In our experiments, the object recognition method used the above components having 210 patterns as primitive knowledge to train the neural network. It was enough to recognize the assembly components with $\varrho_a = 0.2$ (base vigilance), $\varrho_{map} = 0.7$ (vigilance map) and $\varrho_b = 0.9$ parameters, however, the SIRIO system can recognize more complex components (Pena-Cabrera, et al, 2005).

4.2 Grasp

At this stage, the PKB has been acquired and the location information sent to the robot. The motion planning from Home position to zone 1 uses the male component given coordinates provided by SIRIO. The robot uses this information and the F/T sensor readings to grasp the piece and to control the motion in Z direction for two stages:

a) The security stage
In the event that position and orientation of the male component, given by SIRIO, have an error larger than 5 mm in *X* or *Y* axis and 10º around Z direction. Sensing is executed during 10 movements in *Z*- direction with manipulator steps of 0.2 mm. In this stage a collision is possible to occur between gripper and components. The system reacts moving to home position when a force limit in Z direction is reached (4 N). The robot continues its trajectory in Z- direction until a distance of 1 mm component is reached.

b) Grasp Component
This sensing stage begins just before the robot touches the component. The sensor is read every 0.1 mm executed by manipulator, this stage ends when the robot touches the component, in this situation the force magnitude in Z direction is at least 4 N, then the condition to grasp (close gripper) is satisfied.

4.3 Translation

The translation is similar to the grasping operation in zone 1. The path to move the robot from zone 1 to zone 2 (assembly point) is accomplished by using the

coordinates given by the SIRIO system. The possibility of collision with obstacles is avoided using bounded movements.

4.4 Assembly Operation

4.4.1 Neural Network Controller (NNC)

a) ART Models

Several works published in the literature inspired ideas about contact recognition and representation (Xiao & Liu, 1998), (Ji & Xiao, 1999), (Skubic &Volz, 1996), however the fuzzy representation appealed to be suitable to expand the NNC capability and further work was envisaged to embed the automatic mechanism to consider contact states that are actually present in a specific assembly operation. It was believed that by using only useful information, compliance learning could be effective in terms of avoiding learning unnecessary contact information, hence also avoiding unnecessary motions within the motion space.

The Adaptive Resonance Theory (ART) is a well established associative brain and competitive model introduced as a theory of the human cognitive processing developed by Stephen Grossberg at Boston University. Grossberg suggested that connectionist models should be able to adaptively switch between its *plastic* and *stable* modes. That is, a system should exhibit plasticity to accommodate new information regarding unfamiliar events. But also, it should remain in a stable condition if familiar or irrelevant information is being presented. An analysis of this instability, together with data of categorisation, conditioning, and attention led to the introduction of the ART model that stabilises the memory of self-organising feature maps in response to an arbitrary stream of input patterns (Grossberg, 1976).

The theory has evolved in a series of real-time architectures for unsupervised learning, the ART-1 algorithm for binary input patterns (Carpenter & Grossberg, 1987). Supervised learning is also possible through ARTMAP (Carpenter, et al, 1991) that uses two ART-1 modules that can be trained to learn the correspondence between input patterns and desired output classes. Different model variations have been developed to date based on the original ART-1 algorithm, ART-2, ART-2a, ART-3, Gaussian ART, EMAP, ViewNET, Fusion ARTMAP, LaminART just to mention but a few.

b) NNC Architecture

The functional structure of the assembly system is illustrated in figure 8. The Fuzzy ARTMAP (FAM) (Carpenter, et al, 1992) is the heart of the NNC. The controller includes three additional modules. The Knowledge Base that stores initial information related to the geometry of the assembling parts and which is autonomously generated. The Pattern-Motion Selection module keeps track of the appropriateness of the F/T patterns to allow the FAM network to be re-trained. If this is the case, the switch SW is closed and the corresponding pattern-action provided to the FAM for on-line retraining. The selection criterion is given by expression (3), discussed next.

Future predictions will be based on this newly trained FAM network. The Automated Motion module basically is in charge of sending the incremental motion request to the robot controller and handling the communication with the Master Computer.

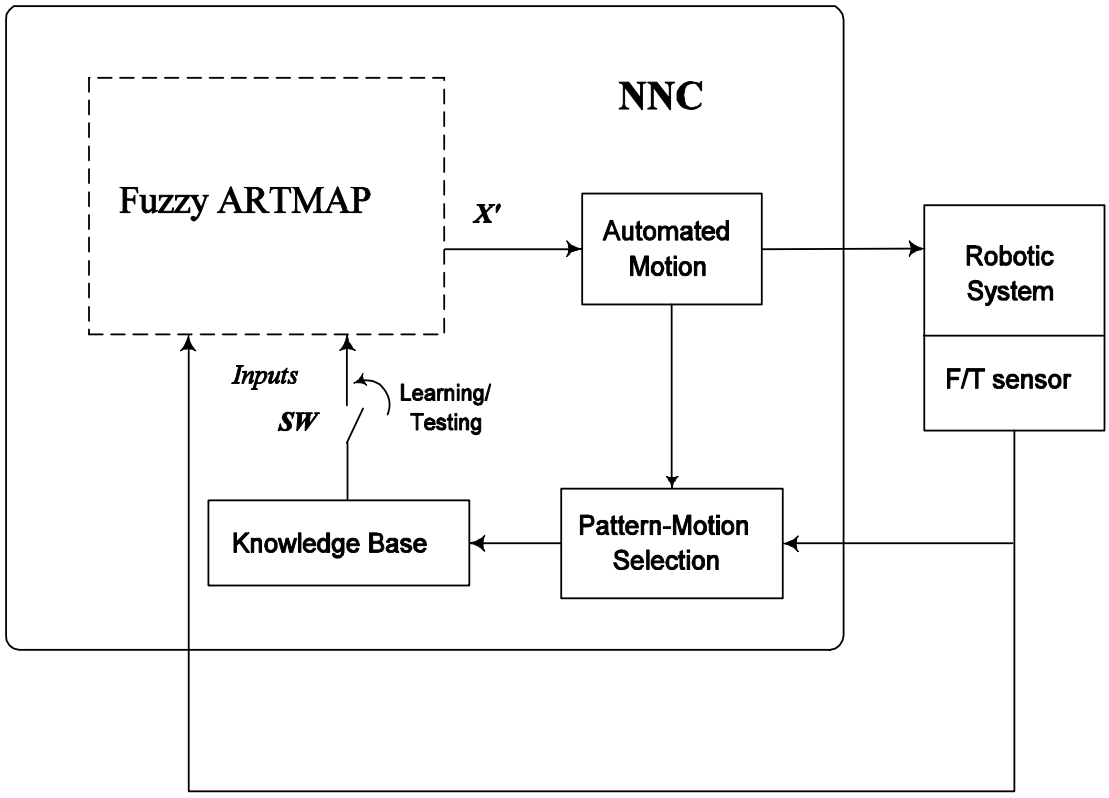


Figure 8. System Structure

c) Knowledge refinement during fine motion

There are potential overtraining problems associated with learning patterns on-line during fine motion and which are solved by the Pattern-Motion Selection module indicated in figure 8. The robot should continue moving in the insertion direction if, and only if, a minimum force value has been reached. In this situation, on-line learning is started to allow the acquisition and learning of the pattern-action pair that produced such contact state and favoured the assembly. In the event of continual learning after having reached this minimum force value, the performance of the NNC might decay. This situation is similar to what is known as overtraining, overfitting or overlearning in ANNs. At this point the learning should be stopped because if the robot learns other patterns under the above circumstances, eventually the minimum force value will be different leading to wrong motions. The same applies to the condition when the end-effector meets a force higher than the force limit. There should not be any further learning during this situation since learning a higher force would probably damage the sensor.

The above situations can be resumed in three fundamental questions:

1) How to recover from errors?
2) What is a good motion?
3) which motions should or should not be learned?

Having an assembly system which is guided by compliant motion, the criterion to decide whether the motion was good enough to be learnt is based on the following heuristic expression:

$$\left(F_{initial} - F_{after}\right) \geq 10 \qquad (4)$$

$F_{initial}$ and F_{after} are a merit figures measured before and after the corrective motion are applied and computed using the following equation as in (Ahn, et al, 1992):

$$F = \sqrt{fx^2 + fy^2 + fz^2 + s(mx^2 + my^2 + mz^2)} \qquad (5)$$

The heuristic expression (5) is used for all tasks; the scale factor s has been included in this equation in order to allow the use of different units or size com-

ponents. In our experiments, the scale factor was selected to be equal to 1 and the expression (4) is used in general for any learn task and means that if the total force after the incremental motion is significantly reduced then that pattern-action will be considered good to be included in the knowledge base.

There will be ambiguous situations in which learning should not be permitted. This applies to patterns in the insertion direction (usually Z direction). Consider downward movements in the Z- direction. At the time the peg makes contact with the female block, there may well be a motion prediction in the Z+ direction, see figure 3. This recovery action will certainly diminish the contact forces and will satisfy the condition given by the expression (4) in order to learn the force-action pair. However, this situation is redundant since it has already been given when the PKB was formed and further learning will corrupt the PKB changing probably the peg's assembly direction in Z+ instead Z-. Similarly, learning should not be allowed when the arm is in free-space. In this situation, $F_{initial}$ and F_{after} will be very similar and again learning another pattern in the Z- direction will be redundant. Both situations were tested experimentally and revealed that an unstable situation may appear if further learning is allowed. After the pattern-action has satisfied expression (4) and the prediction direction is not in the Z direction, the pattern is allowed to be included in the new "expertise" of the robot, PKB, now the Enhanced Knowledge Base (EKB). The above procedure can be better understood with the flowchart of the NNC processing as shown in figure 9.

4.4.2 Compliant motion during peg-in-hole operations

The robot carries out the assemblies with incremental straight and rotational motions of 0.1 mm and 0.1°, respectively. Rotation around the X and Y axes was avoided so that only straight directions were considered which means that only compliant motion in XY plane and rotation around the Z axis was considered. In order to get the next motion direction the forces are read, normalized and classified using the NNC.

Several tests were carried out to assess the compliant motion performance of the NNC using aluminum pegs with different cross-sectional geometry: circular, squared and radiused-square, see figure 2.

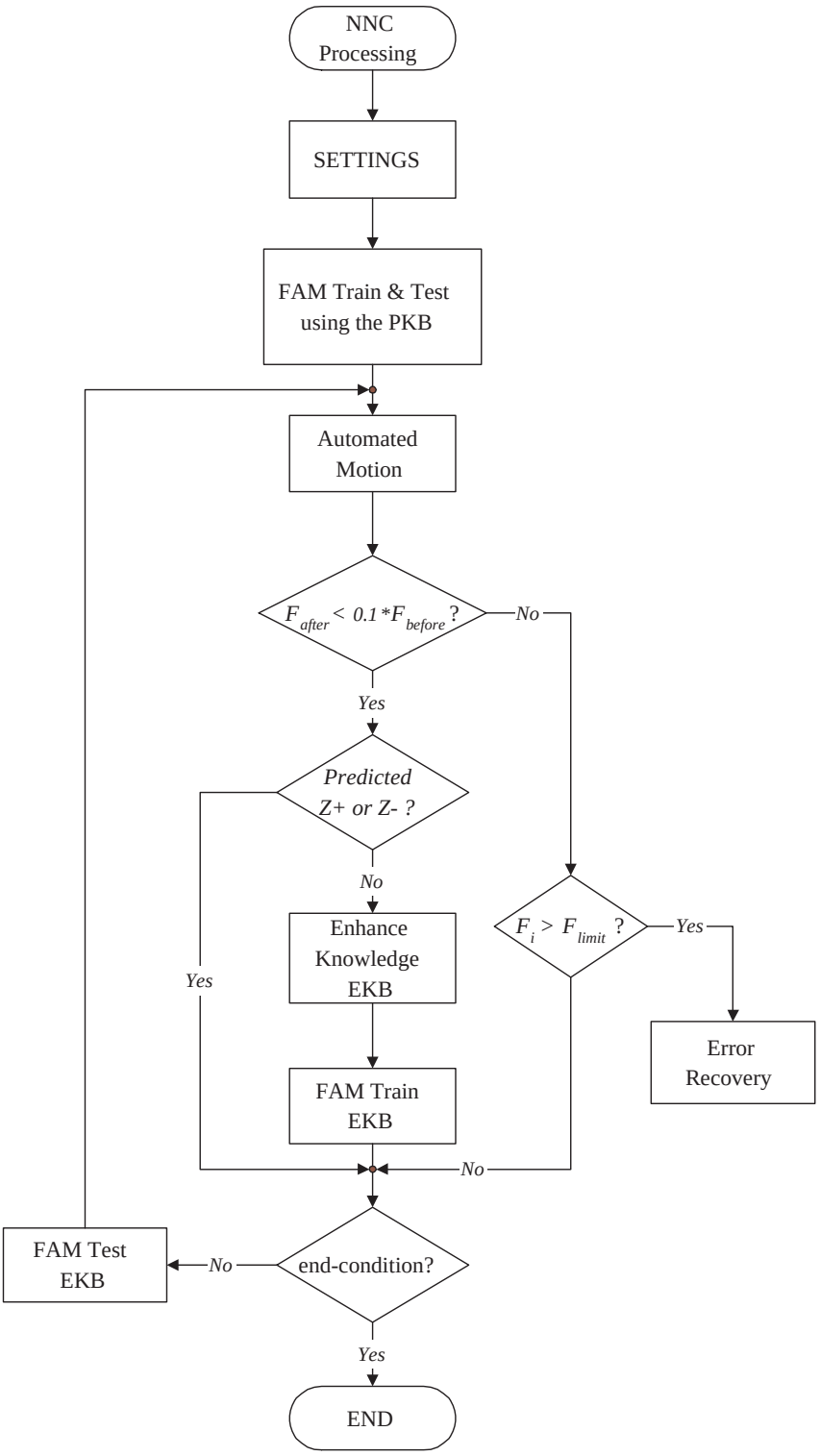


Figure 9. Flowchart of the NNC processing

The assembly was ended when 3/4 of the body of the peg were inside the hole. This represented 140 motion steps in the -Z assembly direction. A typical assembly operation is shown in figure 10.

The Fuzzy ARTMAP network parameters during experiments were set for fast learning (learning rate = 1). The values for the vigilance - in the range (0-1) - were selected based on the fact that it was required for the FuzzyARTMAP network to be as selective as possible to cluster all different patterns and which is achieved by having a high vigilance level for the of ϱ_{map} and ϱ_a; hence, this was the main criterion to select the vigilance and was not related to the task conditions (shape, offset errors). ϱ_b is small since this is increased internally according to the disparity between the input patterns and the previous recognition categories in the match tracking mechanism, for a detailed description of the Fuzzy ARTMAP architecture see (Carpenter, et al, 1992). In our experiments the values for the vigilance were as follows: $\varrho_a = 0.2$ (base vigilance), $\varrho_{map} = 0.9$ and $\varrho_b = 0.9$.

Figure 10. Peg-in-hole operation

5. Assembly Results

5.1 Assembly results using ACQ-PKB

Typical results in a chamfered squared peg insertion using the GVN-PKB are summarised in table 2.

Insertion	Offset (δx, δy, δR_z) (mm, mm, °)	Using GVN-PKB				Using ACQ-PKB			
		New Patterns	Alignment Motions	Total Motions	Time (s)	New Patterns	Alignment Motions	Total Motions	Time (s)
1	(0.7, 0.8, 0.8)	0	23	173	47.08	0	26	166	44.53
2	(-0.8, 1.1, -0.8)	1	24	178	48.19	1	36	176	47.83
3	(-0.7, -0.5, 0.8)	2	65	213	57.78	0	22	162	43.55
4	(0.8. -0.9, -0.8)	0	20	160	43.41	0	25	165	44.56
5	(0.7, 0.8, -0.8)	1	28	174	47.11	0	20	160	43.14
6	(-0.8, 1.1, 0.8)	3	30	170	46.27	1	32	173	46.48
7	(-0.7, -0.5, -0.8)	2	21	171	46.3	0	26	168	45.56
8	(0.8. -0.9, 0.8)	0	17	157	42.58	0	22	162	43.50
9	(0.7, 0.8, 0.8)	0	18	158	42.92	0	27	167	44.80
10	(-0.8, 1.1, -0.8)	3	18	158	42.77	0	28	172	46.22
11	(-0.7, -0.5, 0.8)	4	31	171	46.55	1	19	159	42.78
12	(0.8. -0.9, -0.8)	0	19	159	43.08	0	25	173	46.59
13	(0.7, 0.8, -0.8)	0	68	210	56.98	0	20	162	43.62
14	(-0.8, 1.1, 0.8)	3	38	184	49.91	1	28	168	45.30
15	(-0.7, -0.5, -0.8)	0	21	161	43.66	1	22	162	43.94
16	(0.8. -0.9, 0.8)	0	32	172	46.72	0	20	160	42.94

Table 2. Results using a GVN-PKB and ACQ-PKB

At the start of the operation different positional offsets were given as indicated in the second column. During all insertions the robot's learning ability was enabled. During the first insertion, for instance, the network learned 0 new patterns requiring 140 motions in the Z- direction and 23 motions for alignment to complete the assembly, making a total of 173 motions. The processing time for the whole insertion was 47 seconds.

Using the knowledge acquired from the squared peg insertion, the robot was also able to perform the assembly. For comparison purposes, insertions using the same offset as before were carried out and the results are given in table 2.

From the results given above in table 2, using both, the GVN-PKB and the ACQ-PKB, it can be observed that the number of new patterns using the GVN-PKB was much higher (19) compared with the number of new patterns acquired by using the ACQ-PKB (5). Learning a lower number of new patterns indicates that when using the acquired knowledge the robot needs only few

examples more which are acquired on-line. However, when using the GVN-PKB, the required number of contact force patterns needed for that specific assembly is much higher, which demonstrates a lower compliant motion capability. The robot's behaviour improved over time in terms of the assembly speed and in the number of alignment motions when the ACQ-PKB was used.
A quality measure that helps to assess the robot's dexterity is the force and moment traces during assembly and while in constraint motion. This quality measure can be obtained from the continuous monitoring of the force and torque. The quality measure during experiments using the GVN-PKB and the ACQ-PKB is illustrated in figure 11 for forces and in figure 12 for torques.

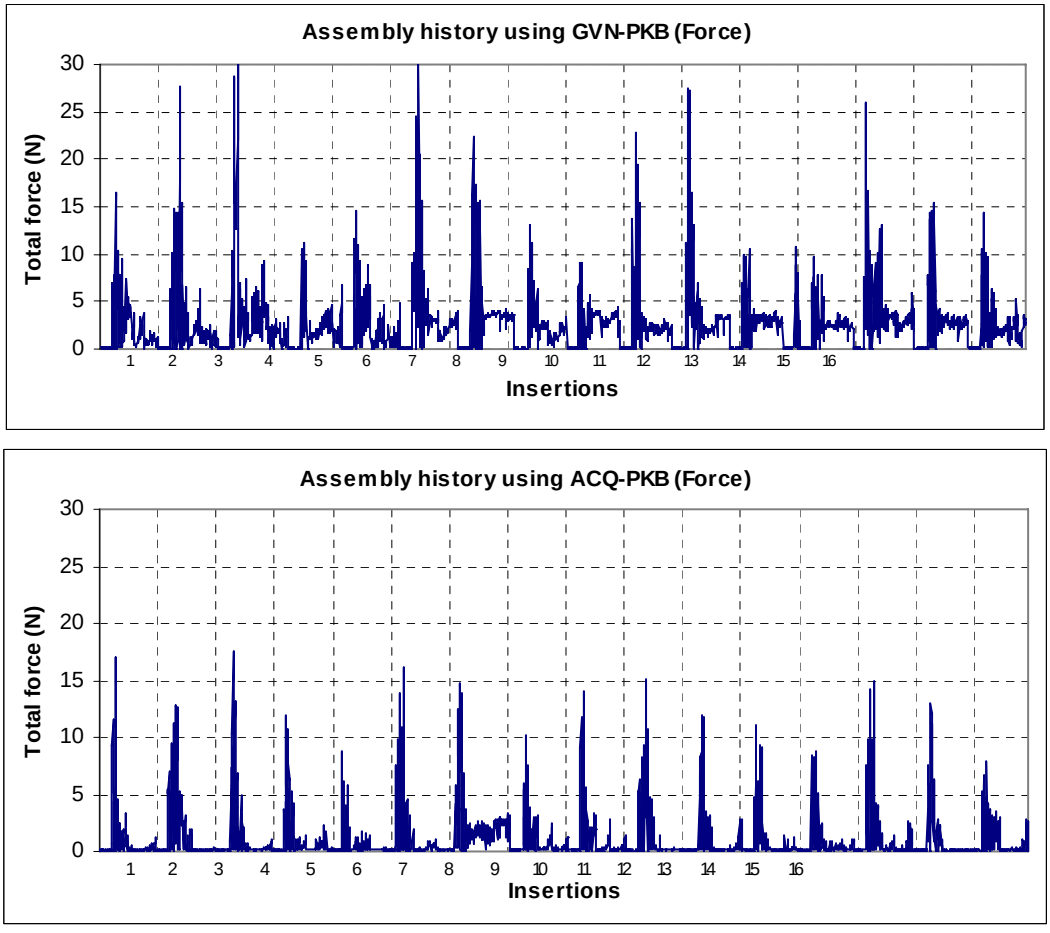


Figure 11. Forces during square chamfered peg insertion

From Figure 11 and Figure 12; it can be observed that when using the ACQ-PKB the magnitude of the forces and torques were significantly lower and in

certain cases they were almost half the value in the same experiments when using the GVN-PKB.

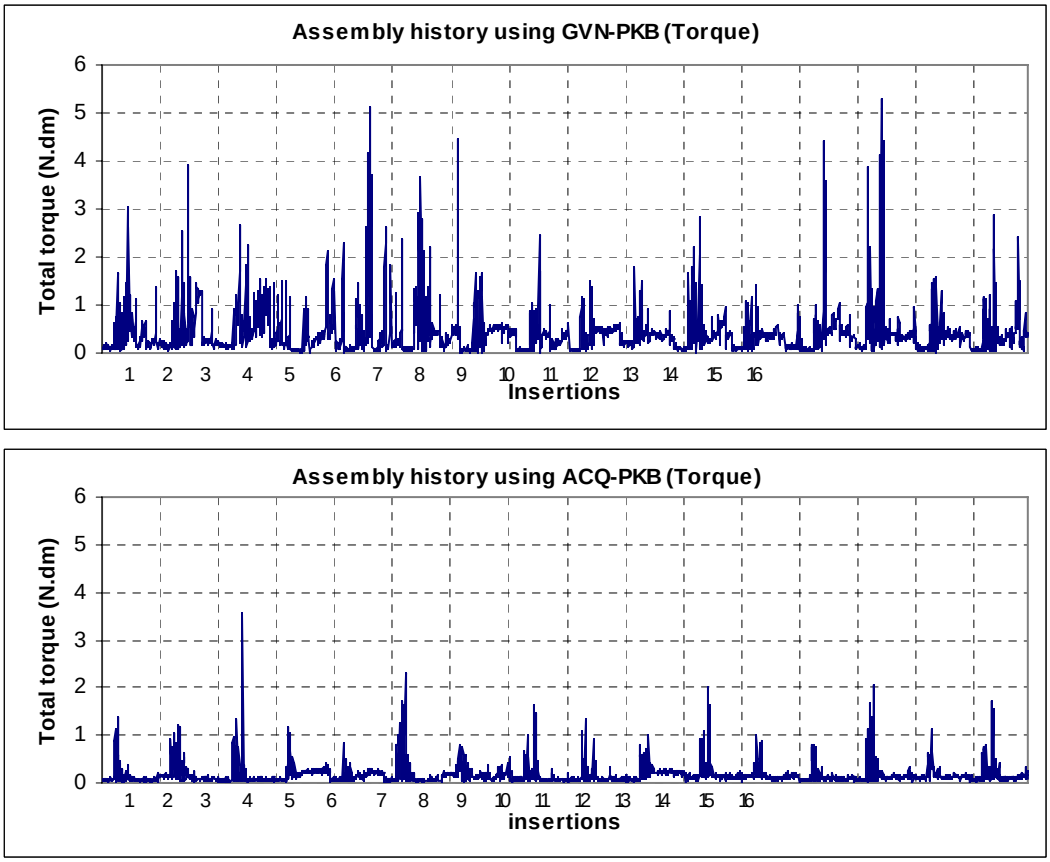


Figure 12. Torque during square chamfered peg insertion

Some forces and torques remain at the end of the insertion when the GVN-PKB is used. These residues are due to the orientation error (Rz) which is not completely recovered. The recovery of the orientation error is illustrated in figure 13, when the ACQ-PKB is used the orientation error is recovered in almost all insertions.

The total distance on XY plane by the robot is showed in figure 14 for both, GVN-PKB and ACQ-PKB. The ideal distance is de minimum distance required to reach the center point of the insertion.

Figure 15 evaluates if the robot reached de center point of the insertion in XY coordinates after the assembly end condition was satisfied, when is used an ACQ-PKB the center point was reached more efficiently than with GVN-PKB.

It was also tested the generalisation capability of the NNC by assembling different components using the same ACQ-PKB. Results are provided in table 3. For the insertion of the radiused-square component, the offsets were the same as before and for the insertion of the circular component a higher offset was used and no rotation was given. The time for each insertion was computed with the learning ability on (L_{on}) and also with learning inhibited (L_{off}); that is, using only the initial ACQ-PKB. The assembly operation was always successful and in general faster in most cases when the learning was enabled compared with inhibited learning.

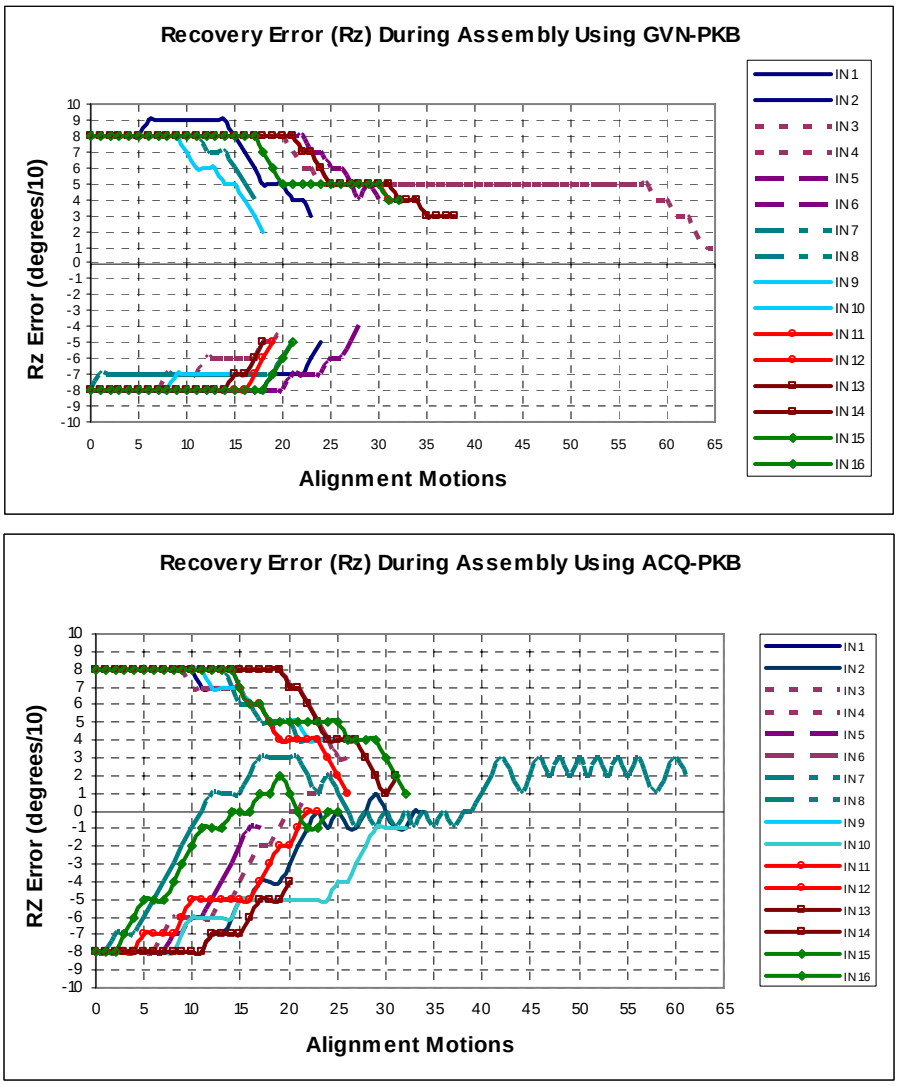


Figure 13. Recovery error (Rz) during assembly

Radiused-square chamfered peg insertion				Circular chamfered peg insertion		
Inser-tion	Offset (dx, dy, dRz) (mm, mm, °)	L_{on} time (s)	L_{off} time (s)	Offset (dx, dy, dRz) (mm, mm, °)	L_{on} time (s)	L_{off} time (s)
1	(0.7, 0.8, 0.8)	45	48	(0.7, 0.8, 0)	42	43
2	(-0.8, 1.1, -0.8)	45	51	(-0.8, 1.1, 0)	41	41
3	(-0.7, -0.5, 0.8)	43	47	(0.8. -0.9, 0)	40	42
4	(0.8. -0.9, -0.8)	50	54	(0.8. -0.9, 0)	41	41
5	(0.7, 0.8, -0.8)	44	44	(-0.8, 1.1, 0)	41	41
6	(-0.8, 1.1, 0.8)	53	51	(0.8. -0.9, 0)	41	42
7	(-0.7, -0.5, -0.8)	54	55	(1.4, 1.6, 0)	45	45
8	(0.8. -0.9, 0.8)	50	49	(1.6. -1.8, 0)	43	45
9	(0.7, 0.8, 0.8)	46	46	(1.4, 1.6, 0)	43	44
10	(-0.8, 1.1, -0.8)	45	55	(-1.4, -1, 0)	42	43
11	(-0.7, -0.5, 0.8)	44	45			
12	(0.8. -0.9, -0.8)	53	51			
13	(0.7, 0.8, -0.8)	43	43			
14	(-0.8, 1.1, 0.8)	53	51			
15	(-0.7, -0.5, -0.8)	44	59			
16	(0.8. -0.9, 0.8)	45	50			

Table 3. Results using an ACQ-PKB

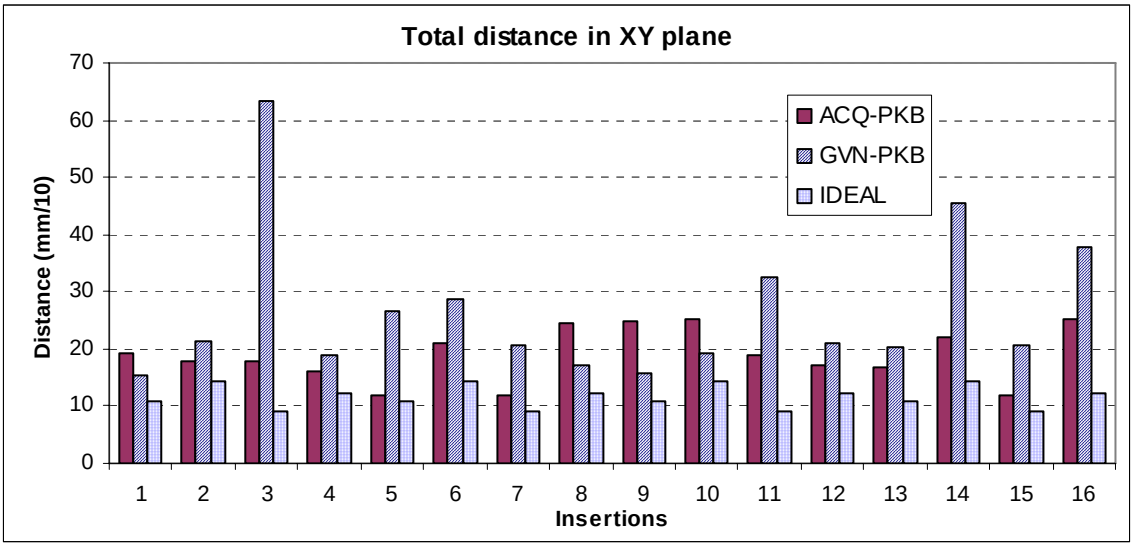


Figure 14. Total distance on XY plane

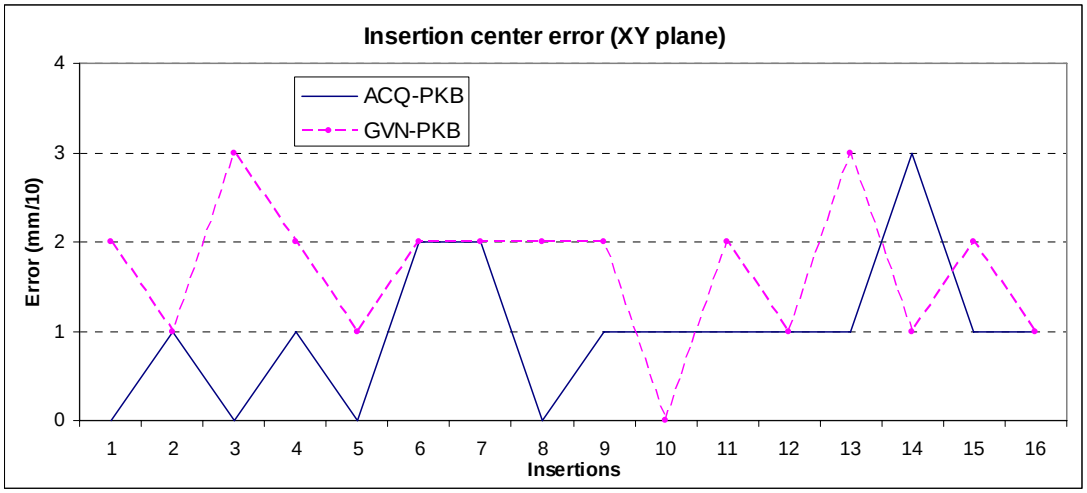


Figure 15. Insertion center error on XY plane

5.2 Whole assembly process results

Several tests were carried out to asses the performance. The diameter of the male components was 24.8 mm whereas the diameter of female components was 25 mm; the chamfer was set to 45° and 5 mm width. Results are contained in table 4. In zone 2 the SIRIO only provides location (X,Y) because the female component orientation was fixed, however an error occurs and it is related to the component's tolerance. The error for the chamfered square component is 0.8°, 0.5° for the chamfered radiused-square and 0.4° for the chamferless square and 0.6° for the chamferless radiused-square. Error recovery is illustrated in figure 18. The assembly operation ends when ¾ of the body of male component is in the hole, this represents 14 mm. The NNC was operated during the first 10 mm (100 manipulator steps), the FuzzyARTMAP parameters were: $\varrho_a = 0.2$, $\varrho_{map} = 0.9$ and $\varrho_b = 0.9$.

Table 4 shows the position errors in zone 2 which is represented in figures 16 and 17 as the trajectory followed by the robot. The minimum time of assembly cycle was **1:11** min, the maximum was **1:24** min and the average time was 1.17 min.

The system has an average angular error of **3.11**° and a maximum linear position error from -1.3 mm to 3.1 mm due to the camera positioning system in Zone 1.

# IN	P	Ch	TC (min)	TA (s)	ZONE 1			Zone 1 Error			ZONE 2		Zone 2 Error		NC
					Xmm	Ymm	RZ°	Xmm	Ymm	RZ°	Xmm	Ymm	Xmm	Ymm	
1	S	Y	1:15	32.5	62.4	144.1	10	0.2	-1.3	0	84.6	102.1	0.3	-1	Y
2	S	Y	1:15	30.4	62.4	45.7	12	1.8	0.2	2	85.6	101.1	-0.7	0	Y
3	S	Y	1:15	31.8	178.7	47.7	23	0.9	-0.8	3	84.7	100.9	0.2	0.2	Y
4	R	Y	1:11	30.1	181.6	147	29	-0.3	-0.7	-1	84.7	100.6	0.2	0.5	Y
5	R	Y	1:14	29.4	62.4	145.1	36	0.2	-0.3	-4	84.9	100.7	0	0.4	Y
6	R	Y	1:19	29.6	67.3	44.8	48	3.1	-0.7	-2	85.3	101.6	-0.4	-0.5	Y
7	C	Y	1:15	29.6	180.6	49.6	57	1	1.1	-3	84.6	102.4	0.3	-1.3	Y
8	C	Y	1:13	30.2	180.6	148	77	-0.7	0.3	7	84.3	101	0.6	0.1	Y
9	C	Y	1:14	30.2	61.5	146	79	-0.7	0.6	-1	83.9	101.6	1	-0.5	Y
10	S	N	1:18	29.9	63.4	45.7	83	-0.8	0.2	-7	85.4	100.5	-0.5	0.6	Y
11	S	N	1:19	30.4	179.6	48.6	104	0	0.1	4	83.2	100.8	1.7	0.3	Y
12	S	N	1:22	34.6	180.6	147	104	-0.7	-0.7	-6	83.2	101.8	1.7	-0.7	Y
13	R	N	1:22	38.3	61.5	146	119	-0.7	0.6	-1	84.8	102.8	0.1	-1.7	Y
14	R	N	1:22	36.8	63.4	43.8	126	-0.8	1.7	-4	83.6	101.8	1.6	-0.7	Y
15	R	N	1:24	36.6	179.6	47.7	138	0	-0.8	-2	83.2	101.7	1.7	-0.6	Y
16	C	N	1:17	30.5	182.6	149	150	1.3	1.3	0	83.7	101.2	1.2	-0.1	Y
17	C	N	1:15	28.3	63.4	146	155	1.2	0.6	-5	84.6	100.7	0.3	0.4	Y
18	C	N	1:15	29.7	64.4	47.7	174	0.2	2.2	4	83.9	101.1	1	0	Y

Table 4. Eighteen different assembly cycles, where IN= Insertion, P=piece, Ch=chamfer present, TC=Assembly cycle time, TA= Insertion time, NC=correct neural classification, S=square, R=radiused-square, C=circle, N=no and Y=yes.

The force levels in chamferless assemblies are higher than the chamfered ones. In the first one, the maximum value was in Z+, 39.1 N for the insertion number 16, and in the chamfered the maximum value was 16.9 N for the insertion number 9.

In chamfered assembly, in figure 16, it can be seen that some trajectories were optimal like in insertions 2, 5, 7, 8 and 9, which was not the case for chamferless assembly; however, the insertions were correctly completed.

In figure 17, each segment corresponds to alignment motions in other directions different from Z-. The lines mean the number of Rz+ motions that the robot performed in order to recover the positional error for female components. The insertion paths show how many rotational steps are performed. The maximum alignment motions were 22 for the chamfered case in comparison with 46 with the chamferless component.

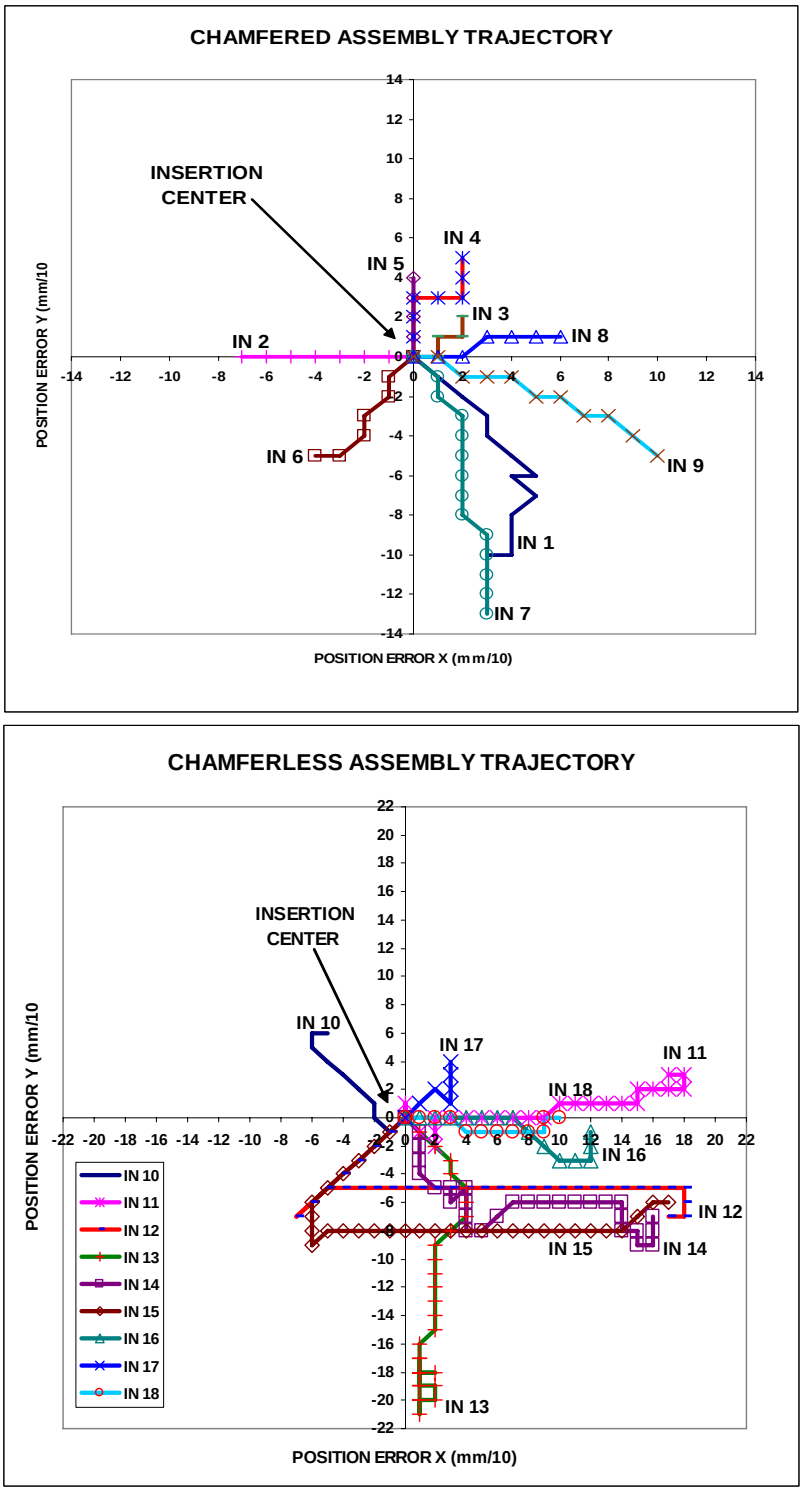
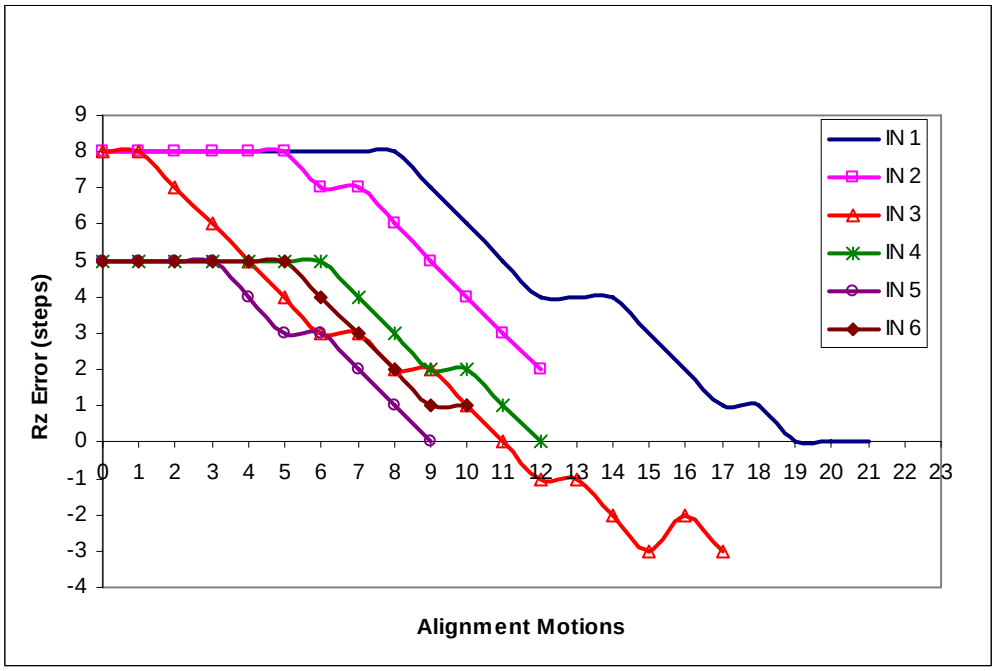


Figure 16. Assembly trajectory in top view for each insertion in zone 2. The trajectory starts with the labels (INx) and ends at 0,0 origins coordinate



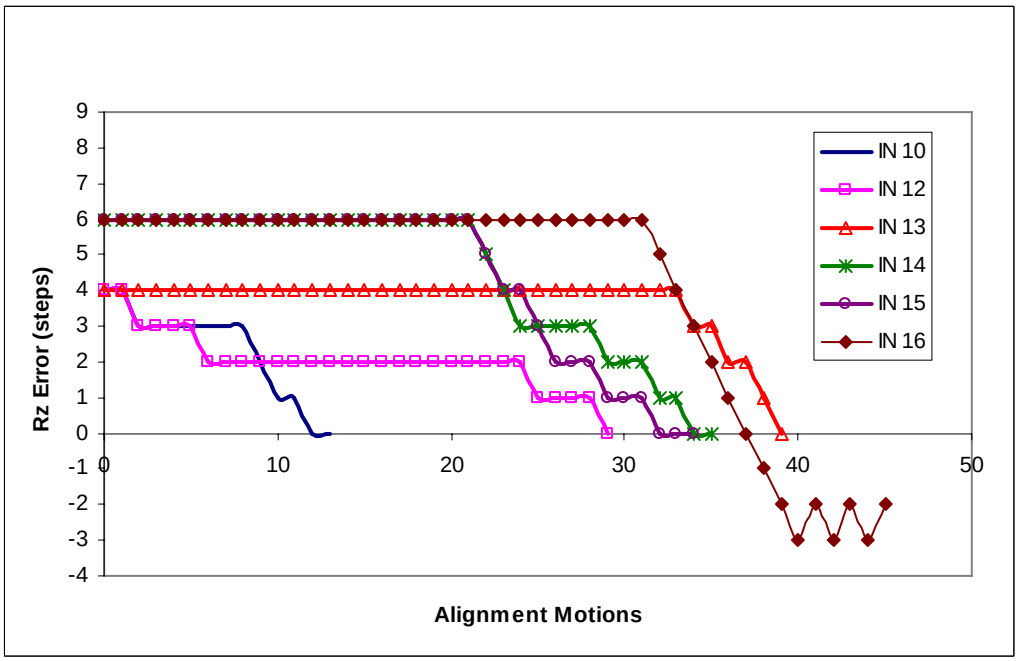


Figure 17. Compliant rotational motions (only Rz+) for each insertion in zone 2, left chamfered assembly, right chamferless assembly

6. Conclusions

A task planner approach for peg-in-hole automated assembly was presented. The proposed methodologies were used to achieve the tasks and tested successfully in the real world operations using an industrial manipulator. The robot is able to perform not only the assembly but also it can start working without initial knowledge about the environment, and it can increase its PKB at every assembly if it is necessary.

The presented approach using the vision and force sensing system has envisaged further work in the field of multimodal learning in order to fuse information and to increase the prediction capability of the network which contributes towards the creation of truly self-adaptive industrial robots for assembly.

All assemblies were successful showing the system robustness against different uncertainties and its generalization capability. The generalization of the NNC has been demonstrated by assembling successfully different component geometry using different mechanical tolerances and offsets employing the same acquired knowledge base.

Initial knowledge is acquired from actual contact states using explorative motions guided by fuzzy rules. The knowledge acquisition stops once the ACQ-PKB is fulfilled. Later this knowledge is refined as the robot develops new assembly tasks.

The dexterity of the robot improves using the ACQ-PKB by observing the magnitude of forces and moments as shown in Figures 11 and 12. Values are significantly lower, hence motions were more compliant in this case indicating that information acquired directly from the part geometry allowed also lower constraint forces during manipulation. Having implemented the knowledge acquisition mechanism, the NNC acquires only real contact force information from the operation. In comparison with our previous results, insertion trajectories improved enormously; we believe that given *a priori* knowledge (GVN-PKB) is fine, but contact information extracted directly form the operation itself provides the manipulator with better compliant motion behaviour.

Results from this work have envisaged further work in the area of multimodal data fusion (Lopez-Juarez, et al, 2005). We expect that data fusion from the F/T sensor and the vision system result in an improved confidence for getting the contact information at the starting of the operation providing also important information such as chamfer presence, part geometry and pose information, which will be the input data to a hierarchical task level planner as pointed out by (Lopez-Juarez & Rios-Cabrera, 2006).

7. References

Ahn, D.S.; Cho, H.S.; Ide, K.I.; Miyazaki, F.; Arimoto, S. (1992). Learning task strategies, in robotic assembly systems. *Robotica* Vol. 10, 10 (409–418)

Asada, H. (1990). Teaching and Learning of Compliance Using Neural Nets. IEEE *Int Conf on Robotics and Automation,* 8 (1237-1244)

Baeten, J.; Bruyninckx, H.; De Schutter, J. (2003). Integrated Vision/Force Robotic Servoing in the Task Frame Formalism. *The International Journal of Robotics Research*. Vol. 22, No. 10-11, 14 (941-954)

Carpenter, G. A.; Grossberg, S. (1987). A Massively Parallel Architecture for a Self-Organizing Neural Pattern Recognition Machine. *Computer Vision, Graphics, and Image Processing,* Academic Press, Inc. 62 (54-115)

Carpenter, G. A.; Grossberg, S.; Reynolds, J. H. (1991). ARTMAP: Supervised Real-Time Learning and Classification of Nonstationary Data by Self-Organizing Neural Network. *Neural Networks,* 24 (565-588)

Carpenter, G.A.; Grossberg, J.; Markunzon, N.; Reynolds, J.H.; Rosen, D.B. (1992). Fuzzy ARTMAP: A Neural Network Architecture for Incremental Learning of Analog Multidimensional Maps. IEEE *Trans. Neural Networks,* Vol. 3, No. 5, 36 (678-713)

Cervera, E.; del Pobil, A. P. (1996). Learning and Classification of Contact States in Robotic Assembly Tasks. Proc *of the 9th Int. Conf. IEA/AIE*, 6 (725-730)

Cervera, E.; Del Pobil, A.P. (1997). Programming and Learning in Real World Manipulation Tasks. In: *Int. Conf. on Intelligent Robot and Systems (IEEE/RSJ,* Proc. 1, 6 (471-476)

Cervera, E.; del Pobil, A. P. (2002). Sensor-based learning for practical planning of fine motions in robotics. *The International Journal of Information Sciences,* Special Issue on Intelligent Learning and Control of robotics and Intelligent Machines in Unstructured Environments, Vol. 145, No. 1, 22 (147-168)

De Schutter, J.; Van Brussel, H. (1988). Compliant Robot Motion I, a formalism for specifying compliant motion tasks. *The Int. Journal of Robotics Research*, Vol. 7, No. 4, 15 (3-17)

Doersam, T.; Munoz Ubando, L.A. (1995). Robotic Hands: Modelisation, Control and Grasping Strategies. In: *Meeting annuel de L'Institute Fanco-Allemand pour les Application de la recherche IAR*

Driankov, D.; Hellendoorn, H.; Reinfrank, M. (1996). *An Introduction to Fuzzy Control*. 2nd ed. Springer Verlag

Erlbacher, E. A. *Force Control Basics*. PushCorp, Inc. (Visited December 14th,

2004). http://www.pushcorp.com/Tech%20Papers/Force-Control-Basics.pdf

Grossberg, S. (1976). Adaptive Pattern Classification and universal recoding II: Feedback, expectation, olfaction and illusions. *Biological Cybernetics,* Vol. 23, 16 (187-202)

Gullapalli, V.; Franklin, J. A.; Benbrahim, H. (1994). Acquiring Robot Skills via Reinforcement Learning. IEEE *Control Systems,* 12 (13-24)

Gullapalli, V.; Franklin, J.A.; Benbrahim, H. (1995). Control Under Uncertainty Via Direct Reinforcement Learning. *Robotics and Autonomous Systems*. 10 (237-246)

Howarth, M. (1998). *An Investigation of Task Level Programming for Robotic Assembly*. PhD thesis. The Nottingham Trent University, UK

Ji, X.; Xiao, J. (1999). Automatic Generation of High-Level Contact State Space. *Proc. of the Int. Conf. on Robotics and Automation*, 6 (238-244)

Joo, S.; Miyasaki, F. (1998). Development of a variable RCC and its applications. *Proceedings of the 1998 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems,* Vol. 2, 7 (1326-1332)

Jörg, S.; Langwald, J.; Stelter, J.; Natale, C.; Hirzinger, G. (2000). Flexible Robot Assembly Using a Multi-Sensory Approach. In: *Proc. IEEE Int. Conference on Robotics and Automation,* 8 (3687-3694)

Kaiser, M.; Dillman, M.R. (1996). Building Elementary Robot Skills from Human demonstration. IEEE *International Conference on Robotics and Automation*, Minneapolis, Minnesota, 6 (2700 – 2705)

Lopez-Juarez, I.; Howarth, M.; Sivayoganathan, K. (1996). Robotics and Skill Acquisition. A. Bramley; T. Mileham; G. Owen. (eds.) In: *Advances in Manufacturing Technology X,* ISBN 1 85790 031 6, 5 (166-170)

Lopez-Juarez, I. (2000). *On-line learning for robotic assembly using artificial neural networks and contact force sensing*. PhD thesis, Nottingham Trent University, UK

Lopez-Juarez, I.; Ordaz-Hernandez, K.; Pena-Cabrera, M.; Corona-Castuera, J.; Rios-Cabrera, R. (2005). On the Design of A multimodal cognitive architecture for perceptual learning in industrial robots. In MICAI 2005: Advances in Artificial Intelligence. LNAI 3789. Lecture Notes in Artificial Intelligence. A Gelbukh, A de Albornoz, and H Terashima (Eds.). 10 (1062-1072). Springer Verlag, Berlin

Lopez-Juarez, I.; Rios-Cabrera, R. (2006). Distributed Architecture for Intelligent Robotic Assembly, Part I: Design and Multimodal Learning. Ad-

vances Technologies: Research-Development-Application. Submitted for publication

Lozano-Perez T.; Mason, M.T.; Taylor R. H. (1984). Automatic Synthesis of Fine Motion Strategies. *The Int. Journal of Robotics Research,* Vol. 3 No. 1, 22 (3-24)

Mason, M. T. (1983). *Compliant motion,* Robot motion, Brady M et al eds. Cambridge: MIT Press

Peña-Cabrera, M.; López-Juárez, I.; Ríos-Cabrera R.; Corona-Castuera, J. (2005). Machine vision learning process applied to robotic assembly in manufacturing cells. *Journal of Assembly Automation*, Vol. 25, No. 3, 13 (204-216)

Peña-Cabrera, M.; Lopez-Juarez, I. (2006). Distributed Architecture for Intelligent Robotic Assembly, Part III: Design of the Invariant Object Recognition System. Advances Technologies: Research-Development-Application. Submitted for publication

Skubic M.; Volz, R. (1996). Identifying contact formations from sensory patterns and its applicability to robot programming by demonstration. *Proceedings of the 1996 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, Osaka, Japan

Skubic, M.; Volz, R.A. (2000). Acquiring Robust, Force-Based Assembly Skills from Human Demonstration. IEEE *Trans. on Robotics and Automation,* Vol. 16, No. 6, 10 (772-781)

Whitney. D.; Nevis, J. (1979). What is the Remote Center Compliance (RCC) and what can it do?. *Proceedings of the 9th Int. Symposium on Industrial Robots*, 18 (135-152)

Xiao J.; Liu, L. (1998). Contact States: Representation and Recognizability in the Presence of Uncertainties. IEEE/RSJ *Int. Conf. Intell. Robots and Sys*

14

Distributed Architecture for Intelligent Robotic Assembly Part III: Design of the Invariant Object Recognition System

Mario Pena-Cabrera and Ismael Lopez-Juarez

1. Introduction

In previous chapter it has been described the overall architecture for multimodal learning in the robotic assembly domain (Lopez-Juarez & Rios Cabrera, 2006). The acquisition of assembly skills by robots is greatly supported by the effective use of contact force sensing and object recognition. In this chapter we will describe the robot's ability to invariantly recognise assembly parts at different scale, rotation and orientation within the work space. The chapter shows a methodology for on-line recognition and classification of pieces in robotic assembly tasks and its application into an intelligent manufacturing cell. The performance of industrial robots working in unstructured environments can be improved using visual perception and learning techniques. In this sense, the described technique for object recognition is accomplished using an Artificial Neural Network (ANN) architecture which receives a descriptive vector called CFD&POSE as the input. This vector represents an innovative methodology for classification and identification of pieces in robotic tasks, every stage of the methodology is described and the proposed algorithms explained. The vector compresses 3D object data from assembly parts and it is invariant to scale, rotation and orientation, and it also supports a wide range of illumination levels. The approach in combination with the fast learning capability of ART networks indicates the suitability for industrial robot applications as it is demonstrated through experimental results.

Robotics field has grown considerably with new technologies, industrial robots today, needs sensorial capabilities to achieve non-structured and more sophisticated tasks; vision systems as a sensorial mode for robots have a growing demand requiring more complex and faster image processing functions in order to implement more sophisticated industrial applications, like assembly automation.

In this sense, vision recognition systems must be capable of perceiving and detecting images and objects, as close as the human vision does; this fact has encouraged research activity to design artificial vision systems based on the neural morphology of the biological human vision system. Now scientists understand better about how computational neural structures and artificial vision systems must be designed following neural paradigms, mathematical models and computational architectures. When a system involves these aspects, it can be referred to as a "Neuro-Vision System" (Gupta and Knopf, 1993), (Peña, 2004), which can be defined as an artificial machine with ability to see our environment and provide visual formatted information for real time applications.

It has been shown by psychological and clinical studies that visual object recognition involves a large activity area on the cerebral cortex when objects are seen the first time and the region's activity is reduced when familiar objects are perceived (Gupta and Knopf, 1993). New objects can also be learned quickly if certain clues are given to the learner. Following this psychological evidence a novel architecture was designed. The architecture is firstly trained with clues representing different objects that the robot is likely to encounter within the working space to form its initial knowledge base. This information then triggers the on-line learning subsystem based on an Artificial Neural Network (ANN), the new image vector descriptors override initial clues, and the robot learns to identify familiar objects and to learn new ones.
The above ideas suggested that it was possible to get fast and reliable information from a simple but focused analysis of what an object might show. The very important aspects of the scene (we have called "clues"), can be used later to retrieve memorized aspects of the object without having to recall detailed features. By using neural networks it is possible to learn manipulative skills which can be used by an industrial manipulator (Lopez-Juarez and M. Howarth, 2000). In someway we humans do that process once an object has been seen and learned for the first time.

The chapter describes a methodology for on-line object recognition, based on artificial neural networks for identification and classification purposes. Robust algorithms for perimeter, centroid calculations, object functions and pose estimation are presented.

2. Related work

Intelligent manufacturing cells using robots with sensorial capabilities are being investigated using Artificial Intelligence techniques like ANN's and Fuzzy Logic among others, since mathematical and control models are simplified.
Acquiring information from multiple sensors in manufacturing systems provides robustness and self-adaptation capabilities, hence improving the performance in industrial robot applications. A few researchers have applied neural networks to assembly operations with manipulators and force feedback. (Vijaykumar Gullapalli, 1994), used BackPropagation (BP) and Reinforcement Learning(RL) to control a Zebra robot, its neural controller was based on the location error reduction beginning from a known location, (Enric Cervera, 1997), employed Self-Organization Map (SOM) and RL to control a Zebra robot, the location of the destination piece was unknown, (Martin Howarth, 1998), utilized BP and RL to control a SCARA robot, without knowing the location of assembly, (Lopez-Juarez, 2000), implemented FuzzyARTMAP to control a PUMA robot also with an unknown location. All of the above authors considered only constraint motion control during assembly; however, to complete the autonomy of the assembly system a machine vision system has also to be considered. Additionally, a new concept was introduced in 1988 called "Robotic Fixtureless Assembly" (RFA) (Hoska, 1998), that eliminates the need of using complex and rigid fixtures, which involves new technical challenges, but allows very potential solutions. Studies of RFA of flexible parts with a dynamic model of two robots which does not require measurements of the part deflections have been done (W. Ngyuen and J.K. Mills, 1996). (Plut, 1996), and (Bone, 1997), presented a grasp planning strategy for RFA. The goal of RFA is to replace these fixtures with sensor-guided robots which can work within RFA workcells. The development of such vision-guided robots equipped with programmable grippers might permit holding a wide range of part shapes without tool changing. Using Artificial Neural Networks, an integrated intelligent vision-guided system can be achieved as it is shown by (Langley et al., 2003). This job can be achieved by using 2D computer vision in different manner so that 3D invariant object recognition and POSE calculation might be used for aligning parts in assembly tasks if an –"adequate descriptor vector"- is used and interfaced in real time to a robot. Many authors had come with descriptor vectors and image transformations, used as general methods for computer vision applications in order to extract invariant features from shapes.

(Alberto S. Aguado et al., 2002), developed a new formulation and methodology for including invariance in general form of the Hough transform, (Chin-Hsiung et al., 2001), designed a technique for computing shape moments based on the quadtree representation of images, (P. J. Best and N. McKay, 1992), describe a method for registration of 3D shapes in minutes, (A. Torralba and A. Oliva, 2002), present a method to infer the scale of the scene by recognizing properties of the scene structure for depth estimation, (Freeman, 1961), introduced the first approach for representing digital curves using chain codes, and showing classical methods for processing chains (Freeman, 1974), (E. Bribiesca, 1999), developed a new chain code for shapes composed of regular cells, which has recently evolved even to represent 3D paths and knots.

Some authors use multiple cameras or multiple views to extract information, performs invariant object recognition and determine object's position and motion, (Stephen A. Underwood, 1975), developed a visual learning system using multiple views which requires deterministic description of the object's surfaces like measurements and interconnections, (Yong-Sheng Chen et al., 2001), propose a method to estimate the three dimensional ego-motion of an observer moving in a static environment, (Hiroshi Murase and Shree K. Nayar, 1995), have worked in visual learning and recognition of 3D objects from appearance, (E. Gonzalez-Galvan et al., 1997), developed a procedure for precision measure in 3D rigid-body positioning using camera-space manipulation for assembly. (Dickmanns, 1998), and (Nagel, 1997), have shown solutions to facilitate the use of vision for real world-interaction, (Hager et al., 1995), and (Papanikolopoulos et al., 1993), use markers on the object to simplify detection and tracking of cues.

Some other authors have contributed with techniques for invariant pattern classification, like classical methods as the universal axis of Lin, and invariant moments of (Hu, 1962), or artificial intelligence techniques, as used by (Cem Yüceer and Kemal Oflazer, 1993), which describes an hybrid pattern classification system based on a pattern pre-processor and an ANN invariant to rotation, scaling and translation, (Stavros J. and Paulo Lisboa, 1992), developed a method to reduce and control the number of weights of a third order network using moment classifiers and (Shingchern D. You and G. Ford, 1994), proposed a network for invariant object recognition of objects in binary images. Applications of guided vision used for assembly are well illustrated by (Gary M. Bone and David Capson, 2003), which developed a vision-guide fixtureless assembly system using a 2D computer vision for robust grasping and a 3D computer

vision to align parts prior to mating, and (Stefan Jörg et al., 2000), designing a flexible robot-assembly system using a multi-sensory approach and force feedback in the assembly of moving components.

3. Invariant Object Recognition

Recognising an object using a vision system involves a lot of calculations and mathematical processes which have to be implemented and tested in order to be used in a real application. In this way, vision systems and solid state technologies has been evolved at the same time, the more faster and sophisticated computers had come with the more complex and better vision systems developed as well as more sophisticated algorithms implementation had became a reality.
Basic ideas were established with approximated representations, architectures and algorithms, which motivated the development of this research area. A general solution to the recognition problem does not exist (J. L. Mundy, 1998). Most classical methods to object recognition use shape analysis and metric feature extraction from digitized images to reach the goal. Recent research, points to use artificial intelligence techniques to look for invariant recognition solutions using vision. In this way artificial neural networks are a well representative technique to this purpose

3.1 Learning

For a better understanding of the recognition problem, it is convenient to explore how humans make the recognition process, sometimes in an automatic manner and many times in short periods of time, as it is the case of a projected image in the retina which can be rotated, moved or even scaled when the object or eyes are moved. Humans can recognize an object within a complex scene with overlapped objects.
Most recognition techniques uses indirect mapping between objects and retinal images, there is a set of object representations in long term memory which have been associated with object physical representations, and information is not just a copy of a pattern of retinal stimulus but a set of features representative of the object with invariant properties, so the object can be recognized from different views. There must be a criteria to decide which is the best object representation within the long term memory when an input object representa-

tion is not exactly the same to the one already memorized. It is difficult to know when the perception process ends and when the recognition process begins.
Human visual process can execute a universal set of routines with simple subprocesses operating with object sketches even in 2 ½ dimensions (S. Ulman, 1984), this activities might involve: contour scanning, region filling and area marking to obtain as the output, basic features from the most important issues of a visual scene as shapes and its spatial relation, this goes to the process of object recognition by way of its parts grouping reconstruction, because objects can be overlapped or occluded. Assuming that memorized parts corresponds to previous visual learning processes, a shape can be addressed and reconstructed with a huge part list, so recognition can be achieved using only the visible parts within the scene.

3.2 Memory and Recall

There are at least two different processes to storage information in the long term memory. The first process is to store a list of situations about the objects including information on how many parts are grouped together, its size, category names and so on. The other process is to store the codification of object appearance. There is evidence that the time a person takes to decide if two 3D objects have the same shape is a linear function of its orientation difference (R.N. Shepard and S. Hurwitz, 1984). This information might be understood as the humans falls in a continuous and smooth rotation mental process which is achieved until the orientation input shape, matches the correct canonical orientation of shape already stored.
A person cannot see an object in a single view from an arbitrary view angle, in fact, first the object is visualized with a canonic orientation to be rotated to the specific orientation, this suggests the idea that long term memory image representations are oriented to a primitive and primary observation stored as canonical shapes, which is an important point to our research inspiration.

3.3 Design considerations and techniques

In order to design and implement a vision system and use the appropriate techniques in the visual process, three aspects might be considered:

- Vision is a computational process
- Obtained description of object is a function of the observer
- Not used information within the scene has to be eliminated

Because vision is a complex process different vision levels are established for better methodology directions:

- Low level processing.- pixel direct working is done in this level to extract properties as: edges, gradients, textures, grey levels, etc.

- Medium level processing.- elements of low level are grouped here to obtain lines, regions of interest in order to use segmentation tech niques.

- High level processing.- it is oriented to interpretation of lower levels Information using models and previous knowledge. It has to deal with recognition and looks for consistency on feature primitive in formation interpretation.

Representation is a formal system to specify features of what is seen with a vision system, two aspects are important:

a) *the model representation*, is the structure used to model the representation
b) *the recognition process* is the way the model and representation are used for recognition.

These aspects have to be generic, time and space efficient, rotation, translation and scaled invariant, and conform robustness and noise and incomplete information tolerance. High level vision systems can be classified in:

- Model based vision systems.- use a geometrical representation and recognition is based on correspondence.
- Knowledge based vision systems.- use a symbolic representation and recognition is based on inferences.

Model based vision systems use predefined geometrical methods to recognize the objects whose description has been obtained from images, its principal components are: feature extraction, modelling, and matching and can use 2D or 3D

models. Because of its robustness and fast processing, manufacturing applications mostly use binary images, becoming the quantization process an important factor because all parametric calculations and description are derived from it. Shape based human recognition takes the information about perimeter as fundamental instead of regions, most 2D models used in real implementations are model and recognition object oriented as a function of its image representation as a 2D matrix array.

Knowledge based vision systems use proposal models for representation; they have a collection set of them representing knowledge about objects and their relationship. Recognition is achieved by way of inferences, from image data and domain knowledge, object identity is obtained, its principal components are:

a) *feature extraction*, important attributes of the image are obtained to be integrated in a symbolic image
b) *knowledge representation*, is constructed with a learning process being stored in the primitive knowledge data base
c) *inference*, a deductive process is achieved from the symbolic image and primitive knowledge data base to get the object identity and location.

4. Visual behaviour

The problems of modelling and understanding visual behaviour and their semantics are often regarded as computationally ill-defined. Cognitive understanding cannot adequately explain why we associate particular meanings with observed behaviours. Interpretation of visual behaviour can rely on simple mappings from recognized patterns of motion to semantics, but human activity is complex, the same behaviour may have several different meanings depending upon the scene and task context. Behaviour interpretation often also requires real-time performance if it is to be correct in the relevant dynamic context, by real time, it is not necessary implied that all computation must be performed at full video frame-rate, as long as the interpretation of behaviour proceeds within some required time constraint, (Shaogang et. al., 2002).
Considering that it is estimated that 60% of sensory information in humans is provided by the visual pathway (Kronauer, 1985), and the biological vision

concerning the pathway is a massively parallel architecture using basic hierarchical information processing (Uhr, 1980), it seems logical to look for an alternative approach with less computational power to better emulate the human visual system and it is given by connectionist models of the human cognitive process, such idea has to be considered to develop machine vision system applications today, as robotic assembly vision guided manufacturing processes.

4.1 Inspiring Ideas

Sensorial capabilities are naturally used by humans and other life animals everyday; providing with sensorial capabilities to a machine is an interesting and actual challenge which means an open research area today. Figure 1, shows a baby grasping an object, even for him this is a natural way to achieve such a task with controlled movements and pointing to grasp targets by way of having real time visual information.

Figure 1. Human grasping action.

Visual information happens to be the 60% of the sensorial information coming in from human environment (Kronauer, 1985) and is mainly used for grasping objects; estimate his 3D position or assembly parts (figure 2).

Figure 2. Grasp, 3D position and assembly actions in humans.

The same action achieved by a robot machine implies different disciplines integration and a robust hardware and software implementation (figure 3).

Sensorial capabilities real time acquired information deals with different sensors, hardware architectures and communication and control approaches and configurations in order to achieve the task. Knowledge can be built either empirically or by hand as suggested by (Towell and Shavlik, 1994). Empirical knowledge can be thought of as giving examples on how to react to certain stimuli without any explanation and hand-built knowledge, where the knowledge is acquired by only giving explanations but without examples. It was determined that in robotic systems, a suitable strategy should include a combination of both methods. Furthermore, this idea is supported by psychological evidence that suggests that theory and examples interact closely during human learning, (Feldman, 1993).

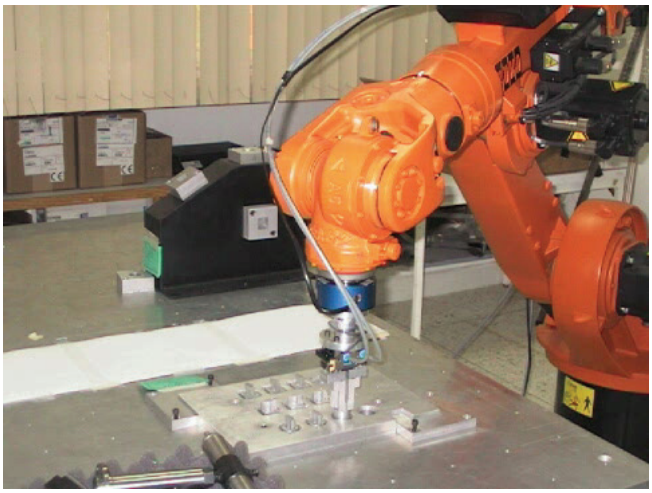


Figure 3. Grasping action with a robot.

Learning in natural cognitive systems, including our own, follows a sequential process as it is demonstrated in our daily life. Events are learnt incrementally, for instance, during childhood when we start making new friends, we also learn more faces and this process continues through life. This learning is also stable because the learning of new faces does not disrupt our previous knowledge. These premises are the core for the development of *Connectionist Models of the Human Brain* and are supported by Psychology, Biology and Computer Sciences. Psychological studies suggest the sequential learning of events at different stages or "storage levels" termed as Sensory Memory (SM), Short Term Memory (STM) and Long Term Memory (LTM).

4.2 Artificial Neural Networks

There are different types of ANN, for this research a Fuzzy ARTMAP network is used, ART stands for Adaptive Resonance Theory, which is a well established associative brain and competitive model introduced as a theory of the human cognitive processing developed by Stephen Grossberg at Boston University. Grossberg resumed the situations mentioned above in what he called the *Stability-Plasticity Dilemma* suggesting that connectionist models should be able to adaptively switch between its plastic and stable modes. That is, a system should exhibit plasticity to accommodate new information regarding unfamiliar events. But also, it should remain in a stable condition if familiar or irrelevant information is being presented. These features suggested the use of this network because of its incremental knowledge capabilities and stability, but mostly because of the fast recognition and clustering responses.
Grossberg identified the problem as due to basic properties of associative learning and lateral inhibition. An analysis of this instability, together with data of categorisation, conditioning, and attention led to the introduction of the ART model that stabilises the memory of self-organising feature maps in response to an arbitrary stream of input patterns (S. Grossberg, 1976). The core principles of this theory and how Short Term Memory (STM) and Long Term Memory (LTM) interact during network processes of activation, associative learning and recall were published in the scientific literature back in the 60's.
The theory has evolved in a series of real-time architectures for unsupervised learning, the ART-1 algorithm for binary input patterns (G. Carpenter, 1987), supervised learning is also possible through ARTMAP (G. Carpenter, 1991), that uses two ART-1 modules that can be trained to learn the correspondence between input patterns and desired output classes. Different model variations have been developed to date based on the original ART-1 algorithm, ART-2, ART-2a, ART-3, Gaussian ART, EMAP, ViewNET, Fusion ARTMAP, LaminART just to mention but a few.

5. Manufacturing

5.1 Intelligent manufacturing systems

Different sensors have been used in manufacturing systems to achieve specific tasks such as robot guiding, soldering, sorting, quality control and inspection.

Integration of new architectures and methods using sensorial modalities in manufacturing cells like vision, force-sensing and voice recognition becomes an open research field today.
Most automated systems integrators and designers had pushed hard to get faster and more accurate industrial robot systems but sensorial capabilities have not been developed completely to provide the required flexibility and autonomy for manufacturing tasks.
Basic requirements within an industrial production environment have to be satisfied to guarantee an acceptable manufacturing process; some factors are the tool or work-piece position uncertainty, which is achieved by using expensive structured manufacturing cells. Other factors are the force-torque and interaction evaluation with the task environment. By using self-adaptive robots with sensorial capabilities and skill learning on-line, great flexibility and adaptability is given to manufactured processes, so the idea of giving machines capabilities like humans in learning and execution tasks becomes real, (L.Wu, 1996).

These ideas points to use in manufacturing applications today, what is called *Intelligent Manufacturing Systems*, and can be thought as a high technology set of tool devices arranged within a working close ambient called manufacturing cell, and having an efficient collective team work, to achieve an autonomous manufacturing process with on line reprogramming facilities and having a manufactured product as it output. Such an intelligent system then, becomes a self-adapting and self-learning system (figure 4).

5.2. Assembly

The success of assembly operations using industrial robots is currently based on the accuracy of the robot itself and the precise knowledge of the environment, i.e., information about the geometry of the assembly parts and their localisation in the workspace. Robot manipulators operate in real world situations with a high degree of uncertainty and require sensing systems to compensate from potential errors during operations. Uncertainties come from a wide variety of sources such as robot positioning errors, gear backlash, arm deflection, ageing of mechanisms and disturbances. Controlling all the above aspects would certainly be a very difficult task; therefore a simpler approach is preferred like using vision-guided robots for aligning parts in assembly tasks.

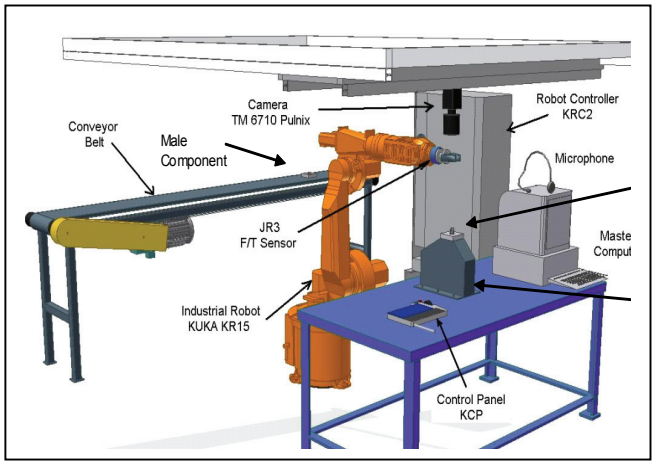


Figure 4. Intelligent Manufacturing System

6. Vision system

Machine Vision systems are mainly used today in applications as inspection, quality control and assembly tasks, they have been adopted now as a necessary technology in modern automated industries based in functional parameters and can be seen as a technology which is connecting cameras with computers for real-time interpretation of industrial behaviour images to acquire manufacturing applications. In this chapter a novel method to this purpose is presented and several tests were carried out to assess a vision-guided assembly process using aluminium pegs with different cross-sectional geometry, they are named: circular, squared and radiused-square (termed radiused-square because it was a square peg with one corner rounded). These components are shown in Figure 5 as well as the peg-in-hole operation in Figure 6. The diameter of the circular peg was 25 mm and the side of the square peg was also 25 mm. The dimensions of the non-symmetric part, the radiused-square, was the same as the squared peg with one corner rounded to a radius of 12.5 mm. Clearances between pegs and mating pairs were 0.1 mm, chamfers were at 45 degrees with 5 mm width. The assembly was ended when 3/4 of the body of the peg were inside the hole. This represented 140 motion steps in the -Z assembly direction.

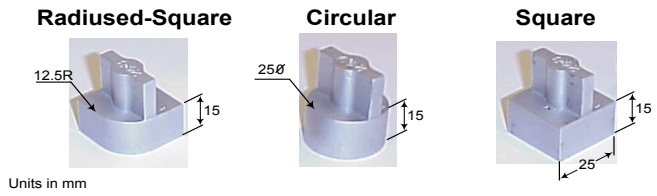


Figure 5. Assembly Components

Figure 6. Peg-in-hole operation

In our experiments, the robot grasps pieces from a conveyor belt and performs an assembly task using a force-sensing system described in (Corona-Castuera & Lopez-Juarez, 2006), the vision system obtains an image to recognize and calculates the object's pose estimation and sends the information to the robot.

6.1 Vision workspace

The vision system was implemented with a high speed camera CCD/B&W, PULNIX 6710, with 640x480 resolution and a PC dedicated computer, the

camera movements above the X-Y plane was implemented with a computer controlled 2D positioning electro-mechanical system (figure 7).

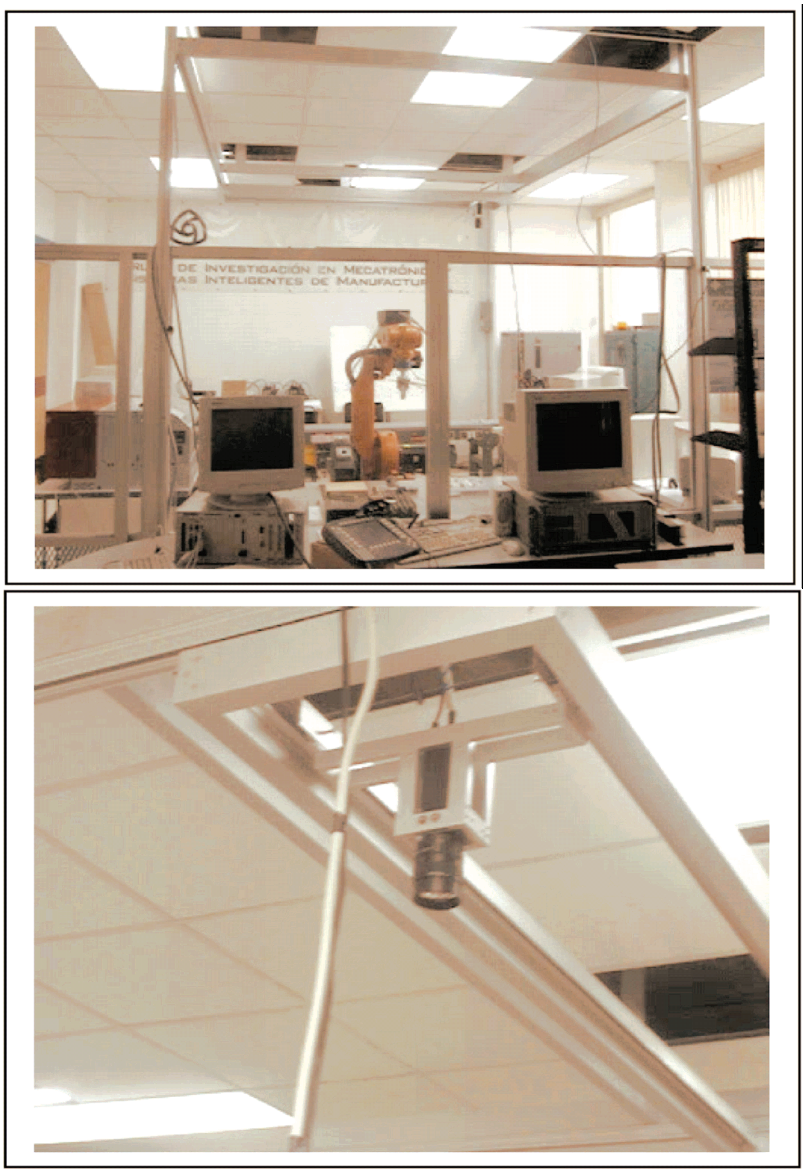


Figure 7. Vision workspace. Overview and 2D close up of positioning system

The vision system interaction schedule is working in a distributed systems as described in (Lopez-Juarez & Rios-Cabrera, 2006) linked to a robotic assembly module and a custom interface with a camera positioning system configured as a monocular dynamic system.

The vision system can get visual information from the manufacturing system workspace. To achieve an assembly task, the robotic assembly system sends commands to the vision system as follows:

$SENDINF#1 Send Information of Zone 1:

zone 1 is the place where the robot grasps the male components. The robot can locate different pieces and their characteristics.

$SENDINF#2 Send information of zone 2:

zone 2 is the place where the robot is performing the assembly task. The assembly system can request information about the female component such as position and shape.

$RESEND#X Resend information of zone X:

This command will be useful when the information received by the assembly system coming from the vision system is incorrect, due to an error in the check sum or any other error.
The communication protocol is as follows:

# Zone	Command	Type	C-Sum

The response from the vision system is a function of the request command from the assembly system, which coordinates the activities of the intelligent manufacturing cell (Corona-Castuera & Lopez-Juarez, 2004).

6.2 Invariant Object Recognition Methodology

The proposed methodology for invariant object recognition is based on the use of canonic shapes within what we have called the Primitive Knowledge Base (PKB). This PKB is conformed at training stage, once having embedded this knowledge, the idea is to improve and refine it on-line, which compares favourably with Gestalt principles such as grouping, proximity, similarity and simplicity. To illustrate the methodology, it will be useful to consider the assembly components used during experiments. The 2D representation of the working assembly pieces is shown in figure 8.

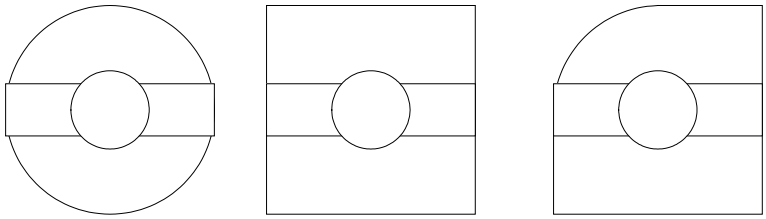


Figure 8. Assembly pieces

These canonical shapes serve as "clues" inserted initially in the PKB which initialise the grouping process (clustering). The knowledge is acquired by presenting multiple instances of the object such as those shown in figure 9 where an example of the circular shape and some of the possible views are illustrated. The following step is to code the object's information to get a *descriptor vector*, so that its description be invariant to location, scaling and rotation, the algorithm is explained in the following section.

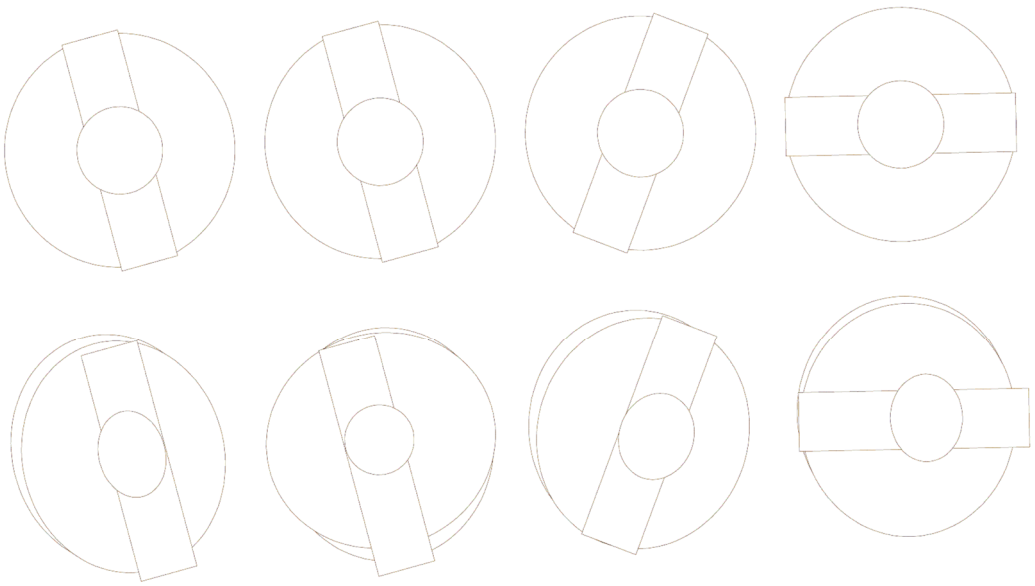


Figure 9. Multiple instances of the circular shape.

Having such a descriptor vector, an ANN can be trained to conform the *descriptor vector families* which can be generated on-line with the vision system.

6.3 Methodology

CFD&POSE methodology steps are:

- Fast acquisition of working visual scene
- Find the region of interest (ROI)
- Calculate the histogram of the image.
- Search for pieces
- Centroid calculation
- Piece orientation
- Calculate boundary object function (BOF)
- Descriptor vector generation and normalization (CFD&POSE).
- Information processing in the neural network

6.3.1 Fast acquisition of working visual scene

Image acquisition is carried out by the vision workspace configuration described previously, comprised with a CCD/B&W digital camera, frame grabber and custom visual C++ based software acquisition.

6.3.2 Finding the region of interest

It is desirable first to segment the region of the whole scene to have only the workpieces *Region of Interest* (ROI). There are two defined regions of interest in the manufacturing cell:

- the assembly workspace (zone 1)
- the identification/grasping workspace (zone 2).

The camera has to be positioned in the vision zone requested by the robot. The 2D positioning system, which uses feedback vision using a searching algorithm, employs two LED's within the scene as a calibration reference in order to reach the exact position of the camera vision system (figure 10). The original image is 480 x 640 pixels, 8-bit greyscale resolution. Image conditioning is carried out avoiding the processing of small objects and finding the initial position of the desired zone. The quantized grey level value of the LEDs in the image, is greater than or equal to a specific gray level value GL regardless of the amount of light in the zone. With this process, most of the objects that can confuse the system are rejected. Then the ROI is first extracted by using the 2D histogram information and initial position reference.

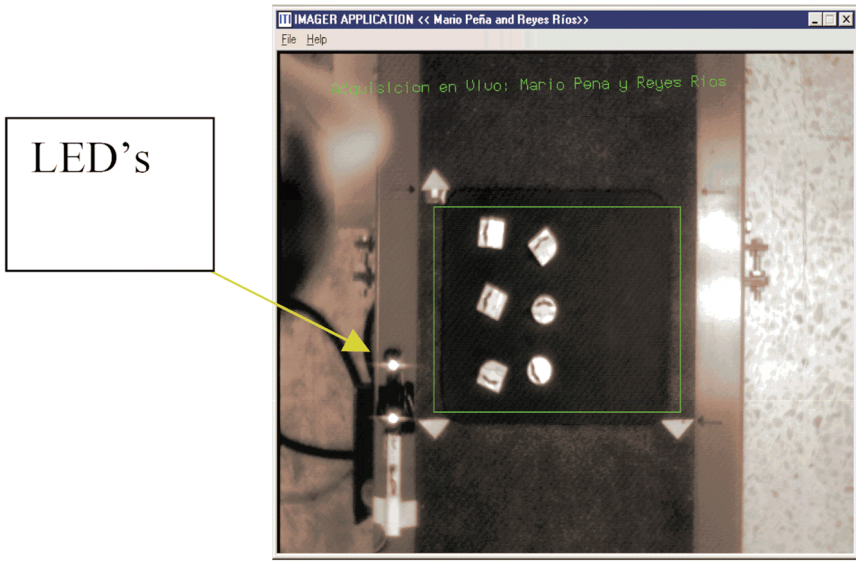


Figure 10. Zone 1 vision workspace

To determine which are the more approximated white blobs within the image, it has to be considered the mark using the following criteria:

- colour GL > 245
- 25 ≤ Perimeter ≤ 35 pixels (i.e., LED measured size)
- Distance between LED's, must be constant (50 mm ± 3 mm).

In the initial position search, only the objects that fulfil all mentioned characteristics are processed, all others are rejected. In this way, initial position is found and ROI is defined as it is shown in figure 10.

6.3.3 Image histogram process

An algorithm using 1D and 2D image histograms is used in order to provide the system of illumination invariance within some specific range. From these histograms, threshold values are used for image segmentation of the background and the pieces within the ROI eliminating the noise that may appear. This dynamic threshold value calculation allows independent light conditions operation of the system. The 1D histogram normally has the aspect shown in figure 11.

The 2 peaks in the histogram represent the background and the pieces in the image. After the histogram calculation, an image binarization is performed using a threshold operator.

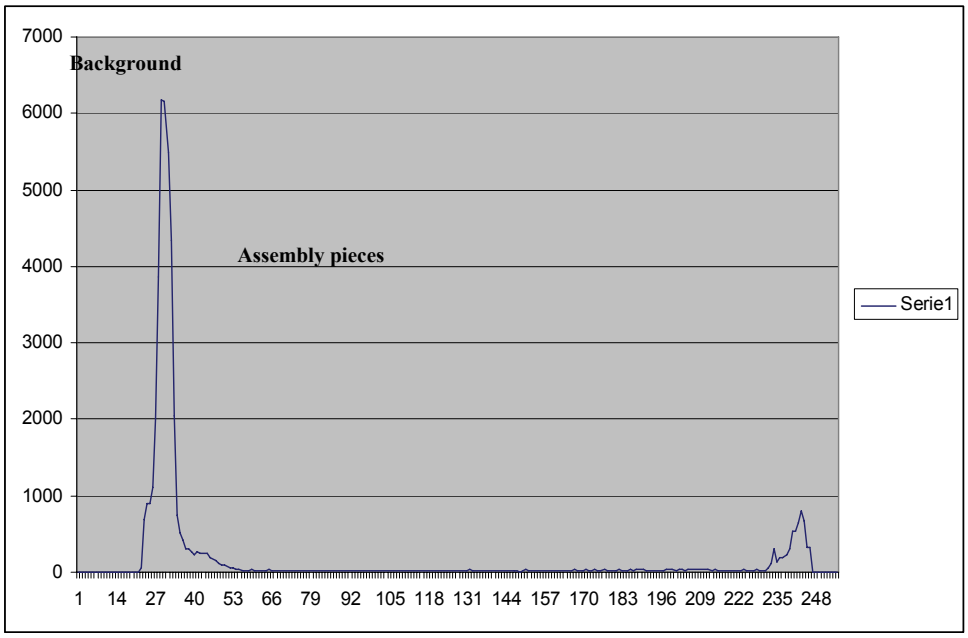


Figure 11. Histogram of the region of interest (ROI).

6.3.4 Search for pieces

For searching purposes, the system calculates the perimeter obtaining:

- number of points around a piece
- group of points coordinates X&Y, corresponding to the perimeter of the piece measured clockwise
- boundaries of the piece 2D Bounding Box (2D-BB)

The perimeter calculation for every piece in the ROI is performed after the binarization. Search is always accomplished from left to right and from top to bottom. Once a white pixel is found, all the perimeter is calculated with a search function (figure 12).

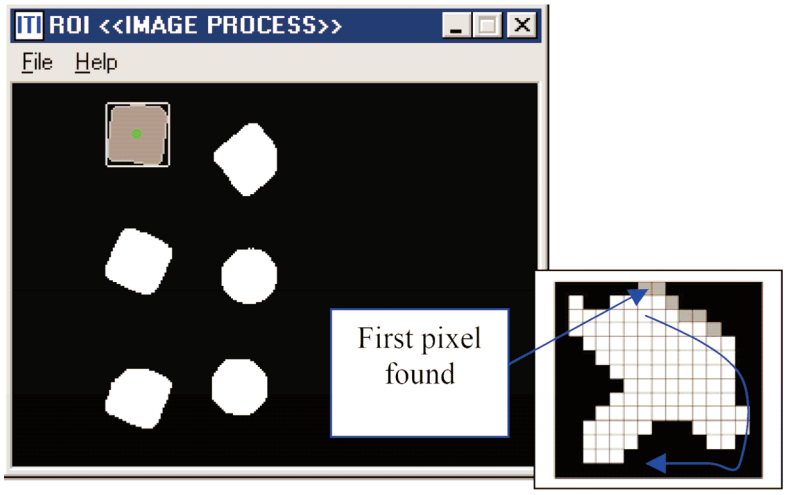


Figure 12. Perimeter calculation of a workpiece

The next definitions are useful to understand the algorithm:

- A *nearer pixel to the boundary* is any pixel surrounded mostly by black pixels in connectivity eight.
- A *farther pixel to the boundary* is any pixel that is not surrounded by black pixels in connectivity eight.
- The *highest and lowest* coordinates are the ones that create a rectangle (Boundary Box).

The search algorithm executes the following procedures once it has found a white pixel:

1. *Searches for the nearer pixel to the boundary that has not been already located.*
2. *Assigns the label of actual pixel to the nearer pixel to the boundary recently found.*
3. *Paints the last pixel as a visited pixel.*
4. *If the new coordinates are higher than the last higher coordinates, it is assigned the new values to the higher coordinates.*
5. *If the new coordinates are lower than the last lower coordinates, it is assigned the new values to the lower coordinates.*
6. *Steps 1 to 5 are repeated until the procedure begins to the initial point, or no other nearer pixel to the boundary is found.*

This technique will surround any irregular shape, and will not process useless pixels of the image, therefore this is a fast algorithm that can perform online classification, and can be classified as linear:

$$O\ (N * 8*4)$$

where *N* is the size of the perimeter, and 8 & 4 are the number of comparisons the algorithm needs to find the pixel farer to the boundary, the main difference with the traditional algorithm consist of making the sweep in an uncertain area which is always larger than the figure, this turns the algorithm into:

$$O(N*M)$$

*N*M*, is the size of the Boundary Box in use, and it does not obtain the coordinates of the perimeter in the desired order.

6.3.5 Centroid calculation

The proposed procedure for centroid calculation is performed at the same time that the coordinates of the perimeter are calculated without using the *N*M* pixels box, (*Boundary Box*).
The coordinates of the centroid (*Xc, Yc*) are calculated with the following procedure:

1. *If a new pixel is found and it has not been added, the value of i, j coordinates from pixel to the left is added, until a new black or visited pixel is found.*
2. *While a new pixel is found repeat step 1.*

Figure 13 demonstrates how the sum is made from right to left as indicated by the black arrows.

The equation (1) is used for centroid calculation in binarized images:

$$X_c = \frac{\sum_{x,y} j}{A}, Y_c = \frac{\sum_{x,y} i}{A} \qquad (1)$$

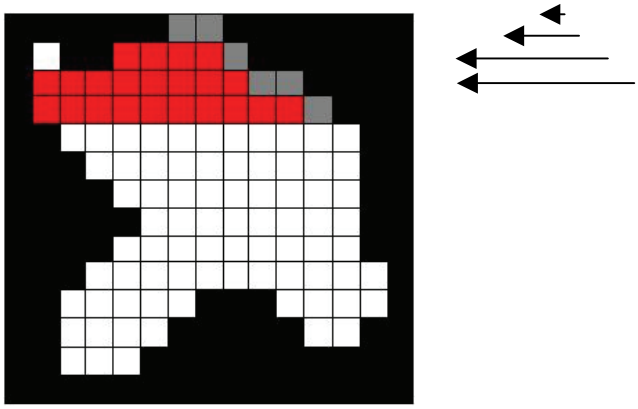


Figure 13. Centroid calculation

Where A is the area or number of pixels that composes the piece.

6.3.6 Piece orientation

The projected shadow by the pieces is used to obtain its orientation. Within the shadow, the largest straight line is used to calculate the orientation angle of the piece using the slope of this line, see figure 14.
The negative image of the shadow is obtained becoming a white object, from which, the perimeter is calculated and also the two most distant points (x_1 y_1, x_2 y_2) are determined.

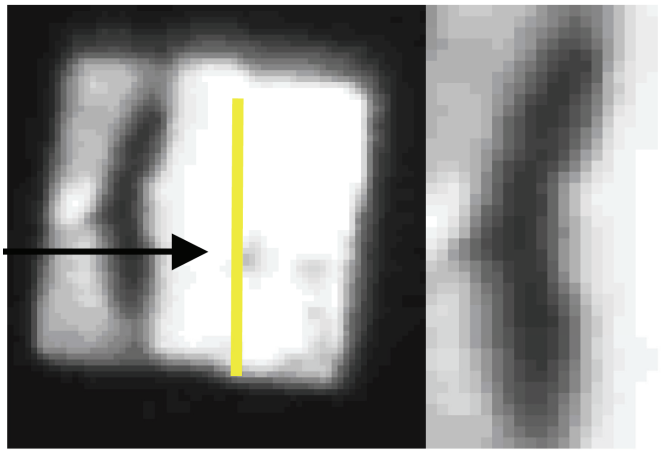


Figure 14. Shadow for the orientation

These points define the largest straight line, the equation for the distance between 2 points is used to verify if it is the largest straight line, and also if it contains the centroid using equation (2).

$$Y_C - y_1 = m(X_C - x_1) \tag{2}$$

The slope is obtained using equation (3):

$$m = \frac{y_2 - y_2}{x_2 - x_1} \tag{3}$$

6.3.7 Boundary Object Function (BOF)

The Boundary Object Function (BOF), is the function that describes a specific piece and it will vary according to the shape. This is illustrated in figure 15.

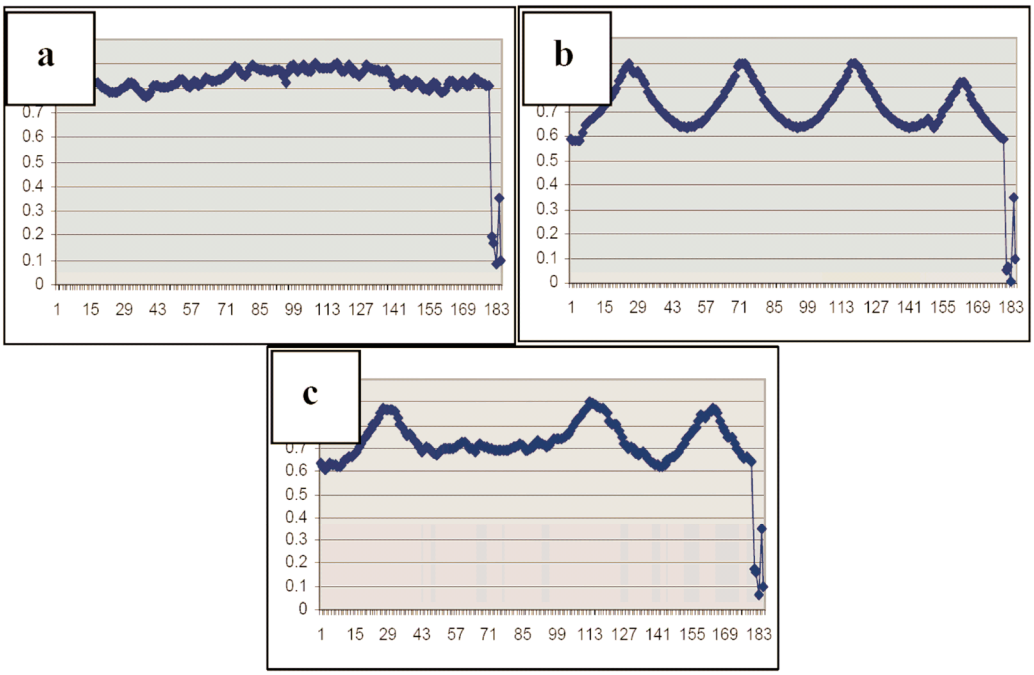


Figure 15. BOF a) circle, b) square, c) radiused-square

The centroid, the coordinates of the perimeter and the distance from the centroid to the perimeter points are used to calculate the BOF.

With the coordinates P_1 (X_1, Y_1) and P_2 (X_2, Y_2), equation (4) is applied:

$$d(P_1, P_2) = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2} \tag{4}$$

6.3.8 Descriptive vector generation and normalization

Once the information has been processed, a descriptive vector is generated. This vector is the input to the neural network. The descriptive vector is called CFD&POSE and it is conformed by:

$$[CFD \ \& \ POSE] = \begin{bmatrix} D_1 \\ D_2 \\ D_3 \\ D_n \\ X_c \\ Y_c \\ \phi \\ Z \\ ID \end{bmatrix} \tag{5}$$

where:

- D_i is the distance from the centroid to the object's perimeter point.
- X_c, Y_c, are the coordinates of the centroid.
- ϕ, is the orientation angle.
- Z is the height of the object.
- ***ID*** is a code number related to the geometry of the components.

6.3.9 Information processing in the neural network

The vision system extends the BOF data vectors to 180, plus 4 more data vectors, centroid (Xc, Yc), orientation, height and ID as showed to conform the descriptor vector which is the input to the FuzzyARTMAP neural network. :

Data	Centroid	Orientation	Height	ID
1-180	181-182	183	184	185

7. Experimental Results

7.1 Training and recognition on-line

In order to test the architecture, experimental work was carried out with the distributed manufacturing sytem using the vision system, to achieve the task and to test the robustness of the ANN, the Fuzzy ARTMAP Neural Network was trained first with 2808 different patterns corresponding to the described working pieces and the learning capability was analyzed. Results regarding the percentage of recognition and the number of generated neurons are shown in figure 16. The graph shows how the system learned all patterns in three epochs, creating only 32 neurons to classify 2808 patterns.

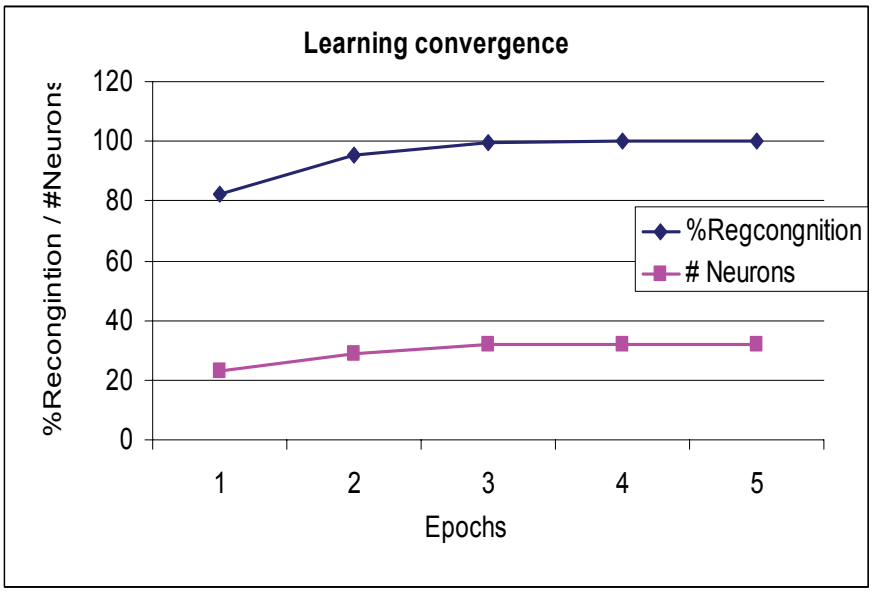


Figure 16. Learning of the neural network

The average time for **training wa**s 4.42 ms, whereas for **testing** was 1.0 ms. Later a normalization procedure was applied to the descriptor vector CFD&POSE so that the experimental work employed only 216 patterns corresponding to 72 square, 72 circle and 72 radiused-square components of the same size. The orientation values were 0, 45, 90, 135, 180, 215, 270 and 315 degrees. With these training patterns set, the system was able to classify correctly

100% of the pieces presented on-line even if they were not of the same size, orientation or locations and for different light conditions. The pieces used to train the neural network are shown in figure 17 and figure 18 which show different graphs corresponding to different descriptor vectors for different positions, sizes and illumination conditions of these components.

Figure 17. Workpieces used to create the initial knowledge base

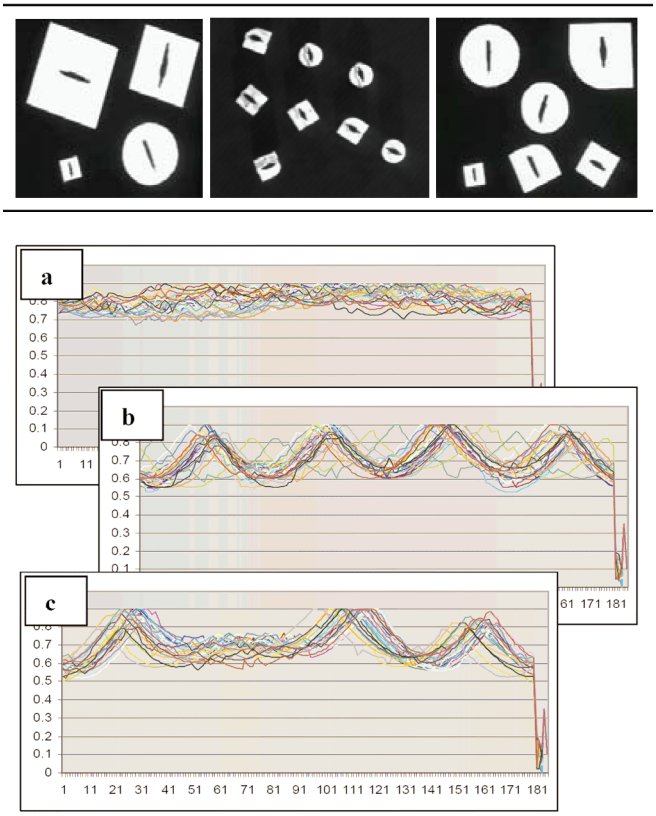


Figure 18. workpieces used to test the system. a) circle, b) square, c) radiused-square several tests with different geometry, positions and light conditions, were carried out on-line.

The normalization of the BOF is done using the *maximum value divisor of the distance* vector method. This method allows having very similar patterns as input vectors to the neural network, getting a significant improvement in the operation system. Figure 19 shows the generated similar patterns using totally different size, location and orientation conditions for working pieces.

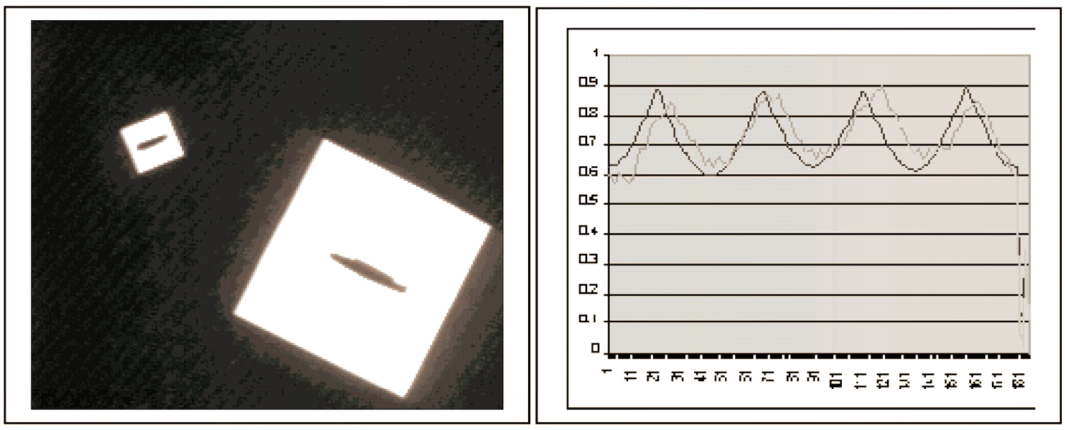


Figure 19. a) Squares b) Similar Patterns

7.2 Assembly cycles in the distributed manufacturing system

In order to test the methodology within the distributed manufacturing system, a complete robotic assembly operation was carried out and the results are given in Table 1. The table shows the results for 18 assembly cycles using the vision system and the assembly system of the cell. This table contains information regarding the type of piece in use, presence or absence of chamfer, total operational time (object recognition, grasping, moving the part from pick up place to assembly point and assembly), the calculated error based on the centroid and rotation angle of the pieces for zone 1 and the offset error in zone 2. Finally, in the last column, the type of geometry recognized on-line by the neural network is provided. The vision system provides the robot with the capability to approach 2 zones:

Zone 1: assembly workpiece (male) area of vision, where the robot picks up the piece after having POSE information of the object, then it grasps the piece and takes it to zone 2.

Zone 2: peg-in-hole assembly (female) area of vision, here the visually guided robot approaches the zone where the female component is located to achieve the assembly task and releasing the control of the operation to the SIEM assembly system (Corona-Castuera & Lopez-Juarez, 2006).

POSE 1 means the location estimation of a workpiece within the zone 1, and
POSE 2 means the location estimation within the zone 2 of the work piece/counterpart.

Grasp testing (zone 1) was accomplished for each geometry; every one was placed three times within the vision area, incrementing 10 degrees its orientation and changing the locations in four defined poses. In the zone 2, the location of the female component was fixed, hence the angle. However, It is important to mention that the POSE was unknown to the assembly controller.

#			TC	TA	ZONE 1			Zone 1 Error			ZONE 2		Zone 2 Error		
IN	P	Ch	(min)	(s)	Xmm	Ymm	RZ°	Xmm	Ymm	RZ°	Xmm	Ymm	Xmm	Ymm	NC
1	S	Y	1:15	32.5	62.4	144.1	10	0.2	-1.3	0	84.6	102.1	0.3	-1	Y
2	S	Y	1:15	30.4	62.4	45.7	12	1.8	0.2	2	85.6	101.1	-0.7	0	Y
3	S	Y	1:15	31.8	178.7	47.7	23	0.9	-0.8	3	84.7	100.9	0.2	0.2	Y
4	R	Y	1:11	30.1	181.6	147	29	-0.3	-0.7	-1	84.7	100.6	0.2	0.5	Y
5	R	Y	1:14	29.4	62.4	145.1	36	0.2	-0.3	-4	84.9	100.7	0	0.4	Y
6	R	Y	1:19	29.6	67.3	44.8	48	3.1	-0.7	-2	85.3	101.6	-0.4	-0.5	Y
7	C	Y	1:15	29.6	180.6	49.6	57	1	1.1	-3	84.6	102.4	0.3	-1.3	Y
8	C	Y	1:13	30.2	180.6	148	77	-0.7	0.3	7	84.3	101	0.6	0.1	Y
9	C	Y	1:14	30.2	61.5	146	79	-0.7	0.6	-1	83.9	101.6	1	-0.5	Y
10	S	N	1:18	29.9	63.4	45.7	83	-0.8	0.2	-7	85.4	100.5	-0.5	0.6	Y
11	S	N	1:19	30.4	179.6	48.6	104	0	0.1	4	83.2	100.8	1.7	0.3	Y
12	S	N	1:22	34.6	180.6	147	104	-0.7	-0.7	-6	83.2	101.8	1.7	-0.7	Y
13	R	N	1:22	38.3	61.5	146	119	-0.7	0.6	-1	84.8	102.8	0.1	-1.7	Y
14	R	N	1:22	36.8	63.4	43.8	126	-0.8	1.7	-4	83.6	101.8	1.6	-0.7	Y
15	R	N	1:24	36.6	179.6	47.7	138	0	-0.8	-2	83.2	101.7	1.7	-0.6	Y
16	C	N	1:17	30.5	182.6	149	150	1.3	1.3	0	83.7	101.2	1.2	-0.1	Y
17	C	N	1:15	28.3	63.4	146	155	1.2	0.6	-5	84.6	100.7	0.3	0.4	Y
18	C	N	1:15	29.7	64.4	47.7	174	0.2	2.2	4	83.9	101.1	1	0	Y

Table 1. 18 assembly cycles using the vision system and the assembly system (Eighteen different assembly cycles, where IN= Insertion, P=piece, Ch=chamfer present, TC=Assembly cycle time, TA= Insertion time, NC=correct neural classification, S=square, R=radiused-square, C=circle, N=no and Y=yes).

The first 9 assembly cycles were done with female chamfered components and the last 9 with chamferless components.

The average time of the **total cycle** is 1:50.6 minutes and the minimum time is 1:46 minutes, the longest time is: 1:58 minutes.

The average of the **error made** in both zones is: 0.8625 mm, the minimum is: 0 mm while the maximum is 3.4 mm.

The average of the **error angle** is: 4.27 ° , the minimum is: 0° and the maximum is 9°.

The figure 20, shows eighteen different X and Y points where the robot might reach the male component showed as error X(mm) and error Y(mm).

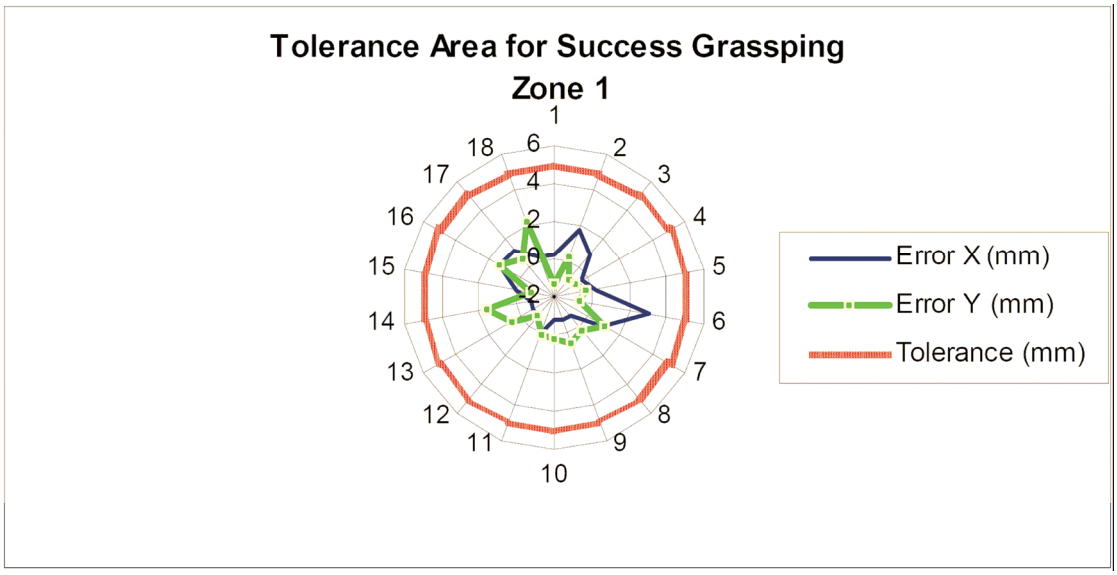


Figure 20. Positional error referenced to real centroid in male component

In the assembly area the robot gets vision guided capabilities to approach the zone to the centre of the workpiece/counterpart, the figure 22 shows eighteen different X and Y points where the robot might reach the female and releases control to force/sensing system.

The 18 assembly cycles were done successfully. The figures 20, 21 y 22 show that all the poses given by the vision system are inside the error limits in both areas: zone 1 and zone 2. This permitted to have a 100% of success in the total assembly cycle operation

Figure 21 shows the angle error for orientation grasping purpose

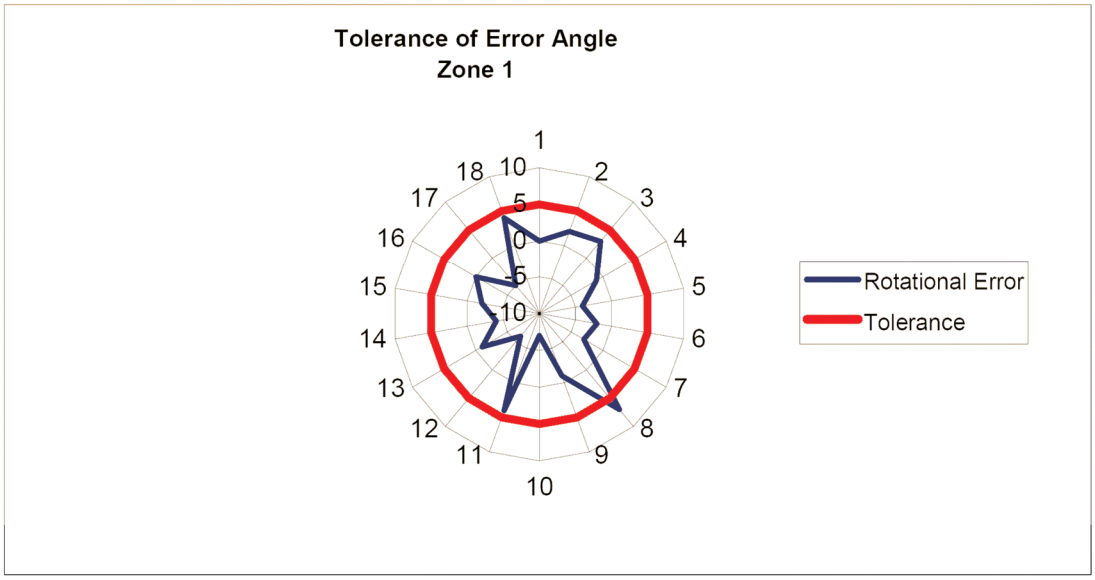


Figure 21. Rotational error for orientation grasp

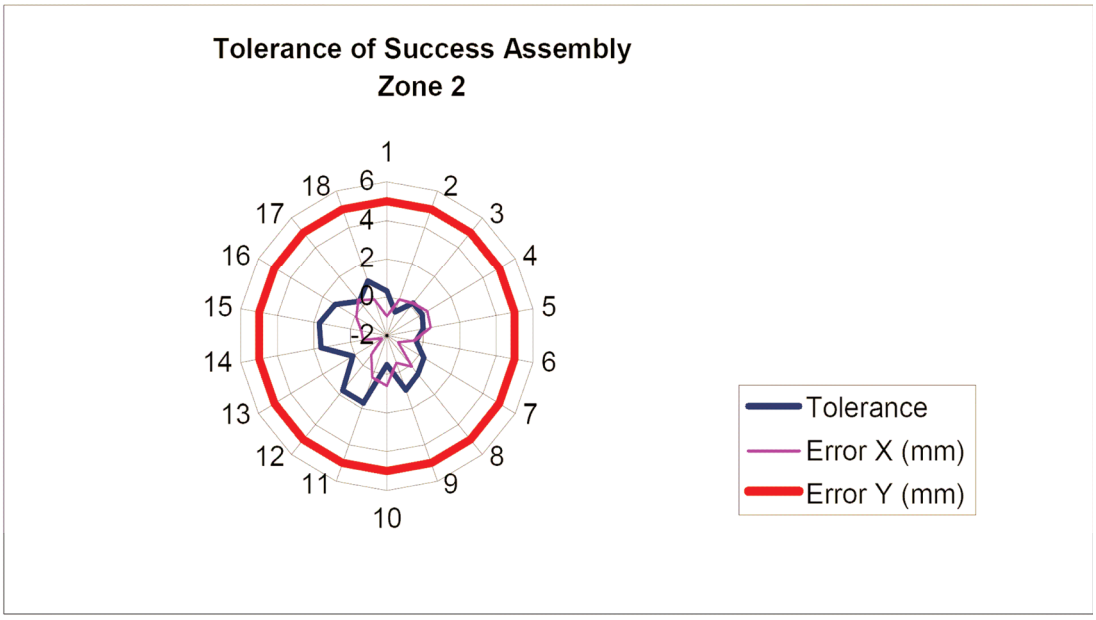


Figure 22. Positional error referenced to real centroid in female component.

8. Conclusions and future work

A novel methodology for fast object recognition and POSE estimation for assembly components in a distributed manufacturing system has been described. Experimental results show the methodology. Issues regarding image processing, centroid and perimeter calculation are illustrated. The methodology was tested on a distributed manufacturing system using an industrial manipulator to perform assembly operations. Results show the feasibility of the method to send grasping and morphologic information (coordinates and classification characteristics) to the robot in real-time. A robust positioning system that corrected errors due to wheel sliding was implemented using visual feedback information. The overall methodology was implemented and integrated in a manufacturing cell showing real performance of industrial processes. Accurate recognition of assembly components and workpieces identification was successfully carried out by using a FuzzyARTMAP neural network model. The performance of this model was satisfactory with recognition times lower than 5 ms and identification rates of 100%. Experimental measurements showed ±3 millimeter of precision error in the information sent to the robot. The orientation angle error for the pieces was up to ±9 degrees, which was still good enough for the robot to grasp the pieces. Future work is envisaged using the intelligent distributed manufacturing system with multimodal and fusion sensor capabilities using the methodology presented in this work. Current work addresses the use of ANN's for assembly and object recognition separately; however work is oriented towards the use of the same neural controller in a hierarchical form for all other different sensorial modalities (Lopez-Juarez & Rios-Cabrera, 2006).

9. References

Aguado A., E. Montiel, M. Nixon . Invariant characterization of the Hough Transform for pose estimation of arbitrary shapes. Pattern Recognition 35 , 1083-1097 , Pergamon, (2002).

Best Paul J.and Neil D. McKay. A Method for Registration of 3-D Shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol 14 No. 2, February (1992).

Bribiesca E. A new Chain Code. Pattern Recognition 32 , Pergamon, 235-251 , (1999).

Bone Gary M. and David Capson. Vision-guided fixturless assembly of automotive components. Robotics and Computer Integrated Manufacturing 19, 79-87 , (2003).

Carpenter Gail A. and Stephen Grossberg. Computer Vision, Graphics, and Image Processing. A Massively Parallel Architecture for a Self-Organizing Neural Pattern Recognition Machine. Academic Press, Inc. Pp. 54-115. 1987.

Carpenter Gail A. and Stephen Grossberg, John H Reynolds. ARTMAP: Supervised Real-Time Learning and Classification of Nonstationary Data by Self-Organizing Neural Network. Neural Networks. Pp 565-588. 1991.

Cem Yüceer adn Kema Oflazer, A rotation, scaling and translation invariant pattern classification system. Pattern Recognition, vol 26, No. 5 pp. 687-710, (1993).

Cervera Enric and Angel P. del Pobil. Programming and learning in real world manipulation tasks. Proc. 1997 IEEE/RSJ Int Conf on Inteligent Robot and Systems, 1:471-476, September (1997).

Chen K., Efficient parallel algorithms for computation of two-dimensional image moments, Pattern Recognition 23, 109-119, (1990).

Chin-Hsiung Wu et al. Anew computation of shape moments via quadtree decomposition. Pattern Recognition 34 , 1319-1330, Pergamon, (2001).

Corona Castuera J. and Ismael Lopez-Juarez, "Intelligent Task Level Planning for Robotic Assembly: Issues and Experiments", Mexican International Conferences on Artificial Intelligence, México 2004, ISBN 3-540-21459-3, Springer-Verlag.

Corona Castuera, J; Lopez-Juarez, I. (2006). Distributed Architecture for Intelligent Robotic Assembly, Part II: Design of the task planner. ADVANCED TECHNOLOGIES: Research-Development-Application. Submitted for publication.

Dickmanns E. "Vehicles capable of dynamic vision: a new breed of technical beings?", Artifical Intelligence, vol 103, pp, 49-76, August (1998).

Freeman H., Computer processing of line drawings images, ACM Comput. Surveys 6 57-97, (1974).

Freeman H., On the encoding of arbitrary geometric configurations, IRE Trans. Electron. Comput. EC-10 , 260-268, (1961).

Gupta Madan M., G. Knopf. Neuro-Vision Systems: a tutorial. A selected reprint Volume IEEE Neural Networks Council Sponsor, IEEE Press, New York, 1993.

Gupta Madan M., G. Knopf. Neuro-Vision Systems: Part 6 Computational Architectures and Applications. A selected reprint Volume IEEE Neural Networks Council Sponsor, IEEE Press, New York, 1993.

Gonzalez-Galvan Emilio J. et al. Application of Precision-Enhancing Measure in 3D Rigid-Body Positioning using Camera-Space Manipulation, The International Journal of Robotics Research, vol 16, No. 2, pp. 240-257, April (1997).

Geoffrey G. Towell; Jude W. Shavlik. Knowledge-based artificial neural networksArtificial Intelligence. Vol. 70, Issue 1-2, pp. 119-166. 1994

Robert S. Feldman, Understanding Psychology, 3rd edition. Mc Graw-Hill, Inc., 1993.

Grossberg Stephen, Adaptive Pattern Classification and universal recoding II: Feedback, expectation, olfaction and illusions. Biological Cybernetics. Vol. 23, pp. 187-202, 1976.

Gullapalli Vijaykumar; Judy A. Franklin; Hamid Benbrahim. Acquiring robot skills via reinforcement learning. IEEE Control Systems, pages 13-24, February (1994).

Hager G., et al, Calibration-free visual control using projective invariance, proceedings ICCV, pp 1009-1015, (1995).

Hiroshi Murase and Shree K. Nayar, Visual Learning and Recognition of 3-D Objects from Appearenace. International Journal of Computer Vision, 14, 5-24 (1995).

Hoska DR. Fixturless assembly manufacturing. Manuf Eng , 100:49-54 , April (1988).

Howarth Martin. An investigation of task level programming for robotic assembly. PhD thesis, The Nottingham Trent University, January 1998.

Hu M.K. Visual pattern recognition by moment invariants, IRE Trans Inform Theory IT-8, 179-187, (1962).

Jörg Stefan et. al. Flexible Robot-Assembly using a Multi-Sensory Approach. In Proc. IEEE, Int. Conference on Robotics and Automation, San Fco. Calif, USA (2000), pp 3687-3694.

Kollnig H. and H. Nagel. 3d pose estimation by directly matching polyhedral models to gray value gradients. International Journal of Computer Vision, vol 23, No. 3, pp 282.302, (1997).

Kronauer R.E. , Y. Zeevi . Reorganization and Diversification of Signals in Vision. IEEE Trans. Syst. Man, Cybern., SMC-15,1,91-101. (1985).

Langley C.S., D. Eleuterio GMT, A memory efficient neural network for robotic pose estimation, In proceedings of the 2003 IEEE International Symposium on Computational Intelligence in Robotics and Automation, No. 1 , 418-423, IEEE CIRA, (2003).

Lopez-Juarez I. Howarth M. "Learning, Manipulative Skills with ART", Proc 2000 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems. October 2000.

Lopez-Juarez I. On-line learning for robotic assembly using artificial neural networks and contact force sensing. PhD thesis, The Nottingham Trent University, (2000).

Lopez-Juarez, I; Rios-Cabrera, R. (2006) Distributed Architecture for Intelligent Robotic Assembly, Part I: Design and Multimodal Learning. ADVANCED TECHNOLOGIES: Research-Development-Application. Submitted for publication.

Mundy J.L. , Object Recognition based on geometry: progres over 3 decades. Phylosophical Transactions: mathematical, physical and engineering sciences, 356: 1213-1231, 1998.

Ngyuen W. and J.K. Mills. Multirobot control for flexible fixturless assembly of flexible sheet metal autobody parts. In proceedings of IEEE International Conference on Robotics and Automation, , pp. 2340-2345, (1996).

Papanikolopoulous N. P. and K. Khosla. Adaptive robotic visual tracking: theory and experiments. IEEE Transactions on Automatic Controller, vol. 38, No. 3, pp 429-445, (1993).

Peña-Cabrera M. et al, "A Learning Approach for On-Line Object Recognition in Robotic Tasks", Mexican International Conference on Computer Science ENC 2004, México, IEEE Computer Society Press. (2004).

Peña Cabrera M. et al, "Un Proceso de Aprendizaje para Reconocimiento de Objetos en Línea en Tareas Robotizadas", 3ra Conferencia Iberoamericana en Sistemas, Cibernética e Informática CISCI 2004, del 21 al 25 de julio del 2004, Orlando, Florida EE.UU.

Philips W., A new fast algorithm for moment computation, Pattern Recognition 26, 1619-1621, (1993).

Plut W.J. and G.M. Bone. Limited mobility grasps for fixturless assembly, In proceedings of the IEEE International Conference on Robotics and Automation , Minneapolis, Minn., pp. 1465-1470, (1996).

Plut W.J. and G.M. Bone. 3-d flexible fixturing using multi-degree of freedom gripper for robotics fixturless assembly. In proceedings of the IIEEE In-

ternational Conference on Robotics and Automation, Alburquerque, NM, pp. 379-384, (1997).

Shaogang Gong, Hilary Buxton, (2002). Understanding Visual Behaviour, Image and Vision Computing, vol 20, No.12, Elsevier Science., (2002).

Shepard R.N. and S Hurwitz. Upwards direction, mental rotation and discrimination of left and right turns in maps. Cognition, 18, 161-193, 1984.

Shingchern D. You , Gary E. Ford, Network model for invariant object recognition. Pattern Recognition Letters 15, 761-767, (1994).

Stavros J. and Paulo Lisboa. Transltion, Rotation , and Scale Invariant Pattern Recognition by High-Order Neural networks and Moment Classifiers., IEEE Transactions on Neural Networks, vol 3, No. 2 , March (1992).

Stephen A. Underwood et al. Visual Learning from Multiple Views. IEEE Transactions on Computers, vol c-24, No. 6 , June (1975).

Torralba A. and A. Oliva. Depth Estimation from Image Structure. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol 24, No. 9 September (2002).

Uhr L. Psychological motivation and underlying concepts . Structured Computer Vision, S. Tanimoto, A. Klinger Ed. , 1-30, (1980).

Ulman S., Visual Routines Cognition, 18: 97-159, 1984.

Wu L., S. L. Oviatt, P. R. Cohen, " Multimodal Integration – A Statical View", IEEE Transactions on Multimedia, vol 1 , Num. 4, pp 334-341, (1999).

Yong-Sheng Chen et al. Three dimensional ego-motion estimation from motion fields observed with multiple cameras. Pattern Recognition 34, 1573-1583, Pergamon , (2001).

15

Assembly Sequence Planning Using Neural Network Approach

Cem Sinanoglu and Huseyin Riza Borklu

1. Introduction

The competition between manufacturing firm's makes it necessary that those firms must supply highly quality goods with shorter time and chipper for survive in the international market. The intensive research in this field aims at augmenting methods and tools for product development and manufacturing. By the use of new and efficient methods it is possible to shorten the time from design to manufacturing and reduce the mistakes originating from humans. Therefore, full automation in the manufacturing processes can be accomplished.

An assembly can be defined as the overall function of individual parts after joining each other's, each of which has an independent function. It is possible to divide an assembly into various subassemblies depending on its complexity levels (Pahl & Beitz, 1988). Although intensive research efforts in the field of assembly sequence planning, there are still some problems to be solved. For example, there exist some shortcomings to support for full automatic assembly sequence planning or to obtain assembly plans for large systems (Singh, 1997).

By the development of efficient assembly planning systems it is possible to provide some advantages in both areas:

- CAD automation as well as
- Manufacturing performed by the use of robots

In order to develop a system that addresses computer-aided assembly sequence planning, these issues should be taken into consideration:

- The connectivity structure of parts and/or subassemblies that are used in the assembly system

- If they have connectivity properties, then what are the theatrical number of their different connection ways
- Finally the choice of the most optimum assembly process among various alternatives

Basic approach for finding the most suitable assembly sequences is to represent assembly system in a space where it is possible to explore all different assembly sequences. Thus, some criterion may be used to obtain these sequences. Then optimum assembly sequence can be selected by the application of some other criterion. This criterion may include: the number of tool changes, part replacement and clamping on the jigs, concurrency in operations, reliable subassemblies, etc (Kandi & Makino, 1996). A new approach may be stated as a rank evaluation from new scanning factor. But this may not exactly show the differences between operation time and costs.
The initial assembly planning system was inquired to user data necessity for operation and it was formed to assembly sequences with given data. It was worked with the user interaction (De Fazio, 1987, Homem de Mello & Sanderson, 1991). The later work concentrated on assembly sequence planning systems based on geometric reasoning capability and full automatic (Homem de Mello, 1991). Then, the system included some modules that allow assembly sequence plans to be controlled and to be tested.
Since many assembly sequences share common subsequences, attempts have been made to create more compact representations that can encompass all assembly sequences. Therefore, the works in this field are graph-based approaches to represent all feasible assembly sequences (FAS).
This paper presents a neural-network-based computational scheme to generate feasible assembly sequences for an assembly product. The inputs to the networks are the collection of assembly sequence data. This data is used to train the network using the Back propagation (BP) algorithm. The neural network model outperforms the feasible assembly sequence-planning model in predicting the optimum assembly sequences.

2. System Definition

This assembly planning system is used to assembly's connection graph (ACG) for the representation product. Parts and relations among these parts are represented by this graph. Contact relations between parts are supplied by scan-

ning module. Scanning module scans various assembly views of any product whose assembly view is produced by *CAD* packet program in the computer environment and determines to contact and interference relations between parts (Sinanoğlu & Börklü, 2004).

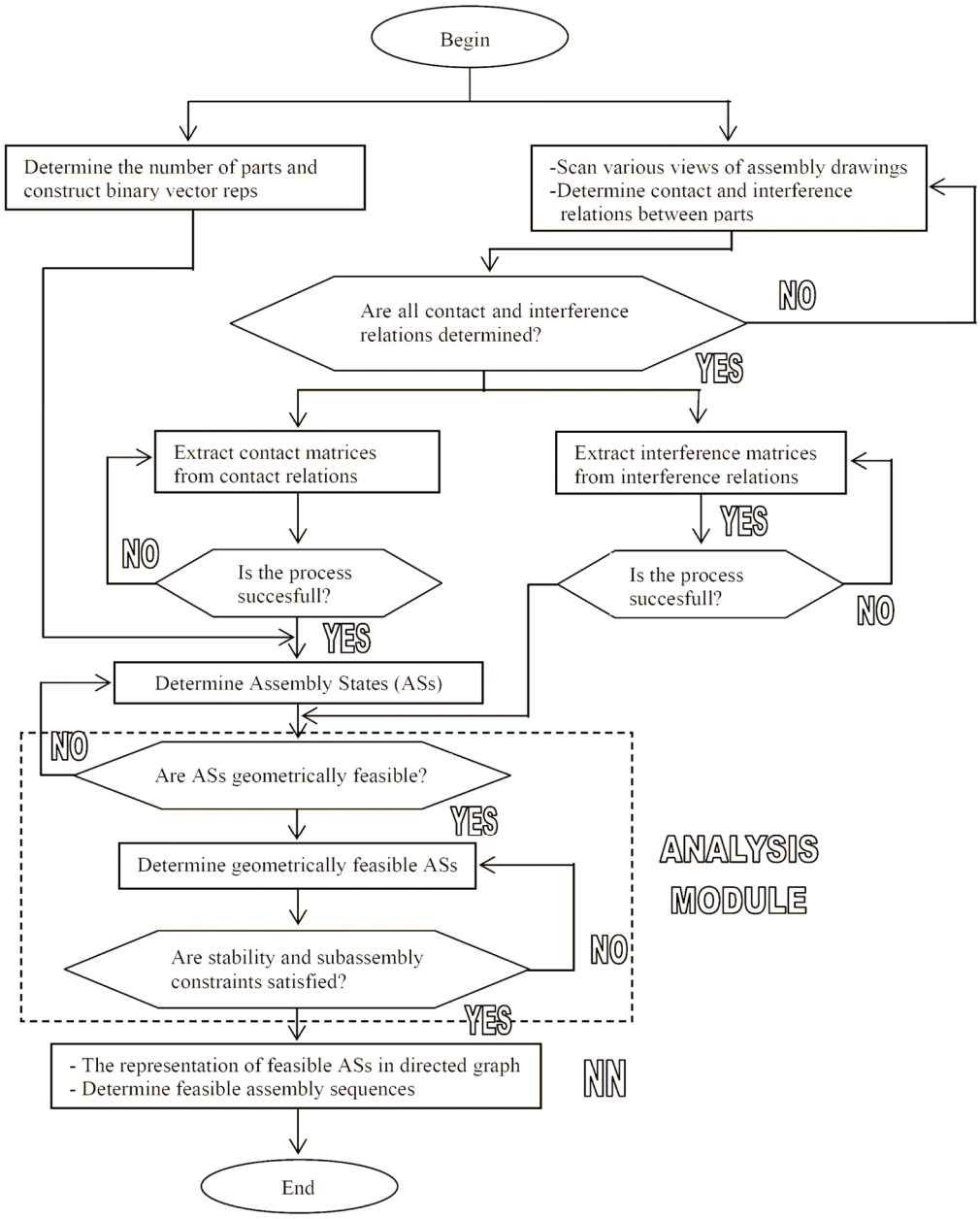


Figure 1. Block diagram of proposed assembly planning system

These relations are formed as a matrix. The system constitutes ACG of product to be assembled, by applying Boolean operators on the elements of the contact matrices according to certain rules. Moreover, interference relations are also accumulated in a form of interference matrices for determination of geometric feasibility of assembly states later (Sinanoğlu & Börklü, 2005).
Block diagram of assembly system is shown in Fig. 1. In the assembly planning system, binary vectors represent assembly states. Therefore, all binary vector representations, whether corresponding to assembly states or not, are produced by the system. By evaluating ACG and binary vector representations simultaneously with the scanning module, vector representations corresponding to assembly states are determined.
Some of the assembly states cannot take part in a feasible assembly sequence. The determination of the assembly states not corresponding to feasible assembly sequence is achieved with the analysis module. The analysis module controls all assembly states according to stability, subassembly and geometric feasibility constraints. Boolean operators apply to the elements of interference matrices determination of geometric feasibility. The feasible assembly states and assembly sequences are represented by a directed graph (Homem de Mello & Arthur, 1990). Assembly states supplying constraints are settled down in the nodes of directed graph hierarchically by the system.

Any path from the root node to terminal node in the directed graph corresponds to feasible assembly sequence. The optimum one is selected from the feasible assembly sequences with optimization module. The neural network approach has been employed for analyzing feasible assembly sequences for sample product. Due to parallel learning structure of the network, the proposed neural network has superior performance to analyze these systems.

3. Artificial Neural Networks

An artificial neural network (or simply a neural network-NN) is a biologically inspired computational structure, which consists of processing elements (neurons) and connections between them with coefficients (weights) bound to the connections.
Training and a recall algorithm are associated to every NN. NN are also called connectionist models because of the main role of the connections the weights are the result of the training phase; they are the "memory" of the structure. The connection weights change during learning (training).

A structure of a typical biological neuron is shown in Fig. 2(a). It has many inputs (in) and one output (out). The connections between neurons are realized in the synapses. An artificial neuron is defined by (Fig. 2(b)):

- Inputs $x_1, x_2, ..., x_n$
- Weights, bound to the inputs $w_1, w_2, ..., w_n$
- An input function (f), which calculates the aggregated net input
- Signal U to the neuron (this is usually a summation function)
- An activation (signal) function, which calculates the activation
- Level of the neuron: $O = g(U)$

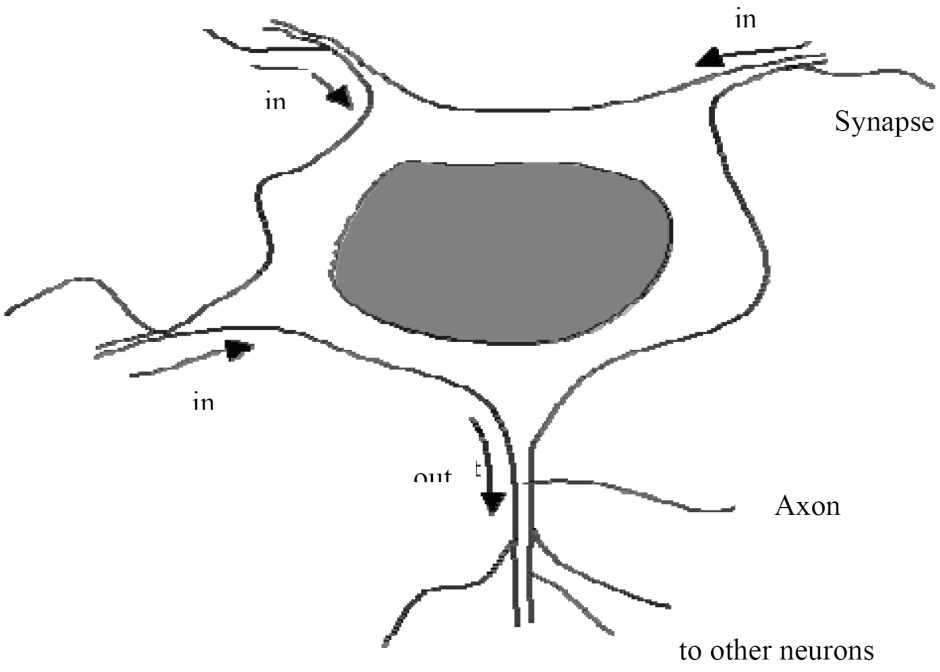


Figure 2(a). Schematic view of a real neuron

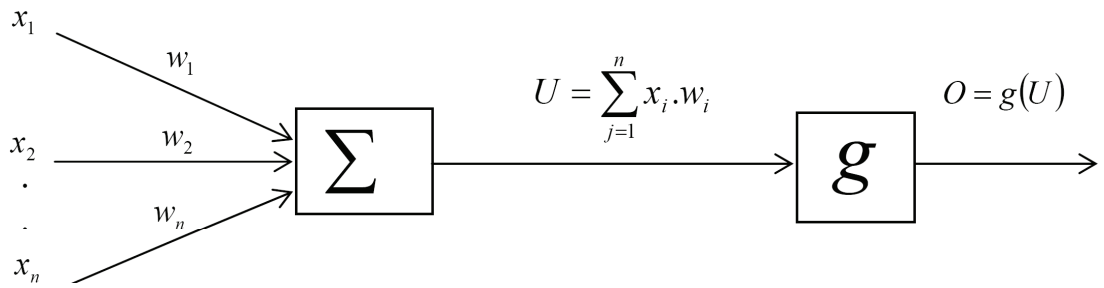


Figure 2(b) Schematic representation of the artificial neural network

Fig. 2(c) shows the currently loaded network. The connections can represent the current weight values for each weight. Squares represent input nodes; circles depict the neurons, the rightmost being the output layer. Triangles represent the bias for each neuron. The neural network consists of three layer, which are input, output and hidden layers. The input and outputs data are used as learning and testing data.

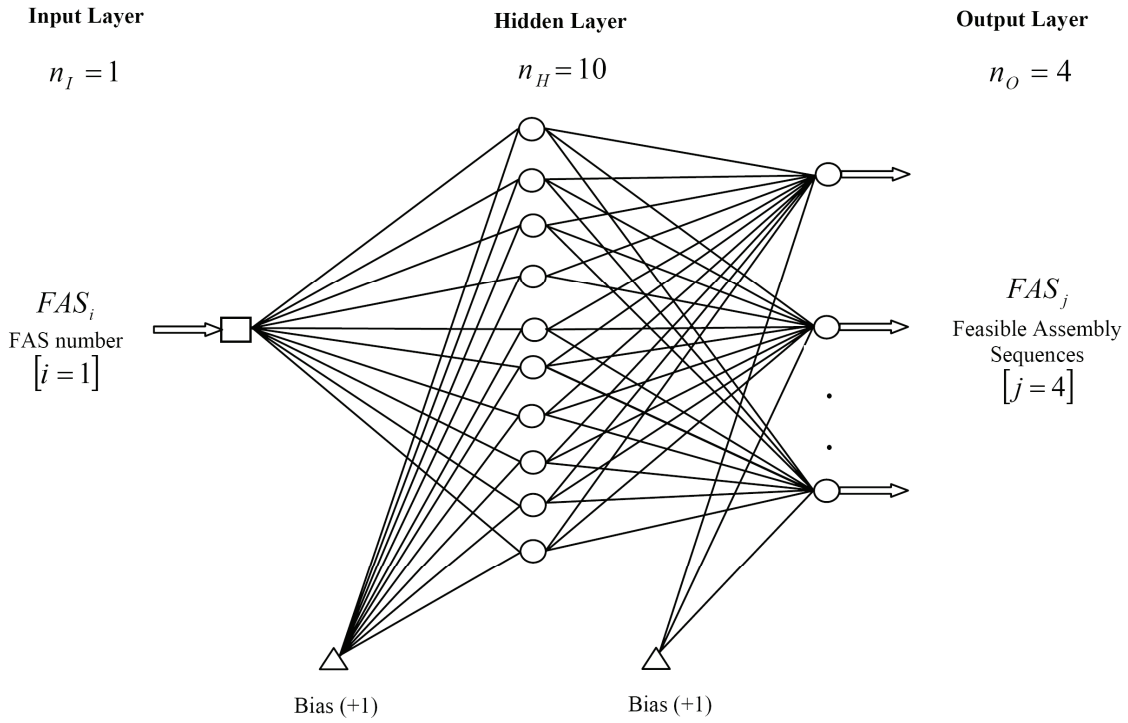


Figure 2(c) Currently loaded network

The most important and time-consuming part in neural network modeling is the training process. In some cases the choice of training method can have a substantial effect on the speed and accuracy of training. The best choice is dependent on the problem, and usually trial-and-error is needed to determine the best method. In this study, logistic function and back-propagation learning algorithm are employed to train the proposed NN.

Back propagation algorithm is used training algorithm for proposed neural networks. Back propagation is a minimization process that starts from the output and backwardly spreads the errors (Canbulut & Sinanoğlu, 2004). The weights are updated as follows;

$$\Delta w_{ij}(t) = -\eta \frac{\partial E(t)}{\partial w_{ij}(t)} + \alpha \Delta w_{ij}(t-1) \tag{1}$$

where, η is the learning rate, and α is the momentum term.

In this study, the logistic function is used to hidden layers and output layers. Linear function is taken for input layer. Logistic function is as follows;

$$y = f(x) = \frac{1}{1+e^{-x}} \tag{2}$$

Its derivative is;

$$\frac{\partial y}{\partial x} = y.(1-x) \tag{3}$$

The linear function is;

$$y = f(x) = x \tag{4}$$

Its derivative is;

$$\frac{\partial y}{\partial x} = 1 \tag{5}$$

Training and structural parameters of the network are given in Table 1.

Proposed Neural Network	η	μ	n_I	n_H	n_O	N	AF
	0.1	0	1	10	4	500000	log*istic*

Table 1. Training and structural parameters of the proposed network

4. Modeling of Assembly System

An assembly is a composition of interconnected parts forming a stable unit. In order to modelling assembly system, it is used ACG whose nodes represent assembling parts and edges represent connections among parts. The assembly process consists of a succession of tasks, each of which consists of joining subassemblies to form a larger subassembly. The process starts with all parts separated and ends with all parts properly joined to form the whole assembly. For the current analyses, it is assumed that exactly two subassemblies are joined at each assembly task, and that after parts have been put together, the remain together until the end of the assembly process.

Due to this assumption, an assembly can be represented by a simple undirected graph $\langle P, C \rangle$, in which $P = \{p_1, p_2, ..., p_N\}$ is the set of nodes, and $C = \{c_1, c_2, ..., c_L\}$ is the set of edges. Each node in P corresponds to a part in the assembly, and there is one edge in C connecting every pair of nodes whose corresponding parts have at least one surface contact.

In order to explain the modeling of assembly system approach better way used for this research, we will take a sample assembly shown as exploded view in Fig. 3. The sample assembly is a pincer consisting of four components that are: bolt, left-handle, right-handle and nut. These parts are represented respectively by the symbols of $\{a\}$, $\{b\}$, $\{c\}$ and $\{d\}$. For this particular situation, the connection graph of assembly has the set of the nodes as $P = \{a, b, c, d\}$ and the set of the connections as $C = \{c_1, c_2, c_4, c_5\}$.

The connections or edges defining relationships between parts or nodes can be stated as: c_1 between parts $\{a\}$ and $\{b\}$, c_2 between parts $\{a\}$ and $\{d\}$, c_3 between parts $\{c\}$ and $\{d\}$, c_4 between parts $\{a\}$ and $\{c\}$ and finally c_5 between parts $\{b\}$ and $\{c\}$.

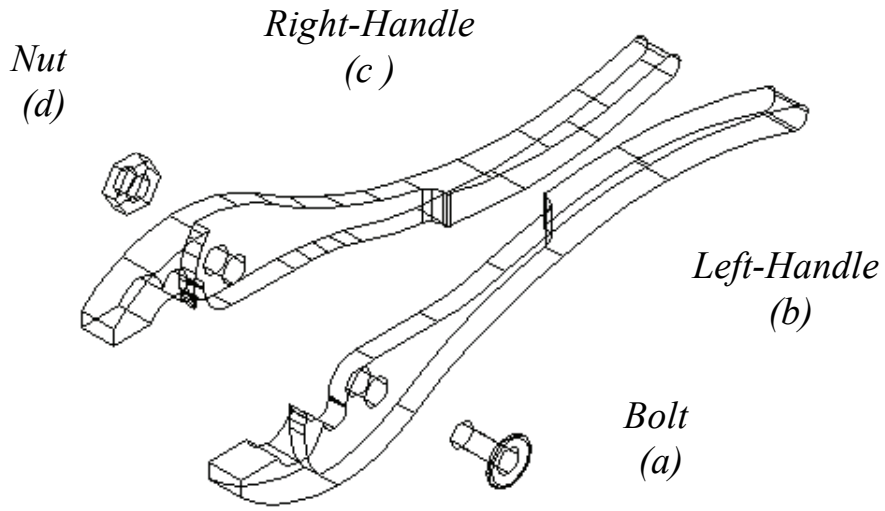


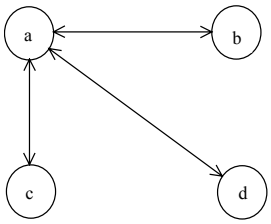
Figure 3. The pincer assembly system

4.1 Definition of Contact Matrices and ACG

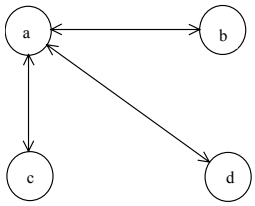
The contact matrices are used to determine whether there are contacts between parts in the assembly state. These matrices are represented by a contact condition between a pair of parts as an $\{A, B\}$. The elements of these matrices consist of *Boolean* values of *true* (1) or *false* (0). For the construction of contact matrices, the first part is taken as a reference. Then it is examined that whether this part has a contact relation in any i axis directions with other parts. If there is, that relation is defined as *true* (1), else that is defined as *false* (0).

The row and column element values of contact matrices in the definition of six main coordinate axis directions are relations between parts and that constitutes a pincer assembly. To determine these relations, the assembly's parts are located to rows and columns of the contact matrices. Contact matrices are square matrices and their dimensions are 4×4 for pincer.

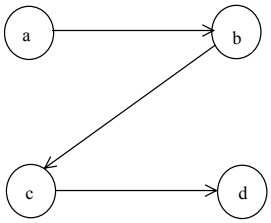
For example, $[a,b]$ element of B contact matrix in i direction is defined to whether there exists any contacts or not between parts $\{a\}$ and $\{b\}$ for the related direction and the corresponding matrix element may have the values of (1) and (0), respectively.



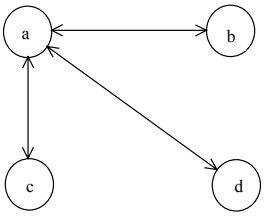
$$B_y = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$



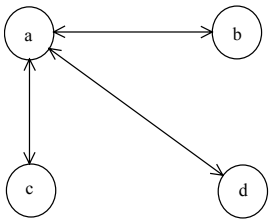
$$B_z = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$



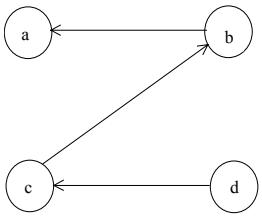
$$B_{-x} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$



$$B_{-y} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$



$$B_{-z} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$



$$B_x = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Figure 4. Contact matrices and their graph representations

In this system, in order to get contact matrices in the direction of Cartesian coordinate axis, assembly view of pincer system was used. These matrices were automatically constructed (Sinanoğlu & Börklü, 2004). Contact matrices of the pincer assembly system are also shown in Fig. 4.

The connection graph can be obtained from contact matrices. To construct ACG, contact conditions are examined in both part's sequenced directions. For instance, in the manner of $\{a,b\}$ sequenced pair of parts, it is sufficient to determine contacts related sequenced direction so that its contact in any direction. Due to this reason, an $[\vee : Or]$ operator is applied to these parts. But it is

also necessary contacts in any direction for inverse sequenced pairs of parts in the ACG. If these values are (1) for every sequenced pair of parts, then there should be edges between corresponding nodes of the ACG. For this purpose, every pair of parts must be determined.

- $\{a,b\}$, $\{b,a\}$ Sequenced pair of parts

To investigate whether there is an edge between $\{a\}$ and $\{b\}$ in ACG or not, it should be searched contact relations for these pairs of parts. Table 2 shows contact relations regarding $\{a,b\}$ and $\{b,a\}$ pairs of parts.

$c_1 \Rightarrow (a \div b)$	x	y	z	$-x$	$-y$	$-z$	$\vee : Or \Rightarrow$	$\wedge : And \Downarrow$
a/b	0	1	1	1	1	1	1	1
b/a	1	1	1	0	1	1	1	1
								1

Table 2. Contact relations of $\{a,b\}$ and $\{b,a\}$ pairs of parts

In this table, $\{a,b\}$ sequenced pair of parts is supplied to at least one contact condition in the related direction of $(0 \vee 1 \vee 1 \vee 1 \vee 1 \vee 1 = 1)$. $\{b,a\}$ pair of parts is also supplied to at least one contact in the related direction of $(1 \vee 1 \vee 1 \vee 0 \vee 1 \vee 1 = 1)$. An $(\wedge : And)$ operator is applied to these obtaining values. Because, these parts have at least one contact in each part sequenced direction, there is an edge between parts in the ACG. This connection states an edge in the ACG shown in Fig. 5.

If similar method is applied to other pairs of parts: $\{a,d\}$, $\{d,a\}$, $\{b,c\}$, $\{c,b\}$, $\{c,d\}$, $\{d,c\}$, $\{a,c\}$ and $\{c,a\}$, the results should be (1). Therefore, there are edges between these pairs in ACG.

The graph representation of this situation is shown in Fig. 5, where there is no edge between parts $\{b\}$ and $\{d\}$. Therefore, these parts do not have any contact relations.
Fig. 5 shows the pincer graph of connections. It has four nodes and five edges (connections). There is no contact between the left-handle and the nut. There-

fore, the graph of connections does not include an edge connecting the nodes corresponding to the left-handle and the nut. By the use of the contact matrices and applying some logical operators to their elements, it is proved that it is supplied to one connection between two part in ACG not all contacts between them are established in every direction.

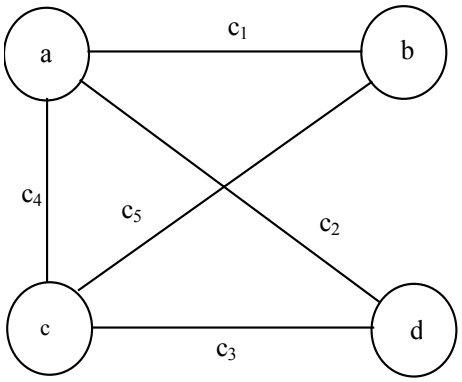


Figure 5. The graph of connections for four-part pincer assembly

5. Determination of Binary Vector Representation and Assembly States (ASs)

The state of the assembly process is the configuration of the parts at the beginning (or at the end) of an assembly task. The configuration of parts is given by the contacts that have been established. Therefore, in the developed approach an L-dimensional binary vector can represent a state of the process $(x = \{x_1, x_2, ..., x_L\})$. Elements of these vectors define the connection data between components. Based upon the establishment of the connections, the elements of these vectors may have the values of either (1) or (0) at any particular state of assembly task. For example, the i^{th} component x_i would have a value of *true* (1) if the i^{th} connection were established at that state. Otherwise, it would have a value of *false* (0). Moreover, every binary vector representations are not corresponding to an assembly state. In order to determine assembly states, the established connections in binary vectors and ACG are utilised together.

There are five edges in the example ACG. Because of that, the elements of vectors are five and the 5-dimensional binary vector of can represent that $[c_1, c_2, c_3, c_4, c_5]$. For instance, the initial state of the assembly process for the product shown in Fig. 3 can be represented by binary vector $[FFFFF]$ whereas the final state can be represented by $[TTTTT]$.

If the first task of the assembly process is the joining of the bolt to nut, the second state of the assembly process can be represented by $[FTFFF]$.

For example, an assembly sequence for pincer system can be represented as follows:

$$([FFFFF],[FTFFF],[TTTFF],[TTTTT]) \quad ([00000],[01000],[11100],[11111])$$

The first element of this list represents the initial state of the assembly process. The second element of the list shows the second connection c_2 between bolt and nut. The third element represents c_1 connection between right-handle and bolt and c_3 connection between right-handle and nut. The last element of the list is $[11111]$ and it means that every connection has been established.

In the developed planning system, first of all binary vector representations must be produced. The purpose of that it is classified to binary vectors according to the number of established connections. Table 3 shows vector representations for pincer assembly in Fig. 3. There are thirty-two different binary vectors. While some of them correspond to assembly state, some of them are not.

To form assembly sequences of pincer system, vector representations corresponds to assembly states must be determined. In order to determine whether the vector is a state or not, it must be taken into consideration established connections in vector representation. And then it is required that establishing connections must be determined to established connections by ACG.
For instance, if the first task of the assembly process is the joining of the bolt to the left-handle, the second state of the assembly process can be represented by $[10000]$. It is seen in Fig. 6 that it does not necessary to establish any connection so that c_1 connection between part $\{a\}$ and $\{b\}$ is establish. Therefore, $[10000]$ vector is an assembly state. Therefore, vectors only one established connection form assembly state.

LEVEL 0	LEVEL 1	*LEVEL* 2			LEVEL 3
00000	10000		11100		11111
		11000	11010	11110	
		10100	11001	11101	
		10010	10110	11011	
		10001	10101	10111	
			10011		
	01000		11100		
		11000	11010	11110	
		01100	11001	11101	
		01010	01110	11011	
		01001	01101	01111	
			01011		
	00100		11100		
		10100	10110	11110	
		01100	10101	11101	
		00110	01110	10111	
		00101	01101	01111	
			00111		
	00010		11010		
		10010	10110	11110	
		01010	10011	11011	
		00110	01110	10111	
		00011	01011	01111	
			00111		
	00001		11001		
		10001	10101	11101	
		01001	10011	11011	
		00101	01101	10111	
		00011	01011	01111	
			00111		

Table 3. Hierarchical levels of binary vector representations for pincer assembly system

Moreover, some of vectors do not correspond to an assembly state. For instance, in the $[10001]$ vector, connections of c_1 between $\{a\}$ and $\{b\}$, c_5 between $\{b\}$ and $\{c\}$ have been established (1). It has been necessary to establish c_4 connection between $\{a\}$ and $\{c\}$ so that these connections have been established (Fig. 6). But this connection has not been established in $[10001]$, $[10001]$ vector is not an assembly state.

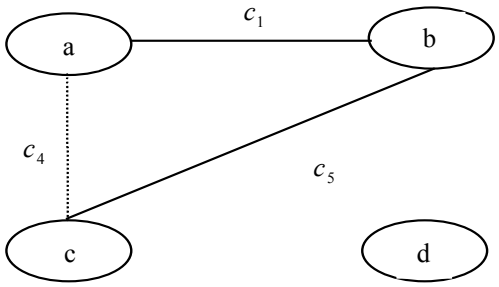


Figure 6 c_1, c_5 and c_4 connections in $[10001]$ vector

There are thirteen assembly states in pincer assembly system. These are;

$[00000],[10000],[01000],[00100],[00010],[00001],[11000],[10100],[01001],[00101],[10011],[01110],[11111]$

6. Productions and Representation of Assembly Sequences

Given an assembly whose graph of connections is $\langle P,C\rangle$, a directed graph can be used to represent the set of all assembly sequences (Homem de Mello & Lee, 1991). The directed graph of feasible assembly sequences of an assembly whose set of parts is P is the directed graph $\langle x_p,T_p\rangle$ in which, x_p is the assembly's set of stable states, and T_p is the assembly's set of feasible state transitions.

In the pincer assembly, $P=\{a,b,c,d\}$ is the assembly's set of parts or set of nodes, $C=\{c_1,c_2,c_3,c_4,c_5\}$ is the assembly's set of connections or set of edges. $\langle x_p,T_p\rangle$ corresponds to directed graph of pincer system. A path in the directed graph of feasible assembly sequences $\langle x_p,T_p\rangle$ whose initial node is $\Theta_I=\{\{a\},\{b\},\{c\},\{d\}\}$ and whose terminal node are $\Theta_F=\{\{a,b,c,d\}\}$. Vector representations of these sets are $[00000]$ and $[11111]$ respectively.

Assembly states not correspond to feasible assembly sequences must eliminate by some assembly constraints. In this study, three assembly constraints are applied to assembly states. These are subassembly, stability and geometric feasibility constraints.

The subassembly constraint defines feasibility of subassembly of set of partitions to established connections in assembly states. In order to form a subassembly of a set of partition, it is not a set of partition contains a pair of part has not contact relation in the ACG. Therefore, in pincer assembly bolt and left-handle has not contact relations. Because of that it is not supplied to subassembly constraint set partitions contains $\{b,d\}$ set of partition.

The second constraints is stability. A subassembly is said to be stable if its parts maintain their relative position and do not break contact spontaneously. All one-part subassemblies are stable.

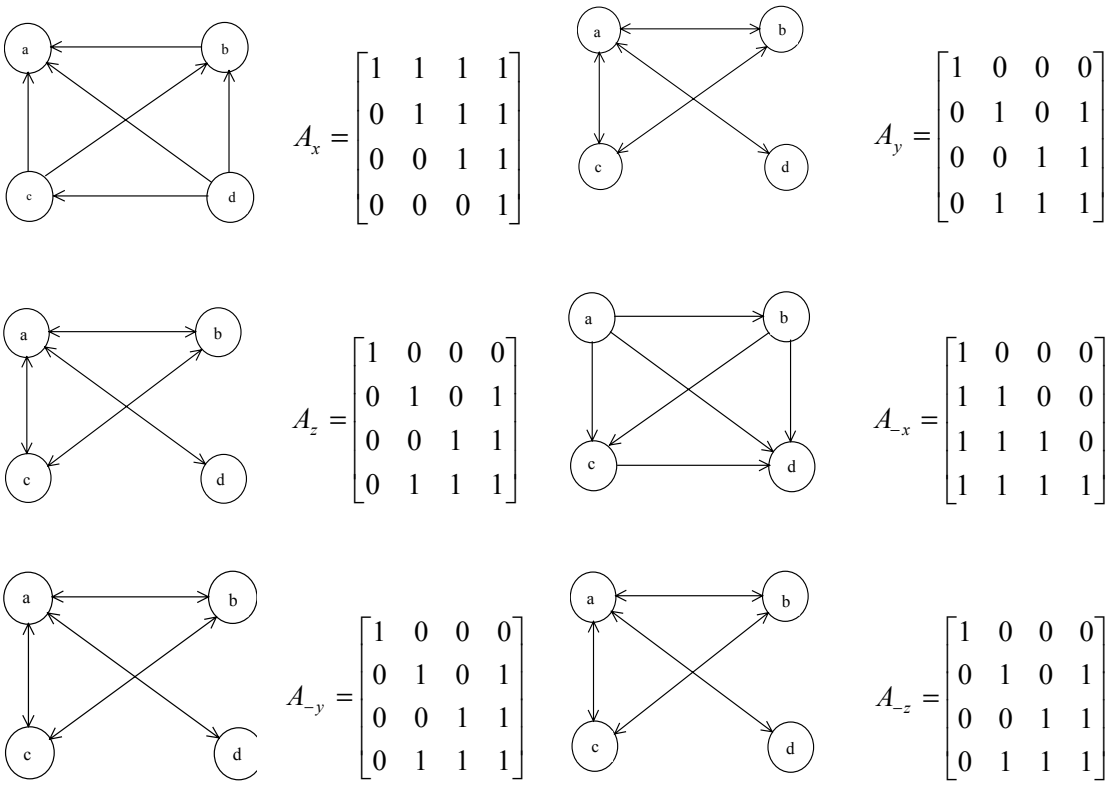


Figure 7. Interference matrices and their graph representations for pincer assembly

The last constraint is geometric feasibility. An assembly task is said to be geometrically feasible if there is a collision-free path to bring the two subassemblies into contact from a situation in which they are far apart.

Geometric feasibility of binary vectors correspond to assembly states are determined by interference matrices. The elements of interference matrices were taken into consideration interference conditions during the joining parts. In the determination of geometric feasibility, it is applied to elements of interference matrices ($\wedge$) and ($\vee$) logical operators. At this operation, it must be utilise established connections and that is joining pairs of part. In order to, whether binary vector representations corresponds to assembly states are geometrically feasible or not, it is necessary to applying Cartesian product between sequenced pairs of parts which are representing established connections and parts which are not in this sequenced pairs.
In the determination of interference matrices elements, it is taken into consideration interference while the reference part is moving with another part along with related axis direction. If it is interference during this transformation motion, interference matrices elements are (0) if not are defined as (1).

For instance, in the A_x matrice the movement of part bolt is interfered to movement along with $\{+x\}$ axis by other parts. Therefore, the first row elements of A_x matrice defined to interference among parts along this axis are (1). But the movement of left-handle along with $\{+x\}$ axis does not interfere any parts (Fig. 3). This interference relation is illustrated to designate (0) value by element of second row and third column in A_x matrice. These matrices are also formed automatically from various assembly views.

Graph representations for the pincer assembly and construction of their interference matrices can be also determined as follows (Fig. 7).

In order to determine whether assembly states are geometrically feasible or not, it is necessary to apply Cartesian product between sequenced pairs of parts which represent established connections and parts which are not in this sequenced pairs of parts. In this situation, different interference tables are obtained and these tables are used to check geometric feasibility.

- $[01000]$ Assembly State

In this assembly state, connection of c_2 between part $\{a\}$ and $\{d\}$ has been established. To determine geometric feasibility of this assembly state, parts

without in established conditions are taken. Those are $\{b\}$ and $\{c\}$. $\{a,d\}$ sequenced pair of part represents established connection c_2. Cartesian product, which is between $\{a,d\}$ and $\{b\}$ is given as follows.

$$(a,d)(b) \Rightarrow (a,b)(d,b)$$

Table 4 shows interference's of $\{a,b\}$ sequenced pair of part.

$c_2 \Rightarrow (a \div d)$	x	y	z	$-x$	$-y$	$-z$	
a/b	1	0	0	0	0	0	$\vee \Rightarrow$
							1

Table 4. Interference of $\{a,b\}$ sequenced pair of part

Another part without parts constituted assembly state is part $\{c\}$. As a result of Cartesian product is $(a,d)(c) \Rightarrow (a,c)(d,c)$. Table 5 shows interference of them.

$c_2 \Rightarrow (a \div d)$	x	y	z	$-x$	$-y$	$-z$	
a/c	1	0	0	0	0	0	
d/c	0	1	1	1	1	1	
$(a/c \wedge d/c)\Downarrow$	0	0	0	0	0	0	$\vee \Rightarrow$
							0

Table 5. Interference relations of $\{a,c\}$ and $\{d,c\}$ pairs

Although it is geometrically feasible (1) to disassemble from $\{a,d\}$ to $\{b\}$, it is not geometrically feasible (0) to disassemble from $\{a,d\}$ to $\{c\}$. As a result of $(\wedge)$ logical operator is (0) $(1 \wedge 0 = 0)$. This result explained that $[01000]$ assembly state is geometrically unfeasible.

Moreover, other assembly states of $[00100]$ and $[00001]$ are geometrically feasible, but $[00010]$ is geometrically unfeasible. Similarly, $[11000]$,$[10100]$ and $[00101]$ assembly states contain two established connection are geometrically feasible, but $[01001]$ is not geometrically feasible. Moreover, $[10011]$ assembly state contains three connections that are geometrically feasible but $[01110]$ is geometrically unfeasible. $[11111]$ vector is also geometrically feasible. The number of nodes is reduced from 15 to 8 in the di-

rected graph by applying assembly constraints. The assembly states supplied to these constraints are as follows:

$$[00000][10000][00100][00001][10100][00101][10011][11111]$$

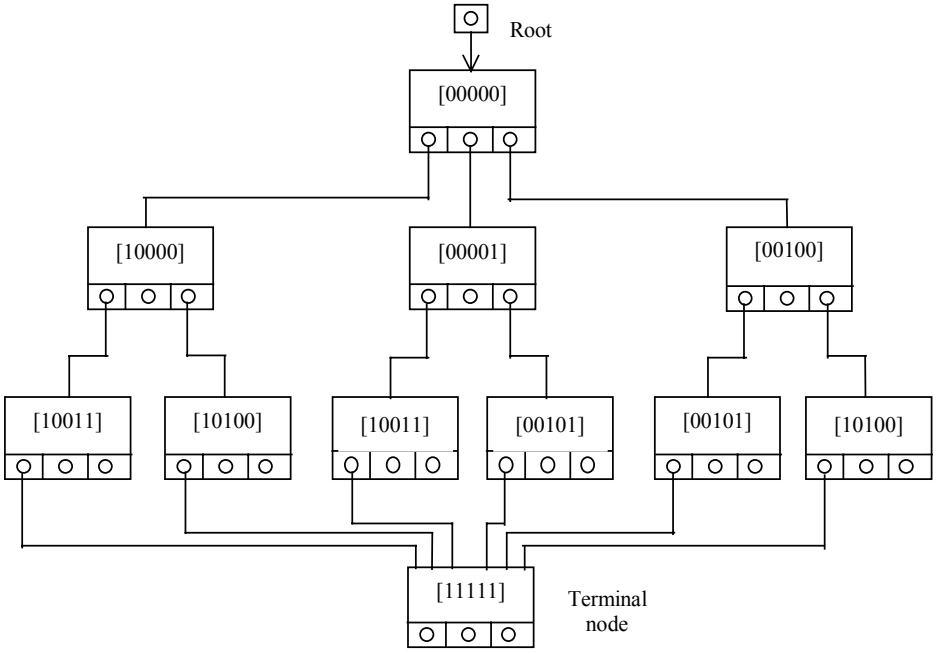


Figure 8. Constrained directed graph for pincer system

Fig. 8 shows the directed graph of feasible assembly sequences after applied constraints. A path in the directed graph of feasible assembly sequences whose initial node is $[00000]$ and terminal node is $[11111]$ corresponds to a feasible assembly sequences for pincer. The feasible assembly sequences for pincer assemblies are as follows:

$$\begin{aligned}
FAS-I &\quad [00000][10000][10011][11111]\\
FAS-II &\quad [00000][10000][10100][11111]\\
FAS-III &\quad [00000][00001][10011][11111]\\
FAS-IV &\quad [00000][00001][00101][11111]\\
FAS-V &\quad [00000][00100][00101][11111]\\
FAS-VI &\quad [00000][00100][10100][11111]
\end{aligned}$$

For example, in the third assembly sequence for pincer system, at first, the left handle is joined to right handle with connection of c_5. After that this subassembly is joined by using the bolt with the connections of c_1 and c_4. Finally, the nut fixes all parts.

7. Optimization of Assembly Sequences

Developed assembly planning system can be determined to find the optimum assembly sequence. In this section, an optimization approach is explained. For this purpose, the pincer assembly system is taken as an example. It has been obtained from feasible assembly sequences in previous sections. In order to optimize the assembly sequence, two criteria are developed, weight and the subassembly's degree of freedom. First certain costs are assigned to edges of directed graph depend on these criteria, and then the total cost of each path from root node to terminal is calculated the minimum cost sequence is selected as an optimum one.

7.1. Optimization of Weight Criterion

In order to determine the optimum assembly sequence, all assembly states in an assembly sequence must be taken into consideration. The heaviest and bulkiest part is selected as a base part and then the assembly sequence continues from heavy to light parts. The parts with the least volume, i.e. connective parts, like bolts and nuts must be assembled last (Bunday, 1984). The weights and volumes of parts were calculated automatically with a CAD program.

Therefore, determination of the costs of assembly states is necessary to obtain an optimum feasible assembly sequence. After that these costs are used as a reference to different assembly states. Calculated weight costs of assembly states in the assembly sequence are compared with reference weights. The difference of weight is multiplied by unit weight value (100). The weights of parts of the pincer system are as follows: Bolt $(0.0163kg)$, left-handle $(0.3843kg)$, right-handle $(0.3843kg)$ and nut $(0.0092kg)$. Using the weight criterion, the total established connection weights of each assembly state in optimum assembly sequence could be determined.

The total weight of assembly states in the optimum sequence according to weight criterion can be defined as;

$$Ow_m = \sum_{i=1}^{n}(W) \tag{6}$$

where W is the weight of assembly states, (n) is the number of the established connections and (m) is the order of assembly states. Ow_m is the required weight of assembly states in the optimum assembly sequence.

In order to determine the optimum assembly sequence, the weights of all assembly states in assembly sequences are calculated. This weight is expressed as;

$$Cw_m = \sum_{i=1}^{n}(W) \tag{7}$$

where W is the weight of assembly states, (n) is the number of the established connections and (m) is the sequenced of assembly states.

After that, (Ow_m) is used as a reference for different assembly states. Calculated weights of assembly states in assembly sequences are compared with reference weights. This weight difference (Dw) is multiplied by unit weight value $(Uwv = 100)$. The result is the weight costs of any assembly states (Wc).

$$Dw_m = Ow_m - Cw_m \tag{8}$$

$$\begin{aligned} &Wc = \sum_{i=1}^{l}((Wc)_m) = \\ &\sum_{i=1}^{l}((Ow_m - Cw_m)xUwv) = \sum_{i=1}^{l}\left(\left(\left(\sum_{i=1}^{n}(W)\right)_O - \left(\sum_{i=1}^{n}(W)\right)_H\right)xUwv\right) \end{aligned} \tag{9}$$

For example, in the second assembly state of the first feasible assembly sequence $([10000])$, c_1 connection between $\{a\}$ and $\{b\}$ is established. The number of the established connection is $n = 1$. The required weight of this state is $Cw_2 = \sum_{i=1}^{1}(W)_H = W(c_1) = 0.4006kg$.

In the second assembly state, the necessary weight is $Ow_2 = 0.7686kg$.

The difference of weight Dw is $Dw_2 = Ow_2 - Cw_2 = 0.7686 - 0.4006 = 0.368kg$. If Dw is multiplied by the unit weight value $(Uwv = 100)$, the weight cost of $([10000])$ will be calculated as follows; $Wc_2 = Dw_2.Uwv = 0.368x100 \cong 37$

The total weight cost of any feasible assembly sequence is expressed as;

$$Wct = \sum_{i=1}^{z} Wc \qquad (10)$$

where z is the total assembly state number of any feasible assembly sequence.

7.2. Optimization of Subassembly Degree of Freedom Criterion

The subassembly degree of freedom criterion is based on the selection of parts with low degrees of freedom. So degree of freedom between the subassembly parts is low, the assembly of these parts can be done more easily. It is a unit cost (unit degree of freedom value, $Udofv$) also used for this criterion. It is "25" and this criterion is more important than the other. Therefore, it is selected as the lower unit cost according to weight criterion, and so that total cost of assembly sequences can be reduced.

It determines degree of interference for pairs of parts connections established along the six main directions of the Cartesian coordinate system. The total degree of freedom $(Tdof)$ for pairs of parts is the product's unit cost.

$$DOFc = Tdof \times Udofv \qquad (11)$$

Therefore, in the directed graph costs of degree of freedom according to this criterion are calculated as the degree of freedom for each path from initial node to terminal node. As a result, the minimum cost of the assembly sequence can be selected as an optimum with respect to the degree of freedom criterion. The total weight cost of any feasible assembly sequence is expressed as;

$$DOFct = \sum_{i=1}^{z} DOFc \qquad (12)$$

where z is the total assembly state number of any feasible assembly sequence.

The total cost of feasible assembly sequence for any product is expressed as a cost function fc;

$$\begin{aligned} Wc &= [(Ow_1 + Ow_2 + ... + Ow_k) - (Cw_1 + Cw_2 + ... + Cw_k)]xUwv \\ &= \left[\left(\sum_{i=1}^{l} Ow_k\right) - \left(\sum_{i=1}^{l} Cw_k\right)\right] xUwv \end{aligned} \tag{13}$$

$$DOFc = TdofxUdofv \tag{14}$$

$$fc = Wct + DOFct \tag{15}$$

(l) is the number of assembly states in the feasible assembly sequence for any product.

For example, in the second assembly state of the first feasible assembly sequence $([10000])$, c_1 connection between $\{a\}$ and $\{b\}$ is established. The number of the established connection is $n = 1$. The degree of freedom of this pair of parts is shown in Table 6.

$c_1 \Rightarrow (a \div b)$	x	y	z	$-x$	$-y$	$-z$
a/b	1	0	0	0	0	0
b/a	0	0	0	1	0	0

Table 6. Degree of freedom between parts $\{a\}$ and $\{b\}$

The total degree of freedom for $([10000])$ is $Tdof = 2$. If this value is multiplied by the $Udofv = 25$ unit freedom cost, the result will be "50". Therefore, the degree of freedom cost $DOFc$ for $[100000000]$ assembly state is "50".

Fig. 9 shows feasible assembly sequences and costs of them for the pincer assembly system. The first and third assembly sequences for the pincer system according to the subassembly degree of freedom criterion have been selected

with an optimum total cost of "300". The weight costs are in parentheses ( ) and the degree of freedom costs are in quotation marks " ".

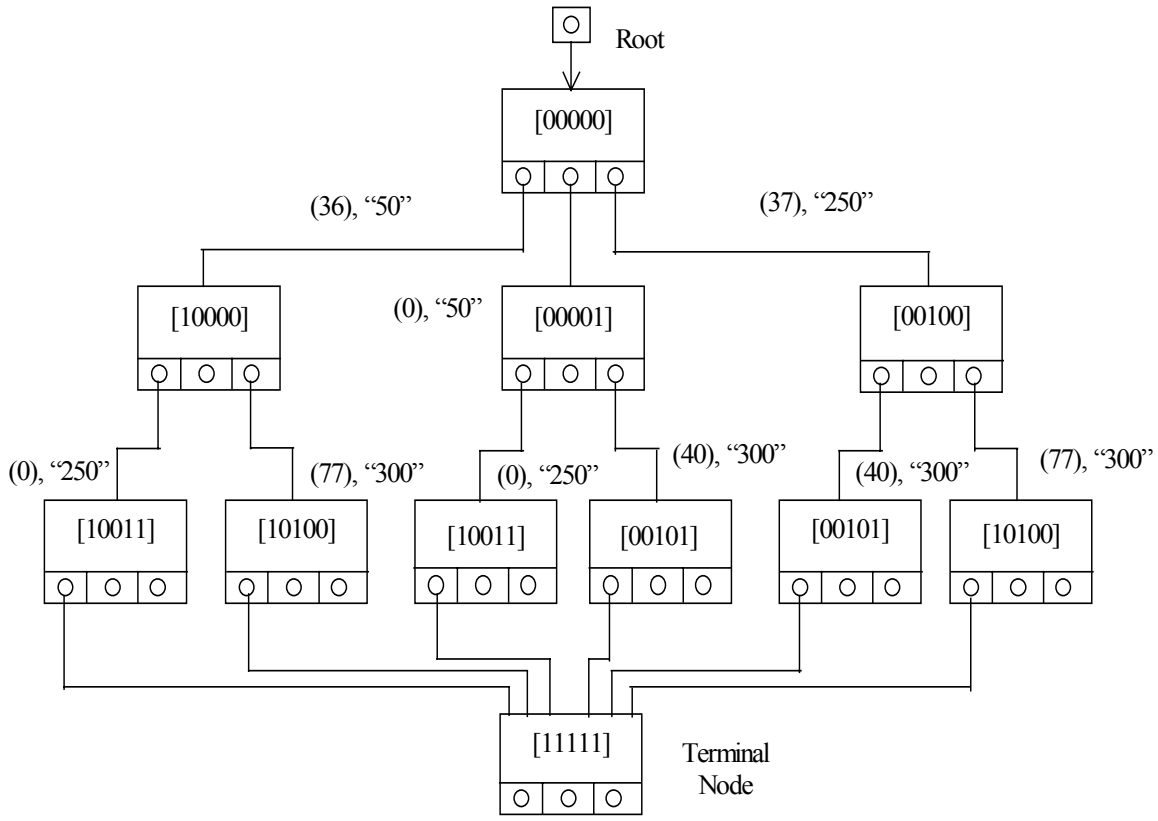


Figure 9. The weight and degree of freedom costs for pincer system

Fig. 9 shows that the third assembly sequence is optimum "0" weight cost and "300" degree of freedom cost. Moreover, the sixth assembly sequence is the least preferable sequence in the feasible assembly sequences. In the optimiza-

tion approach, both optimization criteria indicated that assembly sequence is optimum.

Therefore, the optimum assembly sequence for pincer system is

$$[00000],[00001],[10011],[11111].$$

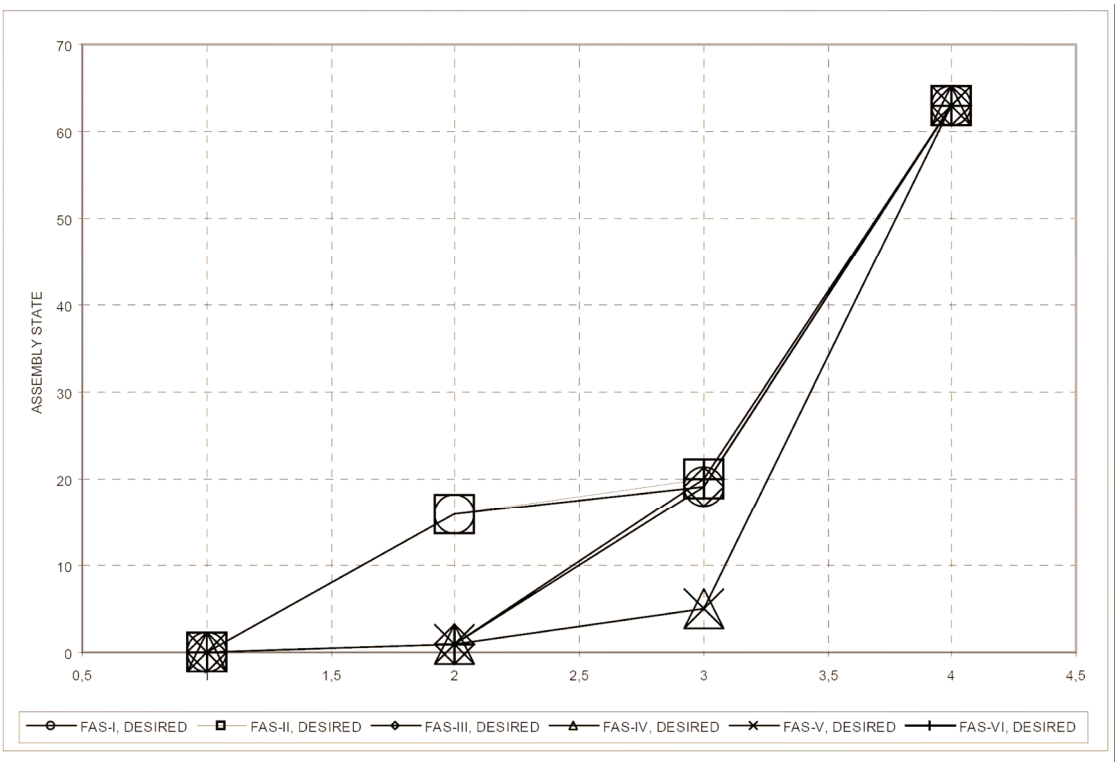


Figure 10(a). The desired feasible assembly sequences for pincer assembly system

Fig. 10(a) (Case 1) shows the desired feasible assembly sequences for pincer assembly system. Fig. 10(b) (Case 2) is also shows these feasible assembly sequences for neural network approach.

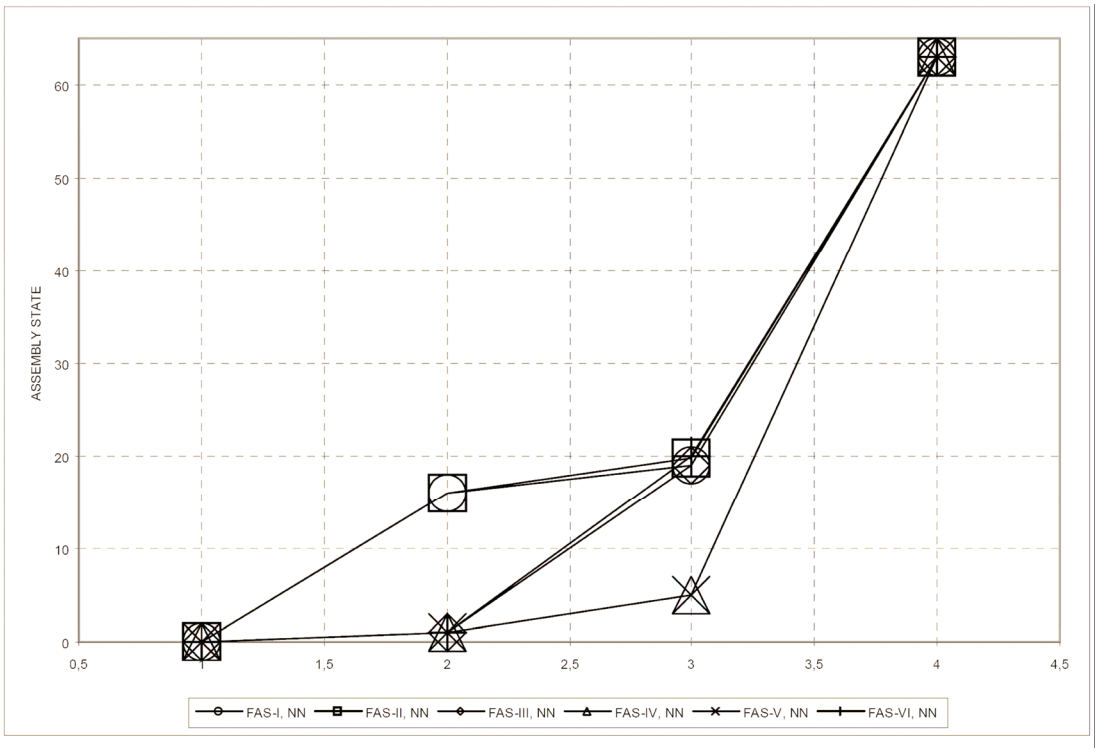


Figure 10 (b). The feasible assembly sequences for proposed neural networks approach.

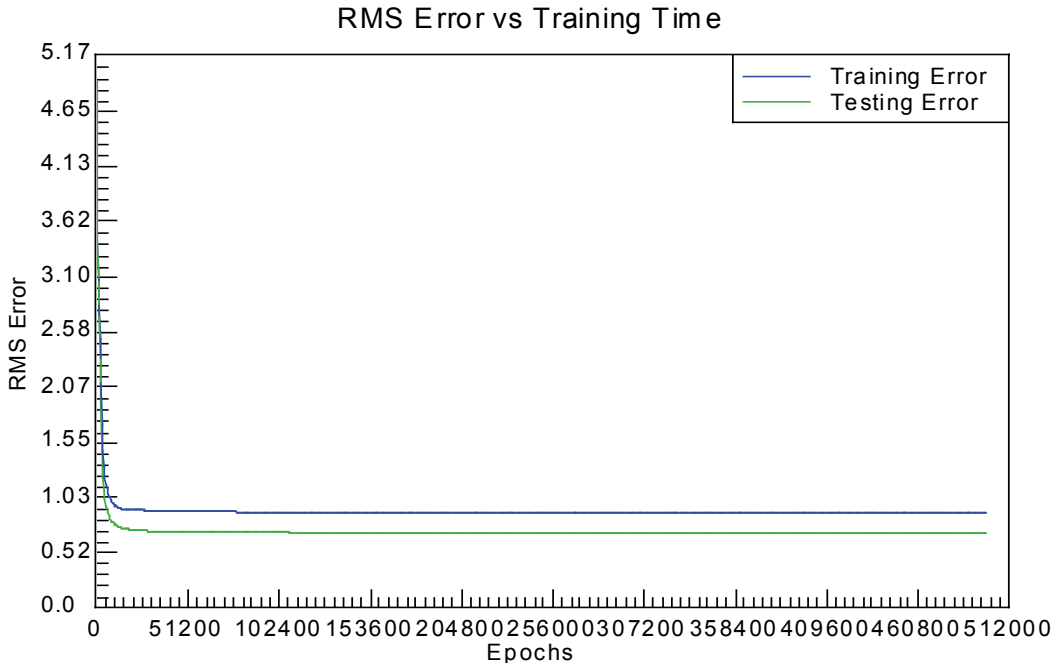
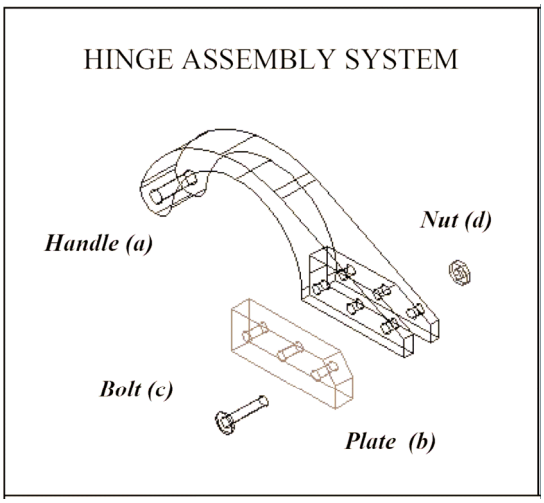


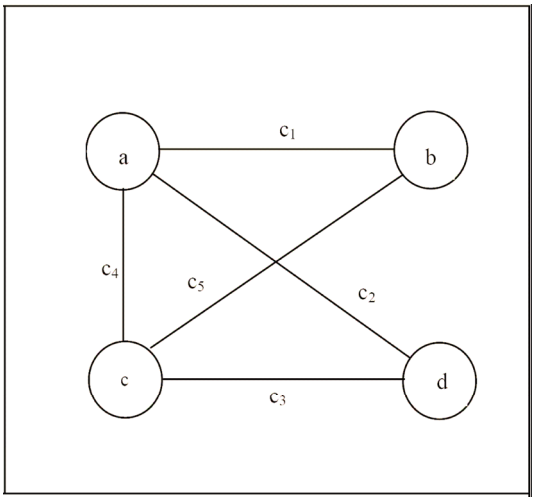
Figure 11. The error convergence graph of the case 2

The error convergence graph of the case 2 is depicted in Fig. 11 during the training of the network. As can be seen from the figure, the error is suddenly reducing to small values. Small epoch can be employed for case 2 (51200 epoch).

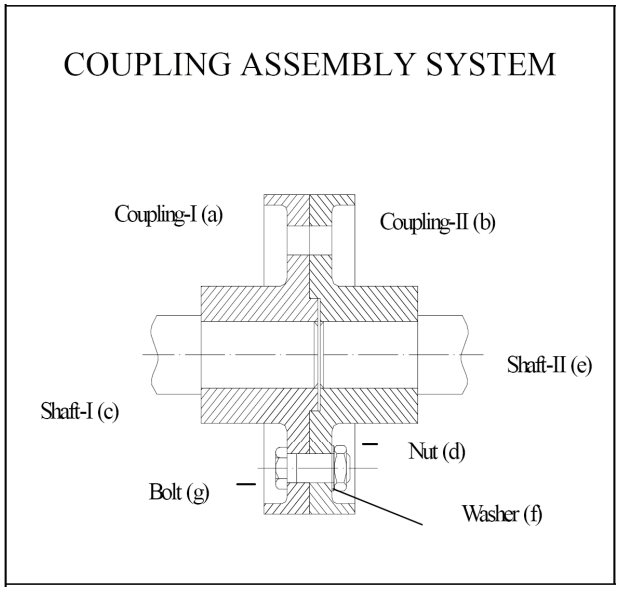
8. Some Assembly Case Studies

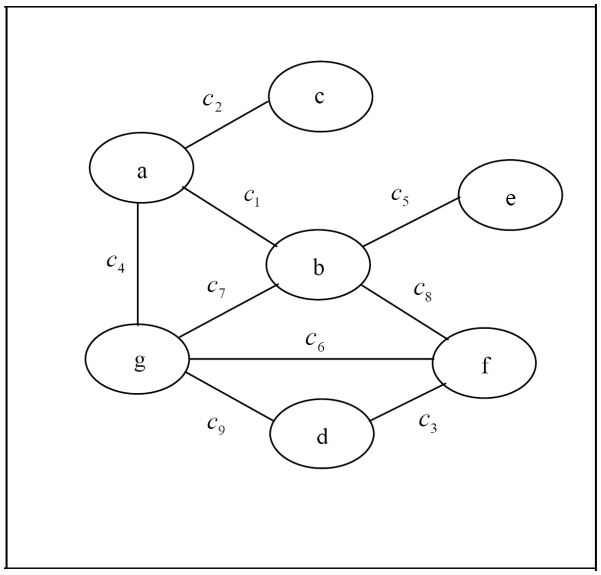
In this work, some sample assembly systems are examined. Among these examples, four-part hinge system and seven-part coupling system have been investigated. Fig. 12 shows this assembly's exploded views and ACG.





Nodes of ACG**:** a ; Handle, b ; Plate, c ; Bolt, d ; Nut





Nodes of ACG**:** a ; Coupling-I, b ; Coupling-II, c ; Shaft-I, d ; Nut, e ; Shaft-II, f ; Washer, g ; Bolt

Figure 12. The hinge and coupling systems and their ACG

The assembly sequences of the hinge system contain four different assembly states. The first one is $[00000]$ and the last is $[11111]$. Fig. 13 shows feasible assembly sequences for hinge system.

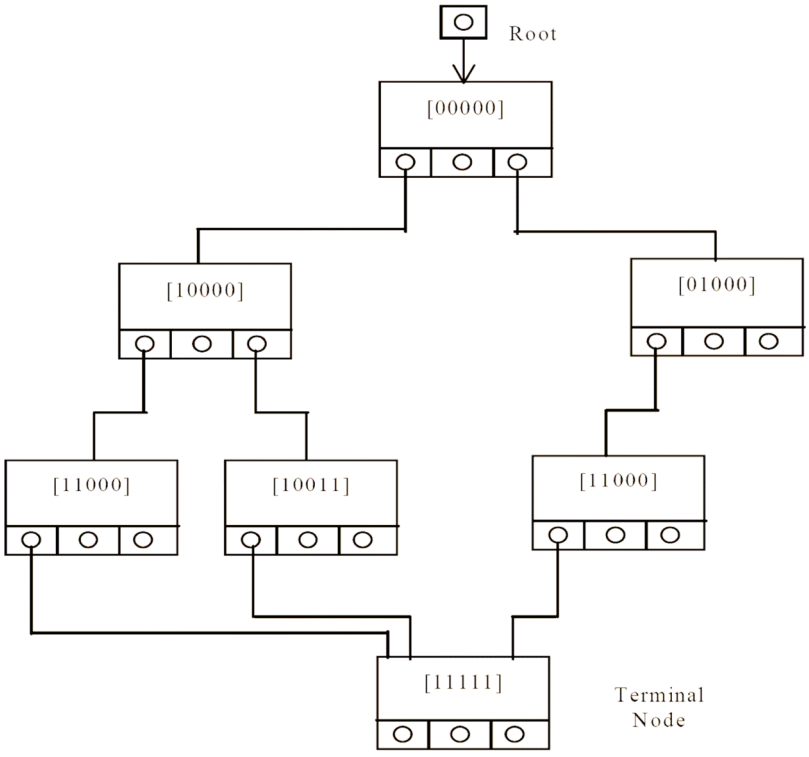


Figure 13. The feasible assembly sequences for hinge system

For instance, in the second assembly sequence, the plate and handle are connected with the connection of c_1. Then using bolt this subassembly is fixed with the connections of c_4, c_5. And the assembly process is completed with the addition of nut.

In the coupling assembly system, $[000000000]$ and $[111111111]$ are the same as all assembly sequences. Some of the feasible assembly sequences for coupling system are shown in Table 7. One of the feasible assembly sequences for coupling system is:

$$[000000000], [010000000], [010010000], [010110000], [111111111]$$

FEASIBLE ASSEMBLY SEQUENCES FOR COUPLING SYSTEM						
000000000	010000000	010010000	011010000			111111111
000000000	010000000	010010000	011010000	011010010	111010010	111111111
000000000	010000000	010010000	011010000	011010010		111111111
000000000	010000000	010010000	010110000			111111111
000000000	010000000	010010000	010110000	010111000		111111111
000000000	010000000	010010000	010110000	010110010		111111111
000000000	010000000	010010000	010110000	010110001		111111111
000000000	010000000	010010000	010011000	110011000		111111111

Table 7. Several assembly sequences of coupling systems

The coupling-I is joined to the shaft-I in the first assembly state; the coupling-II and the shaft-II are added in second assembly state. In the third assembly state, the bolt is used to fix this subassembly, and then the washer and nut is joined to last subassembly.

9. Conclusions

In this chapter, an assembly sequence planning system, based on binary vector representations, is explained and some assembly systems are examined. Among these examples, six assembly sequences for four-part pincer system; three assembly sequences for four-part hinge system; three hundred and seventy three assembly sequences for seven-part coupling system have been investigated. NN application was presented for analyzing assembly sequences on assembly system. As can be depicted from the results, the neural predictor would be used as a predictor for possible assembly system applications. The properties and originalities of the developed system;

Easy Representation

- The representation of assembly models by *ACG*
- The representation of contact and interference relations between parts by interference matrices
- The representation of assembly steps by binary vectors

Automatic Production

- Contact and interference matrices
- *ACG*
- All binary vectors
- Assembly states
- Geometric feasibility of assembly states
- The automatic determination of assembly states that have the conditions of stability and subassembly
- The automatic determination of feasible assembly sequence plans and optimum assembly one using NN

Practical use

- It is possible to obtain appropriate assembly sequence plans after giving various assembly drawings of a product
- It is understood that the determination of assembly sequence plans that save time and cost after the application of the developed systems to various products

10. References

Canbulut, F. & Sinanoğlu, C. (2004). An Investigation on the Performance of Hydrostatic Pumps Using Artificial Neural Network. *JSME International Journal Series C*, Vol. 47, 864-872, ISSN: 1344-7653.

De Fazio, TL. & Whitney, DE. (1987). Simplified Generation of All Mechanical Assembly Sequences. *IEEE Transaction on Robotics and Automation*, Vol. 3, No. 6, 640-658, ISSN: 1552-3098.

Homem de Mello, LS. & Arthur, CD. (1990). AND/OR Graph Representation of Assembly Plans. *IEEE Transaction On Robotics and Automation*; Vol. 6, No. 2, 188-199, ISSN: 1552-3098.

Homem de Mello, LS. & Lee, S. (1991). *Computer Aided Mechanical Assembly Planning*, Kluwer Academic Publishers, ISBN: 0792392051, Massachusetts.

Homem de Mello, LS. & Sanderson, AC. (1991). A Correct and Complete Algorithm for The Generation of Mechanical Assembly Sequences, *IEEE Transaction On Robotics and Automation*, Vol. 7, No. 2, 228-240, ISSN: 1552-3098.

Kandi, S. & Makino, H. (1996). ASPEN: Computer Aided Assembly Sequence Planning and Evaluation System Based on Predetermined Time Standard, *Annals of the CIRP*, Vol. 45, No. 1, 35-39, ISSN: 1726-0604.

Pahl, G. & Beitz, W. (1988). *Engineering Design*, The Design Council, Springer-Verlag, ISBN: 3-540-50442-7, London.

Sinanoğlu, C., Kurban, AO. & Yıldırım, Ş. (2004). Analysis of Pressure Variations on Journal Bearing System Using Artificial Neural Network, *Industrial Lubrication and Tribology*, Vol. 56, No.2, 74-87, ISSN: 0036-8792.

Sinanoğlu, C. & Börklü, HR. (2004). An Approach To Determine Geometric Feasibility To Assembly States by Intersection Matrices in Assembly Sequence Planning. *Journal of Intelligent Manufacturing*, Vol. 15, 543-59, ISSN: 0956-5515.

Sinanoğlu, C. & Börklü, H.R. (2005). An Assembly Sequence Planning System for Mechanical Parts Using Neural Network, *Assembly Automation*, Vol. 25, No. 1, 38-52, ISSN: 0144-5154.

Singh, N. (1997). *Systems Approach to Computer Integrated Design and Manufacturing*. John Wiley & Sons Inc., ISBN: 0-471-58517-3, International Edition.

16

Evolutionary Optimisation of Mechanical Structures or Systems

Marcelin Jean-Luc

1. Introduction: the need for an integrated optimal design process

The research of the best compromise between economic, mechanical and technological imperatives has always been the primary objective of the mechanical engineer. The methods used to achieve these excellence objectives have evolved considerably over the last few years. The author's experience in optimisation began in 1983. At this time, the design stage would come first, then the calculation and finally optimisation afterwards. In practice, and during experience of shape optimisation of mechanical structures, between 1985 and 1990, many extreme cases were encountered. In these cases, the question of optimisation wasn't posed until damage had occurred in service; the author's industrial partners realized, often too late, that their designing left quite a bit to be desired. They would then call for the author's help in using optimisation programs to supply them with an improved shape. These shapes were reached despite technological limitations being very severe at this stage; so severe, in fact, that engineers were powerless to resolve the problem. Innumerable problems such as this were dealt with. Figure 1 exemplifies this very well. In this case, the very localized optimisation of the rear bearing of a hydraulic hammer is presented (the type of which had been sold in most parts of the world). The bearing in question would break after relatively few cycles of operation. The automatic optimisation of the shape of this product would, simply by a small modification of shape (which would be difficult to predict other than by calculation (increased radius, decreased width), considerably improved the mechanical durability of the bearing: the over-stress being reduced by 50%, the objective being the minimisation of the maximum value of the *Von Mises* equivalent stress along the mobile contour, whilst taking into account the technological constraints of the industrial partners.

Such an approach to designing has become unthinkable these days. The economic competitivity has increased, the design and manufacture delays have

been reduced and therefore the numerous overlaps that this approach involves have become prohibitive. In short, optimisation can no longer be separated from the act of designing. It is now admitted that in an integrated design approach, optimisation has to begin from the design stage taking into consideration the constraints both of specification and those induced by different materials. Optimisation is therefore made easier because constraints or limitations can be more easily varied, in accord with all those involved with the project. Examples will be found in (Clozel, 1991), (Guillot et al., 1989), (Hernot et al., 1995). This was not the case in the preceding example, where the optimisation did not take place until after being put into service, and which became an extremely constrained problem.

In this chapter, it will be shown that the integration of optimisation from the design phase is, according to the author, possibly thanks to the new optimisation techniques. A certain number of optimisation methods are popular at the moment, which are known as probabilistic or stochastic. For example, the simulated annealing method or genetic algorithms, whose principle advantages are assured convergence without using derivatives and eventual functions with discrete and non-derivable variables, even though determinist methods of optimisation (called gradient methods) necessitate a calculation resistant to these sensitivities.

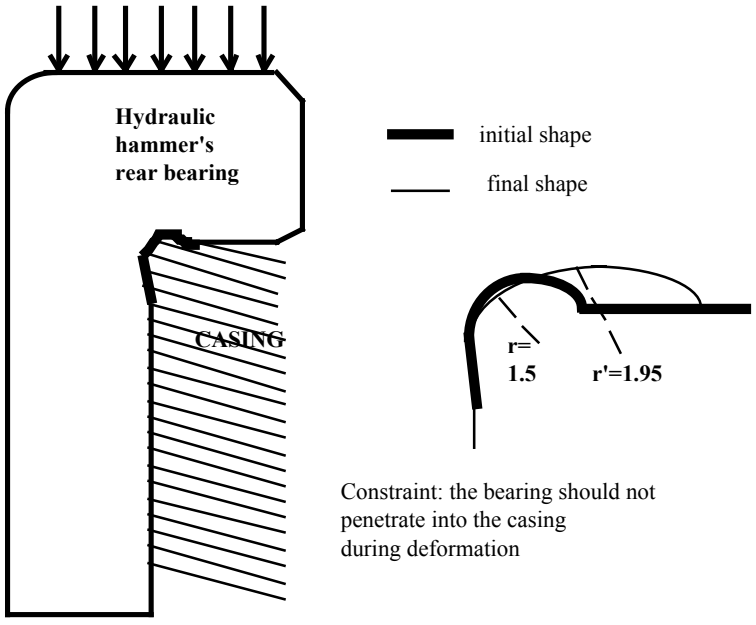


Figure 1. Optimisation of the shape of a hydraulic hammer's rear bearing

Genetic algorithms rely on the natural laws of selection which allow a living organism to adapt to a given environment. From these principles, it seems sensible to apply genetic algorithms to the optimisation of mechanical structures. As will be shown in precise examples, genetic algorithms will allow, from the beginning of the design process, adaption of the mechanical object to it's environment and to the specifications. It will be seen, especially on the example of a stiffened plate (part 3.3), how a product responding to the specifications of stiffness, weight, etc..., can be obtained directly.

After a presentation of the methods and tools used (in part 2), this chapter focuses on applications entering into the field of mechanical technology and the analysis of mechanical systems and processes (part 3). It will be seen in the conclusion (part 4), that the difficulties are more important in the case of an integrated, optimal design process of mechanical systems, because of the complexity of the problems. Nevertheless, it will be seen in this conclusion that integrated optimisation and even alternatives to A.I. (artificial intelligence) techniques can effectively be considered, for precise problems of mechanical technology, such as the optimisation of gears (part 3.1) or the construction of a mechanism (part 3.2). The conclusion supplies possible solutions for the problem in its entirety.

2. The methods used: optimisation tools adapted to mechanical technology

In addition to what has already been mentioned, the author's experience began with the shape optimisation of mechanical structures (2-D and symmetrical), although this was in the context of traditional design. See (Trompette & Marcelin, 1987), (Marcelin & Trompette, 1986), (Marcelin & Trompette, 1988), (Steffen & Marcelin, 1988).
Mathematical optimisation programs were quite difficult to use and not sufficiently versatile to be adapted quickly to new cases. In the opinion of the author, the optimal integrated design could not be achieved with normal mathematical programming techniques, which require a formulation heavily adapted to each particular problem. It will be shown in this book, that stochastic techniques are ideally suited to integrated optimisation and to mechanical technology problems.

Note that the essential characteristics of the problems are as follows:

- the design variables are often a mixture of discrete and continuous values;
- they are often highly constrained by strict technological constraints.

The problem is to maximise a function of n variables. The principle of genetic algorithms is to make a population of individuals evolve according to a replica of Darwinian theories. The starting point is a population of individuals chosen randomly and coded by binary numbers (as an example) called chromosomes. From this point, the algorithm generates, more or less randomly, new populations formed from individuals, increasingly more adapted to a given, well-defined environment. Selections and reproductions are made from the best performing parents of the population from which they come. They are stochastic or deterministic. The creation of these offspring is done by the application of genetic operators (mutation, crossing). It is always stochastic. The new replacement population is created by the selection of the best performing individuals, among either the offspring or the parents of the offspring. The replacement is either stochastic or deterministic. In the books (Goldberg, 1989), (Koza, 1992), (Michalewicz, 1996), (Rumelhart & McClelland, 1986), additional information can be found along with a demonstration of the convergence of the method.

The essential advantage of these methods is that they operate simultaneously on a test space of the solutions. The genetic method differs from the simulated annealing method by the operators which are used to force the evolution of the test population. In all cases, the convergence is always assured towards an extreme. This extreme is not necessarily the absolute extreme, but has more chance of being so, than if a traditional gradient method is used. This is shown in (Goldberg, 1989). In effect, a stochastic method explores a larger solution space. In addition, another essential advantage of these methods lies in the small number of assumptions that are required for the objective function.

2.1 Genetic algorithms

The genetic algorithms used are optimisation algorithms and make up part of the stochastic methods. They were first used in 1975. As their name implies, these algorithms seek the optimal solutions to a given problem by simulating the evolution and adaption of living organisms.
"The individual most able to adapt to a well defined environment has the greatest chance of continuing to survive and transmitting it's qualities to new individuals."

When Charles Darwin claimed his theory of the "Survival of the Fittest", a new century of understanding nature and life began. Over millions of years, life on earth has been in a process of optimal adaption to current environments. A natural selection between different individuals has maintained a changing of their genetics, that way the fittest ones, in the sense being best adapted, have survived and transferred their genetic codes to their descendants by a randomised information exchange. On the other hand, the worst adapted ones have died off.

This process is a process of optimisation. The natural selection is like a search algorithm for finding the best solution of living in a particular, natural environment. Certainly, the system of nature can not easily be transferred to technical systems. But we can have a look at, how the main rules of selection in nature are processed. And we can try to abstract and implement these for solving problems of optimisation with the help of computers. David Goldberg (Goldberg, 1989) described, based on John Holland's work (Holland, 1975), two important subjects:

- explanation of the adaptative processes of life and other natural systems,
- design of artificial systems which behave similarly like natural systems concerning their important mechanisms.

As biological systems obtain such a robustness, efficiency and flexibility, the basic rules of their mechanisms have been very interesting for artificial systems of engineering, computer or business applications. A genetic algorithm, now shortly called G.A., is a stochastic search method with the aim to scan randomly through these areas of the search space where the biggest chance of success seems to be provided. Today, G.As are proved theoretically and empirically for searches in complex spaces. They are simple for computer implementation but very powerful and effective. And there are no fundamental restrictions or limitations about the search space like continuity, existence of derivatives, unimodality, and so on.

The process of optimisation is a performance towards an optimum. This means that there is, on the one hand, the process of improvement and, on the other hand, the reaching of the optimum. Mostly, not only destination of the maximum is important, as supported by many conventional methods, but also a comparison of being or behaving better relatively to others, like the G.A. do this. The following points are figuring out the differences between G.As and other methods:

- as G.As work with a special coding of the parameter set, they can use particular similarities of this coding in a very general way. Conventional methods use the parameters themselves, and often they are restricted by a lot of limitations (continuity, unimodality, ...),
- most algorithms seek point-by-point to find the peaks in the search space. G.As do this in a quasi paralell way with a population of many strings,
- G.As have no use of any auxiliary functions like the derivatives. They use only a so-called payoff information of the objective function,
- G.As work with probabilistic transition rules, not with deterministic rules. But this stochastic effect leads to an improvment in searching, not to a random search.

Conventional search methods are not very robust, but often specially designed for particular problems, they are mostly succesful in many applications. Even sometimes, their performance can be better than these of G.As. But nevertheless, already simple G.As have their own advantages. To apply this process to optimisation, the starting point is a group of solutions to the posed problem. This group is called the initial population, and is chosen randomly from the valid domain. Each individual solution is called a chromosome, which carries information relevant to evaluate the cost function of each chromosome. This information is encoded in symbols (usually natural numbers) of which each symbol is a gene. It is known that each chromosome has the same number of genes. The greater the size of the population, the higher the effectiveness of the search, as the solution space will be explored in a wider manner. It is always limited, however, by the time needed for calculation and the capacity of the operating system. Once the initial population is generated, the algorithm begins to create a new population, equal in size to the initial population, in the case of a simple G.A.. The individuals of the new population will be developed from those of the preceding population, after having been subjected to a number of operators. In this way, each population generates a new one. The algorithm stops after a given number of generations, or by fixing other relevant function-cost criteria. In living organisms, the genetic operators are applied to the chromosomes of the parents; by analogy, the operators used in genetic algorithms are applied to the different codings of the individuals of a generation. In general, the genes can take binary values (0 or 1) to represent the state of a given piece of information. In particular cases, a gene can take a certain number of fixed values.

Genetic algorithms make up a large family of algorithms, each one different

from the other by their degree of simulation of different phenomena, related to the adaption to natural systems.
By limiting to the application of a simple G.A., the three operators which are going to be used will be:

- reproduction
- crossing
- mutation

The main idea of these operators comes from the observation of the interaction in natural surroundings. The problem is always translated into terms of maximising the objective function. The greater the objective function value of an individual relative to other individuals of the same population, the greater its chances of selection. The selection is carried out randomly, respecting the weighting of each individual.

$$P_{sel\ i} = \frac{f_i}{\Sigma_{pop} f}$$

The selection operation is applied as many times as are necessary to complete the population.

Once the selection is finished, the individuals are paired, and a crossing operator is applied to each. The probability of crossing is Pc, which is fixed beforehand. This operator consists of choosing a random position for the two chromosomes, cutting them at that position and changing the two parts beyond that position.

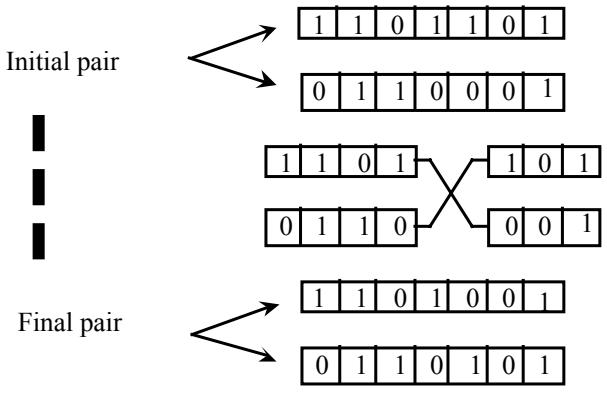


Figure 2. Crossing diagram

Reproduction and crossing are operators which transmit the good qualities of one generation to another. Important information, however, might not be represented by the chromosomes of the mother generation, or those carrying it might be destroyed accidentally by the transmission operators. Mutation consists of changing the values of a certain number of genes chosen randomly from those carried by the whole population. The probability Pm of applying mutation to a gene is fixed beforehand. If a gene is chosen to undergo a mutation, the new value is chosen randomly from among all the possible values which it could take.

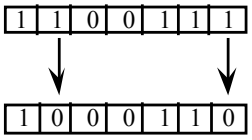


Figure 3. Mutation diagram

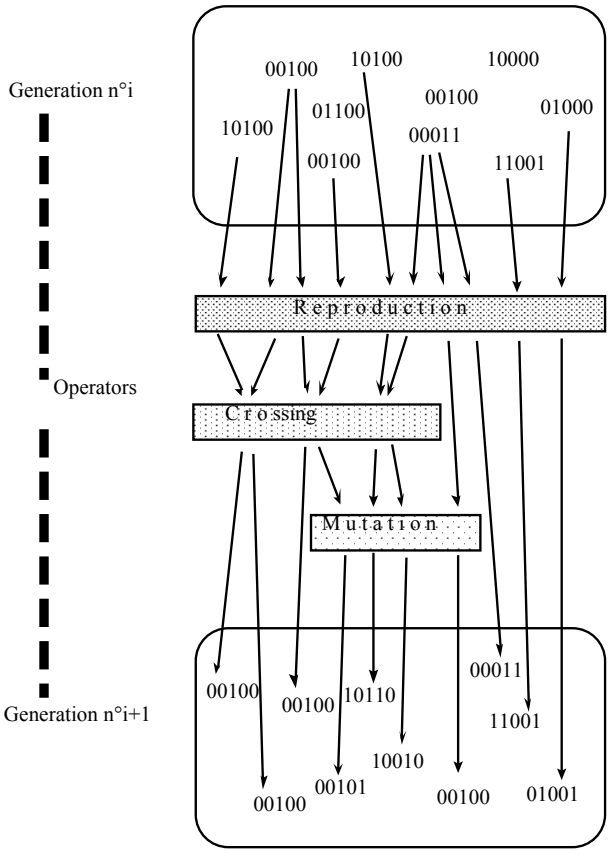


Figure 4. Diagramatic representation of the simple G.A.

The operators mentioned apply themselves to a generation to create another, respecting the sequences already stated. Figure 4 gives an overview of the actions of the three operators on a generation. The genes can take three values: 0, 1 or 2.

2.2 The simulated annealing method

The simulated annealing method, as the name implies, is based upon the metallurgical process of the same name. This "random search" method of minimisation is characterized by accepting the increases of the objective function with a given probability. This allows it to get out of the troughs (unlike deterministic methods) and therefore to escape from local minima. In the metallurgical process of annealing, if a metallic body is heated to it's melting point and then slowly cooled to ambient temperature, then the global energy of the metal will eventually pass through an absolute minimum. The basic algorithm is the "metropolis" algorithm, which is the standard of random research methods. Here is a reminder of this very simple algorithm:

Step 1: choose initial value of X0, evaluate F(X0), for k=0

Step 2: at the k+1 iteration, create a vector X, from Xk; if F(X)<F(Xk) then Xk+1=X, else Xk+1=Xk.

Step 3: if the finishing criteria have not been met, then let k=k+1 and go to step 2. If the criteria have been met then finish.

Several theorems of convergence to a global minimum were established for these methods. Difficulties in escaping from local minima remained however, which is why, of course, simulated annealing is needed.
The metallurgical process of annealing is applied to the optimisation problem. The objective function, F, is equivalent to an energy term. A temperature function, T(k), is introduced, whose purpose is to allow acceptance of the growth, by using the probability: p = exp(-DELTA(F)/T(k)). The principles and details were given in (Bonnemoy & Hamma, 1991). We shall see an application of this method in part 3.2.

2.3 Neural networks

The operation of artificial neural networks, as their name suggests, takes inspiration from that of biological neural networks. A large part of their vocabulary

is therefore been borrowed to describe them. In the author's opinion, this is as far as the similarities go. Detail of the theory, which will be summarized later, can be found in (Jodouin, 1994). The use of neural networks for the simulation or modelisation will be done in two stages: one phase which is called apprenticeship, using finite elements calculations for example in mechanics of structures, followed by a calculation or generalisation phase. In the present case, neural networks should be able to estimate an objective function or a cost function of entry or design variables. It should be noted here that the entry variables will be the binary digits of the chromosomes when using a G.A. or the real values of the design variables when using the simulated annealing method. To describe a neural network, it is sufficient to know the neuron model and the arrangement of the connections between the neurons.

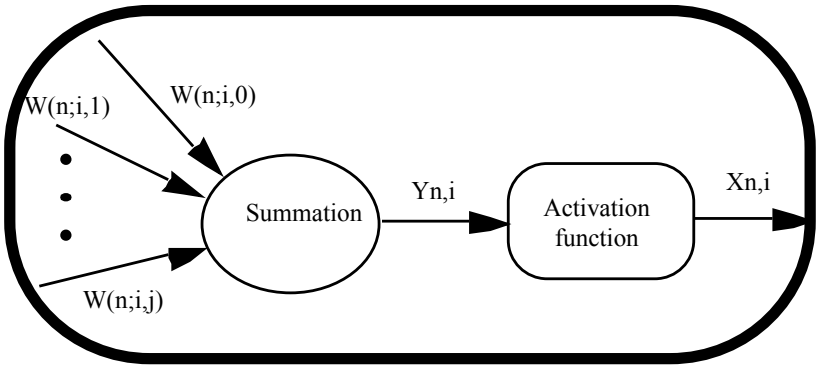


Figure 5. The elementary neuron

A neuron is modeled by two operators (figure 5). Firstly, a summing operator which develops a potential Yn,i, equal to the balanced sum of the cell entries (it is this which will be translated by the optimisation of balanced weights in the apprenticeship phase). Secondly, an operator which calculates the state of the exit value Xn,i = f(Yn,i) of the neuron as a function of its potential (f is called the neuron function, it can be either linear or non-linear). The entries are the exit values of the same layer or of another layer, or eventually the exterior entries themselves.

In the case of the optimisation procedure and binary coding for the chromosomes of the G.A., the exterior entries will be 0s or 1s, which will correspond to the chromosome digits. The function, f, can take different forms depending on

the type of network (figure 6). The most up to date models of connection networks are defined in figure 7. A complex network can be split into many layers and in this sequence of layers, a direction of information transfer can be defined. Furthermore, in the interior of one layer or between two layers, the connections can be partial or total.

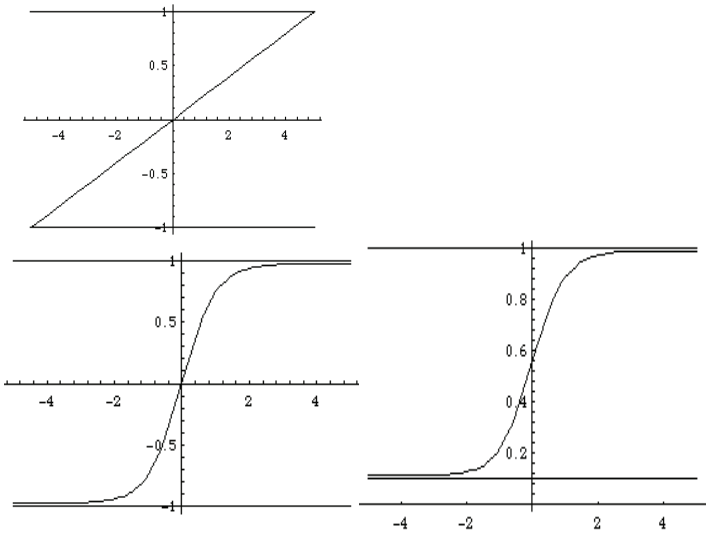


Figure 6. Some neural functions

The apprenticeship phase consists of optimising or adapting, by an apprenticeship rule, modifying the weights at each link. An apprenticeship sample is used to do this, that is, solutions which will be previously determined by finite element analysis for example in mechanics of structures. The principle criterion is to have a minimal error for the evaluation of the function. Local adaption rules (for which the weight optimisation is based on the states of the neurons connected to corresponding links) are distinguished from other rules which are much more difficult to put into use. Among the rules which are called local, and for which details can be found in (Jodouin, 1994), the best known are those which are called supervised or non-supervised, and the iterative rules. Among the non-local rules, are two of the most frequently used. Firstly, there is the Widrow-Hoff rule. This applies in particular to completely interconnected two layer networks. This was generalized in multi-layer networks by the retropropogation algorithm, to the error gradient. The error at the exit of each neuron being expressed as a function of the error calculated for the following layer by a simple differential calculation. Secondly, the Hopfield model

for the adaption rule is based on the minimisation of total energy of the network, which is the sum of the elemental energies which characterize a neuron. Each neuron is updated in a randomly drawn series.

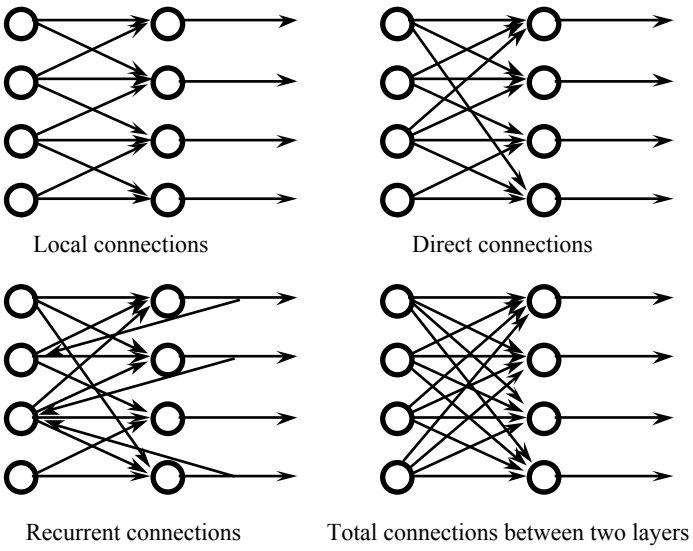


Figure 7. Some models of network connections

To summarize, a neural network works in two phases. In the first place, the apprenticeship phase, during which the adaption function is active. This allows the weight values to be optimised from a set of entry values (the design or conception variables) and from exit values (the objective function(s) or cost function) called the apprenticeship set. In the second place, the calculation or generation mode, during which the values of the weights are fixed. This allows the calculation by the neural network of the exit values as a function of the entry values.

The application of neural networks to modelisation, especially for the simulation of the calculations for the mechanical structures, seems promising from the results obtained. See (Berke & Hajela, 1992), (Szewczyk & Hajela, 1994) and (Hajela & Szewczyk, 1994). The continuation to modelisation seems natural as the action of modeling a process or a behavior, necessitates the knowledge of the principle characteristics of the process or behavior. The network knows how to extract these characteristics and can therefore be memorised easily. On the other hand, this ability to model exploits the adaption qualities of networks, allowing them to improve as they are exploited. In this work, effective

neural networks were used, at the current level of knowledge, and for which apprenticeship and generalisation/calculation algorithms are described in (Jodouin, 1994). This neural network, quite easily programmed, is a three layer network with a sigmoid neural function (figures 6 and 8) called MLP (multilayer perceptron). The MLP has been used in most of our applications.

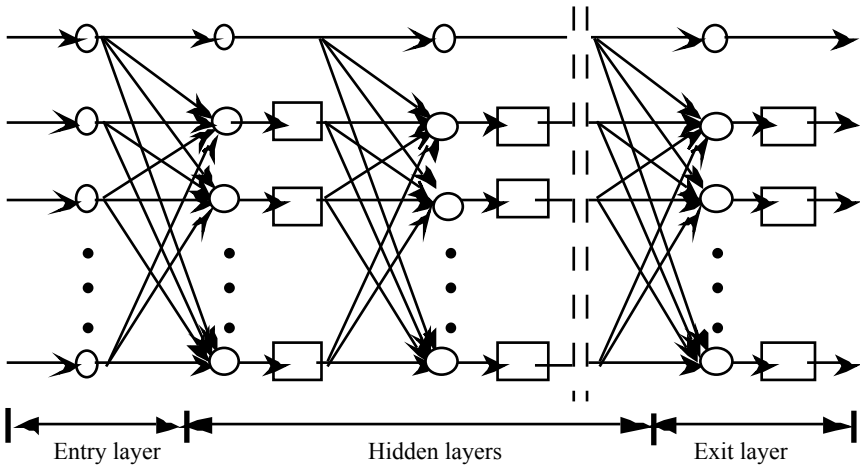


Figure 8. Neural networks used

3. Integrated optimal design of particular mechanical systems and process

3.1 Optimisation of gears

Gears are very complicated components. A large number of dominating factors vary in every case: radius of curvature, unitary loading, pressure, slip speeds, etc.... The design variables are huge in certain cases and take very discrete values (such as the module, the choice of materials). Often several objectives work in competition: balancing the energy transmission in bending and under pressure, optimisation of masses, balancing the slips, to mention but a few. The idea consists of automatically dimensionning a right-sided cylindrical gear or helical gear, so as to find a good compromise between a minimum weight, dynamic performance (energy transmission) and geometric criteria such as balancing the slips. This optimisation problem is very difficult to re-

solve by hand and often leads to compromised solutions that are not entirely satisfactory, so therefore the idea of an automatic optimisation technique is most desirable for this complex problem. In (Daidie, 1993), the authors of the paper propose a classic optimisation technique for gears. Nevertheless, these mathematical optimisation methods depend on the understanding of objective function gradients and are difficult to adapt to gears for three principle reasons:

a) initiall, certain design variables are continuous while others are discrete,
b) the derived programming is quite delicate because the optimising functions often depend implicitly on the design variables,
c) finall, the major flaw is that these methods become blocked at a local extreme (often the only way to pursue the program is to rerun the calculation with a new starting point). Thus all specialists know, the optimisation of gears is acknowledged to have numerous solutions and often it is better to adapt them to given situations. This therefore leads to the use of a genetic algorithm in order to solve the problem. Note that in (Mekhilef & Dupinet, 1993) the researchers use a method of simulated annealing (part 2.2) to solve the problem, and with some success. The problem of gear optimisation is illustrated in figure 9.

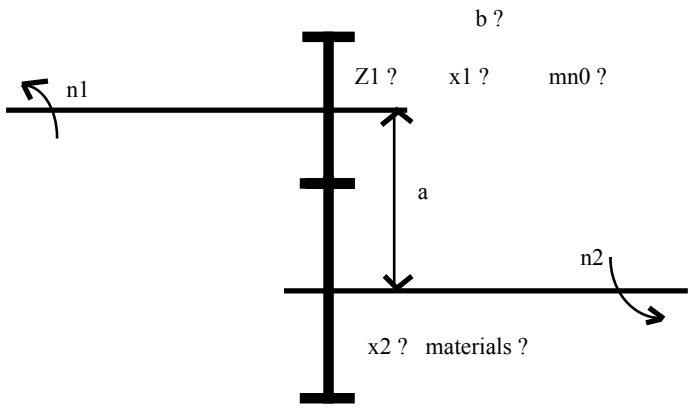


Figure 9. Definition of optimising gearing

There are two main difficulties with this problem. First of all, it is a matter of coding the solution in the form of a chromosome that will be simple and efficient; then there is the matter of finding a good compromise in function of different objectives between the different criteria (weight, power differential, balancing slips).

The coded parameters are restrained so the field of study is not too large and so therefore the chromosome is not too long. So in this way we have not considered all the design parameters in gearing, but only six main parameters: k , z1 , x1 , x2 , m_{n0}, and the material; other parameters such as the helix angle for example are fixed during the course of the optimisation. So in which case, this choice can be modified without any problems (figure 9).

When considering the choice of the objective function, we are using a multi-objective technique where the objective function will infact be a balanced sum of the different functions that we want to obtain, such as for example the minimum weight and minimum difference between slips. In which case we must weight certain objectives in respect to others; the effective choice can easily be reset, in the case of the same script of the different objectives (the shape under which they appear is at the discretion of the researcher in the domaine of optimisation). Above all though, the difficulty consisted of choosing the weighting of the coefficients in respect to their influence (that are of a different nature). This can only be done resulting from numeric experiments, where the goal was to find the best compromise possible between the different objectives. In the first place, the coding of the variables that we have used in the genetic algorithm is as follows: each of the values: k, z1, x1, x2, m_{n0} and materials are written into a binary numeration system. So, six chains are obtained C1, C2, C3, C4, C5, C6, with lengths of 4, 6, 6, 7, 4, 3 respectively.
An example of a genetic identity card (chromosome) for a gearing system is given here.

Genetic identity coding (chromosome) for gearing:

1001 110101 101100 1010100 1001 010
C1 C2 C3 C4 C5 C6

C1 : size coefficient of tooth ‘k’,
C2 : number of teeth ‘Z1’,
C3 : coefficient ‘x1’,
C4 : coefficient ‘x2’,
C5 : real shape module ‘mn0’,
C6 : material chosen from a library of 8 different types.

This coding is limited to a chromosome of a total of thirty genes long which arrange end to end (where the order is not important) the relative information of

the gearing. This coding restrains the admissible domaine of design itself. There are only $2^4 = 16$ possible width 'k' coefficients; only $2^6 = 64$ possibilities for the number of teeth 'z1' (that can vary between 12 and 75 for example); x1 and x2 only vary between -0.5 and 0.5 with two significant numbers; for m_{n0} there are 16 normalised numbers possible; finally, the material for the pinion and gearwheel is the same, and there are eight possible choices from the library of materials. For example the code 001 corresponds to 30CND8, the code 110 to 16NC6, and so on. It therefore follows, that it is possible to modify the structure or the length of the chromosome without too much difficulty.

In the second place, 'multi-objective' functions in the case of gearing are rather complicated. We propose that by the following we can modify at will, in function of the results of diverse numerical experiments. The idea is to build the function as if it were the sum of the weighted representative terms, by the coefficients that we can vary when we wish, more or less according to the importance of such and such a criteria. The function that we have used for the tests that follow, is illustrated below:

$$F=10^{10} - \frac{I_1}{Rap}\left(\frac{b}{b_{max}}\right)\left(\frac{d_1}{d_{1max}}\right)^2 - I_2|g_{s1}-g_{s2}|$$

$$-I_3 \frac{Rap}{Ptrans}\left[\,|P_{rup} - c.P_{trans}|+|P_{pres}-c.P_{trans}|\,\right]$$

I1, I2 and I3: weighting coefficients ,
gs1, gs2: maximum absolute slips,
Rap: ratio of quality against price of material ,
b: width of material,
d1: primitive diameter of pinion.

The presence of the term 10^{10} is due to the fact that the G.A. maximises the functions. To calculate the functions at a minimum, it is possible to look for the maximum of the opposing function plus a very large term. The second term affected by coefficient I1 is a term relating to the minimisation of gear size with relation to a given maximum size. This term is penalized in terms of quality/price of a material. The third term affected by coefficient I2 expresses the equalizing of the absolute slip (very important in reducing wear). Finally, the fourth term, affected by coefficient I3, is a term expressing the search for balance between the transmissible powers under pressure and under flexion, and

also respecting the safety factor with relation to power transmitted. It is possible to add other criteria to this multi-objective function, e.g. a term expressing maximisation of driving relation or a term ensuring imposed distances between axes are respected. After several numeric tests on a basic example, the values of weighting coefficients below were chosen for the following test case: I1=0.2, I2=0.1 and I3=10. This test concerns a helicoidal gear used in a fixed axis gearbox. The parameters of the G.A. are: population size=200 and number of generations=100. The results are compared to a reference solution, optimised using other methods.
The given factors are:

Ptrans=400KW
Nmax(input)=1485 tr/min
transmission relationship u=6
developing circle of teeth ⌀=8°33'
Quality factor Q=6
Life H=200000 hours
functionning with negligible shock.

The following is the best solution of the last generation:

Geometrical analysis :

	mn0	Z1	x1	x2	k	Material
reference solution	5	26	0.44	-0.4	32	16NC6
genetic algorithm	6	23	0.10	0.09	13	16NC6

	reference solution	genetic algorithm
width b (mm)	160	78
d1 (mm)	131.4	139.9
volume bd_1^2	2.7E6	1.5E6
gs1	0.24	0.48
gs2	0.33	0.30
e_e	1.63	1.85
Pflex (kW)	1100	760
Ppres (kW)	1000	740

The objectives have been achieved: a correct balance between absolute slips gs1 and gs2 and powers with a sufficient saftey factor. It is also notable that the

volume of the gear in the genetic solution is clearly less than the volume of the reference solution.
Numberous other tests of this nature have been conducted and the optimisation objectives are always successfully met. For all these tests the material systematically selected (from a list of available materials) is the highest performing, that is 16NC6. In reality, amongst the final solutions, there are other perfectly acceptable solutions using less high perfomance steels, but we have chosen the best one each time.

3.2 Optimisation of mechanisms

We are going to show that the evolutionary methods can also be very efficient for problems of optimisation of mechanisms. Computer Aided Design (CAD) of mechanisms has already been approached in a number of different ways in different papers. There are systems of assistance in mechanism design (Guillot et al., 1989), (Clozel, 1991); certain articles have tried techniques of artificial intelligence (Hernot et al., 1995). Others have attacked the difficult problem of aided choice of mechanism topology taking into account kinematic criteria, cost reducing criteria, and ease of design (Sardain, 1994). The present part concentrates on the problem of optimisation of mechanisms under the following restricting hypothesis: we remain within the scope of fixed topologies and we consider mechanisms which are isostatic or slightly hyperstatic. The principal objective is to minimise the force transmitted in each liason; the design variables are the relative positions of the different liasons in respect to each other; furthermore certain technological limitations on overall size, or the exclusion of certain areas of the layout or space for the predifined liasons must be respected. It is in fact a question of a first approach destined to show that it is possible to optimise mechanisms using methods of optimisation completely random (trial and error) and automatic, without calling on techniques of artificial intelligence (A.I.). The alternatives proposed in place of A.I. are interesting because probabilistic methods allow, at once, a rigorous and robust selection of ideal technological solutions which are compatible with the existing technological limitations. Subsequently, it will then be possible to attempt more complex problems, like the optimisation of mechanism topologies; problems for which stochastic methods are particularly suitable since they are perfectly suited to problems with discret variables. We will show, by means of examples, that when put into action the procedure allows optimisation of mechanisms with a definite efficiency.

Now to study a mechanism (representing a mixer) which creates a transformation of movement which is represented by figure 10. It is composed of 3 solids S1, S2, S3, and a fixed housing S0. S1 is an entry shaft with uniform rotation. It is connected to the housing by means of a horizontal axis pivotal liason (fixed rolling element bearing). It is connected to S2 by means of a free floating rolling element bearing on a horizontal axis. S2 is connected to the exit shaft S3 by means of a ball and socket joint (in effect a basic fixed rolling element bearing on a vertical axis); the fact that S2 and S3 are linked by a ball and socket joint renders the system isostatic. S3 whom is linked to the mixer blades is connected to the housing with a free floating rolling element bearing about a vertical axis. It is assumed that only parts 1 and 3 recieve external forces applied to the mechanism. The objective is once more to minimise the inter efforts unknowns inside the system with a goal of finding the dimensions which have the least cost. In order to be able to calculate the forces transmitted in the different liasons, and for the purposes of this calculation only, it is assumed that all the components of the external forces applied to parts 1 and 3 are 10kN and all the components of moments applied to 1 and 3 are 1kNm. This being, a standard program of static analysis of mechanisms allows to calculate, for a given configuration of the mechanism, the torques transmitted in the different liasons. For the purpose of this test, the objective function is taken as minimising the quadratic sum of all the components of forces and of moments of every liason (for this function to be homogenous, the moments are divided by an equal reference length, 100mm). If analysing only 1 or 2 particular liasons it is possible to limit the objective to only the components in question.

With regards to design variables allowing to define the relative positions of liasons with respect to one and other, it is possible to identify 5 which are independent. These variables, marked X1 to X5 are the following:
X1= 1 (see figure 10), X2= R = O1O2, X3= z(O4), X4= angle A1, X5= angle A2
With regards to limitations on the design variables the following factors are used: the horizontal dimension, L, is fixed at a value of L= 200mm (figure 10); this limitation gives us a relationship allowing to calculate the distance between O2 and the centre of the ball and socket joint in terms of L, X1, X2, X4, and X5.

Otherwise, the design variables are limited in the following manner:

$0<X1<100$, $0<X2<50$, $-100<X3<0$, $0<X4<90°$, $0<X5<45°$

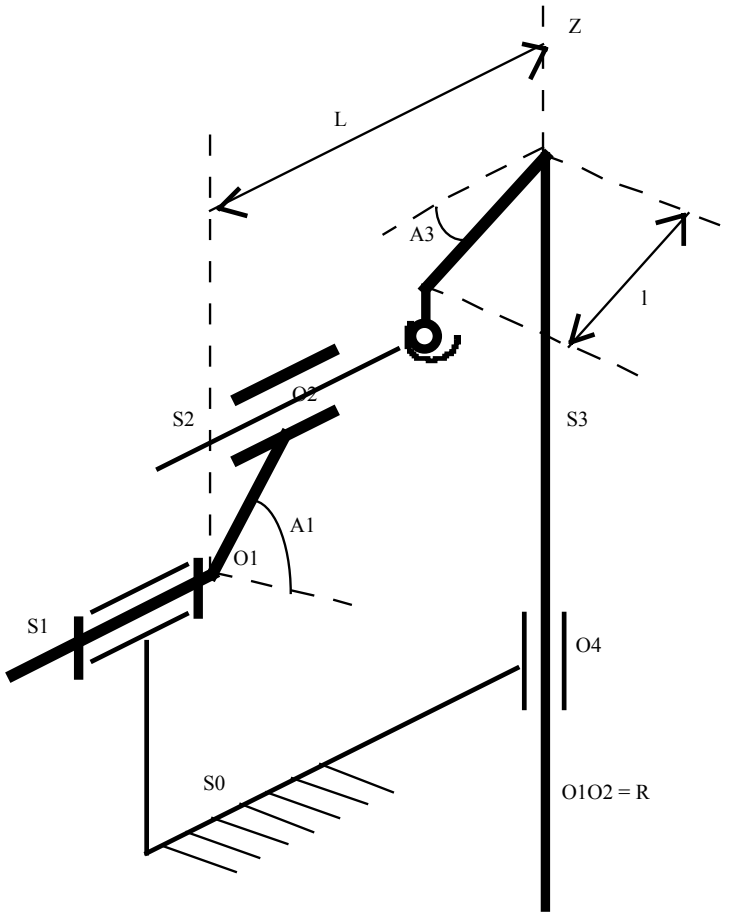


Figure 10. Optimisation of mechanism

For this test we must make an initial optimisation using a G.A., to roughly work out the problem, the more precise optimisation using simulated annealing methods, starting from an initial solution given by the G.A.. So that the optimisation by the G.A. will be effective and as it is only used for the first approximation, we volountarily limit the coding of the five design variables to a binary chromosome of 10 digits in total. The structure of this binary chromosome is as follows:

- the first 2 digits allow the coding of variable X1, the following 2 digits the coding of X2, and so on;
- the 2x5 digits are then put side by side to form a chromosome of 10 digits;

The coding is crude but it is after an initial passage that it is possible to improve using a more precise coding. In the present case the decoding will be done in the following manner:

variable X1:	00	-->	20, 01	-->	50, 10	-->	75, 11	-->	100
variable X2:	00	-->	20, 01	-->	30, 10	-->	40, 11	-->	50
variable X3:	00	-->	-50, 01	-->	-60, 10	-->	- 80, 11	-->	- 100
variable X4:	00	-->	0°, 01	-->	20°,10	-->	50°, 11	-->	90°
variable X5:	00	-->	0°, 01	-->	10°,10	-->	25°, 11	-->	40°

It can be seen that the limitations of the problem, in particular on the design variables are integrated in the coding. It is not as necessary to penalise an objective function that will be of type a - F (a being a very large constant) because the G.A. maximises the functions. For a population of 30 individuals and 50 generations, the G.A. quickly comes to the following solution:

X1=100; X2=50; X3=-50; X4=20°; X5=0°

which corresponds to the chromosome 111000100, and a value of 1.1013E8 for F. This represents a gain of 30% in comparison to an average solution, for example:

X1=20; X2=20; X3=-100; X4=90°; X5=40°;
chromosome=0000111111;

F=1.318E8;

We now bring in optimisation by simulated annealing, starting from the inital solution:

X1=100; X2=50; X3=-50; X4=20°; X5=0°; F=1.013E8;

The improvement is hardly noticable, the solution proposed is:

X1=100; X2=50; X3=-26; X4=5°; X5=2°; F=1.0105E8;
giving a gain of only 0.3%.

We note that the tendency of solutions is to take X1 as large as possible and angles A1 and A2 small, as far as the technology will allow. With regards to components of forces and moments, results for the final solution and some force characteristics are given in the table below. Values for the average solution (chromosome 0000111111) are given in brackets.

for liason 01: Y01 = 200. DaN (753.), M01 = 4974. mmDaN (8467.)
for liason 12: Y12 = 1. DaN (6.5), M12 = 5025. mmDaN (18467.)
N12 = 5028. mmDaN (120522.)
for liason 23: Y23 = 1. DaN (6.5)
for liason 03: Y03 = 0.1 DaN (553.), L03 = 13386. mmDaN (89601.)
M03 = 6. mmDaN (8467.)

One notes a very important reduction of the values of forces and moments. The programming and the implementation of the two methods is simple and does not pose any particular problems given their remarkable effectiveness, and it is this that is their appeal. It is sufficient to call upon the program as many times as is necessary (for the G.A. and the last test done 30x50=1500 times), after having first done the decoding of the proposed chromosome following the rules given earlier.
With regards to simulated annealing, the number of calls made upon the program is over 100000 because there are 5 design variables. The size of this number is the principle disadvantage of this method.

3.3 Optimisation of stiffened plates and shells

To be able to anticipate and to optimise from the design phase, the dimensions and the number of stiffeners or ribs in mechanical structures, is probably one of the greatest problems for mechanical engineers; its resolution from the initial conception makes it possible to eliminate a great deal of ultimate problems and adjustments.

In the following section, the example of hull supports are treated simply as beams. The plates and hulls are treated such that they are thin so that the thickness dimension is much smaller than the other two dimensions. Since the ratio of toughness against weight is very important as well as good behavioural characteristics, these examples can be put to many different uses but are mainly used in industry and civil engineering applications: food tins, car bodywork, planes, ships, liquid and gas tanks, bridges, cooling structures, spatial vessels, petrol tankers, and so on.

The problems studied in this part will be limited to the bending of plates. The plates possess a higher stiffness concerning coplanar strains that are the displacements in the mean plane of the plate and rotations perpendicular to this plane. This said, the stiffness in relation to the perpendicular displacement to the plane of the plate and in relation to the rotations about parallel axes to the

plane of the plate, is a great deal weaker. This part treats the stiffening of plate structures with the addition of beam supports. Essentially, the aim is to optimise the positioning of the supports with the objective being the yield (perpendicular displacement to mean plane of the sheet); this procedure can bring about interesting improvements in a structure behaviour.

One uses a genetic algorithm to optimise the position of a series of supports of equal lengths precisely on a plate in bending. The plate studied in this example had dimensions of 2.4 m long by 1.5 m wide. The plate was completely embedded along one of it's width and was made of steel of elastic modulus E = 2.1 10^4 DaN/mm^2 and Poisson's ratio 0.3. It was 2 mm thick (figure 11). The supports were made from the same material as the plate and were completely integrated. They have a drop height of 8 mm and a width of 5 mm .

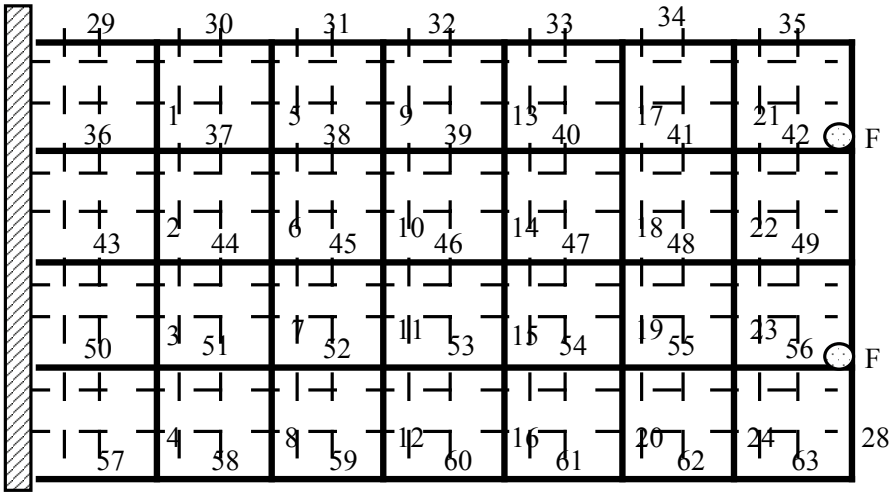


Figure 11. Numeration of the plate supports

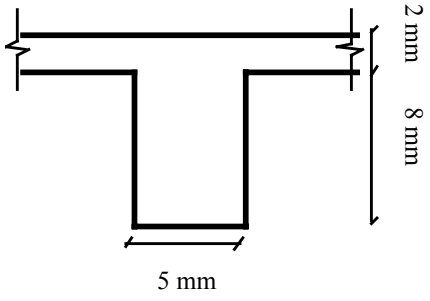


Figure 12. Cross-section of the plate and support

The plate is held horizontal by being embedded along one of it's width. At the opposite end, two identical forces of 144 DaN are applied vertically. In this example, the overall mass of the structure is not taken into account.
So, our objective here is to reduce the maximum bending of the plate while optimising the positioning of a number of stiffeners. The maximum bending is the greatest vertical displacement of the plate. The element distribution represented on figure 11 shows that there are 525 possible positions for the stiffeners, each 100 mm long. Such a large domain of research requires a huge calculation time, so in order to reduce this, the number of possible positions is reduced and regrouped such that a support is made of three aligned consecutively to a length of 300 mm. This process means that there are only 63 possible positions (as shown in figure 11).

A traditional coding technique in a genetic algorithm with such an example consists of building a chromosome with 63 genes, with each one taking a binary number (0 or 1) and decoded as follows:

> if a gene of position 'i' carries binary number 1, stiffener n°1 exists but in the opposite case it doesn't exist. If the genetic algorithm is run with such a coding it will obviously head towards a chromosome where all the genes carry a binary number 1, because the stiffer the plate is, the smaller the displacement will be. Nonetheless, the aim is to limit the number of stiffeners distributed on the plate and in this study it is limited to 14.

One coding technique consists of traditional coding and a system that disposes of any chromosome with any more than 14 stiffeners to multiply their functional values by a number less than unity and function the number of overabundant stiffeners.
This technique though, risks filling the population by ultimately rejected individuals and therefore cost functional values will be uselessly calculated. In our study, another type of coding has been used that builds chromosomes that are of equal length to the total number of stiffeners hoped for, and only containing the number of positions of the different stiffeners that must exist in the configuration. This works so that the chromosomes built are 14 genes long with each gene taking an integer between 1 and 63 inclusive.
So with this coding technique we have individuals with one or several genes carrying the same value which can be interpreted and then decoded in several different ways following the desired design. If there are 'n' genes carrying the same value 'm', we can consider for example that at position 'm', we have:

- just one stiffener 5 mm wide and drop height 8 mm ,
- one stiffener having a drop height of 8 mm and width n5 mm ,
- one stiffener having a width of 5 mm and drop height n8 mm ,
- effectively 'n' identical stiffeners of width 5 mm and drop height 8 mm .

Here we have adopted the fourth possibility. It is important to note that the individuals obtained by permutation of the positions of a chromosome have the same distribution configuration, so the same cost functional value and therefore the program considers them as being the same individual.
Summary: the objective is to minimise the maximum yield (localisation can vary in some cases), the problem was uniquely constrained by a maximum number of stiffeners arranged on the plate.

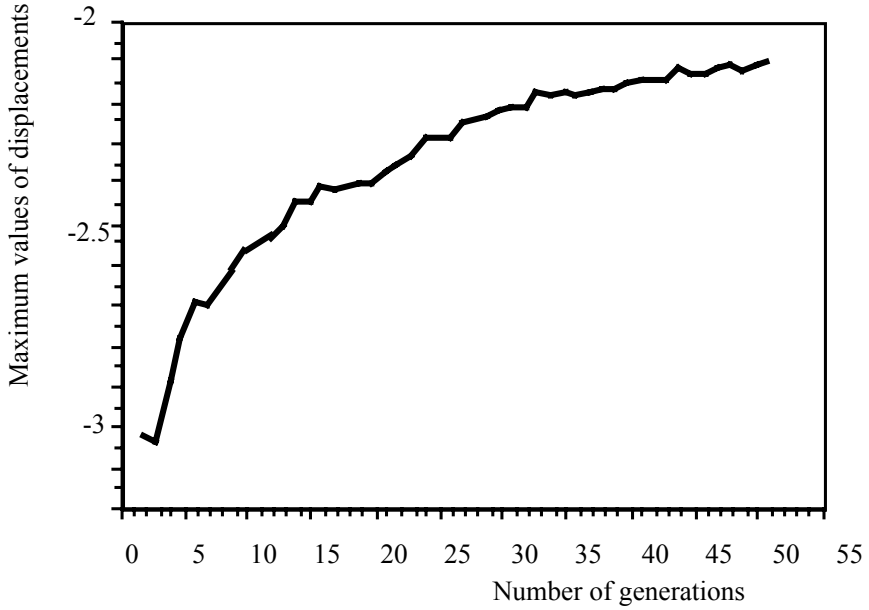


Figure 13. Maximum yield values for each population against revolutions

3.3.1 Without neural networks

On figure 13, we find the converging results obtained by the genetic algorithm for the plate, settled at 50 generations, the number of individuals per population is 50, Pc = 0.6 and Pm = 0.03 (Pc and Pm are probabilities of crossover and

mutation respectively). The curve represents the maximum value of yield for each population function to the number of generations: this curve is the mean of the result of 9 runs of the program with the same parameters. The maximum yield of the unstrengthened plate with the limiting conditions already mentioned, including the loading , is -4.6702 mm.

The optimum solution found by the genetic algorithm is represented graphically on figure 14. The corresponding cost function value is f max = -1.8243 mm.

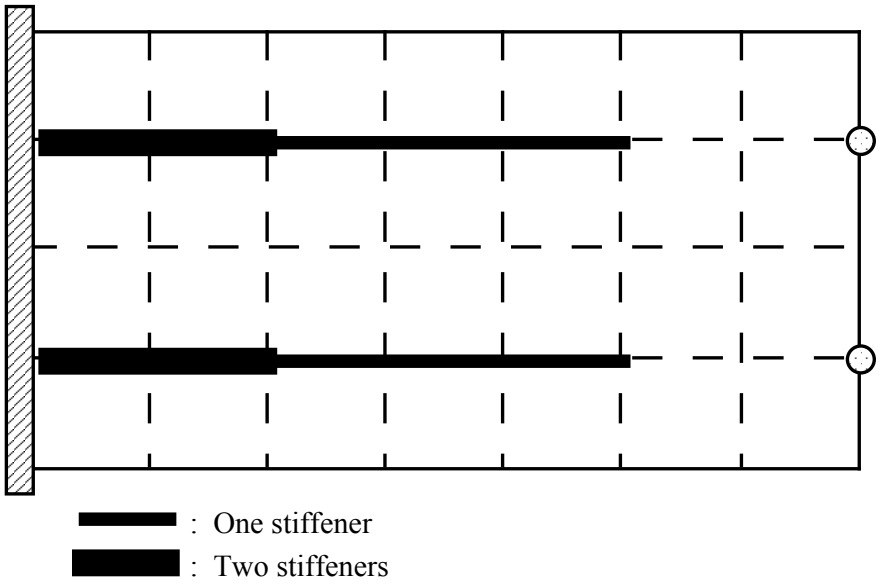


Figure 14. Optimum position of stiffeners

3.3.2 With neural networks

Initiating a neural network with 100 chromosomes, we arrive at a relatively effective neural network, since the total errors made are no greater than 7 % as shown by the examples below with, in the following order, the chromosome, the maximum yield evaluated by the neural network and the percentage error with respect to calculations from the finite elements. These errors are tolerated by the genetic algorithm (so much so that they are always of the same sign). The genetic algorithm re-runs on the same solutions as previously, but around twenty times quicker.

3	12	12	27	29	31	34	36	39	40	50	51	54	57	-3.093	0.08%
3	4	5	19	19	21	22	30	31	34	53	54	55	59	-3.980	0.15%
2	3	5	17	24	42	43	47	47	48	51	56	59	60	-3.557	0.6%
1	13	14	21	32	38	39	44	49	53	55	56	62	63	-3.995	0.7%
1	13	14	16	18	19	23	25	29	39	42	45	52	55	-4.019	0.17%
2	7	9	10	12	19	34	34	34	40	43	44	53	59	-3.499	6.3%
8	12	14	15	17	24	24	35	37	41	52	56	57	58	-3.635	5.8%
12	31	32	36	40	42	44	44	50	50	50	55	59	61	-2.996	6.5%
4	7	9	11	18	18	25	29	32	35	39	42	51	51	-4.015	6.34%
7	8	12	27	28	38	39	46	49	50	51	54	54	54	-3.439	6.9%

4. General conclusions and synthesis

4.1 Conclusions on neural networks

In the course of our numeric experimentation on neural networks they have seemed to present some limitations. These limitations are not to do with data processing: currently, thanks to the improvements in computers it is possible to use neural networks of a significant size (more than 1000 neurons and 1000 weights). It is more the absence of an established theoretical knowledge of the functioning of the networks that renders their use delicate. Various problems have needed to be dealt with, such as the necessity to study the feasibility of every application before the numerable numeric experiments, the uncertain sizing of a network, or the absence of theory for anticipatory calculation of errors. For the modelling of mechanical structures it is reasonable to wonder if the use of a simple method derived from the Rayleigh-Ritz method (well known for vibrations) would not be more suitable for the problems considered here. This idea can be illustrated using a well known example in statics.

Suppose that one wanted to optimise, for example, the number and situation of stiffeners on a given plate. One can begin, as with the neural networks, to evaluate by finite element methods, a number of representative solutions. In a neuromimetic strategy these solutions act as the "learning". In a Rayleigh-Ritz type strategy the solutions are used to find the stiffness corresponding to a

new configuration of stiffeners without having to re-do the finite element calculations. The new solution is searched in the form of a weighted sum of test solutions previously calculated. To find the weighting coefficients, or weights (as with neural networks), mechanics offers a reliable theory: the weights are obtained by minimising the total potential energy of the plate in question. It is also possible to take the minimisation of the error, or of the residue on the equilibrium equation as criteria. This strategy has been developed in (Marcelin, 2001). The method then appears to be more a Galerkin weighted residue method. One noted a certain similarity between these strategies and the neural networks, the difference being that mechanics offers a rigorous error criteria. In their favour neural networks have the advantage of having a better capability to adapt. Moreover there is nothing to stop us operating a neural network using the error crieria of mechanics to optimise and control the weights. It is proposed to test these strategies in the near future. Part of this work has been given over to neural networks which present a number of intrinsic qualities. These qualities may eventually make the networks superior to conventional mechanical methods discussed earlier. The first quality which comes to mind is parallelism. The networks are made up of elementary units who can calculate simultaneously (one of the reasons for the superiority of the brain). They are also very capable of adapting. Finally, they can resolve the imprecise, recognize the vague, and so, prove to be highly robust. The new strategy presented here consisted of calculating the objective values of the initial population of the G.A. by finite element methods and, after, doing the learning stage of a neural network which takes over the calculation of the objective functions of following populations. This strategy has also been developed in (Marcelin, 1999). The neural network generally used in mechanics and that been used here is the Multi-Layered Perceptron (MLP); the learning of this network is effectuated while minimising the error at the exit of the network (this error being defined as the square of the difference between the desired value and the value given by the MLP). This strategy has proved very effective since the error given by the MLP has no influence over the convergence of the G.A..

Nevertheless the MLP has presented a number of difficulties:

- determining the necessary number of layers and neurones;
- difficulties due to optimisation of network parameters by means of gra dient analysis (starting points, focusing on local maximums,...).

4.2 Evolutionary optimisation: an alternative to A.I. techniques ?

The stochastic methods are based on the natural laws of evolution and of adaption of space which allow living organisms to adapt to a given environment. As is shown in literature, more and more abundant, on the subject, it seemed astute and "natural" of the numberous authors to apply these laws of evolution, that is to say adapt, to artificial stuctures. The principal advantages of these methods are an assured convergence without the use of differentiation, and for the eventual functions with discrete variables. The major inconvenience however is the number of calculations, but this may be relieved, as we have seen, by the use of neural networks.
The problem posed is to adapt mechanical structures to their technological environment. To optimise these mechanical structures, it is not often obvious to use the deterministic methods of classical optimisation, methods of gradient. These methods require a reliable calculation of sensitivity, which can be difficult for certain problems. Furthermore, in mechanical technology, the problems are essentially with discrete variables (since optimal components are normally selected from catalogues), and until now authors have tended more towards the use of artificial intelligence (A.I.) to find solutions to these problems. Nevertheless, in certain applications, such as optimisation of gears (part 3.1), stochastic methods of optimisation are quite well suited. The CAD of mechanical systems has already been discussed in a number of manners, often calling on techniques of A.I. The facility to implement the stochastic methods, as their versitility and adaptability suggest, at least for the examples considered here, can be used as an alternative to classical techniques of A.I.. It is this we have tried to demonstrate in this work by means of numerous examples: optimisation of gears; optimisation of mechanisms; optimisation of topology of stiffeners on plates.

4.3 Towards an optimal integrated design for mechanical systems

As was shown in part 3, it is possible to tend towards an optimal integrated design for mechanical structures. Currently the implementation of an optimal integrated design for mechanical systems, that is to say, taking into account a maximum of information from the beginning (know-how, ability, optimisation constraints), proves difficult due to the fact that the necessary specialist software, in most cases, functions independently from other programs. This can be illustrated by the example of a gear box. A program of mechanical analysis is

used initially to ensure a sound structure from the start, afterwards, specialist software is used for calculations of gears, bearings, shafts, ...The same applies for finite element calculations to control the shape and strength of certain components. Currently, even if each stage of the problem is presented in terms of optimisation as is seen in part 3.1 (dealing with gears), the problems remain, most of the time, bound to a specific order. Research in integrated design is orientated towards the use of common databases at different stages of the design. The goal of this work is to propose a fundamentally different approach, allowing at once, an optimisation which is both global and almost automatic. It should be made clear that given here is the point of view of a mathematician. The principle of the proposed method is to use a neural network as a global calculation program and to couple this network with stochastic methods of optimisation. Bearing in mind that the new strategy proposed consists of three stages: first, defining the parameters of the mechanism taking stock of all design variables, as well as assesment of desired objectives and technological limitations; secondly, the "learning" of the neural network with the goal of having a "mega-program" of analysis and calculation (perfectly adapted to the task in hand), including a knowledge of all the programmes which will be used in the design process; finally, use of this "mega-program" for totally automatic optimisation, without the need for human intervention, thanks to stochastic methods; the method used here is that of G.A.. The expected result is a play of optimal design variables. This strategy has been developed in (Marcelin, 1998).

5. References

Berke, L. & Hajela, P. (1992). Applications Of Artificial Neural Nets In Structural Mechanics. *Structural Optimisation,* Vol 4, 1992, P. 90-98.

Bonnemoy, C. & Hamma, S. (1991). La Méthode Du Recuit Simulé: Optimisation Globale Dans Rn. *Apii,* Vol 25, N°5, 1991, P. 477-496.

Clozel, P. (1991). Mecamaster: Outil De Conception Mécanique Et De Cotation 3d Pour Les Bureaux D'études, *Actes De La 10ème Conférence Micad Publiés Par Les Editions Hermès*, 1991, P. 196

Daidie, A. (1993). Dimensionnement Optimal D'un Train D'engrenages A L'aide D'un Logiciel Cao, *Actes Colloque Primeca, Ecole Centrale Paris, 24, 25-26 Novembre 1993.*

Goldberg, D. E. (1989). *Genetic Algorithm In Search, Optimisation And Machine Learning.* Addison - Wesley, 1989.

Guillot, J.; Rousselot, J.Y. & Vignat, J.C. (1989). Conception Assistée Par Ordinateur D'ensembles Mécaniques Avec Recherche D'une Bonne Solution: Le Logiciel Sicam *Micad, Paris,* 1989.

Hajela, P. & Szewczyk, Z. (1994). Neurocomputing Strategies In Structural Design-On Analysing Weigths Of Feedforward Neural Networks. *Structural Optimisation,*Vol 8, 1994, P. 236-241.

Hernot, X.; Daidie, A.; Silberberg, Y. & Guillot, J. (1995). Conception Intégrée De Mécanismes, *Revue Internationale De Cfao Et D'infographie,* Vol 10 - N°1-2, 1995, P.57-70.

Holland, J. H. (1975). *Adaptation In Natural And Artificial Systems,* Ann Arbor, The University Of Michigan Press, 1975.

Jodouin, J.F. (1994). *Les Réseaux Neuromimétiques,* Hermès, 1994.

Koza, J.R. (1992). *Genetic Programming: On The Programming Of Computers By Means Of Natural Evolution.* Mit Press, Massachussets, 1992.

Marcelin, J.L. & Trompette, Ph. (1986). On The Choice Of The Objectives In Shape Optimisation. A Nato Advanced Study Institute, Computer Aided Optimal Design, Troia (Portugal), June 29-July 11, 1986, Actes P. 247-261

Marcelin, J.L. & Trompette, Ph. (1988). Optimal Shape Design Of Thin Axisymmetric Shells. *Engineering Optimisation*, Vol. 13, Pp. 109-117, 1988

Marcelin, J.L. (1998). Optimisation Intégrée De Mécanismes A L'aide De Réseaux Neuromimétiques Et De Méthodes D'optimisation Evolutionnaires. *Revue Internationale De Cfao Et D'informatique Graphique*, Vol.13, N° 3, Pp. 265-281, 1998

Marcelin, J.L. (1999). Evolutionary Optimization Of Mechanical Structures: Towards An Integrated Optimization, *Engineering With Computers,* Vol. 15, 1999, P. 326-333.

Marcelin, J.L. (2001). Genetic Optimization Of Stiffened Plates And Shells, *Inter. Journal For Numerical Methods In Engineering,* Vol. 51, 2001, P.1079 - 1088 .

Mekhilef, M. & Dupinet, E. (1993). "Optimisation D'un Train D'engrenages En Variables Mixtes", *Actes Colloque Primeca, Ecole Centrale Paris, 24, 25-26 Novembre 1993.*

Michalewicz, Z. (1996). *Genetic Algorithms+Data Stuctures = Evolution Programs* Springer Verlag, New York, 3rd Edition, 1996.

Rumelhart, D.E. & Mcclelland, J.L (1986). *Parallel Distributed Processing,* Vol. 1 And 2, Mit Press, Cambrigde, Massachussets, 1986.

Sardain, P. (1994). Un Environnement Cao Pour La Synthèse De Mécanismes Articulés , *Revue Internationale De Cfao Et D'infographie,*Vol 9 - N°1-2, 1994, P.135-154.

Steffen, V. & Marcelin, J.L. (1988). On The Optimisation Of Vibration Frequencies Of Rotors. *Int. Journal Of Modal Analysis*, July 1988, P.77-80.

Szewczyk, Z. & Hajela, P. (1994). Neurocomputing Strategies In Structural Design- Decomposition Based Optimisation. *Structural Optimisation,* Vol 8, 1994, P. 242-250.

Trompette, P. & Marcelin, J.L. (1987). On The Choice Of Objectives In Shape Optimisation. *Engineering Optimisation,* Vol. 11, 1987 P. 89-102.

17

Improving Machining Accuracy Using Smart Materials

Maki K. Rashid

1. Introduction

Both economical and ecological factors might encourage conventional machines to continue in service by healing tool vibration problems. Higher productivity in automated manufacturing system brought to the attention the importance of machine tool error elimination. Various factors might affect the machining process (Merritt, 1965), some of them are non-measurable, and others might change in real-time. However, the wider use and availability of suitable and economical microcontrollers encouraged the use of intelligent control scheme to overcome such time dependent problem. Large magnitude of excitation forces with a tiny relative motion between cutting tool and working piece promote the use of smart material actuators that interfaced with microcontrollers to counteract such motion errors (Dold, 1996). Rigid fixture is a requirement to minimize displacements of cutting tools from its nominal position during machining. However, the reconfigurable manufacturing era encourage the use of small fixtures with lower mass (Gopalakrishnan, et al., 2002) and (Moo n& Kota, 2002).

Previous dynamic modeling of a smart toolpost (Frankpitt, 1995) is based on linear piezo-ceramic actuator. The system is either modeled as lumped single rigid mass incorporating tool carrier (holder), tool bit, and piezo-actuator. Or by using an effective mass, stiffness, and, damping coefficients for the most dominant mode of vibration. The fundamentals of this model are incorporated to design an adaptive controller using the measured current, and, voltage applied to the actuator as a control signals. Based on identical principles (Eshete, 1996) and (Zhang et al., 1995) a mathematical model is derived for smart tool post using PMN ceramic material. A control system, and real time microprocessor implementation was examined in (Dold, 1996]. Sensitivity analysis for the toolpost design modifications and interfacing parameters on tool dynamic response require further elaboration. No conclusions are drawn related to better design and selection of actuator, tool holder and tool bit stiffness ratios. In

case of a future geometrical change, the validity of the lumped masses in system modeling is questionable. Nature and type of signals that control smart material actuator and how can affect toolpost dynamic response suffer from scarcity of information. Recently a systematic engineering approach is used to investigate an optimum fixture–workpiece contacts property (Satyanarayana & Melkote, 2004), machining fixtures dimension (Hurtado & Melkote, 2001) and structural stiffness in toolpost dynamic (Rashid, 2004) by using the finite element approach.

Present analysis investigates the capability of smart material in tool error elimination using finite element modeling. This incorporates structural stiffness evaluations for toolpost actuator, tool holder, holder fixture, and tool bit. Radial tool movement relative to the workpiece is regarded as a main source for cutting tool error. Considerations are given for evaluating lumped mass modeling, effectiveness of dynamic absorber in case of PWM voltage activation and effect of toolpost stiffness ratios on error elimination. Awareness is given for the model to be capable of handling large variations in design parameters for future toolpost development in the case of limited space and weight requirements. Other issues are related to the effectiveness of dynamic absorber presence, switching rate and voltage modifications to minimize tool error.

2. Toolpost FEM Model

In this work the Lead Zirconate Titanate (PZT), is the intelligent material for the investigated smart toolpost actuator. This encouraged by the well-developed theoretical analysis of this material and its common use. Two models are applied for obtaining the toolpost results. The first is shown in Fig. 1 (a) represented by actuator, tool carrier (holder), diaphragm support and tool bit as a spring buffer between tool carrier and the axially actuated cutting force at tool tip (radial to the work piece). The second model in Fig. 1 (b) is added to it the dynamic absorber as a disk supported by a diaphragm. In this work 8-node isoparametric solid element is used for domain discretization. The FEM model is tested in terms of mesh refinement, and, the results compared to a similar verified analytical work. Maximum difference between calculated values throughout verifications is within 8%.

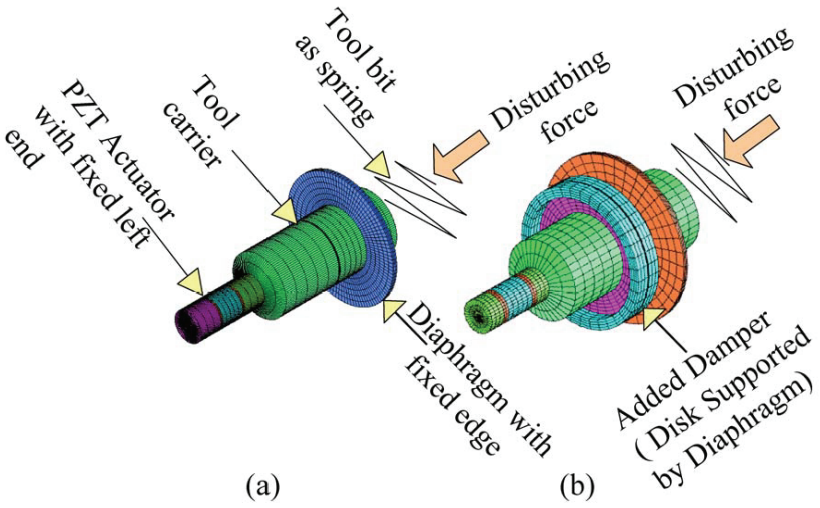


Figure 1. Toolpost Models

Conventional stacked PZT actuator incorporates polarized ferroelectric ceramic in the direction of actuation, adhesive, supporting structure, and electrodes wired electrically as shown in Fig. 2.

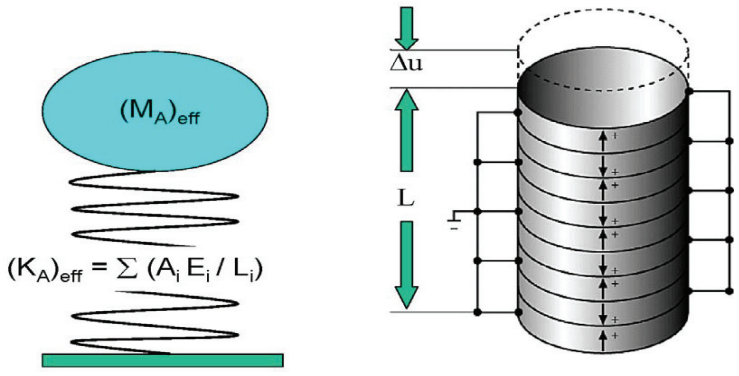


Figure 3. PZT Stacked Actuator

Modeling of active materials and toolpost are based on the general constitutive equations of linear piezoelectricity and the equations of mechanical and electrical balance (Piefort, V. 2001). The equations are thus expressed as

$$\begin{aligned} \{T\} &= [c^E]\{S\} - [e]^T\{E\} \\ \{D\} &= [e]\{S\} + [\varepsilon^S]\{E\} \end{aligned} \tag{1}$$

The momentum balance equation is

$$\rho\{\ddot{u}\} = \nabla.\{T\} \tag{2}$$

Moreover, the electric balance equation is

$$\nabla.\{D\} = 0 \tag{3}$$

Known $\{S\} = \nabla^S.\{u\}, \qquad \{E\} = -\nabla\phi$

Where $\{T\}$ represents the stress vector, $\{S\}$, the strain vector, $\{E\}$, the electric field, $\{D\}$, the electric displacement, $[c^E]$, the elastic coefficients at constant $\{E\}$, $[\varepsilon^S]$, the dielectric coefficients at constant $\{S\}$, and $[e]$, the piezoelectric coupling coefficients. $\{u\}$ is the mechanical displacement vector and $\{\ddot{u}\} = \partial^2\{u\}/\partial t^2$ is the acceleration. ϕ is the electric potential (voltage). The boundary conditions are expressed in Fig. 1, where zero displacements are assigned to actuator left end and, fixed outer edge for supporting diaphragm. Problem description is finalized by assigning voltage to actuator electrodes and applying force at tool tip.

The unknowns are the displacements vector u_i and the electric potential values ϕ_i at node i. The displacement and voltage fields at arbitrary locations within elements are determined by a linear combination of polynomial interpolation or shape functions N_u and N_ϕ respectively. The nodal values of these fields are used as coefficients. The displacement field $\{u\}$ and the electric potential ϕ over an element are related to the corresponding node values $\{u_i\}$ and $\{\phi_i\}$ by the mean of the shape functions$[N_u]$, and $[N_\phi]$

$$\begin{aligned} \{u\} &= [N_u]\{u_i\} \\ \phi &= [N_\phi]\{\phi_i\} \end{aligned} \tag{4}$$

The dynamic equations of a piezoelectric continuum derived from the Hamilton principle, in which the Lagrangian and the virtual work are properly adapted to include the electrical contributions as well as the mechanical ones (Piefort, 2001 et al., 1990). Taking into account the constitutive Eqs. (1) and substituting the LaGrangian and virtual work into Hamilton's principle to yields variational equation that satisfy any arbitrary deviation of the displacements $\{u_i\}$ and electrical potentials $\{\phi_i\}$ compatible with the essential boundary conditions, and then incorporate Eq. (4) to obtain

$$\begin{aligned} [m_{uu}]\{\ddot{u}_i\} + [c_{uu}]\{\dot{u}_i\} + [k_{uu}]\{u_i\} + [k_{u\phi}]\{\phi_i\} &= \{f_i\} \\ [k_{u\phi}]^T\{u_i\} + [k_{\phi\phi}]\{\phi_i\} &= \{q_i\} \end{aligned} \tag{5}$$

$[m_{uu}]$, $[k_{uu}]$ and, $[c_{uu}]$ are the mechanical mass, stiffness and damping matrices, respectively. $[k_{u\phi}]$ is the piezoelectric coupling matrix. $[k_{\phi\phi}]$ is the dielectric stiffness matrix. $\{f_i\}$ and $\{q_i\}$ are the nodal mechanical force and electric charge vectors, respectively. $\{u_i\}$ and, $\{\phi_i\}$ are the nodal displacement and potential vectors, respectively. For the sake of brevity, (Zienkiewicz & Taylor, 2000) discuss the scheme by which the elemental contributions are assembled to form the global system matrices.

3. Lumped Versus FEM Modeling

Lumped mass modeling for PZT actuator and tool carrier produce simple closed form solutions that are of interest to the designer and modeler (Frankpitt, 1995 and, Piefort, 2001). However, model validity of such representation for different design applications deserves more attention. In some applications, smart materials are used simultaneously in sensing and actuation. Displacement sensing at different locations is dependent on system dynamic, design geometry and system rigidity. Controller effectiveness relies on a valid dynamic system representation and the limits of legitimacy of such model.
A comparative result for a deviation in natural frequency of lumped mass versus continuous system is discussed for a single actuator as a first step toward an integrated tool post.

3.1 Comparative Results for Actuator Modeling

Before solving the time-dependent equation of motion for the smart toolpost, the mode shapes and the resonant frequencies of undamped system are obtained by using Eigenvalue analysis. The Eigenvalue problem is carried using a reduced matrix system obtained by matrix condensation of structural and potential degrees of freedom. Free vibration implies

$$\{[K^*] - \omega^2[m_{uu}]\}\{U_i\} \tag{6}$$

Where ω is the natural frequency, the new stiffness matrix $[K^*]$ indicates that structure is electromechanically stiffened. The modal analysis is based on the orthogonality of natural modes and expansion theorem (Zienkiewicz, and Taylor, 2000 a & b). Usually the actuator is composed off several PZT layers, electrodes, adhesive, and supporting structure as shown in Fig. 2. The effective stiffness of the actuator (***STIFA***) is the stiffness summation of all individual layers neglecting all piezoelectric effects.

$$(K_A)_{eff} = STIFA = \sum(\frac{A_i E_i}{L_i}) \tag{7}$$

For comparison the effective actuator mass assumed to be 20 or 30% of the layers masses as indicated in Fig. 3.

$$(M_A)_{eff} = (0.2 \quad or \quad 0.3)\sum(A_i \rho_i L_i) \tag{8}$$

Then

$$\omega_{Lumped} = \sqrt{\frac{(K_A)_{eff}}{(M_A)_{eff}}} \tag{9}$$

The FEM solution of the first natural frequency for short circuit and open circuit actuator are compared to the lumped mass frequency as obtained from Eq. (9) and the ratio is plotted in Fig. 3.

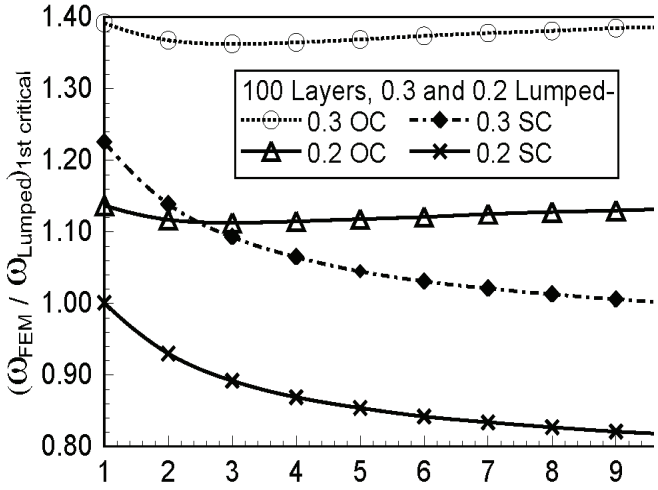


Figure 3. First critical frequency ratio /FEM/Lumped) versus layers thickness ratios for short circuit (SC) and open circuit (OC).

PZT8 properties from (Berlincourt, & Krueger, 2000) are used in FEM calculations. Plotted results in Fig. 3 are not incorporating stiffness variation resulted from actuator fabrication. Short circuit actuator shows a decrease in natural frequency, which indicates actuator stiffness reduction. Actuator short and open circuit conditions maps the two stiffness extremes and such data provide designers quick tool for estimating natural frequencies in early stages of design.

3.2 Comparative Results for Toolpost Model Incorporating Dynamic Absorber

In lumped modeling shown in Fig. 4 the tool carrier is considered as a rigid mass added to it one third of the PZT actuator mass and assigned (M_T). The dynamic absorber is the second mass (M_d) of the two-degree of freedom system and compared to the FEM solution to investigate lumped model validity of such system. A close form solution is obtained for the two-degree of freedom system incorporating the piezoelectric coupling effects (Frankpitt, 1995, and, Abboud, Wojcik, Vaughan, Mould, Powell, & Nikodym, 1998). Nevertheless, there solution does not answer the significant deviation between FEM and

lumped mass solutions in the case of no pizo effects. The supporting diaphragm stiffness (K_D) is calculated as a plate with central hole fixed at both inner and outer edges (Roark, and Young, 1975) then, added to actuator stiffness to form a cushion for tool carrier.

The actuator stiffness (K_A) is calculated as in Fig. 2. Then the dynamic absorber diaphragm stiffness for dynamic absorber (K_d) is considered as a plate with central hole fixed at both inner and outer edges

From Fig. 4 the equations of lumped mass and stiffness matrices for a two-degree of freedom system is:

$$[M]=\begin{bmatrix} M_T & 0 \\ 0 & M_d \end{bmatrix} \qquad (10)$$

$$[K]=\begin{bmatrix} K_A+K_D+K_d & -K_d \\ -K_d & K_d \end{bmatrix}$$

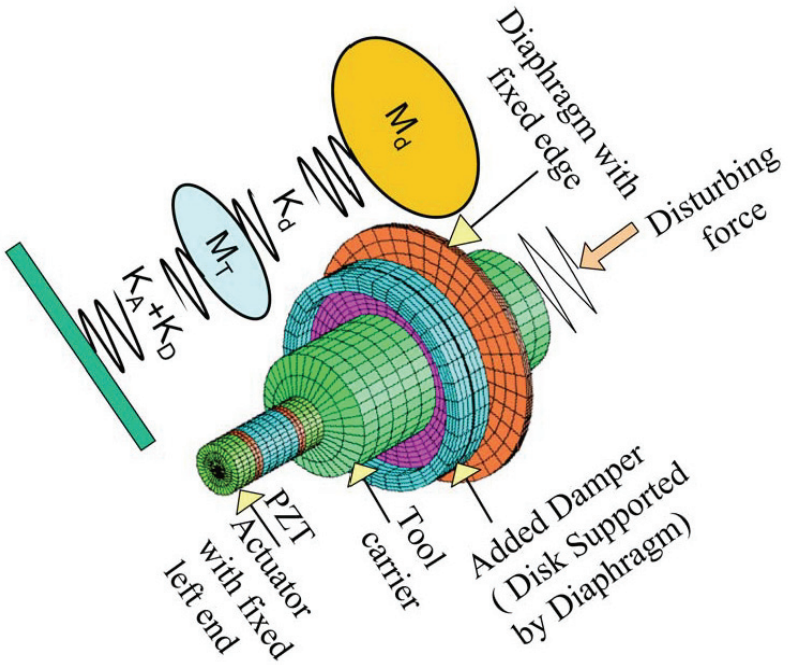


Figure 4. Tool post with dynamic absorber

Then the dynamic equation of motion and its characteristic equations for undamped free vibration can be derived as

$$[M]\{\ddot{u}\}+[K]\{u\}=\{0\}$$
$$-\omega_i^2[M]+[K]=0 \qquad (11)$$

Two natural frequencies are calculated from Eq. (11). Then lumped model frequencies (ω_{Lumped}) compared with the first three natural frequencies of the FEM model (ω_{FEM}) taking into consideration the mode shape and the Eigenvalue results. Three frequency ratios are compared namely $(\omega_{FEM})_{1st}/(\omega_{Lumped})_{1st}$ for 1st critical, $(\omega_{FEM})_{2nd}/(\omega_{Lumped})_{2nd}$ for 2nd critical, and $(\omega_{FEM})_{3rd}/(\omega_{Lumped})_{2nd}$ for 3rd critical.

Figure 5 show such variation of frequency ratios on log-log plot against the ratio of diaphragm support stiffness to actuator stiffness for a unit ratio between tool carriers to actuator stiffness (K_T/K_A). In general, the FEM model predicts lower natural frequencies for the toolpost and this deviation increases with the increase in the ratio of diaphragm support to actuator stiffness (K_D/K_A).

Increasing the ratio of tool carrier to actuator stiffness (K_T/K_A) ten times as in Fig. 5 yields a closer FEM solution to the lumped model at low diaphragm support to actuator stiffness ratio as shown in Fig. 6. However, the deviation again increases with the increase in diaphragm support to actuator stiffness ratio (K_D/K_A).

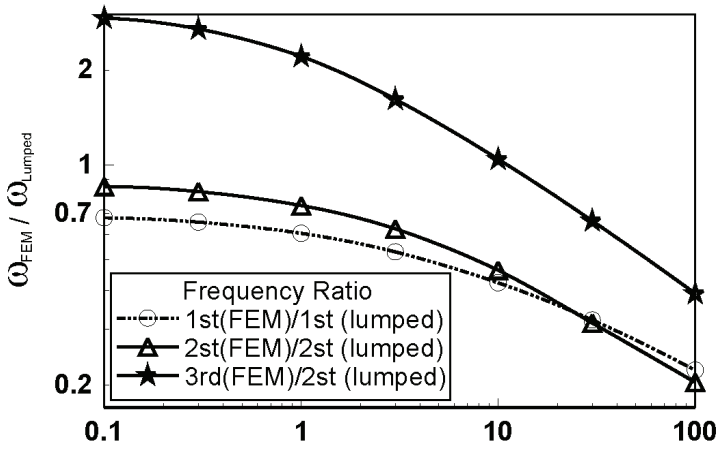


Figure 5. Frequency ratio of REM to lumped masses against diaphragm support to actuator stiffness (K_T/K_A=1.0, open circuit)

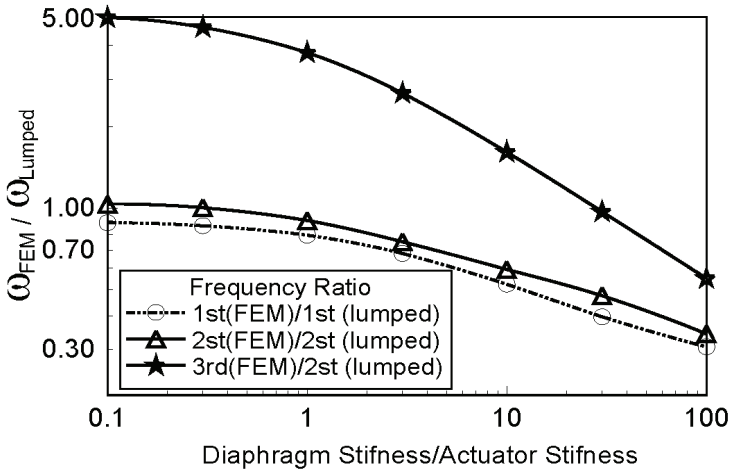


Figure 6. Frequency ratio of FEM to lumped masses against diaphragm support to actuator stiffness ratio (K_T/K_A = 10.0, open circuit)

Although the validity of lumped mass modeling can be defined in, a specific region but the broad requirement of design applications would limit the use of such narrow domain. As noticed, the critical frequencies are quite dependent on stiffness ratio and the FEM third critical can be the same as 2nd critical frequency of lumped mass modeling at high diaphragm stiffness ratio.

4. Results of Estimated Static Force Availability for Error Elimination

Elimination of error in tool positioning under static condition relies on PZT actuator capability in resisting axial tool force within the range of motion. To have initial guessing for the generated force a displacement curve is developed for the investigated PZT toolpost under static condition. Figure 7 shows such force-displacement characteristics at different levels of voltage intensity and for specified values of tool tip to actuator stiffness ratio (TIP-Ratio), diaphragm to actuator stiffness ratio (D-Ratio), and, tool carrier to actuator stiffness ratio (T-Ratio).

Calculations conducted in this work proved the importance of increasing tool tip to actuator stiffness, tool carrier to actuator stiffness and, reducing diaphragm to actuator stiffness ratios for a better utilization of actuator operating range. Figuring out an appropriate actuator for specific application is by relating the cutting force value to the information given in Fig. 7. However, such

information does not predict the required dynamic actuator voltage during service. Smart material data, toolpost dimensions and, actuator layers thicknesses are given in Table 1 for both static and transient force-displacements calculations.

Item	Value	Unit
Cylindrical PZT-8 Stack		
PZT Thickness	0.09e-	m
Electrode Thickness	0.03e-	m
Structural support	0.03e-	m
Adhesive Thickness	10.0e-	m
Number of layers	500	
Effective Radius	5.0e-3	m
Steel Cylindrical Tool Carrier (holder)		
Radius	10.0e-3	m
Length	65.35e-	m
Steel Tool Bit Effective Length		
Assumed Effective	20.0e-3	m
Steel Diaphragm		
Thickness	0.5e-3	m
Outside Radius	20.0e-3	m

Table 1 Toolpost dimension

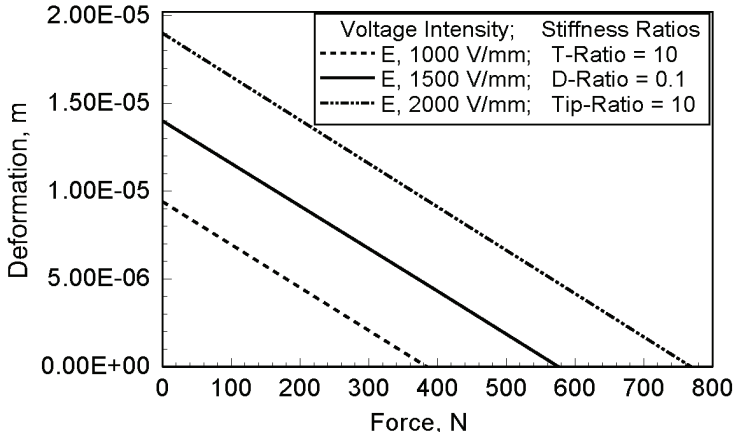


Figure 7. Tool load versus deformation for different PZT voltage intensity and fixed structural stiffnesses

5. Toolpost Time Response Due to Combined Effect of Voltage and Force Activation:

Evaluation of switching effects and system damping on toolpost response during error elimination are quantified by solving Eq. (5) in time domain for the system shown in Fig. 1. The PZT stack pattern is given in Table 1 that incorporates PZT layers, supporting structure, and electrodes for alternating poling direction. A thin layer of glue bonds wafers to one another. Because of this arrangement, the mechanical properties act in series. To reduce computational time the PZT stack is treated as a monolithic layer and precautions are taken accordingly for electric field intensity and other factors for multi-layer.

5.1 Voltage Switching Methodology

Deviation in position between tool tip and workpiece can be minimized by appropriate voltage activation to the PZT actuator. The easy way of activating smart material for vibration suppression is by using Pulse Width Modulation (PWM). It is a common technique available with the microcontroller units (MCU) to govern the time average of power input to actuators. Our main concern is the time dependent response accompanying the tool error suppression in using the PWM for smart material actuator. Voltage activation for smart material might either based on a piezo stack with force sensing layer or using an appropriate type of displacement sensor to detect tool carrier motion. In both methods sensing location should reflect cutting tool position error correctly. Switching circuits (Luan, and Lee, 1998) are not of our concern; however, the required voltage level and the resulted motion are among the targeted results in this work.

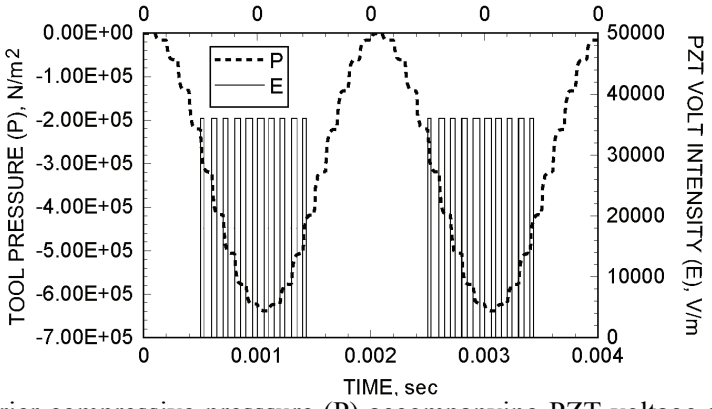


Figure 8. Tool carrier compressive presssure (P) accompanying PZT voltage activation intensity (E) plotted on a common time axis.

Figure 8 shows two cycles of voltage activation for the PZT actuator using PWM to oppose the compressive time dependent cutting force. The waveform of the compressive cutting force is used as a reference for the PWM voltage with a chance to incorporate the time delay. All present work results are assuming harmonic force actuation.

5.2 Solution Scheme for the Toolpost Time Response

The classical Newmark algorithm (Zienkiewicz, and Taylor, 2000b) solves the system of equations for such a nonlinear problem. Time step-by-step integration is used for solving Eq. (5) for the system shown in Fig. 1. This scheme assumes the system-damping matrix as linear combination of stiffness and mass matrices (Rayleigh damping) (Bathe, 1982):

$$[c_{uu}] = \alpha[m_{uu}] + \beta[k_{uu}] \tag{12}$$

Both α and β are constants to be determined from two proposed modal damping ratios (ξ_i) (1% and 5%) for first and second natural frequencies respectively which are obtained from the FEM model and the equation of modal damping as given in (Bathe, K.J. 1982).

5.3 Results for the Tool Time Transient Response

Synchronization of voltage activation with tool radial force can be reached either through a sensing layer in actuator stack or by using a displacement sensor for detecting tool carrier movements. The effective use of any of these techniques requires a profound investigation for toolpost dynamic behavior as related to its structural stiffness properties.

Tool dynamic and structural design for a reconfigurable machine tool (Gopalakrishnan, Fedewa, Mehrabi, Kota, & Orlandea, 2002, and, Moon & Kota, 2002] elevated new design challenges. Among them are methods for reducing tool holder size or developing a special tactics in using smart actuators for reaching targeted precision.

Tool cutting force predictions in dynamic calculations involve some difficulties due to the number of involved variables and the dynamic nature of the problem. In general approximate static force relation (Frankpitt, 1995) in terms of

depth of cut d (mm), cutting speed V (mm/s), feed f (mm/rev), and, coefficients describing nonlinear relationships ($\kappa, \lambda, and, \gamma$) can be used as first guess to express the general trends,

$$F_r(N) = K_r d^{\lambda} V^{\gamma} f^{\kappa}(t); \quad Where \quad K_r \text{ is a general constant.}$$

$$F_r = K_r d^{\lambda} V^{\gamma} f^{\kappa}(t) \qquad K_r \quad \text{a general constant} \tag{13}$$

The factors K_r, λ, γ and, κ are to be calibrated for each tool-workpiece. These constants are assigned to a specific material combinations, process types, tool-wear condition, workpiece hardness, tool geometry, and speed. Fluctuation of the cutting force is inherent and associated with cutting tool motion. Such randomness can vary with different cutting processes and material combinations.
For present results, toolpost dimension and, material are given in Table 1. Previous work (Rashid, M. K. 2004) indicated the use of few PWM, cycles per force period produced unfavorable switching dynamic excitation. Twenty PWM cycles for each force period produce good results more than forty has a little effect. In all calculations a value of ten is assigned to tool bit to actuator stiffness ratio (TIP-Ratio) and tool carrier to actuator stiffness ratio (T-Ratio). On the contrary, the diaphragm to actuator stiffness ratio (D-Ratio) assigned a low value of one tenth. The importances of such ratios are related to the force availability for error elimination and accurate displacement detection.
Figure 5 shows a tiny difference between resonant frequencies obtained from both FEM and lumped model solutions in case of existence of low diaphragm to actuator stiffness (D-Ratio) and high tool bit to actuator stiffness (TIP-Ratio) ratios. Under such conditions incorporating a classical dynamic absorber to a toolpost excited by harmonic inputs should attenuate vibration error. Our main concern is the effectiveness of such dynamic absorber for activated actuator by a PWM voltage instead of a continuous harmonic input voltage as the case in this work.

From classical dynamic absorber theory and for optimum damping, the applied force frequency must be tuned to absorber natural frequency. Also a mass ratio of 0.25 must be secured between dynamic absorber and tool carrier. Then the natural frequency ratio of absorber to tool carrier based on classical dynamic absorber under pure harmonic inputs and optimum-damping condition is obtained from Eq. (14). This natural frequency ratio is enforced to the FEM model by adjusting damper diaphragm stiffness in Fig. 4. Damper effec-

tiveness on error elimination is then compared to other toolpost design parameters under the condition of PWM voltage activation as shown in Figs. 9-13. Graph legends terminology of Figs. 9-13 are given in table 2.

$$\text{Natural frequency ratio of absorber to tool carrier} = \frac{1}{1+(M_d/M_T)} \tag{14}$$

No-A	No dynamic absorber
Y-A	Yes absorber is incorporated
Low-D	Low Damping
Hi-D	High Damping (10 x Low-D)
M-Sw	Modified mean voltage during Switching
Un-Sw	Un-modified mean voltage during Switching
No-volt	No voltage applied to actuator
Y-volt	Yes voltage applied to actuator

Table 2.

Figure 9 shows a significant error reduction can be attained by modifying the mean voltage of the PWM during the force actuation period. A single scheme is used for conducting voltage modification based on harmonic sine wave of the tool actuation force and described by the following set of equations:

If $|\sin \omega t| < 0.2$ then multiply present mean voltage by four,
If $|\sin \omega t| > 0.2$ and $\sin \omega t < 0.6$ do not change the mean voltage,
If $|\sin \omega t| > 0.6$ then multiply present mean voltage by (0.65).

Applying smart material actuator with unmodified mean voltage might deteriorate the error elimination process as shown in Figs. 9-13. Utilizing smart material for tool error elimination require assurance for both force sensing direction and proper voltage modification to reach the targeted beneficial results. Dynamic absorber effectiveness in error elimination is frequency dependent. Absorber presence in Figs. 9, 11 and 13 aggravated the error elimination improvement made by voltage modification. In all of these results, the dynamic

absorber natural frequency is tuned according to Eq. (14). Figures 9 and 11 are plotted for 2-cycles to improve comparison among error results. In Fig. 10, a small improvement is resulted due to the dynamic absorber presence but it is not a solid case to measure on. Figure 13 demonstrate a counteracting effect for the dynamic absorber even with existence of the applied modified voltage to the smart material actuator. The use of high damping (Hi-D) with ten folds the low damping (Low-D) does not have same effectiveness of using smart material actuator with properly modified mean voltage during the PWM. Conducted calculations demonstrated no significant effects for the time delay between applied voltage and activation force if the delay controlled to be within 10% of the force period.

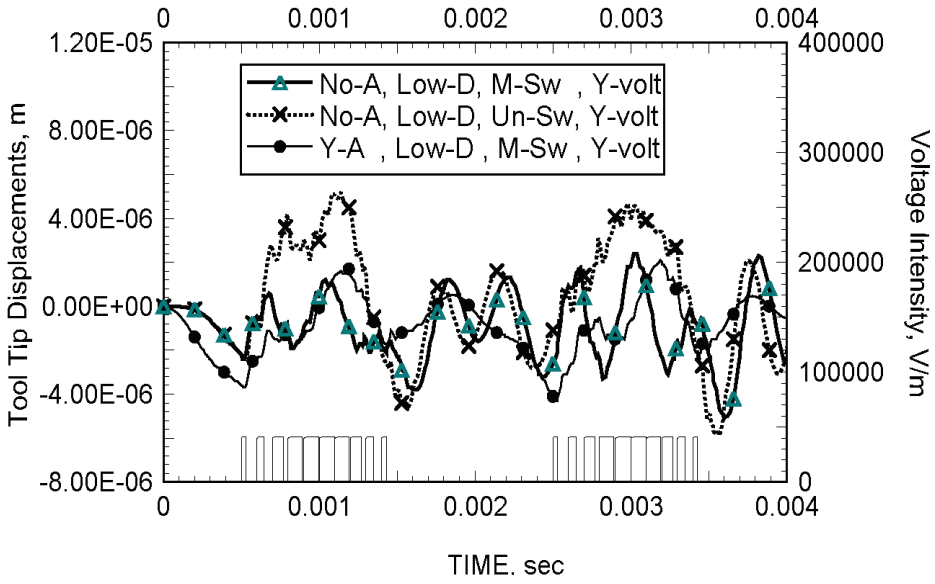


Figure 9. PZT Voltage Intensity and Tool tip displacements Versus time at 500 Hz

The estimated radial cutting force value from Eq. (13) and the static force-displacement relationship shown in Fig. 7 are important in initial guessing for the required applied voltage. But the final magnitude of dynamic applied voltage is deduced from the associated error resulted from the modification methodology for the mean voltage during PWM.

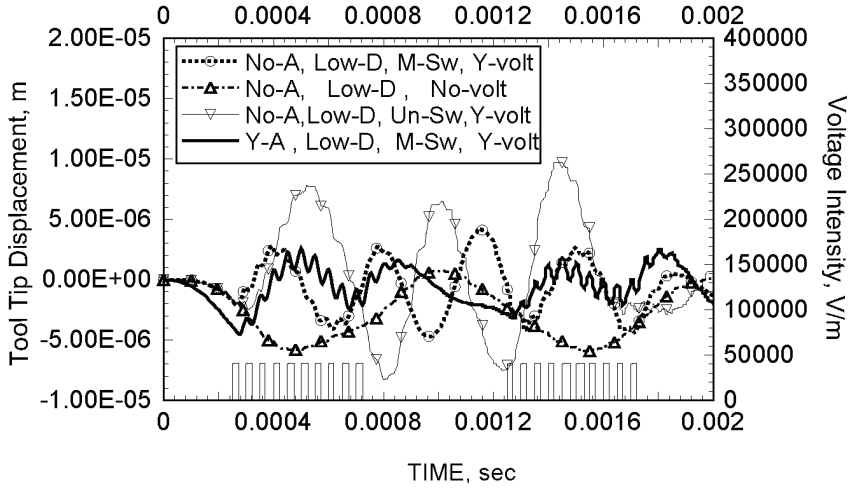


Figure 10. PZT Voltage Intensity and Tool tip displacements Versus time at 1000 Hz

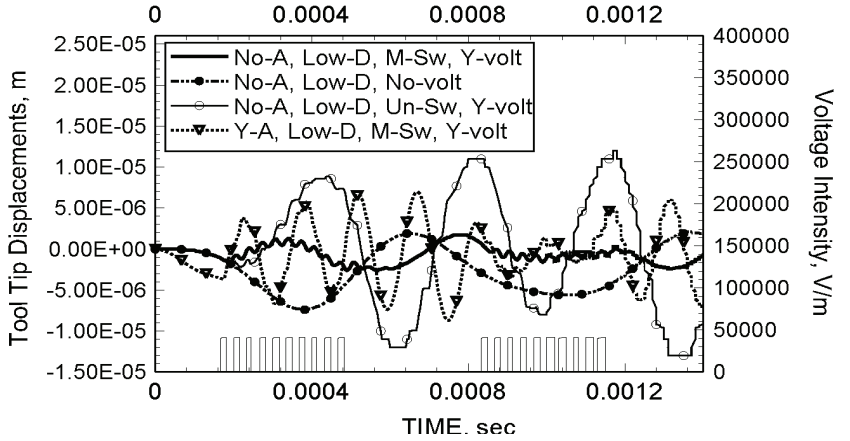


Figure 11. PZT Voltage Intensity and Tool tip displacements Versus time at 1500 Hz

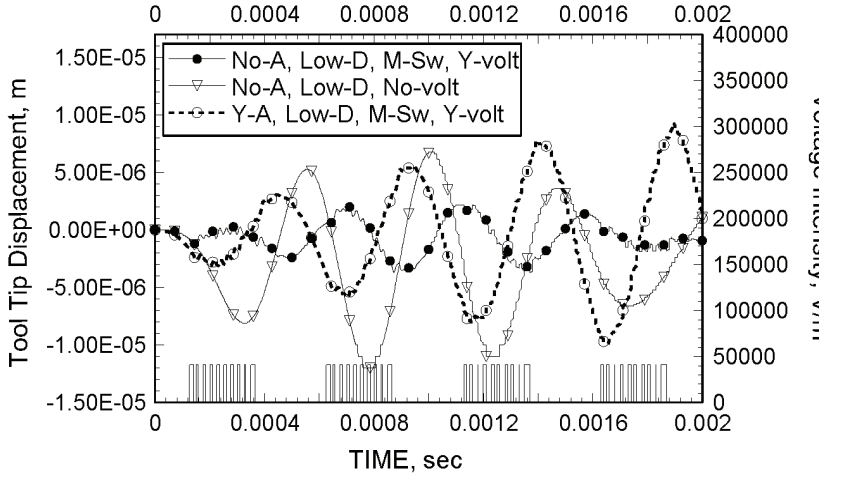


Figure 12. PZT Voltage Intensity and Tool tip displacements Versus time at 2000Hz

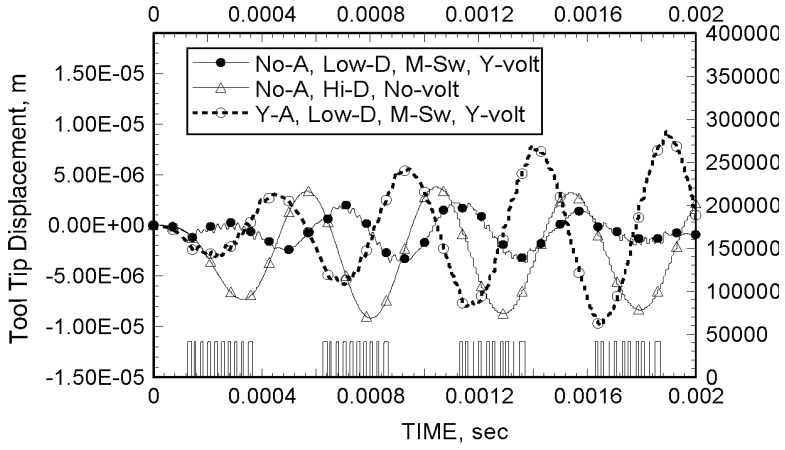


Figure 13. PZT Voltage Intensity and Tool tip displacements Versus time at 2000Hz

6. Conclusions

Attenuating tool vibration error in old turning machines can reduce industrial waste, save money and, improve design flexibility for new cutting tools. Using smart materials in curing machine tool vibration require special attention. The modification of the applied mean voltage during PWM plays a major rule in the effective use of smart materials in tool error elimination. The use of the dynamic absorber showed a slight error reduction in some cases and was not effective in the others. Increasing damping does not show a significant error variation in comparison to the use of smart actuator with modified mean voltage. The FEM solution provided the valid range for the lumped mass modeling to improve both dynamic system modeling and controller design. Tool bit and tool carrier (holder) to actuator stiffness are preferred to be high when both space and weight limitations does not exist. Error elimination requires at least twenty PWM cycles for each disturbing force period to reduce switching transient effects. A reasonable time delay of less than 10% between displacement sensing and actuation has no significance on error elimination. There is a significant difference between the dynamic and the static prediction of the required actuator voltage for error elimination.

7. References

Abboud, N. N., Wojcik, G. L., Vaughan, D. K., Mould, J., Powell, D. J., and Nikodym, L. (1998), Finite Element Modeling for Ultrasonic Transonic Transducers, *Proceedings SPIE Int. Symposium on Medical Imaging 1998*, San Diego, Feb 21-27: 1-24.

Allik, H., and Hughes, T. J. R. (1970), Finite element method for piezoelectric vibration, International Journal for Numerical Methods in Engineering, 2: 151–157.

Bathe, K.J. (1982) ,*Finite Element Procedures in Engineering Analysis*,, Prentice-Hall Inc.: 511-537.

Berlincourt, D. and, Krueger, H. A., (2000), Properties of Morgan ElectroCeramic Ceramics, Technical Publication TP-226, Morgan Electro Ceramics.

Dold, G. (1996), Design of a Microprocessor-Based Adaptive Control System for Active Vibration Compensation Using PMN Actuators, MS Thesis, University of Maryland at College Park.

Eshete, Z. (1996), In Process Machine Tool Vibration Cancellation Using Electrostrictive Actuators, Ph.D. Thesis, University of Maryland at College Park.

Frankpitt, B.A. (1995), A Model of the Dynamics of a Lathe Toolpost that Incorporates Active Vibration Suppression, Institute for System Research, University of Maryland at College Park.

Gopalakrishnan, V., Fedewa, D., Mehrabi, M.G., Kota, S., and Orlandea, N. (2002), Design of Reconfigurable Machine Tools, *ASME J. Manuf. Sci. Eng*., Technical Briefs, 124, Nov.: 483-485.

Hurtado, J. F., and Melkote, S. N, (2001), Improved Algorithm for Tolerance-Based Stiffness Optimization of Machining Fixtures, *ASME J. Manuf. Sci. Eng*., 123, Nov.: 720-730.

Lerch, R., (1990), Simulation of piezoelectric devices by two- and three-dimensional finite elements, *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 37(3): 233–247.

Luan, J. and Lee, F. C. (1998) ,Design of a High Frequency Switching Amplifier for Smart Material Actuators with Improved Current Mode Control,. *PESC* 98, Vol. 1: 59-64.

Merritt, H. E., (1965), Theory of self-Excited Machine-Tool Chatter, *Journal of Engineering for Industry*, November, pp. 447-454.

Moon, Y. and, Kota, S. (2002), Design of Reconfigurable Machine Tools, *ASME J. Manuf. Sci. Eng.*, Technical Briefs, 124, Nov.: 480-483.

Piefort, V. (2001), Finite Element Modeling of Piezoelectric Active Structures, Ph.D. Thesis, ULB, Active Structures Laboratory- Department of Mechanical Engineering and Robotics.

Rashid, M. K. (2004), Smart Actuator Stiffness and Switching Frequency in Vibration Suppression of a Cutting Tool, *Smart Materials and Structures*, Vol.13: 1-9.

Roark, R. J., and Young, W. C. (1975), *Formulas for Stress and Strain (Fifth Edition)*, McGraw-Hill Book Company, New York, Table 24-1j: 337.

Satyanarayana, S., and, Melkote, S. N. (2004), Finite Element Modeling of Fixture–Workpiece Contacts: Single Contact Modeling and Experimental Verification, *International Journal of Machine Tools and Manufacture*. Volume 44, Issue 9, July : 903-913.

Tzou, H. S., and Tseng, C. I. (1990), Distributed piezoelectric sensor/actuator design for dynamic measurement/control of distributed parameter systems: a piezoelectric finite element approach, *Journal of Sound and Vibration*, 138(1):17–34.

Zhang, G., Ko, W., Luu, H., and Wang, X.J. (1995), Design of a smart Tool Post for Precision Machining, *Proceedings of the 27th CIRP International Seminar on Manufacturing Systems*, Ann Arbor, MI, May: 157-164.

Zienkiewicz, O. C., and Taylor, R. L. (2000a), *The Finite Element Method*, Fifth edition Vol.1: The Basis, Butterworth-Heinemann.

Zienkiewicz, O. C., and Taylor, R. L., (2000b), *The Finite Element Method*, Fifth edition Vol.2: Solid Mechanics, Butterworth-Heinemann:.423-424.

18

Concurrent Process Tolerancing Based on Manufacturing Cost And Quality Loss

M. F. Huang and Y. R. Zhong

1. Introduction

In manufacturing practice, actual dimensions are impossible as well as unnecessary to determine exact values. Under stable fabrication conditions, the processed dimensions often vary within certain controlled ranges. Tolerances are specified to control the actual dimensions of processed features within allowable variation zones for product functional requirements and manufacturing costs (Zhang, 1996; Ngoi and Teck, 1997; Lee and Tang, 2000; Fang and Wu, 2000; Huang et al., 2001; Huang and Gao, 2003; Chen et al., 2003).

The contemporary practice of tolerance design has two sequential phases: Product tolerance design and process tolerance design (Ngoi and Teck, 1997).

In product tolerance design, designers use their knowledge and expertise to determine the assembly critical tolerances by computation or design handbooks. These tolerances will then be allocated to component design tolerances (blueprint tolerances) in terms of component structures, assembly restrictions, and given design criteria. If a mathematical model is used, the objective function is usually to minimize manufacturing costs or to maximize weighted component tolerances. The constraints are often tolerance stack-up and economical tolerance ranges of each component part (Swift et al., 1999; Ngoi and Min, 1999; Ngoi and Ong, 1999; Huang and Gao, 2002). Swift *et al* (1999) presented a tolerance optimization model in assembly stacks based on capacity design. In their research, systematic analysis for estimating process capability levels at the design stage is used in conjunction with statistical methods for optimization of tolerances in assembly stacks. Ngoi and Min (1999) presented a new approach for optimum tolerance allocation in assembly. Their method allows all blueprint (BP) tolerances to be determined while ensuring that all as-

sembly requirements are satisfied. Ngoi and Ong (1999) presented a complete tolerance charting in the assembly phase. Their method integrates product tolerance design and process tolerance design. The objective is to maximize the summation of weighted process tolerances. Huang and Gao (2002) presented a discrete hierarchy optimal approach for allocating the optimum component tolerance based on estimated process capability. They minimize the total manufacturing cost by using a cost-tolerance function.

In process tolerance design, manufacturing engineers develop component process planning to determine manufacturing methods, machine tools, fixtures, cutting tools, cutting conditions, manufacturing routines, and process tolerances. At this stage, BP tolerances are the most important factors. If they are too tight and cannot guarantee the economic fabrication for components by using selected process planning, more precise machine tools, special fixtures, and expensive measurements should be introduced (Wu et al., 1998). This inevitably increases the manufacturing cost of the product. The manufacturing engineers may ask for revision of BP tolerances or of the process plan. In process tolerance design, the most popular methods are also the optimal design for minimum manufacturing cost or maximum process tolerances. Huang *et al.* (2002) presented an optimal planar tolerance design approach to allocate dimensional and orientation geometric tolerances. A special relevance graph (SRG) was used to represent the relationships between manufactured elements and their size and tolerance information. In addition, the SRG is also applied for the geometric dimensions and tolerances. A linear programming model was established to solve the problem. Huang and Gao (2003) presented a nonlinear programming model for optimal process tolerance balancing. A linear programming model to determine process dimensions and process tolerances was used in Ji (1993) and Ngoi and Teck (1993). Similar methods to determine optimum process tolerances were proposed by Wei and Lee (1995) and Chang *et al*., (2000).

Though the above methods have been used successfully to distribute both component design tolerances and process tolerances in two different phases, they over-emphasize manufacturing factors and seldom consider quality aspects. Systematically, product satisfaction conflicts with manufacturing cost. In other words, a better product satisfaction requires smaller tolerances and a higher manufacturing cost. Taguchi quality loss is a useful monetary specification to evaluate the quality factors (Taguchi et al., 1989; Taguchi, 1993; Jeang, 1998). Therefore the best policy is to consolidate manufacturing cost and quality loss in the same optimization objective to best balance quality satisfaction

and tolerances (Taguchi, 1993; Huang and Gao, 2002). Using this method, the research work has been carried out in product design and component process planning stages, respectively. Lee and Tang (Lee and Tang, 2000) presented an optimization model for controlling dimensional tolerances of components with multiple functional characteristics by minimizing the sum of manufacturing cost and quality loss. Jeang (1998) introduced a mathematical optimization model to integrate manufacturing cost and quality loss for tolerance charting balancing during machining process planning. Jeang (1997) also discussed a set of models to determine the optimal product tolerance and to minimize combined manufacturing and related costs.

Although tolerance assignment in the product design and process planning stages is often interdependent and interactive and affects overall production costs and product satisfaction, research into these areas is often conducted separately (Ngoi and Teck, 1997). There are some inherent shortcomings in this method. Firstly, in product tolerance design, designers are unable to allocate the real optimal BP tolerances to components because there is no manufacturing information available at this stage. Secondly, in process tolerance design, manufacturing engineers develop process planning in terms of the component information obtained from mechanical drawings, technical notes, and others such as title bars. They are less concerned with functional roles of components than with their manufacturing capabilities. This sequential tolerance design method would result in some problems in cooperation, continuity, and consistency between two separate design stages. Therefore, rework or redesign cannot be avoided.

Until recently, the concurrent tolerancing method has attracted the attention of some engineers (Zhang, 1996; Ngoi and Teck, 1997; Fang et al., 1998; Fang and Wu, 2000; Huang et al., 2001, Huang and Gao, 2003; Chen et al., 2003). Zhang (1996) first systematically presented mathematical methods for concurrent tolerancing and developed a general model of optimal tolerancing that supports concurrent engineering. Ngoi and Teck (1997) proposed a concurrent tolerancing method for product design in which the assembly tolerance can be allocated to the component design tolerance in an early stage of product design. Fang *et al.* (1998) proposed a concurrent tolerancing method to determine the optimum process tolerances with manufacturing cost and quality loss being considered simultaneously. But only a single assembly critical tolerance is related. Fang and Wu (2000) proposed a mathematical model to minimize the cost of sum machining. The constraints include assembly functional requirements, machining methods, stock remove tolerances, and economically attain-

able accuracies. Huang *et al.* (2001) proposed a special relative hierarchical hypergraph (SRHG) to represent the assembly. Through use of SRHG, assembly and process tolerance chains can be generated automatically. The method can allocate required assembly tolerances to process tolerances concurrently. Huang and Gao (2003) and Chen *et al.* (2003) proposed a concurrent method to allocate the optimal process tolerances in early product design stages. Here, a nonlinear optimization model is established to minimize the total manufacturing cost.

So far no design method has been presented to directly allocate multiple correlated critical tolerances to their process tolerances in a concurrent design environment. Therefore, the purpose of this paper is to introduce a concurrent optimal tolerancing method to realize this goal. To implement optimal robust tolerance design from product design stage to manufacturing stage, we first derive the quality loss function of multiple correlated critical tolerances in terms of manufacturing tolerances. A nonlinear optimization model is then given to minimize the summation of total component manufacturing cost and product quality loss. Finally the optimal processes are obtained by solving the model.

This chapter is divided into the following sections. Section 2 discusses the models for converting the geometrical tolerances with fixed tolerance zones into equivalent bilateral sized dimensions and tolerances. In section 3, we discuss the methods to present concurrent dimensional and geometrical tolerance chains. Section 4 further describes integrated concurrent dimensioning and dimensioning. In Section 5 we derive the quality loss of multiple correlated critical dimensions in terms of the process tolerances. In Section 6 we develop the optimal tolerance design model, whereas Section 7 examines the implementation for a specific example. The concluding remarks are given in Section 8.

2. Models for interpretation of geometrical tolerances

Geometric tolerances are usually expressed as graphical symbols, which can contain nominal sizes, tolerance values, and data (references). In order to deal with geometric tolerances in integrated tolerance charts, their geometrical characteristics must be addressed first. Generally geometric tolerances can be classified into five types: individual form, profile, orientation, location, and runout. There are fourteen geometric tolerances items altogether but only

those items with fixed tolerance zones will directly affect tolerance chains. Consequently only four geometrical tolerances in the total fourteen can be included in the integrated tolerance chains. These items — profile, position, symmetry, and concentricity — can be converted into the equivalent bilateral dimensional tolerances. The remaining items are treated as additional tolerance constraints (He & Gibson, 1992; Ngoi & Tan, 1995; Ngoi & Soew, 1996; Tseng & Kung, 1999).

2.1. Profile of a line (surface)

Profile of a line (surface) defines a permitted variation zone of a line (surface) relative to the corresponding theoretical geometry. It can be used to specify the geometrical requirements of an individual and a relevant feature in terms of different graphical notations in mechanical drawing. When profile of a line (surface) tolerance is used to denote an individual feature, then this item doesn't contribute to tolerance stack-up. Thus it can be treated as additional tolerance constraints. However, when profile of a line (surface) tolerance is used to specify a relevant feature, this item possesses a fixed tolerance zone. Thus it can be treated as equivalent bilateral dimensional tolerance. Figure 1 is the interpretation of the relevant profile of a surface. The relationship between profile of a line (surface) and their pertinent processed working dimensions and tolerances can be expressed as:

$$GL \pm TGL = \sum_{i=1}^{n} \xi_i WD_i \pm TWD_i \qquad (1)$$

Where GL and TGL is the nominal dimension the tolerance between the controlled line (surface) and the data (reference), respectively. WD_i and TWD_i is the ith working dimension and tolerance, respectively. ξ_i is the unit vector of WD_i, n is the total number of working dimensions and tolerances.

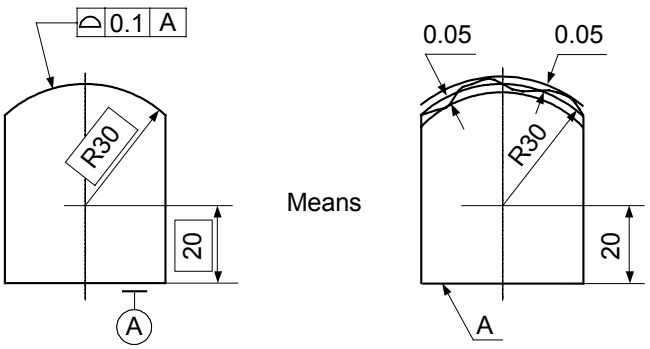


Figure 1. Interpretation of profile of a relevant surface

2.2 Position

Position tolerance defines the true position of a feature with respect to the references or the data. Because position tolerance holds a fixed tolerance zone with respect to the data, it can be transformed into equivalent bilateral dimensional tolerance. All the pertinent dimensions and tolerances in determining position of the controlled feature with respect to the data will be the link members of the position tolerance. Figure 2 is the interpretation of position tolerance. The transform model between position tolerance and their pertinent processed working dimensions and tolerances is:

$$GP \pm TGP = \sum_{i=1}^{n} \xi_i WD_i \pm TWD_i \qquad (2)$$

Where GP and TGP is the nominal dimension and position tolerance from the controlled feature to the data, respectively. WD_i and TWD_i is the ith working dimension and tolerance, respectively. ξ_i is the unit vector of WD_i, n is the total number of working dimensions and tolerances. In Figure 2 the position tolerance value is specified when the controlled hole is under the maximum material condition.

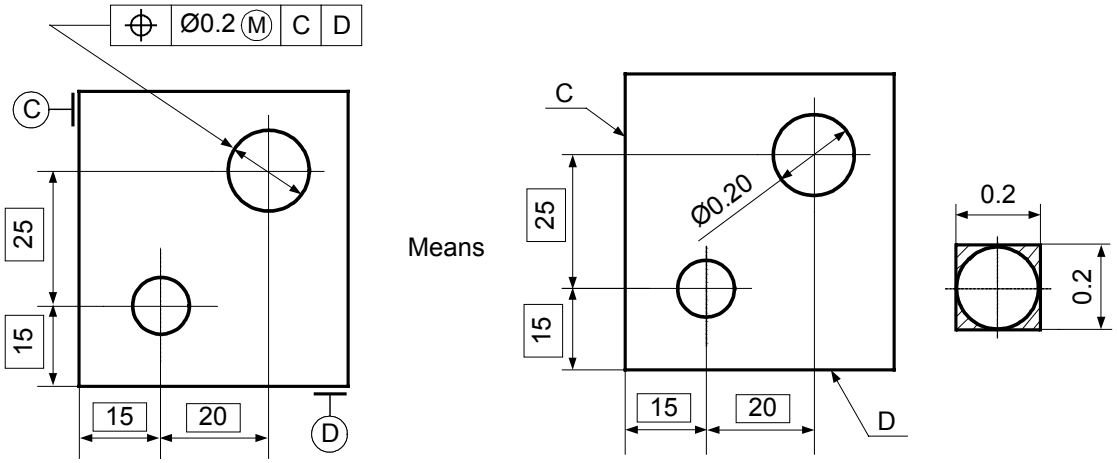


Figure 2. Interpretation of position

2.3 Concentricity

Concentricity tolerance expresses the requirement that the controlled axis should locate within the given allowable cylinder zone whose axis is the datum axis. Thus all the pertinent dimensions contribute to the dimension be-

tween the controlled axis and the datum axis will be the link members of this specification. Figure 3 shows a simple example for interpretation of concentricity into its equivalent bilateral dimensional tolerance. The model for interpretation of concentricity is:

$$GA \pm TGA = \sum_{i=1}^{n} \xi_i WD_i \pm TWD_i \tag{3}$$

Where *GA* and TGA is the nominal dimension and concentricity tolerance between the controlled axis and the datum axis, respectively. Generally this dimension is zero. WD_i and TWD_i is the working dimension and tolerance for the *i*th link member of *GA*, respectively. ξ_i is the unit vector of WD_i. n is the number of link members.

2.4 Symmetry

Symmetry tolerance presents the requirement that the controlled centre relevant feature such as the centre line of a hole, or the centre plane of a slot should locate within the given zone with respect to the datum. So all the related dimensions contribute to the dimension for determining the location of the controlled feature with respect to the datum will be the link member of this specification. Figure 4 gives a simple example for interpretation of symmetry into its equivalent dimensional tolerance specification. The model for interpretation of symmetry is:

$$GB \pm TG\mathrm{B} = \sum_{i=1}^{n} \xi_i WD_i \pm TWD_i \tag{4}$$

Where *GB* and *TGB* is the nominal dimension and symmetry tolerance between the controlled center features with respect to the datum, respectively. Generally, this dimension is zero. WD_i and TWD_i is the working dimension for the *i*th link member of *GB*, respectively. ξ_i is the unit vector of WD_i. n is the number of link members.

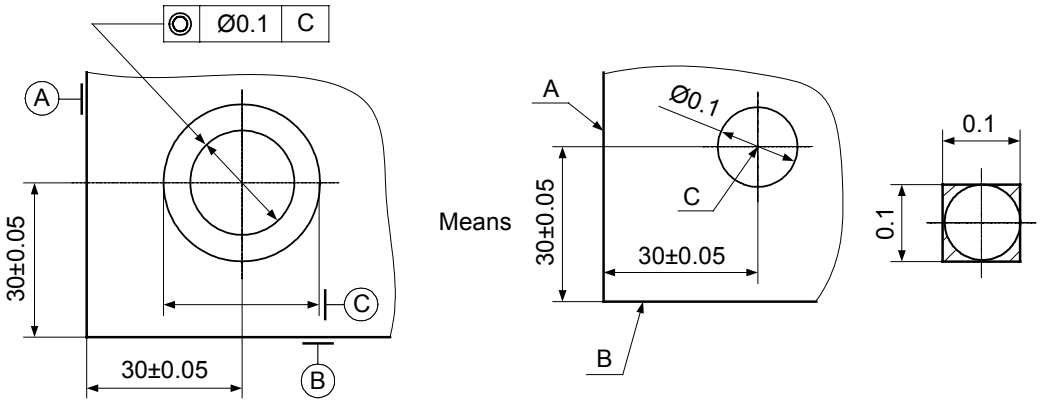


Figure 3. Interpretation of concentricity

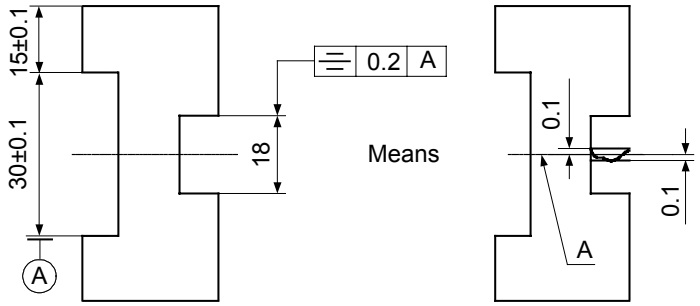


Figure 4. Interpretation of symmetry

3. Concurrent dimensional and geometric tolerance chains

In a concurrent tolerancing environment one of the most important issues is presentation of the concurrent integrated dimensional and geometric tolerance (DGT) chains. In a conventional system, tolerance design is being executed in two separate sequential stages: BP tolerance design and process tolerance design. Unlike the methods presented by several researchers (Ngoi & Tan, 1995; Zhang, 1996; Huang et al., 2001; Huang and Gao, 2002; Gao and Huang, 2003; Chen et al., 2003), this paper presents a general methodology for concurrent allocation of the required assembly functional DGTs to the component process ones.

In the stage of product design, let all the required assembly functional dimen-

sions and tolerances be the set $S_{AD} = \{L_{ADi} \pm T_{ADi}/2,\ i = 1, \ldots, n\}$, where n is the number of functional dimensions and tolerances, L_{ADi} is the ith assembly functional dimension, T_{ADi} is the tolerance of L_{ADi}. Also all the assembly functional geometric tolerances which can be modeled as equivalent dimensions and tolerances be the set $S_{AG} = \{L_{AGi} \pm T_{AGi}/2,\ i = 1, \ldots, m\}$, where m is the number of functional geometric tolerances which can be treated as equivalent bilateral dimensional tolerances, L_{AGi} is the ith equivalent assembly functional dimension, T_{AGi} is the geometric tolerance of L_{AGi}. And all the assembly functional geometric tolerances which can be modeled as additional tolerance constraints be the set $S'_{AG} = \{L_{Agi}(T_{AGi}),\ i = m+1, \ldots, m+\beta\}$, where β is the number of functional geometric tolerances which can be treated as additional tolerance constraints, T_{AGi} is the geometric tolerance treated as additional tolerance constraint, $L_{AGi}(T_{AGi})$ is the ith equivalent assembly functional dimension.

For simplicity the set notation is introduced as $S_{AF} = \{S_{AD}, S_{AG}, S'_{AG}\} = \{L_{AFi} \pm T_{AFi}/2,\ i = 1, \ldots, n+m,\ L_{AGi}(T_{AGi}),\ i = n+m+1, \ldots, n+m+\beta\}$. Where $\{L_{AFi} \pm T_{AFi}/2,\ i = 1, \ldots, n\}$ corresponds to $S_{AG} = \{L_{ADi} \pm T_{ADi}/2,\ i = 1, \ldots, n\}$, $\{L_{AFi} \pm T_{AFi}/2,\ i = n+1, \ldots, n+m\}$ corresponds to $S_{AG} = \{L_{AGi} \pm T_{AGi}/2,\ i = 1, \ldots, m\}$, and $\{L_{AFi}(T_{AFi}),\ i = n+m+1, \ldots, n+m+\beta\}$ corresponds to $S'_{AG} = \{L_{AGi}(T_{AGi}),\ i = m+1, \ldots, m+\beta\}$.

In a given assembly assume that all the component functional dimensions and tolerances be the set $S_{CD} = \{L_{CDj} \pm T_{CDj}/2,\ j = 1, \ldots, r\}$, where r is the number of functional dimensions and tolerances of all the components, L_{CDj} is the jth component functional dimension, T_{CDj} is the tolerance of L_{CDj}. And all the component functional geometric tolerances which can be converted into the equivalent bilateral dimensional tolerances be the set $S_{CG} = \{L_{CGj} \pm T_{CGj}/2,\ j = 1, \ldots, p\}$, where p is the number of functional geometric tolerances, which can be treated as the equivalent bilateral dimensional tolerances of the components, T_{CGj} is the jth component functional geometric tolerance, L_{CGj} is the nominal dimension of T_{CGj}. Also the functional component geometric tolerances which can be treated as the additional tolerance constraints be the set $S'_{CG} = \{L_{CGj}(T_{CGj}),\ j = p+1, \ldots, p+\delta\}$, where δ is the number of the functional geometric tolerances which can be treated as the additional tolerance constraints of the components, T_{CGj} is the jth component functional geometric tolerances which is treated as the additional tolerance constraint, $L_{CGj}(T_{CGj})$ is the nominal dimension of T_{CGj}.

The set notation is introduced as $S_{CF} = \{S_{CD}, S_{CG}, S'_{CG}\} = \{L_{CFj} \pm T_{CFj}/2,\ j = 1, \ldots, r+p,\ L_{CFj}(T_{CFj}),\ j = r+p+1, \ldots, r+p+\delta\}$. Where $\{L_{CFj} \pm T_{CFj}/2,\ j = 1, \ldots, r\}$ corresponds to $S_{CD} = \{L_{CD\,j} \pm T_{CD\,j}/2,\ j = 1, \ldots, r\}$, $\{L_{CF\,j} \pm T_{CF\,j}/2,\ j = r+1, \ldots, r+p\}$ corresponds to $S_{CG} = \{L_{CGj} \pm T_{CGj}/2,\ j = 1, \ldots, p\}$, and $\{L_{CFj}(T_{CFj}),\ j = r+p+1, \ldots, r+p+\delta\}$ corresponds to $S'_{CG} =$

$\{L_{CGj}(T_{CGj}), j = p+1, \ldots, p+\delta\}$.

Using of the assembly drawing, the required functional nominal dimensions of the assembly can be expressed as the related component BP nominal dimensions:

$$L_{AFi} = \sum_{j=1}^{r+p} \alpha_{ij} \xi_{ij} K_{ij} L_{CFij} \quad i = 1, \cdots, n+m \tag{5}$$

where α_{ij} is the BP dimension selection coefficient. When the functional component BP dimension L_{CFij} is the link member of dimension L_{AFi}, $\alpha_{ij} = 1$, otherwise, $\alpha_{ij} = 0$. ξ_{ij} is the unit vector for L_{CFij}. $K_{ij} = \partial L_{AFi} / \partial L_{CFij}$ is the dimension coefficient of L_{CFij}, $0 \le K_{ij} \le 1$, $L_{CFij} \in S_{CF}$. L_{AFi} is an assembly functional dimension, $L_{AFi} \in S_{AF}$.

With above dimensional equations, a set of assembly functional DGT inequalities can be derived to represent the relationship between the assembly functional tolerances and the component functional BP tolerances. The general formulation with the worst-case model is:

$$T_{AFi} \ge \sum_{j=1}^{r+p} \alpha_{ij} K_{ij} T_{CFij} \quad i = 1, \cdots, n+m \tag{6}$$

where T_{CFij} is the tolerance of component functional dimension L_{CFij}, $T_{CFij} \in S_{CF}$, T_{AFi} is the tolerance of the required assembly functional dimension L_{AFi}, $T_{AFi} \in S_{AF}$.

In the stage of process planning, the task of tolerancing, however, is to allocate the obtained component functional BP DGTs to the pertinent process tolerances. In most cases, because the design data, the measurement data, and the process data do not always coincide with each other, the tolerance stack-up is inevitable. Assume that there are ϕ manufactured components in an assembly and the subscription variable u denotes the sequence number of the component, thus $u \in [1, \ldots, \phi]$. The subscription variable v denotes the sequence number of the operations related to each component, thus $v \in [1, \ldots, \theta_u]$. Where θ_u is the total operations of the uth component. Let processing working dimensions and tolerances of the uth component be the set $S_{MD\,u} = \{L_{MD\,u\,v} \pm T_{MD\,u\,v}/2, u = 1, \ldots, \phi, v = 1, \ldots, f_u\}$, where f_u is the number of process dimensions and tolerances of the uth component. Let processing geometric tolerances of the uth component that can be treated as equivalent bilateral dimensional tolerances be the set S_{MG}

$_u = \{L_{MG\,u\,v} \pm T_{MG\,u\,v}/2,\ u = 1, \ldots, \phi,\ v = 1, \ldots, g_u\}$, where g_u is the number of geometric tolerances that can be interpreted as equivalent bilateral process dimensional tolerances related to the uth component. Let processing geometric tolerances of the uth component that can be treated as additional processing tolerance constraints be the set $S'_{MG\,u} = \{L_{MG\,u\,v}(T_{MG\,u\,v}),\ u = 1, \ldots, \phi,\ v = g_u+1, \ldots, g_u+\varepsilon_u\}$, where ε_u is the number of geometric tolerances that can be interpreted as additional processing tolerance constraints related to the uth component, $T_{MG\,u\,v}$ is the component BP geometric tolerances, $L_{MG\,u\,v}(T_{MG\,u\,v})$ is the process dimension of tolerance $T_{MG\,u\,v}$.

The set notation related to the uth component is introduced as $S_{CP\,u} = \{L_{CP\,u\,v} \pm T_{CP\,u\,v}/2,\ u = 1, \ldots, \phi,\ v = 1, \ldots, f_u+g_u,\ L_{CP\,u\,v}+T_{CP\,u\,v},\ v = f_u+g_u+1, \ldots, f_u+g_u+\varepsilon_u\}$. Where $\{L_{CP\,u\,v} \pm T_{CP\,u\,v}/2,\ v = 1, \ldots, f_u\}$ corresponds to $S_{MD\,u} = \{L_{MD\,u\,v} \pm T_{MD\,u\,v}/2,\ v = 1, \ldots, f_u\}$, $\{L_{CP\,u\,v} \pm T_{CP\,u\,v}/2,\ v = f_u+1, \ldots, f_u+g_u\}$ corresponds to $S_{MG\,u} = \{L_{MG\,u\,v} \pm T_{MG\,u\,v}/2,\ v = 1, \ldots, g_u\}$, and $\{L_{CP\,u\,v}+T_{CP\,u\,v},\ v = f_u+g_u+1, \ldots, f_u+g_u+\varepsilon_u\}$ corresponds to $S'_{MG\,u} = \{L_{MG\,u\,v}+T_{MG\,u\,v},\ v = g_u+1, \ldots, g_u+\varepsilon_u\}$.

Using the process planning of each related components, the required nominal functional BP dimensions can be expressed as the process dimensions:

$$L_{CFj} = \sum_{v=1}^{\theta_u} \alpha_{uv} \xi_{uv} K_{uv} L_{CPuv} \qquad u = 1, \cdots, \varphi \tag{7}$$

where α_{uv} is the process dimension selection coefficient. For the given process planning of the uth component, when a process dimension L_{CPuv} is the link member of dimension L_{CFj}, $\alpha_{uv} = 1$, otherwise, $\alpha_{uv} = 0$. ξ_{uv} is the unit vector of L_{CPuv}. $K_{uv} = \partial L_{CFj} / \partial L_{CPuv}$ is the dimension coefficient of L_{CPuv}, $0 \le K_{uv} \le 1$. L_{CPuv} is the vth process dimension of the uth component, $L_{CPuv} \in S_{CP\,u}$. L_{CFj} is the component functional dimension, $L_{CFj} \in S_{CF}$.

With above equation, the allocation of the component functional BP DGTs to the process DGTs can be formulated by following inequalities with the worst-case model:

$$T_{CFj} \ge \sum_{v=1}^{\theta_u} \alpha_{uv} K_{uv} T_{CPuv} \qquad u = 1, \cdots, \varphi \tag{8}$$

where T_{CPuv} is the vth process DGT specification corresponds to process dimension L_{CPuv} of the uth component, T_{CFj} is the component functional BP DGT corresponds to BP dimension L_{CFj}.

In a conventional tolerancing system, the process tolerances are acquired by allocation of the functional component BP DGT specifications to the process ones. The disadvantages of this method are that the obtained process tolerances are just under the constraints of BP tolerances and process accuracies. Moreover, component BP tolerances are first determined in the product tolerance design stage. In this stage, the assembly functional DGT specifications cannot be allocated to the relevant component functional BP DGTs in an optimal way without manufacturing information. Therefore some process DGT specifications obtained in the process stage will be beyond the economical bounds and the manufacturing costs will increase unnecessarily.

In concurrent tolerance design, the assembly functional DGT specifications can be directly expressed as the process DGT specifications through using the process planning information of each related component. When the design criterions such as maximum total manufacturing tolerances or minimum manufacturing costs have been presented, the optimal process tolerances can be obtained through establishing and solving an optimization model. Therefore by substituting Equation (7) into (5), the concurrent integrated dimension chains are obtained as:

$$L_{AFi} = \sum_{u=1}^{\varphi} \sum_{v=1}^{\theta_u} \alpha^*_{uv} \xi^*_{uv} \lambda^*_{uv} L_{CPuv} \quad i = 1, \cdots, n+m \tag{9}$$

where α^*_{uv} is the concurrent dimension selection coefficient. For the given process planning of the uth component, when a process dimension L_{CPuv} is the link member of dimension L_{AFi}, $\alpha^*_{uv} = 1$, otherwise, $\alpha^*_{uv} = 0$. ξ^*_{uv} is the unit vector of dimension L_{CPuv}. $\lambda^*_{uv} = \partial L_{AFi} / \partial L_{CPuv}$ is the dimension coefficient of L_{CPuv}, $0 \le \lambda^*_{uv} \le 1$.

With above equation, the concurrent integrated DGT chains, which will be used for directly allocating of the assembly functional DGTs to the component process DGTs, are formulated as:

$$T_{AFi} \ge \sum_{u=1}^{\varphi} \sum_{v=1}^{\theta_u} \alpha_{Tuv} \lambda_{Tuv} T_{CPuv} \quad i = 1, \cdots, n+m \tag{10}$$

The concurrent integrated DGT chains are main constraints and the technical bridge to link substantially the assembly functional DGT specifications and the component process DGT specifications. The approaches used in this paper for establishing the concurrent DGT chains are divided into three steps. First, the

assembly functional product DGT chains will be formulated by using the related mechanical structures of the components and the assembly constraints as the input data. The assembly functional DGTs are expressed as the related functional component BP DGTs by using the integrated tolerance charts in product tolerance design stage. Second, in terms of the given process planning of each component, the component functional BP DGT specifications will be formulated by process DGTs. In this stage, the pertinent structures and the processing plans of the components are used as the input data. Finally, when each component functional BP DGT equation is substituted into the required assembly functional product DGT chains, the required concurrent integrated DGT chains are obtained.

4. Concurrent integrated dimensioning and tolerancing

In assembling a complex product, normally several critical dimensions evaluate the functional performance requirements. These critical dimensions are controlled simultaneously within certain variation ranges for the best working performances. Let the critical dimension vector $y = [y_1\ y_2\ \dots\ y_p]^T$, and the deviation vector $w = [w_1\ w_2\ \dots\ w_p]^T$, $w_i = y_i - y_{0i}$, $i = 1, 2, \dots, p$, where y_{0i} is the nominal/target value of y_i. In a concurrent design environment, the assembly restrictions, topological relationships, and nominal dimensions of the main component have been determined by the assembly structure design. Let $x = [x_1\ x_2\ \dots\ x_n]^T$ be the vector of component design dimensions. These dimensions include sized dimensions and geometrical dimensions. For the geometrical dimensions with fixed tolerance zones, their dimensions and corresponding tolerances can be converted into equivalent bilateral sized dimensions and tolerances. The remaining geometric tolerances are treated as additional tolerance constraints. For simplicity, we denote both sized dimensions and equivalent bilateral sized dimensions as component design dimensions and process dimensions in their different design and manufacturing stages. Therefore, x_j ($j = 1, 2, \dots, n$) is the combination of a set of pertinent process dimensions of a component. Let the process dimension vector $z_j = [z_{j1}\ z_{j2}\ \dots\ z_{jm_j}]^T$, ($j = 1, 2, \dots, n$), where m_j is the number of the operations related to dimension x_j. Finally the assembly functional equations (Zhang, 1996) are expressed:

$$y_i = f_i(x) \qquad i = 1,2,\dots,p \tag{11}$$

In process planning, the machining equations (Zhang, 1996) are generally expressed as:

$$x_j = g_j(z_j) \quad j = 1,2,\ldots,n \tag{12}$$

Since there is no need or way for critical dimensions to be controlled in the exact nominal/target value, a rational variation zone should be assigned for each design dimension. From Equation (11), the actual critical dimension deviations due to their design dimension deviations are expressed as:

$$w_i = y_i - f_i(\bar{x}) = \sum_{j=1}^{n} \left. \frac{\partial f_i(x)}{\partial x_j} \right|_{\bar{x}} \Delta x_j \tag{13}$$

where $f_i(\bar{x})$ is the nominal value obtained by evaluating the assembly functional Equation (1) with its nominal design dimension vector $\bar{x}$. Δx_j is the algebraic difference between x_j and $\bar{x}_j$.

In tolerance design, accumulated design tolerances must be less than or equal to their critical tolerance, so Equation (13) needs some adjusting. For worst-case tolerance stack-up, each differential coefficient is positive, therefore, absolute value of each differential coefficient is required. w_i and Δx_j are replaced by t_i and tx_j. Where t_i and tx_j are respectively the tolerance of critical dimension y_i and design dimension x_j. With these substitutions, Equation (13) changes into inequality:

$$t_i \geq \sum_{j=1}^{n} \left| \left. \frac{\partial f_i(x)}{\partial x_j} \right|_{\bar{x}} \right| tx_j \tag{14}$$

Similarly, from Equation (12) the actual design dimension deviations due to their process dimension deviations can be expressed as:

$$x_j - g_j(\bar{z}_j) = \sum_{k=1}^{m_j} \left. \frac{\partial g_j(z_j)}{\partial z_{jk}} \right|_{\bar{z}_j} \Delta z_{jk} \tag{15}$$

where $g_j(\bar{z}_j)$ is the nominal value obtained by evaluating the machining Equation (12) with its nominal process dimension vector $\bar{z}_j$. Δz_{jk} is the algebraic difference of z_{jk} and $\bar{z}_{jk}$.

When component design tolerances are allocated to process tolerances, Equa-

tion (15) changes into inequality:

$$tx_j \geq \sum_{k=1}^{m_j} \left| \frac{\partial g_j(z_j)}{\partial z_{jk}} \bigg|_{\bar{z}_j} \right| t_{jk} \tag{16}$$

where t_{jk} is jk-th process tolerance of design dimension z_{jk}.

Assume that all process dimensions are of normal distributions. Because design dimensions are functions of process dimensions and assembly critical dimensions are functions of design dimensions, according to statistical theory, both critical dimensions and design dimensions are of normal distributions. From Equation (11), we get variance equations:

$$\mathrm{var}(w_i) = \sum_{j=1}^{n} \left(\frac{\partial f_i(x)}{\partial x_j} \bigg|_{\bar{x}} \right)^2 \mathrm{var}(\Delta x_j) \quad i = 1,2,\cdots,p \tag{17}$$

where variance var(Δx_j) is obtained from Equation (13) and expressed as:

$$\mathrm{var}(\Delta x_j) = \sum_{k=1}^{m_j} \left(\frac{\partial g_j(z_j)}{\partial z_{jk}} \bigg|_{\bar{z}_j} \right)^2 \mathrm{var}(z_{jk}) \quad k = 1,2,\cdots,n \tag{18}$$

where var(w_i), var(Δx_j), and var(z_{jk}) are variances of w_i, Δx_j, and z_{jk}, respectively. Equations (14) and (16) reveal the worst-case tolerance stack-up effect related to two stages, respectively. In Equation (14), component design stack-up tolerance must be less than or equal to functional critical tolerances. Similarly in Equation (16), component process stack-up tolerance must be less than or equal to design tolerances. As discussed above, interdependent tolerancing is divided into two separate stages. In initial product design, designers care more about product satisfaction than about subsequent production capabilities and costs. On the other hand, process planners are more concerned about component manufacturing capabilities than their functional roles in assembly. This conventional method can obtain only the optimum solutions within two separate stages. The best policy is to integrate the two stages into one.

In concurrent engineering, however, the two separate phases are integrated into only one stage (Zhang, 1996; Ngoi and Teck, 1997). This makes it easy for design and manufacturing to collaborate. Essentially, the product designer can consider more fabrication issues when initially designing the product, while

manufacturing engineers can cope with the manufacturing problems based on the component functional roles. This balances the different targets related to product satisfaction and production costs. Mathematically, by substituting machining equation into functional equations the concurrent design equation can be obtained as:

$$t_i \geq \sum_{j=1}^{n} \sum_{k=1}^{m_j} \left| \frac{\partial f_i(x)}{\partial x_j} \right|_{\bar{x}} \left\| \frac{\partial g_j(z_j)}{\partial z_{jk}} \right|_{\bar{z}_j} \left| t_{jk} \quad i = 1,2,\ldots,p \right. \tag{19}$$

5. Quality loss of multiple correlated critical dimensions

High quality and low cost are two fundamental requirements for product design and manufacturing. In an assembly, critical tolerances must be guaranteed for functional requirements. It is well known that the tighter tolerance is, the higher the cost is, and vice versa. For a selected machining operation, if process tolerance becomes smaller and smaller until it reaches a certain value, it will result in the infinite theoretical manufacturing cost. To simplify computation, let best product performance be the point where tolerance is zero. At that point, the theoretical manufacturing cost is infinite. For a single critical dimension case, when critical dimension deviates from its target, the symmetric quadratic Taguchi quality loss function is (Taguchi et al., 1989):

$$L(y) = k(y - \bar{y})^2 \tag{20}$$

where y and $\bar{y}$ are respectively the actual and target values of critical dimension, and k is a positive constant coefficient

To determine the value of k, provided that when dimension y deviates from its target in value w, will cause the loss of A\$. Thus the following equation will be satisfied:

$$k = A / w^2 \tag{21}$$

where $w = y - \bar{y}$.

For a p-dimensional multivariate vector w, Le and Tang (2000) presented a general formula to evaluate the total quality loss due to w:

$$L(w) = w^T K w \tag{22}$$

where K is a $p \times p$ symmetric constant matrix. $k_{ij} = k_{ji}$, for $i \neq j$, $i, j = 1, 2, \ldots, p$. If $p(p+1)/2$ set of product quality deviations and corresponding quality losses are available. The elements of K are related by:

$$\sum_{i=1}^{p}\sum_{j=1}^{p} k_{ij} w_i^{(k)} w_j^{(k)} = A_k \qquad k = 1,2,..., p(p+1)/2 \tag{23}$$

Since manufacturing dimension distribution is dependent upon the related manufacturing process random factors such as machine tools, fixtures, tool wearing, system vibration, temperature fluctuation, operators, and measurement devices, etc, each actual process dimension z_{jk} is obviously a random variable. In terms of Equations (12) and (11), design dimension x_j is the combination of process dimension z_{jk} and critical dimension y_i is the combination of design dimension x_j, so design dimension x_j and critical dimension y_i are also random variables. The distribution of critical dimension y_i is finally dependent upon the density distribution functions of pertinent process dimensions. The product quality loss is determined by all critical dimension distributions. For a batch of products, average quality loss rather than individual loss should be considered. When a product has only a single critical dimension y, let the density function of y be function $\psi(w)$, the average loss of a batch product could be obtained by integration:

$$E(L(w)) = \int_{-\infty}^{+\infty} \psi(w) k w^2 dw \tag{24}$$

As for the multiple critical dimensions, the expectation loss is obvious the summation of individual contributions derived from Equation (24):

$$E(L(w)) = \sum_k \Psi(w^{(k)})(w^{(k)^T} K\, w^{(k)}) \tag{25}$$

where

$$\sum_k \Psi(w^{(k)}) = 1 \tag{26}$$

For the design vector x, the density function is continuous within an interval. Expected quality loss function is (Lee and Tang, 2000):

$$E(L(w)) = \text{Trace } [KV(w)] \tag{27}$$

where $V(w)$ is the variance-covariance matrix of the parameter vector w expressed by:

$$V(w)=\begin{bmatrix} \mathrm{var}(w_1) & \mathrm{cov}(w_1,w_2) & \cdots & \mathrm{cov}(w_1,w_p) \\ \mathrm{cov}(w_1,w_2) & \mathrm{var}(w_2) & \cdots & \vdots \\ \vdots & \cdots & \ddots & \vdots \\ \mathrm{cov}(w_1,w_p) & \cdots & \cdots & \mathrm{var}(w_p) \end{bmatrix} \tag{28}$$

where variance var(w_i) is determined by Equation (17). The covariance between the i-th and the l-th critical dimensions is:

$$\mathrm{cov}(w_i,w_l)=\sum_{k=1}^{n}\left.\frac{\partial f_i(x)}{\partial x_k}\right|_{\bar{x}}\left.\frac{\partial f_l(x)}{\partial x_k}\right|_{\bar{x}}\mathrm{var}(\Delta x_k) \tag{29}$$

For tolerance design, each dimension variance should be expressed as the function of its dimension tolerance. Under stable machining conditions and for large mass production, it is obviously that process dimensions are normally distributed. Therefore when component design tolerances are expressed as process tolerances in the process planning stage, the relation between design tolerance and process variance is:

$$t_j=\frac{2}{C_j}[\mathrm{var}(\Delta x_j)]^{1/2}=\frac{2}{C_j}\left[\sum_{k=1}^{m_j}\left(\left.\frac{\partial g_j(z_j)}{\partial z_{jk}}\right|_{\bar{z}}\right)^2\mathrm{var}(z_{jk})\right]^{1/2} \tag{30}$$

where t_j is bilateral tolerance of design dimension x_j. C_j is a constant factor depending on the probability distribution of the dimension variations concerned. C_j = 1/3 for normally distributed process dimensions with 99.73% probability. When the above equation is substituted into Equations (17) and (29), the variance and covariance of critical dimensions can be expressed by:

$$\begin{aligned}\mathrm{var}(w_i)&=\frac{1}{4}\sum_{j=1}^{n}C_j^2\left(\left.\frac{\partial f_i(x)}{\partial x_j}\right|_{\bar{x}}\right)^2 t_j^2 \\ &=\frac{1}{4}\sum_{j=1}^{n}C_j^2\left(\left.\frac{\partial f_i(x)}{\partial x_j}\right|_{\bar{x}}\right)^2\sum_{k=1}^{m_j}\left(\left.\frac{\partial g_j(z_j)}{\partial z_{jk}}\right|_{\bar{z}}\right)^2 t_{jk}^2\end{aligned} \tag{31}$$

$$\begin{aligned}\operatorname{cov}(w_i, w_l) &= \frac{1}{4}\sum_{j=1}^{n} C_j^2 \left(\left.\frac{\partial f_i(x)}{\partial x_j}\right|_{\bar{x}} \right)\left(\left.\frac{\partial f_l(x)}{\partial x_j}\right|_{\bar{x}} \right) t_j^2 \\ &= \frac{1}{4}\sum_{j=1}^{n} C_j^2 \left(\left.\frac{\partial f_i(x)}{\partial x_j}\right|_{\bar{x}} \right)\left(\left.\frac{\partial f_l(x)}{\partial x_j}\right|_{\bar{x}} \right) \sum_{k=1}^{m_j} \left(\left.\frac{\partial g_j(z_j)}{\partial z_{jk}}\right|_{\bar{z}} \right)^2 t_{jk}^2\end{aligned} \tag{32}$$

where $\boldsymbol{t}_j$ is an m_j-th process tolerance vector, i.e. $\boldsymbol{t}_j = [t_{j1}\ t_{j2}\ \ldots t_{jk} \ldots t_{jm_j}]^T$, and $k = 1, 2, \ldots, n$.

6. Optimal tolerance assignment

To implement robust tolerance design, the best balance should be made between product satisfaction and manufacturing cost. In a concurrent tolerancing environment, the product quality loss is expressed as the function of pertinent process tolerances. In the optimum model, the objective is to minimize the summation of product manufacturing cost and quality loss:

$$\min \sum_{j=1}^{n}\sum_{k=1}^{m_j} c_{jk}(t_{jk}) + E(L(w)) \tag{33}$$

where $c_{jk}(t_{jk})$ is manufacturing cost of jk-th process operation, and $E(L(w))$ is expected quality loss function of the product.

To determine the manufacturing cost, cost-tolerance functions can be used. With regard to cost-tolerance function, several types of models have been presented (Zhang, 1996; Fang and Wu, 2000). Regression techniques are often applied to the acquired discrete cost-tolerance data and determine the unknown constant coefficients for each model. The models with the highest regression precision are used as cost-tolerance functions. Based on this method, Fang *et al* presented a set of cost-tolerance functions suitable for middle quantitative production in manufacturing enterprises. The one suitable for planar features is (Fang and Wu, 2000):

$$c_{jk}(t_{jk}) = 50.261\exp(-15.8903 t_{jk}) + t_{jk}/(0.3927 t_{jk} + 0.1176) \tag{34}$$

In actual manufacturing, each process dimension z_{jk} has an economical toler-

ance range. It can be expressed mathematically by:

$$t_{jk}^{-} \le t_{jk} \le t_{jk}^{+} \tag{35}$$

where t_{jk}^{-} and t_{jk}^{+} is respectively the lower and upper bounds of process tolerance t_{jk}

In a concurrent tolerancing environment, the complete optimization model can be introduced as:

$$\min \sum_{j=1}^{n} \sum_{k=1}^{m_j} c_{jk}(t_{jk}) + E(L(w))$$

s.t.

$$ty_i^{-} \le \sum_{j=1}^{n} \left| \frac{\partial f_i(x)}{\partial x_j} \right|_{\bar{x}} \left| \sum_{k=1}^{m_j} \left| \frac{\partial g_j(z_j)}{\partial z_{jk}} \right|_{\bar{z}_j} \right| t_{jk} \le ty_i^{+}$$

$$t_{jk}^{-} \le t_{jk} \le t_{jk}^{+}$$

where $ty_i^{-}$and ty_i^{+} are the lower and upper bounds of assembly critical tolerance ty_i, respectively. They are given as input data in terms of product quality and manufacturing cost. The optimum ty_i is determined by solving the optimal model.

Two kinds of constraints are proposed for the optimal model. The first are concurrent design equations. These equations present the tolerance stack-up effects between assembly critical tolerances and pertinent manufacturing tolerances by worst-case or statistical model. In concurrent design equation critical tolerance must be greater than or equal to its pertinent sum manufacturing tolerance. The second constraints are process capabilities. According to selected fabrication methods and machining tools, each processed tolerance should specify an economical variation range.

7. A practical example

Figure 5 shows a wheel assembly with pure size dimensions. For simplicity, we do not consider the geometric tolerances and their conversion in this example. Also the process dimensions obey normal distributions. Assume that nominal design dimensions have already been assigned based on the requirements in size, strength, structure, assembly, and maintenance, etc, they are: x_1 = 9, x_2 =

20, $x_3 = 9$, $x_4 = 12$, $x_5 = 38.2$, $x_6 = 12$, $x_7 = 62.4$ (unit: mm). Two critical dimensions $y_1 = 0.2 \pm 0.080 \sim 0.140$, and $y_2 = 0.2 \pm 0.075 \sim 0.130$. y_1 is the critical axial gap between bush 7 and frame 9. y_2 is another critical axial gap between nut 8 and frame 9. It is not difficult to formulate the assembly functional equations using the method presented by Huang *et al.* (2001).

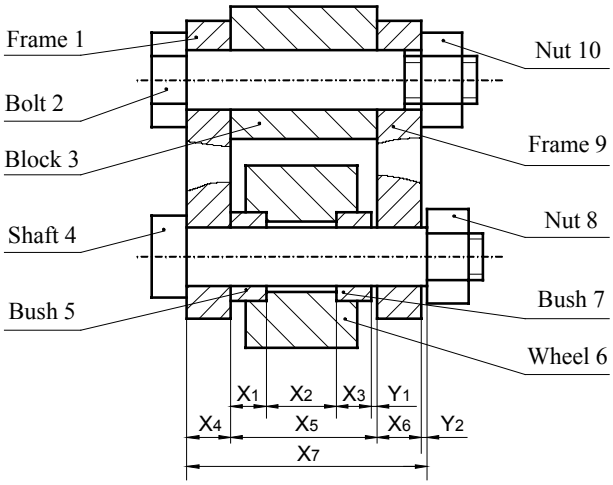


Figure 5. Wheel assembly

$$y_1 = -x_1 - x_2 - x_3 + x_5$$
$$y_2 = -x_4 - x_5 - x_6 + x_7$$

According to Equation (13), the deviation equations of critical dimensions are:

$$w_1 = y_1 - \bar{y}_1 = -\Delta x_1 - \Delta x_2 - \Delta x_3 + \Delta x_5$$
$$w_2 = y_2 - \bar{y}_2 = -\Delta x_4 - \Delta x_5 - \Delta x_6 + \Delta x_7$$

With Equation (14), the functional tolerance inequalities by worst-case model are:

$$ty_1 \geq tx_1 + tx_2 + tx_3 + tx_5$$
$$ty_2 \geq tx_4 + tx_5 + tx_6 + tx_7$$

Provided that the manufacturing process takes place under stable conditions, each process dimension will be of normal distribution. For simplicity, assume that the distribution center of each process dimension is just equal to its nominal value. Each critical dimension variance can be expressed as the function of its design tolerance:

$$\text{var}(w_1) = \frac{1}{36}(tx_1^2 + tx_2^2 + tx_3^2 + tx_5^2)$$

$$\text{var}(w_2) = \frac{1}{36}(tx_4^2 + tx_5^2 + tx_6^2 + tx_7^2)$$

Similarly, the covariance of the two correlated critical dimensions can be expressed as the function of the pertinent design tolerances:

$$\text{cov}(w_1, w_2) = -\frac{1}{36}tx_5^2$$

The critical tolerance ranges of y_1 and y_2 in Figure 5 are determined both by performance satisfaction and manufacturing cost of this assembly. To finally determine the optimum tolerance of these two critical dimensions and then allocate them to the related process dimensions, quality loss and manufacturing cost must be determined first. Provided that when critical dimension y_1 and y_2 deviate from their target (nominal) vector with values $\underline{w}^{(1)} = [\underline{w}_1^{(1)}\ 0]^T = [0.160, 0]^T$, $\underline{w}^{(2)} = [0\ \underline{w}_2^{(2)}]^T = [0, 0.150]^T$, or $\underline{w}^{(3)} = [\underline{w}_1^{(3)}\ \underline{w}_2^{(3)}]^T = [0.140, 0.130]^T$will result in product failure and cause a quality loss of \$300. The constant matrix K can thus be decided by Equation (22):

$$k_{11} = A_1/(\underline{w}_1^{(1)})^2 = 300/0.16^2 = 11718.75$$

$$k_{22} = A_2/(\underline{w}_2^{(2)})^2 = 300/0.15^2 = 13333.33$$

$$\begin{aligned} k_{12} = k_{21} &= (A_3 - A_1(\underline{w}_1^{(3)})^2/(\underline{w}_1^{(1)})^2 - A_1(\underline{w}_2^{(3)})^2)/(\underline{w}_2^{(2)})^2)/(2\underline{w}_1^{(3)}\underline{w}_2^{(3)}) \\ &= (300 - 300\times 0.14^2/0.16^2 - 300\times 0.13^2/0.15^2)/2\times 0.14\times 0.13 \\ &= -4258.81 \end{aligned}$$

With this, total expected loss is:

$$E(L(w)) = \text{Trace}\ [KV(w)]$$

$$= \frac{1}{36}[k_{11}tx_1^2 + k_{11}tx_2^2 + k_{11}tx_3^2 + k_{22}tx_4^2 + (k_{11} - 2k_{12} + k_{22})tx_5^2 + k_{22}tx_6^2 + k_{22}tx_7^2]$$

Figure 6 shows the related structure and design dimension for each machining part. For the corresponding process plan, look at the economical process tolerance bounds for each machining part in Table 1.

Using the method presented by Huang *et al.* (2001), the machining equations are obtained from given component process plans:

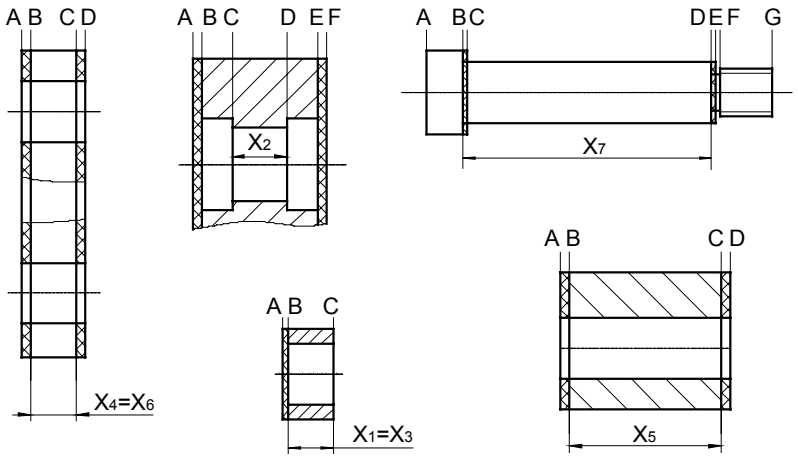


Figure 6. Process plan of related parts

$$x_1 = z_{11} - z_{12}$$
$$x_2 = z_{24} - z_{23} - z_{25}$$
$$x_3 = z_{31} - z_{32}$$
$$x_4 = z_{44}$$
$$x_5 = z_{54}$$
$$x_6 = z_{64}$$
$$x_7 = z_{74} - z_{73}$$

The design tolerance inequalities are:

$$tx_1 \geq t_{11} + t_{12}$$
$$tx_2 \geq t_{23} + t_{24} + t_{25}$$
$$tx_3 \geq t_{31} + t_{32}$$
$$tx_4 \geq t_{44}$$
$$tx_5 \geq t_{54}$$
$$tx_6 \geq t_{64}$$
$$tx_7 \geq t_{73} + t_{74}$$

Part	No	Process name	Measure reference	Machined plane	Process dimension z_{kq}	Process tolerance - t_{kq}	Tolerance bound t_{kq}^{-}~t_{kq}^{+} -(μm)
Bush 5 and 7	11	Parting-off	*C*	*A*	z_{11} = 11	t_{11}	27~70
	12	L plane by FL	*A*	*B*	z_{12} = 2	t_{12}	10~25
Wheel 6	21	R plane by RL	*A*	*F*	z_{21}= 36	t_{21}	54~140
	22	L plane by FL	*F*	*B*	z_{22}= 34	t_{22}	54~140
	23	L pole by L	*B*	*C*	z_{23}= 6	t_{23}	12~30
	24	R plane by FL	*B*	*E*	z_{24}= 32	t_{24}	25~62
	25	R pole by L	*E*	*D*	z_{25}= 6	t_{25}	12~30
Frame 1 and 9	41	R plane by RM	*A*	*D*	z_{41}= 16	t_{41}	43~110
	42	L plane by RM	*D*	*B*	z_{42}= 16	t_{42}	43~110
	43	R plane by FM	*B*	*C*	z_{43}= 14	t_{43}	43~110
	44	L plane by FM	*C*	*B*	z_{44}= 12	t_{44}	27~70
Block 3	51	R plane by RM	*A*	*D*	z_{51}= 42.2	t_{51}	62~160
	52	L plane by RM	*D*	*B*	z_{52}= 42.2	t_{52}	62~160
	53	R plane by FM	*B*	*C*	z_{53}= 40.2	t_{53}	39~100
	54	L plane by FM	*C*	*B*	z_{54}= 38.2	t_{54}	39~100
Shaft 4	71	Step by RL	*G*	*C*	z_{71} = 80.4	t_{71}	54~140
	72	Step by RL	*G*	*E*	z_{72} = 18	t_{72}	27~70
	73	Step by FL	*G*	*D*	z_{73} = 20	t_{73}	21~52
	74	Step by FL	*G*	*B*	z_{74} = 82.4	t_{74}	35~87
	75	Truncation	*G*	*A*	z_{75} = 90	t_{75}	54~140

Table 1. Axial process plan for related parts.

Notes: FM stands for finish milling, RM stands for rough milling, FL stands for finish lathing, RL stands for rough lathing, L stands for lathing, R stands for right and L stands for left.

The component design tolerance can be formulated as the function of its related process tolerances with Equation (20):

$$tx_1^2 = t_{11}^2 + t_{12}^2 \qquad tx_4^2 = t_{44}^2$$
$$tx_2^2 = t_{23}^2 + t_{24}^2 + t_{25}^2 \qquad tx_5^2 = t_{54}^2$$
$$tx_3^2 = t_{31}^2 + t_{32}^2 \qquad tx_6^2 = t_{64}^2$$
$$tx_7^2 = t_{73}^2 + t_{74}^2$$

In a concurrent tolerancing environment, when machining equations are substituted into assembly functional equations, product quality loss is finally obtained as:

$$\begin{aligned}
&E(L(w)) \\
&= \frac{1}{36}[k_{11}tx_1^2 + k_{11}tx_2^2 + k_{11}tx_3^2 + k_{22}tx_4^2 + (k_{11} - 2k_{12} + k_{22})tx_5^2 + k_{22}tx_6^2 + k_{22}tx_7^2] \\
&= 325.52(t_{11}^2 + t_{12}^2 + t_{23}^2 + t_{24}^2 + t_{25}^2 + t_{31}^2 + t_{32}^2) + 370.37t_{44}^2 + 33569.7t_{54}^2 + 370.37t_{64}^2 \\
&\quad + 370.37(t_{73}^2 + t_{74}^2)
\end{aligned}$$

In this example, we only consider the manufacturing cost of process dimensions that are involved in assembly functional equations. The reason is that the other process dimensions can use the most economical tolerances, and manufacturing costs of these operations are minimal. Furthermore, these process dimensions don't contribute to quality loss. The manufacturing cost of these considered operations is:

$$\begin{aligned}
C_M &= \sum_{j=1}^{n}\sum_{k=1}^{m_j} c_{jk}(t_{jk}) \\
&= c_{11} + c_{12} + c_{23} + c_{24} + c_{25} + c_{31} + c_{32} + c_{44} + c_{54} + c_{64} + c_{73} + c_{74}
\end{aligned}$$

The summation of C_M and $E(L(w))$ is:

$$\begin{aligned}
&C = C_M + E(L(w)) \\
&= c_{11} + c_{12} + c_{23} + c_{24} + c_{25} + c_{31} + c_{32} + c_{44} + c_{54} + c_{64} + c_{73} + c_{74} \\
&\quad + 325.52(t_{11}^2 + t_{12}^2 + t_{23}^2 + t_{24}^2 + t_{25}^2 + t_{31}^2 + t_{32}^2) + 370.37t_{44}^2 + 33569.7t_{54}^2 + 370.37t_{64}^2 \\
&\quad + 370.37(t_{73}^2 + t_{74}^2)
\end{aligned}$$

Finally, the entire optimization problem is formulated as:

$$\begin{aligned}
&\min\{c_{11} + c_{12} + c_{23} + c_{24} + c_{25} + c_{31} + c_{32} + c_{44} + c_{54} + c_{64} + c_{73} + c_{74} \\
&\quad + 325.52(t_{11}^2 + t_{12}^2 + t_{23}^2 + t_{24}^2 + t_{25}^2 + t_{31}^2 + t_{32}^2) + 370.37(t_{44}^2 + t_{64}^2 + t_{73}^2 + t_{74}^2) \\
&\quad + 33569.7t_{54}^2\}
\end{aligned}$$

where

$$c_{jk} = c_{jk}(t_{jk}) = 5.0261\exp(-15.8903t_{jk}) + t_{jk}/(0.3927t_{jk} + 0.1176)$$

Subjected to:
The concurrent tolerance stack-up constraints by worst-case model:

$$0.160 = t_1^- \le t_{11} + t_{12} + t_{23} + t_{24} + t_{25} + t_{31} + t_{32} + t_{54} \le t_1^+ = 0.280$$
$$0.150 = t_2^- \le t_{44} + t_{54} + t_{64} + t_{73} + t_{74} \le t_2^+ = 0.260$$

where $t_1^- = 0.160$, $t_1^+ = 0.280$ is the lower and upper tolerance bound of critical dimension y_1, $t_2^- = 0.150$, $t_2^+ = 0.260$ is the lower and upper tolerance bound of critical dimension y_2, respectively.
The economical process tolerance ranges for each process operation are as follows:

$$0.018 = t_{11}^- \le t_{11} \le t_{11}^+ = 0.043$$
$$0.010 = t_{12}^- \le t_{12} \le t_{12}^+ = 0.025$$
$$0.012 = t_{23}^- \le t_{23} \le t_{23}^+ = 0.030$$
$$0.025 = t_{24}^- \le t_{24} \le t_{24}^+ = 0.062$$
$$0.012 = t_{25}^- \le t_{25} \le t_{25}^+ = 0.030$$
$$0.018 = t_{31}^- \le t_{31} \le t_{31}^+ = 0.043$$
$$0.010 = t_{32}^- \le t_{32} \le t_{32}^+ = 0.025$$
$$0.018 = t_{44}^- \le t_{44} \le t_{44}^+ = 0.043$$
$$0.025 = t_{54}^- \le t_{54} \le t_{54}^+ = 0.062$$
$$0.018 = t_{64}^- \le t_{64} \le t_{64}^+ = 0.043$$
$$0.021 = t_{73}^- \le t_{73} \le t_{73}^+ = 0.052$$
$$0.035 = t_{74}^- \le t_{74} \le t_{74}^+ = 0.087$$

The proposed optimization model is solved by the nonlinear optimal method. In order to test the validity of the proposed approach, a similar optimal model is also introduced. This model removes the quality loss from objective function. The constraints are the same for these two different models. The optimization results of the two models are given in Table 2 for comparison. Obtained process tolerance t_{11}, t_{12}, t_{23}, t_{25}, t_{31}, t_{32}, t_{44}, and t_{64} are the same for both approaches. But t_{24}, t_{54}, t_{73}, and t_{74} are different. For the proposed method, these tolerances are of smaller values to maintain less quality loss.

Method	t_{11}	t_{12}	t_{23}	t_{24}	t_{25}	t_{31}	t_{32}	t_{44}	t_{54}	t_{64}	t_{73}	t_{74}	total
C_M+C_L	43	25	30	25	30	43	25	43	25	43	21	35	388
C_M	43	25	30	49	30	43	25	43	35	43	52	87	505

Table 2. The comparison results of the two methods (unit: μm).

8. Concluding remarks

This paper has presented a robust optimization method in a concurrent tolerancing environment. This method can determine multiple correlated critical tolerances and directly allocate them to process tolerances by using component process plans.

In a concurrent environment, the product tolerance design and process tolerance design can be integrated into one stage. Tolerance design has been extended directly from the product design to the manufacturing stage. The necessity of redesign and rework between product tolerance design and process tolerance design has been eliminated, increasing the design efficiency. In a conventional tolerance design, the optimal model is established for two separate stages, and the optimum solutions are for different stages but not for the entire product design process.

Though Lee and Tang (2000) in their research introduced a method to implement tolerance design for products with correlated characteristics, they only dealt with tolerancing problems within the product design stage. The basic method they used has now been extended profoundly to the concurrent environment to determine multiple correlated critical product tolerances and then allocate them directly to pertinent process tolerances.

The purpose of this paper is to propose a robust optimum tolerance design method in a concurrent environment to balance the conflict design targets between manufacturing tolerances and product satisfaction. The design targets are quantified in monetary ways in the optimization objective function. The focus is on establishment of quality loss of product with multiple correlated critical tolerances in a concurrent tolerance design environment. The paper presents an approach to provide the product quality loss function, which is finally expressed as the function of process tolerances.

A wheel assembly example presented by Huang and Gao (2003) has also been applied. The simulation results show the validity of the proposed method. If cost-tolerance function and related information of product quality loss are available, the rational tolerances can be obtained in actual design and production.

Acknowledgements

This research is sponsored by the National Natural Science Foundation of China (Grant No. 50465001) to M. F. Huang. The authors would like to thank Dr. M. Chen, the professor of Department of Mechanical and Industrial Engineering, Concordia University, Montreal, Canada, for his constructive comments on the earlier version of this paper.

9. References

Chang, C. L., Wei, C. C. and Chen, C. B. (2000). Concurrent maximization of process tolerances using grey theory. *Robotics and Computer Integrated Manufacturing*, Vol. 16, No. 2-3, April-June 2000, ISSN 0736-5845, 103-107.

Chen, Y. B., Huang, M. F., Yao, J. C. and Zhong, Y. F. (2003). Optimal concurrent tolerance based on the grey optimal approach. *The International Journal of Advanced Manufacturing Technology*, Vol. 22, No. 1-2, 2003, ISSN 0268-3768, 112–117.

Diplaris, S. C. and Sfantsikopoulos, M. M. (2000). Cost-Tolerance Function. A New approach for cost optimum machining accuracy. *The International Journal of Advanced Manufacturing Technology*, Vol. 16, No. 1, 2000, ISSN 0268-3768, 32-38.

Fang, H. F. and Wu, Z. T. (2000). Concurrent tolerance design and methods of technology economy assessment in process route. *Chinese Journal of Mechanical Engineering* (*Chinese*), Vol. 36, No. 4, Apr. 2000, ISSN 0577-6686, 74-77.

Fang, H. F., He, Y. and Wu, Z. T. (1998). Concurrent tolerancing based on the Taguchi's quality loss", *Mechanical design* (*Chinese*), Vol. 15, No. 3, 1998, ISSN 1001-2354, 22-24.

Gao, Y. and Huang, M. (2003). Optimal process tolerance balancing based on process capabilities. *The International Journal of Advanced Manufacturing Technology*, Vol. 21, No. 7, 2003, ISSN 0268-3768, 501-507.

He, J. R. and Gibson, P. R. (1992). Computer-aided geometrical dimensioning and tolerancing for process-operation planning and quality control. *The International Journal of Advanced Manufacturing Technology*, Vol. 7, No. 1, 1992, ISSN 0268-3768, 11-20.

Huang, M., Xu, Z., Gao, Y. and Li, Z. (2001). Optimal assembly tolerance allocation using hierachical hypergraphs, Proceedings of 17th International Conference on Computer-aided Production Engineering, CAPE 2001, Bin

H., pp. 411-414, ISBN 1-86058-365-2, Wuhan, China, May 2001, IMechE, Professional Engineering Publishing Limited, London.

Huang, M. F. and Gao,Y. S. (2003). Optimal concurrent tolerancing based on sequential process capabilities. *China Mechanical Engineering* (*Chinese*), Vol.14, No. 5, March 2003, ISSN 1004-132X, 385-389.

Huang, M. F. and Gao, Y. S. (2002). A discrete optimal tolerancing approach based on the process capabilities. *Journal of Huazhong University of Science and Technology* (*Chinese*), Vol. 30, No. 4, April 2002, ISSN 1000-8616, 19-21.

Huang, M., Gao, Y., Xu, Z., and Li, Z. (2002). Composite planar tolerance allocation with dimensional and geometric specifications. *The International Journal of Advanced Manufacturing Technology*, Vol. 20, No. 5, 2002, ISSN 0268-3768, 341-347.

Huang, M. F, Zhong, Y. R., and Xu, Z. G. (2005). Concurrent process tolerance design based on minimum product manufacturing cost and quality loss. *The International Journal of Advanced Manufacturing Technology*, Vol. 25, No. 7-8, 2005, ISSN 0268-3768, 714-722.

Huang, M. F. and Zhong, Y. R. (submitted). Dimensional and geometrical tolerance balancing in concurrent design. *The International Journal of Advanced Manufacturing Technology*, (submitted), ISSN 0268-3768, 341-347.

Jeang, A. (1998). Tolerance chart optimization for quality and cost. International Journal of Production Research, Vol. 36, No. 11, Nov. 1998, ISSN 0020-7543, 2969-2983.

Jeang, A. (1997). An approach of tolerance design for quality improvement and cost reduction. *International Journal of Production Research*, Vol. 35, No. 5, May 1997, ISSN 0020-7543, 1193-1211.

Ji, P. (1993). A tree approach for tolerance charting. *International Journal of Production Research*, Vol. 31, No. 5, May 1993, ISSN 0020-7543, 1023-1033.

Lee, C. L. and Tang, G. R. (2000). Tolerance design for products with correlated characteristics. *Mechanism and Machine Theory*, Vol. 35, No. 12, Dec. 2000, ISSN 0094-114X, 1675-1687.

Ngoi, K. B. A. and Tan, C. K. (1995). Geometrics in computer-aided tolerance charting. *International Journal of Production Research*, Vol. 33, No. 3, March 1995, ISSN 0020-7543, 835-868.

Ngoi, B. K. A. and Soew, M. S. (1996). Tolerance control for dimensional and geometrical specifications. *The International Journal of Advanced Manufacturing Technology*, Vol. 11, No. 1, 1999, ISSN 0268-3768, 34-42.

Ngoi, B. K. A. and Teck, O. C. (1997). A tolerancing optimization method for

product design. *The International Journal of Advanced Manufacturing Technology*, Vol. 13, No. 4, 1997, ISSN 0268-3768, 290-299.

Ngoi, B. K. A. and Min, O. J. (1999). Optimum tolerance allocation in assembly. *The International Journal of Advanced Manufacturing Technology*, Vol. 15, No. 9, 1999, ISSN 0268-3768, 660-665.

Ngoi, B. K. A. and Ong, J. M. (1999). A complete tolerance charting system in assembly. *International Journal of Production Research*, Vol. 37, No. 11, July 1999, ISSN 0020-7543, 2477-2498.

Ngoi, K. B. A. and Teck, O. C. (1993). A complete tolerance charting system. *International Journal of Production Research,* Vol. 31, No. 2, Feb. 1993, ISSN 0020-7543, 453-469.

Swift, K. G., Raines, M. and Booker, J. D. (1999). Tolerance optimization in assembly stacks based on capable design, Proceedings of the Institution of Mechanical Engineers, part B (Journal of Engineering Manufacture), Vol. 213, No. 7, 1999, ISSN 0954-4054, 677-693.

Taguchi, G., Elsayed, E. A. and Hsiang, T. C. (1989). *Quality engineering in production system*. McGraw-Hill, New York.

Taguchi, G. (1993). *On robust technology development*. ASME Press, New York.

Tseng, Y. J. and Kung, H. W. (1999). Evaluation of alternative tolerance allocation for multiple machining sequences with geometric tolerances. *International Journal of Production Research,* Vol. 37, No. 17, Nov. 1999, ISSN 0020-7543, 3883-3900.

Wei, C. C. and Lee, Y. C. (1995). Determining the process tolerances based on the manufacturing process capability. *The International Journal of Advanced Manufacturing Technology*, Vol. 10, No. 6, 1995, ISSN 0268-3768, 416-421.

Wu, C. C., Chen, Z. and Tang, G. R. (1998). Component tolerance design for minimum quality loss and manufacturing cost. *Computers in Industry*, Vol. 35, No. 4, April 1998, ISSN 0166-3615, 223-232.

Zhang G. (1996). Simultaneous tolerancing for design and manufacturing. *International Journal of Production Research,* Vol. 34, No. 12, Dec. 1996, ISSN 0020-7543, 3361-3382.

19

Optimize Variant Product Design Based on Component Interaction Graph

Elim Liu and Shih-Wen Hsiao

1. Introduction

Dominating markets with a single product is increasingly difficult, and instead numerous industries are evolving towards mass customization, meaning the production of individually customized and highly varied products or services (Pine, 1993). This proliferation of models allows consumers to find a product that best suits their individual needs. The need for increasing product variety and shorter development time brings more complexity to the company than ever. Corporations are striving to balance customer satisfaction and cost savings, and product design is becoming essential for accomplishing this. Since developing an entirely different product is often uneconomical. A better method is to develop a product architecture that enables a company to offer highly differentiated products that share a substantial fraction of their components. Therefore, introducing product variety within a robust architecture offers one means of enhancing mass customization. Besides, an increase in product variety brings an increase in the volume of information exchanged between customers, designers and marketing department. Due to such increased information processing load, information technology is needed to tackle this problem. This chapter investigates the product variety design methodologies through the computational design optimization methods, and developing product architecture under the support of information technologies. It aims at providing product designers a rational and systematic methodology in dealing with product variety from both qualitative and quantitative viewpoints.

2. Related Literature

The issue of product variety has attracted growing research interest during recent years. In1993, Pine (1993) began discussing the need for product variety in

increasingly competitive markets. Cohen (1995) proposed using Master House of Quality for planning product variety. Suh (1990) viewed product variety as the proper selection of design parameters that satisfy variant functional requirements. Ulrich (1995) examined the relationships between product architecture and product variety, component standardization, modularity, and product development. Erens (1996) developed product variety under functional, technology, and physical domains. Fujita and Ishii (1997) formulated the task structure of product variety design, and Martin and Ishii (1996, 1997, 2002) proposed DFV (Design for Variety), which is a series of methodologies with quantifying indices for reducing the influence of product variety on product life-cycle cost, and thus helping design teams to develop decoupled product architectures. These studies have established a basis for product variety management. However, many investigations have agreed that the key to efficiently designing and delivering multiple products is developing a good product architecture (Meyer 1993, Sawhney 1998, Ulrich& Eppinger 2000). The advantages of developing product architecture is that it enables a company to offer two or more products that are highly differentiated yet share a substantial fraction of their components. The collection of components shared by these products is called a product platform (Ulrich& Eppinger 2000). Erens (1996) defined a product platform as "An architecture concept of compromising interface definitions and key-components, addressing a market and being a base for deriving different product families." Robertson and Ulrich (1998) proposed a method of balancing distinctiveness with commonality within product architecture through identifying the importance of various factors going into this tradeoff. Fujita et al., (1998, 1999) utilized optimization techniques to identify the optimum architecture of a module combination across products in a family of aircraft. Moreover, Yu et al., (1998) defined product family architecture based on customer needs by using the target value of product features for calculating probability distributions. Additionally, Simpson, et al., (1999) used the Product Platform Concept Exploration Method (PPCEM) to design a common product platform. This platform uses the market segmentation grid to help identify suitable scale factors of the platform that are "scaled" or "stretched" to satisfy various requirements.

Although most studies focus on optimizing product structure, some studies have noticed that investigating the physical arrangement and interaction among components is the key for stable product architecture. For example, the component-based DSM (design structural matrix) method has been applied to

explore alternative architectures through clustering high interactive components and arranging them in chunks (Pimmler & Eppinger 1994, Wei 2001). Moreover, Sosa et al., (2000) applied DSM to analyze the different types of interaction between modular and integrative systems, and Salhieh & Kamrani (1995) used the similarity matrix for integrating components into modules. These studies represent component relationships in terms of similarity or reciprocal interaction rather than information flows. However, during the embodiment design stage, variant designs of a single component can lead to numerous other components also requiring modification. The hierarchical structure of component interactions first must be identified, after which the influence of variety and subsequent design changes can be estimated. To deal with this problem, this chapter illustrated two methodologies via identifying component design constraint flows to build up feasible product architecture.

3. Product Design Based on Component Interaction Graph

3.1 Product design rational

Studies of product design have observed that designs are always completed through iteration. Design iteration occurs when a new requirement is inputted into the design task, resulting in the related components needing to be redesigned, and leading to the specifications of the other components that interact with the redesigned components having to change their specifications to fit the redesign. Therefore, the design process becomes iterative, and so tremendous design efforts are required. This problem becomes particularly important in planning product architectures; products must be designed to meet various customer needs, yet also share as many components as possible to minimize costs. This study attempted to solve this problem by modeling component sequential flow using ISM, interpretive structural modeling. ISM is an algebraic technique for system representation and analysis that was first introduced by Warfield (1973). ISM reduces complex system interactions to a logically oriented graph.

This study applies and modifies ISM to establish a hierarchical component interaction structure, which can help designers to determine component commonality, variety, and design priorities.

3.2 Computational procedure of ISM

Phase1: Incidence matrix construction

First, a system is decomposed into a set of components that form a square matrix. The procedure begins with paired comparisons to identify whether a direct influence exists from component i (row) to j (column). The incidence matrix $A=[a_{ij}]$ thus is defined as

$$a_{ij}=\begin{cases}1 & \text{if a direct influence exists from component i to component j}\\ 0 & \text{otherwise}\end{cases}$$

Fig. 1(a) represents the incidence matrix of an example system containing seven components. For example, the second row of the matrix indicates that component 2 directly influences components 1, 5, and 6.

Phase 2: Reachability matrix deduction

The reachability matrix R is deducted from incidence matrix A if a Boolean n-multiple product of $A+I$ uniquely converges to R for all integers $n>n_0$, where n_0 is an appropriate positive integer, I is a Boolean unity matrix, and + is addition in Boolean sense (Warfield, 1995). Matrix R represents all direct and indirect linkages between components. Figure 1(b) represents the reachability matrix R derived from matrix A, in which an entry $r_{ij}=1$ if component j is reachable by i, although the path length may be one or more.

$$A=\begin{array}{c} \\ c_1\\ c_2\\ c_3\\ c_4\\ c_5\\ c_6\\ c_7\end{array}\begin{array}{c}\begin{array}{ccccccc}c_1 & c_2 & c_3 & c_4 & c_5 & c_6 & c_7\end{array}\\ \begin{bmatrix}0&0&0&0&0&0&0\\1&0&0&0&1&1&0\\1&0&0&0&1&0&1\\0&0&0&0&1&0&0\\0&0&1&0&0&0&1\\0&1&1&0&0&0&0\\0&0&1&0&0&0&0\end{bmatrix}\end{array}$$

(a) Original incidence matrix A

$$R=\begin{array}{c} \\ c_1\\ c_2\\ c_3\\ c_4\\ c_5\\ c_6\\ c_7\end{array}\begin{array}{c}\begin{array}{ccccccc}c_1 & c_2 & c_3 & c_4 & c_5 & c_6 & c_7\end{array}\\ \begin{bmatrix}1&0&0&0&0&0&0\\1&1&1&0&1&1&1\\1&0&1&0&1&0&1\\1&0&1&1&1&0&1\\1&0&1&0&1&0&1\\1&1&1&0&1&1&1\\1&0&1&0&1&0&1\end{bmatrix}\end{array}$$

(b) Reachability matrix R

Figure 1 a-b. Stepwise procedure of ISM

Phase 3: Cluster retrieval

A technique for cluster retrieval is inserted in the ISM process to identify components that influence one another and form a loop (Roberts, 1997). The reachability matrix R multiplies the transposed matrix of R, say R^t; thus in $R \bullet R^t$, components i and j mutually interact if $r_{ij}\, r_{ji} = 1$. Figure 1(c) displays the output matrix of $R \bullet R^t$, in which clusters of components can be identified easily by rearranging component order. Figure 1(d) reveals four clusters in the system, namely: {1}, {2,6}, {3,5,7}, and {4}.

$$R \bullet R^t = \begin{array}{c} \\ c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \\ c_6 \\ c_7 \end{array} \begin{array}{c} \begin{array}{ccccccc} c_1 & c_2 & c_3 & c_4 & c_5 & c_6 & c_7 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix} \end{array}$$

(c) Output matrix of $R \bullet R^t$

$$R \bullet R^t = \begin{array}{c} \\ c_1 \\ c_2 \\ c_6 \\ c_3 \\ c_5 \\ c_7 \\ c_4 \end{array} \begin{array}{c} \begin{array}{ccccccc} c_1 & c_2 & c_6 & c_3 & c_5 & c_7 & c_4 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

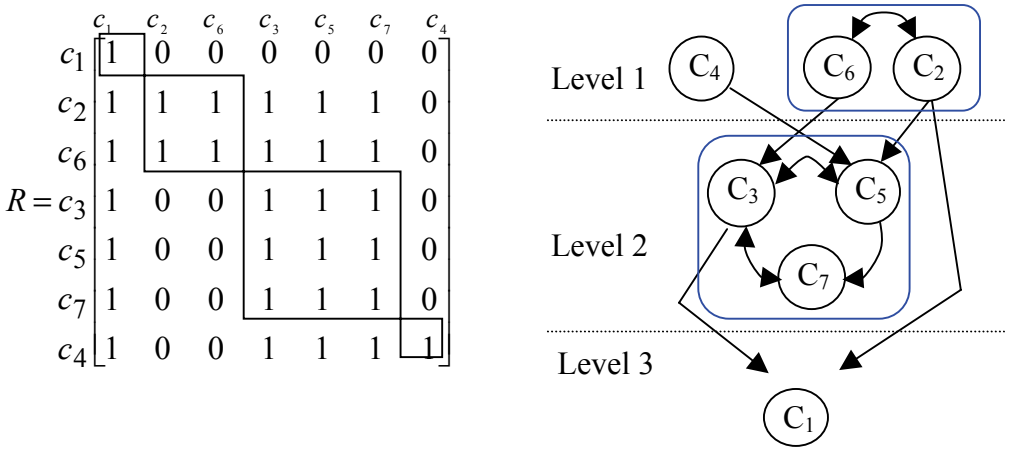
(d) Retrieval of clusters

Figure 1 c-d. Stepwise procedure of ISM

Phase 4: Obtained hierarchy graph

Following cluster retrieval, the order of reachability matrix R is rearranged (as shown in Fig. 1(e)), and the clustered components are integrated and treated as a single entity. The hierarchy graph then is obtained by identifying a set of components in matrix R that cannot reach or be reached by other components outside the set itself, removing the set from the original matrix R, and then repeating this process for remaining matrix until a unique set of nodes that no other nodes can reach is obtained. For example, in Fig. 1(e), c_1 first is identified as an "exit", since it can not reach to other components; meanwhile, {c_2, c_6} and c_4 were separated as "entrances", because they can not be reached by other nodes. In this example, three levels of nodes were obtained (illustrated in Fig.1 (f)). The oriented links then connected the nodes from source to sink one based

on the incidence matrix. Notably, the rounded rectangles in Fig.1 (f) indicate the retrieved clusters, in which the information flow forms a loop.



(e) Rearranged matrix R

(f) Hierarchical interaction graph of the system

Figure 1 e-f. Stepwise procedure of ISM

3.3 Analysis procedure

The Analysis procedure comprises three main phases: market planning, QFD and the ISM approach. Figure 2 presents the flow diagram for linking these phases. The first phase begins with product market planning which clarifies the various requirements of different markets. The second phase involves the QFD analysis, during which the variant requirements are related to physical components with specific values to identify relationship degree, yielding the relative importance of each component towards the market variations. Finally, the inner interactions between physical components are further examined via ISM analysis, with component design priority being represented using a hierarchical graph. The result obtained from QFD is incorporated into the hierarchical graph to identify the component to be redesigned in the influential path, deriving new products that satisfy market niches by redesigning finite components.

4. Case Study for Variant Design Based on Component Interaction Graph

4.1 Case background

This study illustrated the design of a family of 1.5-liter automatic drip coffee makers from an electronic appliances manufacturer (Company X). Ninety-five percent of the products of this company are original design manufactured (ODM), and are mainly exported to America, Europe, and Japan. Company X aims to provide product varieties to simultaneously meet the requirements of each segmented market, as well as to develop product architectures in mass customization. Components of the original product are listed in Table 1.

4.2 Analysis procedure

Phase 1 : Market Planning

The market planning aims at two different markets (spatial variety) with two different launch times (temporal variety), concurrently developing four products, as illustrated in Fig. 3. The launch time of the "current" products is planned for after three months, while that of "future" products is planned for after eight months.

Phase 2: Identify the exterior drivers of variation

To emphasize market differentiation, the QFD matrix lists the differences in customer requirements rather than common requirements. In the case, how to maintain coffee temperature is the key driver for spatial market differentiation, because the weather in Market 2 is much colder than that of Market 1. Table 1 illustrates the mapping from requirements into components, in which the values 9, 5, 3, 1, and 0 indicate the mapping relationships ranging from very strong, through to strong, ordinary, weak, and none, respectively. Table 1 demonstrates that the most important component for Keeping coffee temperature is the Carafe. Furthermore, the key drivers for temporal market differentiation are Ease of cleaning, Comfortable to use, and Fashionable style. These requirements are listed in Table 2, along with their relative importance. The critical components for these requirements include the Housing, Top cover, and Carafe. The QFD results are input into the product design, as described in Section *4.3*.

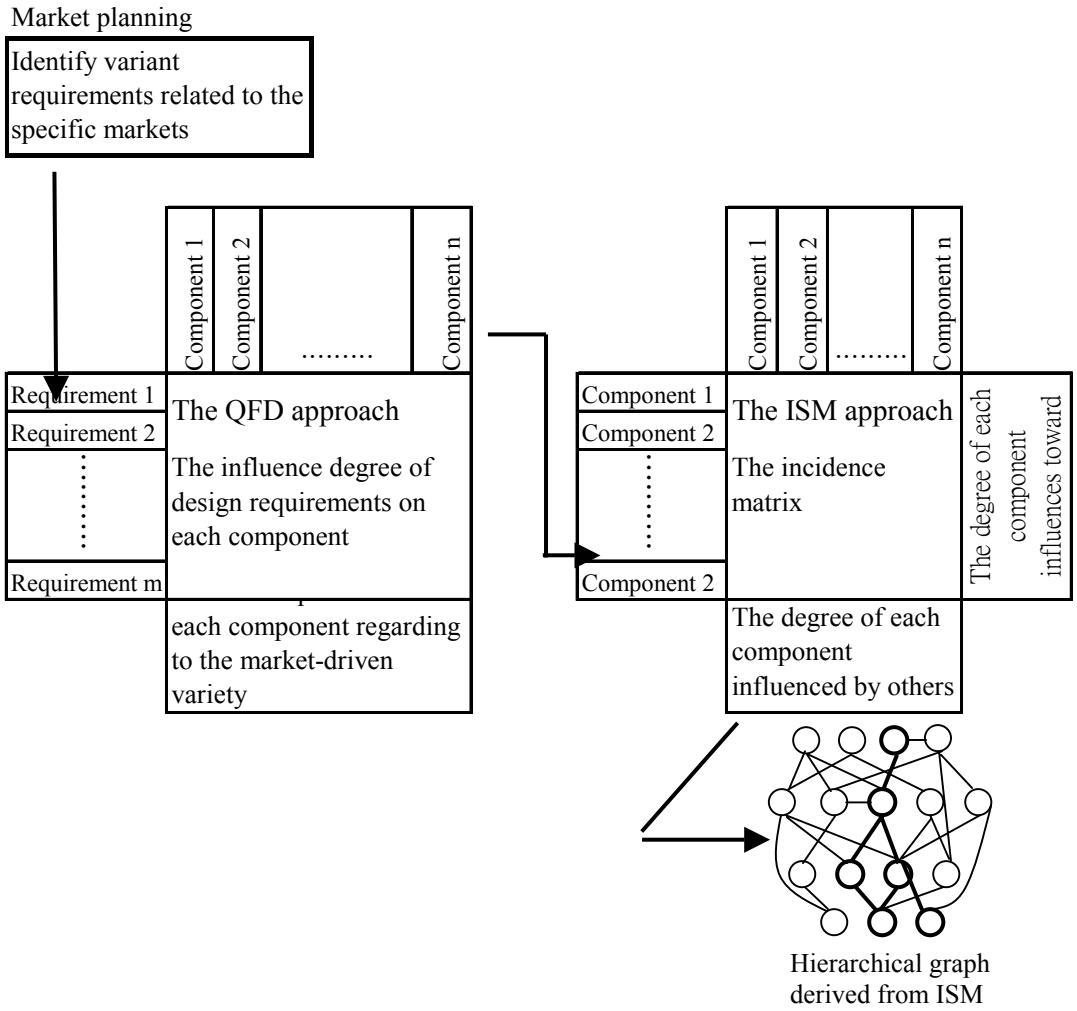


Figure 2. Flow diagram of the analysis phases

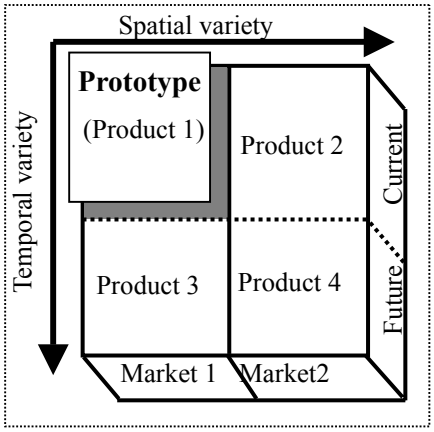


Figure 3. Market planning of the coffee maker

Component / Spatial differentiation requirement	top Cover	top Cover base	spout	spout seat	top cover base	water tank cover	water tank	base	silicone ring	water outlet pipe	pipe connection seat	base cover	packing valve	heating element	switch	hot plate ring	hot plate	cup bank	carafe handle cover	carafe handle	carafe	carafe cover	housing	filter holder packing valve	filter holder
	1	2	3	4	5	6	8	11	12	13	14	15	16	18	19	20	21	22	23	24	25	26	27	28	30
Keep coffee temperature	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	1	3	0	0	0	**9**	3	0	0	0

Table 1. QFD matrix of the spatially differential requirements

Component No. / Temporal differentiation requirement	1	2	3	4	5	6	8	11	12	13	14	15	16	18	19	20	21	22	23	24	25	26	27	28	30	Rwlative Weight
Ease of cleaning	**1**	3	0	0	3	1	3	0	0	0	0	0	0	0	0	0	0	0	1	0	**5**	1	**1**	0	5	3
Comfortable to use	**9**	0	1	1	1	3	0	0	0	0	0	0	0	0	3	0	0	0	1	3	**3**	1	**9**	0	3	5
Fashionable style	**3**	0	0	0	0	1	3	3	0	0	0	0	0	0	1	0	0	0	3	3	**9**	1	**5**	0	0	2
Total	**54**	9	5	5	14	20	15	6	0	0	0	0	0	0	17	0	0	0	14	21	**48**	10	**58**	0	30	

Table 2. QFD matrix of the temporally differential requirements

Phase 3: Identify the interior hierarchical interactions

In this approach, senior design engineers of company X perform the incidence matrix by investigating the relationships between each pair of components. Table 3 lists the original incidence matrix. The cells in the incidence matrix are marked with "1" if the components in rows constraint the specifications of the components in columns. The related design constraints are documented in the form *d (i, j)*, where *i* denotes the source component providing a constraint to component *j*. For example, *d (4, 5)* indicates that the Top Cover Base (component 5) should fit the diameter of the Spout Seat (component 4). This incidence matrix is then manipulated through the ISM procedures illustrated in Section *3.2*. Fig.4 shows the hierarchical graph of the design constraint flow derived through ISM. In this graph, the circles represent components, the oriented lines are design constraints provided by the source components, and the rounded

rectangles indicate that a set of mutually interactive components, which are integrated as a module. These modules and other components then are further grouped into chunks according to the frequency of their interactions. Table 4 lists the incidence matrix after appropriate rearrangement of the order. Four chunks are formed in the product, namely C1 housing chunk, C2 water tank chunk, C3 base chunk, and C4 carafe chunk. The precedence of the four chunks is determined by the inter-chunk interactions.

Part Name	No.	1	2	3	4	5	6	8	11	12	13	14	15	16	18	19	20	21	22	23	24	25	26	27	28	30
top cover	1		1																							
top cover set	2	1																								
spout	3		1		1																					1
spout seat	4			1		1																				
top cover base	5	1																						1		1
water tank cover	6							1																		
water tank	8						1																			
base	11							1					1			1	1	1						1		
silicone ring	12								1		1	1														
water outlet pipe	13								1	1		1														
pipe connection seat	14									1	1			1												
base cover	15								1																	
packing valve	16											1														
heating element	18															1		1								
switch	19								1																	
hot plate ring	20																	1								
hot plate	21																1									
cup bank	22																				1					
carafe handle cover	23																				1					
carafe handle	24																		1	1			1			
carafe	25																	1	1		1		1	1		
carafe cover	26																				1					
housing	27					1																				
filter holder packing valve	28																									1
filter holder	30					1																		1		

Table 3. The original incidence matrix of coffee maker components

4.3 Design procedure

The results of the analysis illustrated in previous section are applied in the product design; four products were designed concurrently to satisfy requests of different markets. The design procedure is demonstrated in the following paragraphs.

Chunk	module/component	No.	1	2	3	4	28	30	5	27	8	6	11	15	19	21	20	14	12	13	16	18	22	23	24	26	25
C1		1		1																							
	Top cover module	2	1																								
		3		1		1		1																			
	Spout module	4			1				1																		
	holder packing valve	28						1																			
		30							1	1																	
		5	1					1		1																	
	Filter module	27							1																		
C2		8									1																
	Tank module	6									1																
C3		11								1	1			1	1	1	1										
		15											1														
	Base module	19											1														
		21															1										
	Heating plate module	20														1											
		14																		1	1	1					
		12											1						1		1						
		13											1						1	1							
	Water pipe module	16																	1								
	Heating element	18													1	1											
C4		22																							1		
		23																							1		
		24																						1	1		1
	Carafe outfit module	26																								1	
	Carafe	25								1						1							1		1	1	

Note: Grayed cells indicate the inter-chunk interactions.

Table 4. Incidence matrix after appropriate rearranging the order

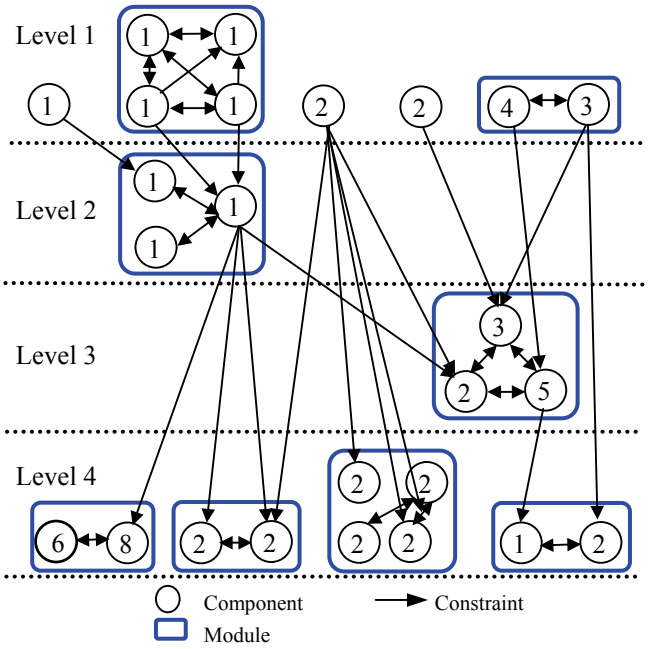


Figure 4. Hierarchical graph of component interaction

Phase 1: Design for spatial variety

Table 1 indicates that Carafe (part No.25) design is essential for maintaining coffee temperature. Therefore, the Carafe is redesigned to meet the requirement: the wall should be thickened and use heat insulation material, the shape slenderized and the top narrowed to reduce heat loss. To identify the influence of the new Carafe design, Fig.5 (extracted from Fig.4) shows the incidence diagram of the Carafe. In this figure, the design constraints the Carafe exports to the sink nodes are listed below:

***d(25, 21)*:** The Heating Plate module should fit the diameter of Carafe base (fixed).

***d(25, 22)*:** The Cup Bank should fit the diameter of Carafe body (changed).

***d(25, 24)*:** The Carafe Handle should fit the arc and weight of Carafe body (changed).

***d(25, 26)*:** The Carafe Cover should fit the diameter of Carafe rim, and the re quested thermal condition (changed).

***d(25, 27)*:** The Filter Module should fit the Carafe height (changed).

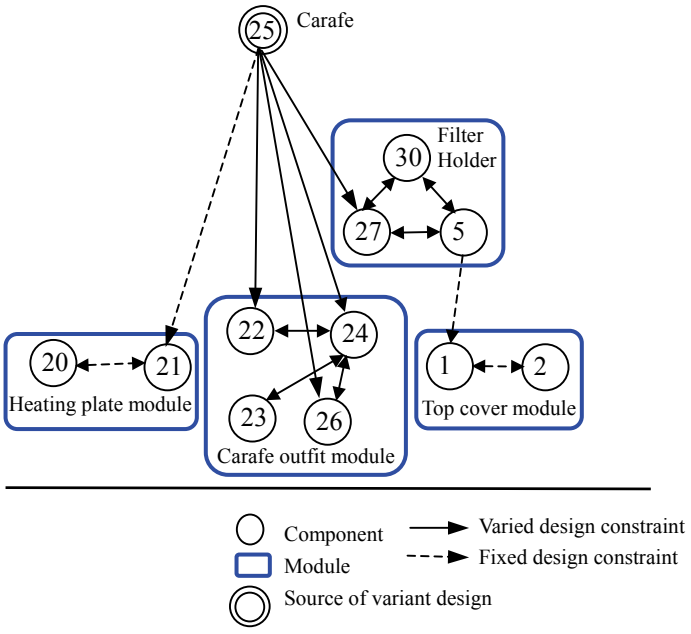


Figure 5. Incidence diagram of carafe (component 25)

The constraint of *d(25, 21)* is fixed (represented as dotted line in Fig.5), and thus parts 20, 21 are left unchanged. However, constraints *d(25, 22)*, *d(25, 24)*, *d(25, 26)*, and *d(25, 27)* are changed (represented as solid lines in Fig. 5) owing to the new carafe specification, resulting in the design of the Filter Module and Carafe Outfit Module having to be changed to match the altered conditions. In the Carafe Outfit Module, the components are redesigned to fit the new Carafe. However, the design change of the Filter Module must refer not only to the Carafe, but also to other components that provide constraints on the Filter Module, as shown in Fig.6. Thus in redesigning the Filter Module, the constraint from the Carafe becomes the source of variant design (represented as solid line in Fig.6), while the others are fixed constraints (represented as dotted lines in Fig. 6) listed below:

***d(11, 27)*:** The Housing should fit the Base.

***d(28, 30)*:** The Filter Holder should fit the Filter Holder Packing Valve diameter.

***d(4, 5)*:** The Top Cover Base should fit the Spout Seat diameter.

***d(3, 30)*:** The Filter Holder should fit the Spout shape.

Under these constraints, the design of Filter Module (parts 27, 5, and 30) is changed from V-shaped to U-shaped to fit the new Carafe design.
Furthermore, constraint from the Filter Module is:

***d(5, 1)*:** The Top Cover should fit the Basket Holder rim diameter.

Since the specification of the Basket Holder rim is fixed, component 1 and 2 need not change their design. Consequently, Table 5 lists the design solution driven by spatial market differentiation.

No.	Redesigned component
25*	Thermal carafe
22*	Cup bank of thermal carafe
23*	Handle cover of thermal ca-
24*	Handle of thermal carafe
26*	Cover of thermal carafe
27*	U-shaped housing
5*	U-shaped cover base
30*	U-shaped filter baseket

Table 5. List of variant components for Market 2

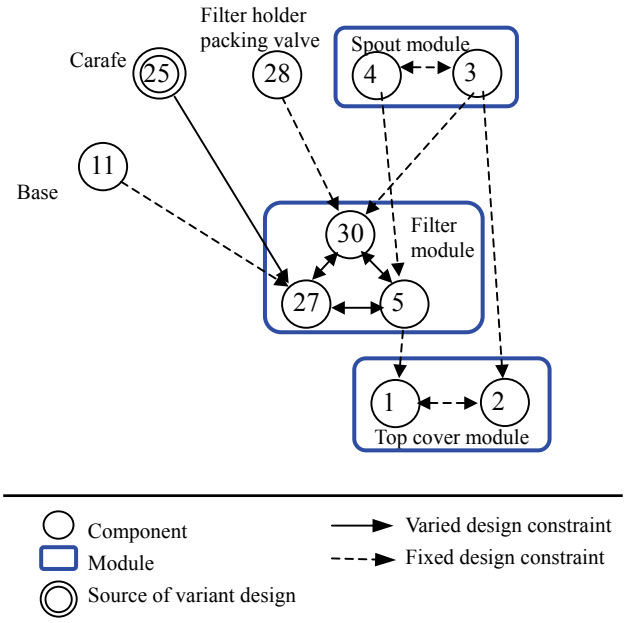


Figure 6. Constraint flow diagram of the filter module

Phase 2: Design for temporal variety

Table 2 indicates that the critical components for realizing temporal variety are the Housing (part 27), Top Cover (part 1), and Carafe (part 25). According to the hierarchical graph in Fig. 4, for these three components, the Carafe occupies the upper level in the interaction hierarchy. This arrangement means that the Carafe design should be addressed first, followed by that of the Housing and finally, the Cover. However, the incidence and costs involved in carafe redesign are quite high. The strategy of Company X thus is to "over design" this component; that is, to improve the quality of the current specifications capable of handling future market requests. Therefore, the Carafe is upgraded for easy cleaning, pouring and dishwasher-safe in both the "current" and "future" versions. Therefore, according to the design priority, the product variety should focus on redesigning the Housing (part 27). To facilitate usability, the design team tends to substitute swing-out housing for the fixed housing. This change divides the component into two new parts; namely, the Swing-out Filter Housing and the Support. The Swing-out Filter Housing is further differentiated into either U-shaped or V-shaped. The original design constraints of the Housing are laid on the two new parts, respectively (see Fig. 7). Thus the shape of the Swing-out Housing must fit the Carafe; and the design of the Support must

fit the Base. The variant design of the Housing directly influences the Top Cover (part 1); meanwhile, for convenient to use, the Top Cover is changed from a lift up to a fixed design.
Finally, Table 6 lists the variant design driven by temporal market differentiation.

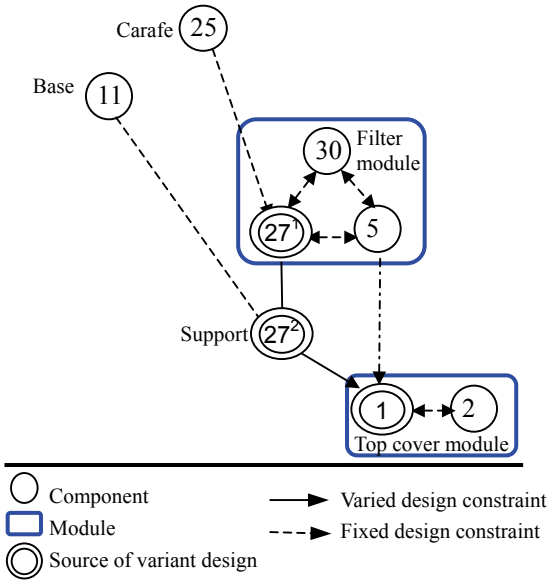


Figure 7. Constraint flow diagram of the new design

No.		Redesigned component
27*[1]		Swing out filter basket
27*[2]		Support
1*		Fixed top cover

Table 6. List of variant components for Future market

4.4 Result

Table 7 lists the components of the four products derived via the proposed methodology. Among these components, most of the variety occurs in chunks 1 and 4, while chunks 2 and 3 remain virtually unchanged, and thus are considered "platforms" of this product architecture. Moreover, the design team further suggested that components of the upper levels of chunk 3, including Water Pipe Module, Heating Element, Base Module, and Heating Plate Module, should be standardized to reduce the redesign effort and production cost.

Chunk	Component	No.	Product 1	Product 2	Product 3	Product 4
	Fastened top cover	1	V	V		
	Lifted-up cover	1*			V	V
	Top Cover base	2	V	V	V	V
	Spout	3	V	V	V	V
	Spout seat	4	V	V	V	V
	Filter holder packing valve	28	V	V	V	V
	V-shaped filter holder	5	V		V	
	U-shaped filter holder	5*		V		V
	V-shaped filter basket	30	V		V	
	U-shaped filter basket	30*		V		V
	V-shaped fixed housing	27	V			
	U-shaped fixed housing	27*		V		
	V-shaped swing out filter housing	27*			V	
	U-shaped swing out filter housing	27*				V
C1	Support	27*			V	V
	Water tank cover	6	V	V	V	V
C2	Water tank	8	V	V	V	V
	Silicone ring	12	V	V	V	V
	Water outlet pipe	13	V	V	V	V
	Pipe connection seat	14	V	V	V	V
	Packing valve	16	V	V	V	V
	Heating element	18	V	V	V	V
	Base	11	V	V	V	V
	Base cover	15	V	V	V	V
	Switch	19	V	V	V	V
	Hot plate ring	20	V	V	V	V
C3	hot plate	21	V	V	V	V
	Glass carafe	25	V		V	
	Thermal carafe	25*		V		V
	Cup bank of glass carafe	22	V		V	
	Cup bank of thermal carafe	22*		V		V
	Handle cover of glass carafe	23	V		V	
	Handle cover of thermal carafe	23*		V		V
	Handle of glass carafe	24	V		V	
	Handle of thermal carafe	24*		V		V
	Cover of glass carafe	26	V		V	
C4	Cover of thermal carafe	26*		V		V

Table 7. Components list of the product family

4.5 Comparison of existing and proposed designs

A team of engineers and managers of Company X estimated the sales volume, marketing, variable (raw material/ production prices) and fixed (engineering/ injection mold) costs for the proposed designs, and compared these estimates to those for products designed independently. Table 8 lists the comparison.

The profit is calculated using the following function:

$$P_i = S_i(PR_i - VC_i) - FC_i - MC_i \quad (1)$$

Where P_i, S_i, PR_i, VC_i, FC_i, MC_i are the profit, sales volume, price, variable cost, fixed cost, and marketing cost of product *i*, respectively.

Table 8 illustrates that the primary cost difference between the two design strategies lies in the fixed cost. The proposed designs significantly reduced the fixed cost for developing new products through sharing most components. The second row from the bottom shows that the profits associated with independently developing products 2 and 4 is minus 73% and 65% of current product, respectively. Therefore, the best decision seems to be not to develop any product in Market 2. However, the proposed designs generate a total profit 127% in current markets and 541% in future markets higher than if product 1 was the only product launched. The result shows the potential savings and profit available using this methodology.

	No product family design				Product family design using this methodology			
% of current product	Product 1	Product 2	Product 3	Product 4	Product 1	Product 2	Product 3	Product 4
Sales volume	100	80	100	80	100	80	100	80
Price	100	120	115	130	100	120	115	130
VC	100	145	105	145	100	140	100	140
FC	100	100	100	100	100	40	10	10
MC	100	200	100	100	100	200	100	100
Profit	100	-73	208	-65	100	27	334	207
Total profit	current=	27	future=	143	current=	127	future=	541

Note: VC, FC, MC are the variable, fixed and marketing costs, respectively.

Table 8. Comparison of independently developed and the proposed designs

4.6 Design strategies based on the component interaction graph approach

The hierarchical graph could optimize variant design in the following design strategies:

1. Design strategies for the source components:

- a: Differentiated customer requirements directly drive design variation of these components. And since the incidence and effort for the design changes are relative huge, this variation must be obvious and valuable to customers. To achieve stable product architecture, "over-design" of dominant components might be unavoidable for extending component lifecycle.
- b: If the components are remained unchanged, they should be considered to be standardized or fixed specification, and become core platform of the product family.

2. Designs of the sink components are more likely to change to comply with both the altered design constraints and the requests of customer requirements. However, the cost and incidence of altering these components is relatively low. Furthermore, the redesign should incorporate the constraints provided by the source components.

3. Components that with their specification flow forms an interaction loop are likely to be further modularized or integrated.

5. Product Design Based on Analytic Network Process

5.1 The rational for using the ANP approach for optimizing product design

The approach in sections 3 and 4 illustrated a product variety design based on the component relationship structure graph. The graph forms design constraint flows from source components to the sink ones. However, in some ill-structured product architectures, the component relationships may form a

complicated network, and the ISM approach may not be applied successfully. Therefore, we developed an integrated approach via analytic network process (ANP) (Saaty, 1996) technique to fix this problem. The differences between ISM and ANP are that (1). ANP treats component relationship as relative importance (from 0 to 1) rather than binary; (2).ANP mathematically copes with the network structure well, while the ISM, hierarchical structure.

ANP is a general form of the widespread multi-criteria decision technique, AHP (analytic hierarchy process) (Saaty, 1990). AHP employs unidirectional hierarchical relationship among levels, while ANP enables consideration of the interrelationships among the decision levels and attributes. The distinguishing features of ANP make it suitable for dealing with the hierarchical mappings as well as component coupling problems in determining the influence of variety on each design element. In this approach, the analysis result of ANP is then input to the goal programming (GP, Dantzig 1963) models for determining the standardized and variant parts of product architecture. The GP model handles multiple objectives and minimizes deviation from desired goals, and thus provides a feasible and consistent solution for optimizing product family design. Although many researchers use mathematic models, such as (Reiter & Rice 1966, Ringuest & Graves 1989) most methodologies are assumed independent among design alternatives. In this study, we integrated ANP and GP approaches for accommodating interdependence among design alternatives that is first applied in the product variety optimization problem.

5.2 Computational procedure of the ANP

The procedure of optimizing design variety via the ANP was summarized as follows: The first step was to estimate the qualitative changes in customer requirements (CRs) in each future market compared to the current product. The importance of the CRs were compared and calculated, corresponding to the first step of the matrix manipulation concept of ANP. The CRs were then deployed into engineering characteristics (ECs) by comparing the ECs with respect to each CR. The ECs were further deployed into components by comparing the relative contributions of components to each EC. Finally, the interdependence priorities of the components were further examined by analyzing the couplings among components. The supermatrix utilized to model the procedure in matrix notation, which is formed from four submatrices, is

constructed as follows:

$$\begin{array}{c} \\ \text{Goal(G)} \\ \text{Customer Requirements(CRs)} \\ \text{Engineering Characteristics(ECs)} \\ \text{Components (C)} \end{array} \begin{array}{c} \begin{array}{cccc} \text{G} & \text{CRs} & \text{ECs} & \text{C} \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 \\ W1 & 0 & 0 & 0 \\ 0 & W2 & 0 & 0 \\ 0 & 0 & W3 & W4 \end{bmatrix} \end{array} \qquad (2)$$

where W1 denotes a matrix representing the relative importance of CRs for satisfying each specified market goal; W2 represents the mappings of the CRs to each ECs, W3 representing the impact of ECs to each component, and W4 denoting the coupling relationship among components.

Using the above notations, the priorities of the components (Wc) were calculated by multiplying W4 and W3. The overall priorities of the components (W^{ANP}) that reflect the degree of required changes of components in response to the niche of each market, then were calculated by multiplying Wc, W2, and W1.

6. Case Study for Optimizing Product Variety Using the ANP Approach

This section presented an illustrative example of a water cooler family design (Martin & Ishii, 2002). The proposed methodology was further demonstrated using a stepwise form.

6.1. Survey customer requirements and segment the future markets

Product variety planning begins with surveying customer requirements. Figure 8 illustrated three future markets defined by the design team, along with the desired product features in these envisioned markets.

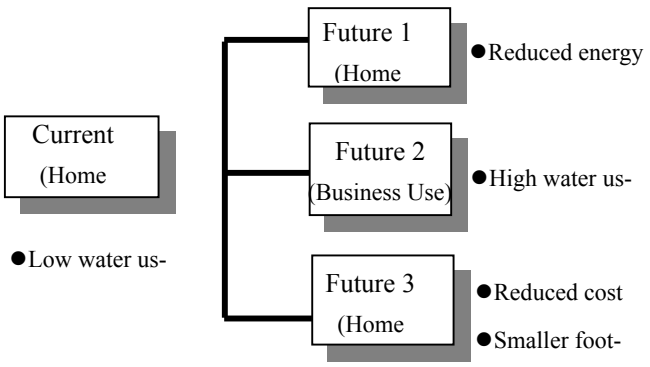


Figure 8. Market planning of the water cooler for three envisioned markets.

6.2 The ANP approach

Phase 1. Estimate relative importance of CRs in each market

For this water cooler example, the main CRs were Fast Cooldown, High Capacity, Low Energy Consumption, Compact, Rapid Pouring, and Low Cost. Figure 8 depicts the desired product features for each market. According to the planning, the design team estimated the range of changes of the CRs for each market using using Saaty's 1-9 scales (Saaty, 1990) pairwise comparisons as shown in Table 9. To avoid comparison inconsistencies, a consistency ratio measured the probability that the pairwise comparison matrix was randomly filled. The upper limit for the consistency ratio was 0.1, which signified that up to 10% chance was tolerable for the comparison conducted in random manner. The procedure was applied in each market. The resulting relative weights of CRs compose W1, as shown in Eq. (3).

$$W1 = \begin{array}{c} \begin{matrix} \text{M1} & \text{M2} & \text{M3} \end{matrix} \\ \begin{bmatrix} 0.071 & 0.300 & 0.045 \\ 0.071 & 0.300 & 0.045 \\ 0.643 & 0.033 & 0.045 \\ 0.071 & 0.033 & 0.409 \\ 0.071 & 0.300 & 0.045 \\ 0.071 & 0.033 & 0.409 \end{bmatrix} \end{array} \begin{array}{l} \text{Fast cooldown} \\ \text{High capacity} \\ \text{Low energy use} \\ \text{Compact} \\ \text{Rapid pouring} \\ \text{Low cost} \end{array} \tag{3}$$

where M1, M2, M3 represent Markets 1, 2, and 3, respectively.

Future Market 3		Fast cooldown	High capacity	LEC	Compact	Rapid pouring	Low cost	Relative weight
Fast cooldown		1	1	1	1/9	1	1/9	0.045
High capacity			1	1	1/9	1	1/9	0.045
Low energy consumption (LEC)				1	1/9	1	1/9	0.045
Compact					1	9	1	0.409
Rapid pouring						1	1/9	0.045
Low cost							1	0.409
Consistency Ratio=	1.60228E-09							

Table 9. Pairwise comparison matrix of CRs for the goal of Market 3.

Phase 2. Translating CRs into ECs

The ECs used in the product design include Cool Down Time (min), Cool Water Volume (gal), Power Consumption (W), Width, Depth (in), Volume Flow Rate (gal/min), and Cost ($). If a CR was fulfilled via two or more ECs, the design team was required to conduct a pairwise comparison to assess the relative importance of the ECs with respect to the CR. Table 10 maps the relations between CRs and ECs. For example, In column 5 of Table 10, two ECs (Width and Depth) specify the request of Compact specification of equal importance, thus, their weighted values were both 0.5.

W2	Fast cooldown	High capacity	LEC	Compact	Rapid pouring	Low cost
Cool down time(min)	1.000	0.000	0.000	0.000	0.000	0.000
Cold water volume(gal)	0.000	1.000	0.000	0.000	0.000	0.000
Power consumption(W)	0.000	0.000	1.000	0.000	0.000	0.000
Width(in)	0.000	0.000	0.000	0.500	0.000	0.000
Depth(in)	0.000	0.000	0.000	0.500	0.000	0.000
Volume flow rate(gal/min)	0.000	0.000	0.000	0.000	1.000	0.000
Cost($)	0.000	0.000	0.000	0.000	0.000	1.000

Table 10. Matrix W2, the mappings of CRs to the relative ECs.

Phase 3. Deploying the ECs to product components

Again, the design team performed AHP to evaluate the relative importance of the components' contribution to each EC, and the aggregation of relative importance weights for components in each EC formed matrix W3, as shown in Table 11. In which the zeros were assigned to the cells if the EC had no effect on the components.

Phase 4. Examining inner dependences among components

In this case, the components are seriously coupled. The degree of the coupling relations between components was identified using a series of pairwise comparisons. Table 12 displays the inner dependence matrix of components with the Fan as controlling component, in which Plumbing and Insulation were excluded because of not impacting the Fan. The schema was performed in each component, and obtained the resulting eigenvectors as shown in Table 13. The matrix indicated the inner dependence among components, in which zeros indicated the eigenvectors of the unrelated components.

W3	Cool down time	Cold water volume	Power consumption	Width	Depth	Volume flow rate	Cost
Fan	0.115	0.000	0.143	0.000	0.000	0.000	0.000
Heat Sink	0.231	0.000	0.000	0.000	0.000	0.000	0.071
TEC	0.115	0.000	0.429	0.000	0.000	0.000	0.000
Power Supply	0.038	0.000	0.429	0.000	0.000	0.000	0.071
Chassis	0.000	0.000	0.000	0.500	0.500	0.000	0.214
Plumbing	0.000	0.000	0.000	0.000	0.000	0.900	0.000
Reservoir	0.231	1.000	0.000	0.000	0.000	0.100	0.214
Insulation	0.038	0.000	0.000	0.000	0.000	0.000	0.000
Fascia	0.231	0.000	0.000	0.500	0.500	0.000	0.429

Table 11. Aggregation of relative importance for components in each EC

Fan	Fan	HS	TEC	PS	Chassis	Reservoir	Fascia	Relative Weights
Fan	1	4	9	6	4	9	9	0.477
Heat Sink(HS)		1	3	3/2	1	5	5	0.154
TEC			1	1/2	1/3	3/2	1	0.048
Power Supply (PS)				1	2/3	3	5/2	0.096
Chassis					1	5	4	0.148
Reservoir						1	4/3	0.034
Fascia							1	0.042
Consistency Ratio=	0.013							

Table 12. Pairwise comparison matrix with the Fan as controlling component.

W4	Fan	HS	TEC	PS	Chassis	Plumbing	Reservoir	Insulation	Fascia
Fan	0.477	0.087	0.000	0.059	0.097	0.000	0.000	0.066	0.125
Heat Sink	0.154	0.498	0.064	0.000	0.145	0.000	0.021	0.000	0.021
TEC	0.048	0.086	0.625	0.092	0.000	0.000	0.056	0.038	0.000
Power Supply	0.096	0.000	0.125	0.673	0.074	0.000	0.000	0.000	0.048
Chassis	0.148	0.167	0.000	0.093	0.276	0.000	0.239	0.000	0.262
Plumbing	0.000	0.000	0.000	0.000	0.000	0.664	0.118	0.000	0.142
Reservoir	0.034	0.067	0.121	0.000	0.253	0.165	0.448	0.373	0.000
Insulation	0.000	0.035	0.064	0.000	0.026	0.050	0.118	0.523	0.021
Fascia	0.042	0.061	0.000	0.082	0.129	0.121	0.000	0.000	0.381

Table 13. Aggregation interdependence matrix among components.

Phase 5. Synthesis the overall priorities of components

According to Eq.(2), the interdependent priority of the components, W_c, was calculated as

$$Wc = W4 \times W3 \qquad (4)$$

The overall priorities of the components regarding the goals of the three markets were calculated as follows:

$$W^{ANP} = W_c \times W2 \times W1 = \begin{array}{c} \begin{matrix} \text{M1} & \text{M2} & \text{M3} \end{matrix} \\ \begin{bmatrix} 0.082 & 0.042 & 0.089 \\ 0.056 & 0.059 & 0.077 \\ 0.216 & 0.064 & 0.032 \\ 0.244 & 0.035 & 0.078 \\ 0.107 & 0.150 & 0.231 \\ 0.067 & 0.241 & 0.100 \\ 0.113 & 0.249 & 0.153 \\ 0.040 & 0.076 & 0.039 \\ 0.074 & 0.082 & 0.198 \end{bmatrix} \end{array} \begin{array}{l} \text{Fan} \\ \text{Heat sink} \\ \text{TEC} \\ \text{Power Supply} \\ \text{Chassis} \\ \text{Plumbing} \\ \text{Reservoir} \\ \text{Insulation} \\ \text{Fascia} \end{array} \quad (5)$$

where M1, M2, M3 represent Markets 1, 2, and 3, respectively.
The ANP result revealed the priority for redesigning components to satisfy market goals. For example, in Market 1, the first component requiring redesign was Power Supply, with a relative importance value of 0.244, whereas Reservoir and Chassis were identified as the most important components in Markets 2 and 3 with relative importance values of 0.249 and 0.231, respectively.

6.3 Optimization

The optimization of the product architecture is to achieve a stable product platform that enable variant products to be highly differentiated yet share as many substantial portions of their components as possible, thus reducing the manufacturing and design costs.

Phase 1: Platform component selection

There are two considerations in selecting the platform components. First, components with high engineering costs should be the initial focus. Second, a product platform stresses on component commonality; therefore, the components with low W^{ANP} factors -which are less sensitive and more stable in response to the changing environment, are suitable as platform items. Therefore, a weighted GP (Schniederjans, 1995) algorithm is utilized for selecting platform components that satisfy two goals: (1) high engineering cost, and (2) control the W^{ANP} weight loss under a tolerable ratio. Furthermore, to consider the relative importance of different markets and to regulate the possible incommensurability problem of different goals (Ringuest & Graves, 1989), the general GP is as follows:

$$\min \quad \omega_1^{cost}\left(\frac{d_1^-}{\sum_{i=1}^{n} c_i}\right)+\omega_2^{ANP}\left(\frac{d_2^+}{\lambda}\right)$$

subject to

$$\sum_{i=1}^{n} c_i x_i + d_1^- - d_1^+ = \sum_{i=1}^{n} c_i, \tag{6}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\sigma_j w_{ij}^{ANP} x_i + d_2^- - d_2^+ = \lambda,$$

$$\sum_{j=1}^{m}\sigma_j = 1,\ x_i \in \{0,1\},\ \textbf{\textit{i=1,2,…n;}}$$

$$\textbf{\textit{j=1,2,…m}}\ d_1^-, d_1^+, d_2^-, d_2^+ \geq 0,\ ;\ \lambda \leq 1\ \textbf{\textit{;}}$$

where ω_1^{cost}, ω_2^{ANP} denote the importance weights, d_1^-, d_1^+, d_2^- and d_2^+ denote the negative and positive deviation variables of the goals, respectively; x_i is the binary variable representing whether the ith component is assigned as a platform item (if x_i =1) or not (when x_i =0), c_i denotes the engineering cost of the ith components, σ_j denotes the relative importance of market j, w_{ij}^{ANP} represents the ith component weight in the jth market, and λ is a controllable variable indicating the tolerable ratio of weight loss.

Phase 2: Variant component selection

This phase considered the distinctiveness of each product for satisfying specific market needs. Therefore, certain components were selected redesigned achieve the distinctiveness under limited design budget. Therefore, the GP was employed to satisfy two goals: (1) select the components with high W^{ANP} factors, and (2) control the cost under a budget. Following the same principle of regulation incommensurability, the general GP is as follows:

To select the redesigned components for market *j*:

$$\min \quad \omega_1^{ANP}\left(\frac{d_1^-}{\sum_{k=1}^{n} w_{jk}^{ANP}}\right)+\omega_2^{budget}\left(\frac{d_2^+}{B_j}\right)$$

subject to

$$\sum_{k=1}^{n} w_{jk}^{ANP} x_k + d_1^- - d_1^+ = \sum_{k=1}^{n} w_{jk}^{ANP}, \qquad (7)$$

$$\sum_{k=1}^{n} c_k x_k + d_2^- - d_2^+ = B_j,$$

$x_k \in \{0,1\},\ d_1^-, d_1^+, d_2^-, d_2^+ \geq 0,\ j=1,2,\ldots m;\ k=1,2,\ldots n$

$k \neq i$ if the ith component has been assigned as a platform item

where ω_1^{ANP} and ω_2^{budget} denote the importance weights, and d_1^-, d_1^+, d_2^- and d_2^+ represent the negative and positive deviation variables of the first and second goals, respectively; x_k represents a binary variable representing whether the kth component is assigned as a redesigned item (if x_k =1) or not (x_k =0). Notably, the variable x_k should not contain components that have been determined as platform items. w_{jk}^{ANP} is priority rating of the kth component in the jth market, c_k denotes engineering cost of the kth component, and B_j represents design budget of the jth market.

6.4 Result

Table 14 lists the engineering cost for redesigning each component. The data and the W^{ANP} weight in Eq.(5) is input into the GP models via LINDO software. The platform components selected by the GP under variant weight loss (variable λ) are shown in Table 14. After examining the solutions, the design team strategically set the weight loss at 20%, yielding Fan, Heat Sink, and Insulation as the components shared across the product family. Furthermore, the GP model of Eq.(7) was applied for selecting the redesign components in the three envisioned markets , yielding the result listed in Table 15, in which the GP solutions identified the focuses for redesign as being TEC, Power Supply, Plumb-

ing and Reservoir in Market 1; Chassis, Plumbing and Reservoir in Market 2; and Power Supply, Chassis, Plumbing and Fascia in Market 3.

Variable		Component	Redesign cost$		GP solutions				
x_1		Fan	10,000			V	V	V	V
x_2		Heat Sink	200,000		V	V	V	V	V
x_3		TEC	20,000				V	V	V
x_4		Power Supply	3,000						
x_5		Chassis	1,000					V	V
x_6		Plumbing	2,000						
x_7		Reservoir	10,000						
x_8		Insulation	3,000			V	V		V
x_9		Fascia	2,000						
λ					10%	20%	30%	40%	50%

Table 14. Platform components selected under variant weight loss (λ).

Variable	Component	GP Solutions		
		Market 1	Market 2	Market 3
x_3	TEC	V		
x_4	Power Supply	V		V
x_5	Chassis		V	V
x_6	Plumbing	V	V	V
x_7	Reservoir	V	V	
x_9	Fascia			V

Table 15. Components selected for redesign in three markets.

7. Conclusion

This chapter illustrates the authors' current studies on managing product variety in different degrees of product architecture maturity (Liu & Hsiao 2005, Hsiao & Liu 2005). We suggested that the occasion in implementing the first

approach (sections 3, 4) is when product architecture is under constructed; the interactions of components have not been investigated. Applications of the approach to product architecture provide the hierarchical graph of component interactions. Furthermore, the methodology provides the following advantages for developing product family:

1. The methodology clarifies the specification flow between components rather than merely symmetric relationship "similarity" or "correlation". Thus the necessary information is provided for determining not only clustering but also precedence among components.

   The incidence matrix with the documented design constraints provides a computable way for design knowledge representation.

2. The hierarchical graphical diagram provides designers with a user-friendly display for clarifying the influence of each component variation.

3. The hierarchical structure along with the QFD analysis helps product family developers to identify whether the components should be standardized, altered, or modularized.

Furthermore, the occasion in implementing the second approach (sections 5, 6) is when component interactions have been clearly defined and formed complicated networks, a flexible and comprehensive decision support system is needed in trading off the product variety, standardization, and resource utilization. In the approach, The interdependent nature inherent in the product design process was considered using the ANP approach. The use of ANP weights, and resource limitations in the multi-objective goal programming provided feasible and more consistent solutions, thus yielding the optimal solutions in determining the platform component as well as the variant components focused on during the redesign phases. The application of the methodologies presented in this chapter can easily be extended to include additional decision criteria, such as the manufacturability, sustainability, and assembly in designing product families. Subsequent research will address these points.

8. References

Cohen, L. (1995). Quality Function Deployment: How to Make QFD Work for

You, Addison-Weesley Reading, ISBN: 0201633302, Massachusetts.

Dantzig, G.B. (1963). Linear programming and Extensions. The RAND Corporation, ISBN: 0691080003, West Sussex.

Erens, F.J. (1996).The synthesis of variety: developing product families, Dissertation Eindhoven University of Technology.

Fujita, K. & Ishii, K. (1997). Task structuring toward computational approaches to product variety design. Proceedings of the 1997 ASME Design Engineering Technical Conferences, Paper No. DETC97/DAC- 3766. Sacramento, California. September 1997.

Fujita, K.; Akagi, S.; Yoneda, T. & Ishikawa, M. (1998). Simultaneous optimization of product family sharing system structure and configuration. Proceedings of 1998 ASME Design Engineering Technical Conferences, Paper No. DETC98/DFM- 5722. Atlanta, Georgia. September 1998.

Fujita, K.; Sakaguchi, H. & Akagi, S. (1999). Product variety deployment and its optimization under modular architecture and module commonalization. Proceedings of the1999 ASME Design Engineering Technical Conferences, Paper No. DETC99/DFM-8923. Las Vegas, Nevada. September 1999.

Hauser, J. & Clausing, D. (1988). The House of Quality. Harvard Business Review, Vol.66, No.3, 63-73, ISSN: 0017-8012.

Hsiao, S.-W. & Liu, E. (2005). A Structural component-based approach for designing product family. Computers in Industry, Vol.56, No.1, 13-28, ISSN: 0166-3615.

Liu, E. & Hsiao, S.-W. (2005). ANP-GP Approach for Product Variety Design. International Journal of Advanced Manufacturing Technology. Published On-line First, ISSN: 0268-3768.

Martin, M.V. & Ishii K. (1996). Design for variety: A methodology for understanding the costs of product proliferation. Proceedings of The 1996 ASME Design Engineering Technical Conferences and Computers in Engineering Conference, Paper No. 96-DETC/DTM-1610, Irvine, California, August 1996.

Martin, M.V. & Ishii, K. (1997). Design for variety: Development of complexity indices and design chart. Proceedings of 1997 ASME Design Engineering Technical Conferences. Paper No. DETC97/DFM- 4359. Sacramento, CA. September 1997.

Martin, M.V. & Ishii, K. (2002). Design for variety: developing standardized and modularized product platform architectures. Research in Engineer-

ing Design, Vol. 13, No. 4 , 213-235, ISSN: 0934-9839.

Meyer, M.H. & Utterback, J.M. (1993). The Product family and the dynamics of core capability, Sloan Management Review, Vol. 34, No.3, 29-47, ISSN: 0019-848X.

Pimmler, Y.U. & Eppinger, S.D. (1994). Integration Analysis of Product Decomposition. Proceedings of the ASME Design Theory and Methodology Conference, 68: 343-351. Minneapolis, MN. September 1994.

Pine, B.J. (1993). Mass Customization: The New Frontier in Business Competition, Harvard Business School Press, ISBN: 0875843727, Boston.

Reiter, S & Rice, D.B. (1966). Discrete optimizing solution procedures for linear and nonlinear integer programming problems. Management Science, Vol. 12, No.11, 829-850. ISSN: 0025-1909.

Ringuest, J.L. & Graves, S.B. (1989). The linear multi-objective R&D project selection problem. IEEE Transactions on Engineering Management, Vol.36, No.1, 54-57. ISSN: 0018-9391.

Roberts, F.S. (1997). Discrete Mathematical Models, Prentice-Hall Englewood Cliffs, ISBN: 013214171X, New Jersey.

Robertson, D. & Ulrich, K.T. (1998). Planning for product platforms. Sloan Management Review, Vol.39, No.4, 19-31, ISSN: 0019-848X.

Saaty, T.L. (1990). The Analytic Hierarchy Process. RWS Publications; 2nd edition, ISBN: 0962031720, Pittsburgh.

Saaty, T.L. (1996). Decision making with dependence and feedback : the analytic network process : the organization and prioritization of complexity. RWS Publications, ISBN: 0962031798, Pittsburgh.

Salhieh, S.M. & Kamrani, A.K. (1999). Macro level product development using design for modularity. Robotics and Computer Integrated Manufacturing, Vol.15, No.4 , 319-329, ISSN: 0736-5845.

Sawhney, M.S. (1998). Leveraged high-variety strategies: from portfolio thinking to platform thinking. Jounal of the Academy of Marketing Science, Vol.26, 54-61, ISSN: 0092-0703.

Schniederjans, M.J. (1995). Goal programming: Methodology and applications. Kluwer Academic Publishers, ISBN: 0792395581, Boston.

Simpson, T.W.; Maier, JRA. & Mistree, F. (1999). A Product Platform Concept Exploration Method for Product Family Design. Proceedings of the 1999 ASME Design Engineering Technical Conference, Paper No. DETC99/DTM- 8761. Las Vegas, Nevada. September 1999.

Sosa, M.E.; Eppinger, S.D. & Rowles, C.M. (2000). Designing modular and in-

tegrative systems. Proceedings of the ASME 2000 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Paper No. DETC2000/DTM-14571. Baltimore, Maryland. September 2000.

Suh, N.P. (1990). Principles of Design, Oxford University Press, ISBN: 0195043456, New York.

Ulrich, K.T. (1995) The role of product architecture in the manufacturing firm. Research Policy, Vol. 24, 419-440. ISSN:0048-7333.

Ulrich, K.T. & Eppinger, S.D. (2000). Product Design and Development, McGraw-Hill, ISBN: 0092-0703, New York.

Warfield, J.N. (1973). On Arranging Elements of a Hierarchy in Graphic Form. IEEE Transactions on Systems, Man, and Cybernetics, Vol.3, No.2, 121-132. ISSN: 10834427.

Warfield, J.N. (1995). A Science of Generic Design: Managing Complexity Through System Design (2nd edition), Iowa State Press, ISBN: 0813822475, Iowa.

Wei, M.J.V.; Greer, J.L.; Campbell, M.I.; Atone, R.B. & Wood, K.L. (2001). Interfaces and product architecture. Proceedings of the ASME 2001 International Design Engineering Technical Conferences And Computers and Information in Engineering Conference, Paper No. DETC01/DTM-21689. Pittsburgh, Pennsylvania. September 2001.

Yu, J.S. Gonzalez-Zugasti, J. P. & Otto, K.N. (1998). Product Architecture Definition Based on Customer Demands. Proceedings of 1998 DETC ASME Design Theory and Methodology Conference. Paper No. DETC98/DTM-5679. Atlanta, Georgia. September 1998.

20

Applying a Hybrid Data Mining Approach in Machining Operation for Surface Quality Assurance

Tzu-Liang (Bill) Tseng, Yongjin Kwon and Ryan B. Wicker

1. Introduction

Conventionally, the quality of a machined product has been measured based on the specifications, once the machining process is complete. However, this post-process inspection has several shortcomings: (1) it is difficult to isolate the causes of the defect; (2) the manufacturing cost has already been incurred when a non-conformance is detected; (3) rework of any scope increases the manufacturing cost and can be very difficult to accomplish; and (4) there could be a significant time lag between the detection of the defects and subsequent corrective actions. Today, efforts of manufacturers are shifting from the post-process inspection to improved monitoring of the manufacturing processes, utilizing sensors and other measurement devices, to effectively control the process. Improvements in machining precision can only be accomplished by the development of manufacturing systems that are capable of monitoring processes. Process monitoring reduces scrap, rework, lead-time, and conventional non value-added inspection activities, thereby, increases the system's productivity. The monitoring has to be based on sound, reliable process control algorithms. Computer numerical control (CNC) of machine tools do help to produce consistent part quality. However, in most cases, CNC machines don't utilize sensor data to compensate for anomalies generated by the cutting processes (e.g., tool wear, chatter, incorrect machine setup, etc.). If sensors such as cutting force, vibration and spindle motor current were integrated into CNC machine tools, the control functions should be able to interpret and respond to sensory data as the process continues. However, when many process variables need to be considered, it becomes rather difficult to predict quality attributes in machining (i.e., surface roughness).

To solve the aforementioned prediction problems, especially with the consideration of negative information and data to improve prediction accuracy, two data mining approaches have been developed. Here, negative information or

data represent the set of data points that do not conform to the conventional modeling techniques but it can be used to facilitate quality prediction. The approaches involve both individual and population based paradigms, such as a rough set theory (RST) and SVMs with the negative information & data training. To validate the proposed approach, one case of the perdition problem related to the surface roughness is applied. Literature review suggests that the hybrid approach of combined individual and population based paradigms has not been widely applied, thus making this research novel. By using the hybrid approach, the following objectives can be attained: (1) search the minimal number of features, or rules, for decision making in prediction; (2) aggregate the weight of the feature and frequency of the object to search the optimal rules; (3) simultaneously identify the outcomes and significant features in prediction; and (4) achieve a high prediction accuracy through the application of negative information & data. In this context, this study uses a hybrid data mining approach to identify variables affecting the quality characteristic of CNC machining operations. Instead of predicting exact surface roughness values, the focus is on the prediction of quality acceptance in machined parts. The hybrid approach is an effective tool for multi-attribute classification problems. This can be instrumental in constructing intelligent control systems, especially when a clear delineation within variables as to how they affect the surface roughness is difficult to achieve.

2. Theoretical Background of Data Mining and Hybrid Approach

Data mining is a process of extracting and refining knowledge from large databases (Berry & Linoff, 1997; Dhar & Stein, 1997; Cheung *et al*., 1996). It is a process that uses a variety of data analysis tools to discover the patterns and relationships in the data. The extracted information can be used to predict, classify, model, and summarize the data being mined. Data mining, a major step in knowledge discovery from databases, involves the application of specific algorithms for identifying interesting structures in data, where the structure designates patterns, statistical, or predictive models from the data as well as the relationships among parts of the data (Fayyad & Uthurusamy, 2002). Data mining is also an emerging area of computational intelligence that offers new theories, techniques, and tools for processing large volumes of data. The growing volume of data available in digital format has accelerated this interest.

Basically, data mining approaches can be categorized into two different cases. One is called "individual based" while the other is called "population based" paradigm (Kusiak, 2001(a); Kusiak & Kurasek, 2001). There are fundamental differences between the two approaches. The individual based approach generates a number of models (usually in the form of decision rules) capturing relationships between the input features and the decision. In other words, the individual based approach identifies unique features of an object and finds whether they are shared with other objects. The population based approach creates a model based on a training data set. The model normally uses a predetermined set of features (Kusiak, 2001(a); Kusiak & Kurasek, 2001). For example, the rule induction approach follows a data object paradigm, while neural networks and SVMs can be viewed as a single model that is formed for the entire population (training data set). The models (rules) created by the rule induction are explicit. The "population based" tools determine features that are common to a population (training data set) (Kusiak, 2001(b); Kusiak & Kurasek, 2001). The deficiency of the individual based approach for prediction is that low accuracy decision rules cannot be used, and the quality rules with a high accuracy do not guarantee to be used since the condition part of the rule should match with the input domain of the testing data sets. Consequently, the limitations of the rule based prediction can be easily observed, and the population based data mining approaches are able to counteract this deficiency.

In general, the material for learning is given in a positive form. This type of information will help organize the core of the target knowledge. Instead of this type of information, negative information will help sharpen the edge or extent of the target knowledge. Hence, it is expected that the negative information will have an effect of minimizing the chance of making errors and thus making the learning faster (Kurosu & Ookawa, 2002). In the data mining domain, negative information/data, which is defined as information/data, misclassified the outcomes from the testing data set and is normally discarded (Chen et. al. 2004 (b)). However, the information/data is possible to be re-used for the training purpose and contains a positive impact on the prediction accuracy (Chen *et al*., 2004 (b)). To date, there is little literature related to using data mining approaches to predict surface roughness with the consideration of utilizing negative information/data. To conduce the individual and population based data mining approaches that are to be applied in the preferred supplier prediction, the two classification approaches: RST and SVMs are reviewed.

2.1 Rough Set Theory

RST has been applied to address a variety of problems (Ziarko, 1993), including (1) representation of uncertain or imprecise knowledge; (2) empirical learning and knowledge acquisition from experience; (3) knowledge analysis; (4) analysis of conflicting data; (5) quality evaluation of the available information with respect to its consistency and presence or absence of repetitive data patterns; (6) identification and evaluation of data dependencies; and (7) approximation of pattern classification. In RST, data is expressed in a decision table, in which each row represents an object and each column represents an attribute. Formally, the decision table is represented by an information function (Pawlak, 1991) in the form of:

$$S = <U,\ Q,\ V,\ f> \qquad (1)$$

where U = a finite set of objects, Q = finite set of attributes, $V = \underset{q \in Q}{\cup} V_q$ and V_q = domain of the attribute q, and $f : U \times Q \to V$ = the total decision function such that $f(x,q) \in V_q$ for every $q \in Q,\quad x \in U$.

The main theme of RST is concerned with measuring what may be described as the "ambiguity" inherent in the data. The essential distinction is made between objects that may definitely be classified into a certain category, and those that may possibly be classified. Considering all decision classifications yields to what is referred to as the "quality of approximation" that measures the proportion of all objects for which definite classification may be achieved. A rough set can be described as a collection of objects that in general cannot be precisely characterized in terms of their values of their sets of attributes, but can be characterized in terms of lower or upper approximations. The upper approximation includes all objects that possibly belong to the concept, while the lower approximation contains all objects that definitely belong to the concept. As each object is characterized with attributes, discovering dependencies between attributes and detecting main attributes is of primary importance in RST. Attribute reduction is one unique aspect of the rough set approach. A reduct is a minimal sufficient subset of attributes, which provides the same quality of discriminating concepts as the original set of attributes. For example, let's consider the five objects in Table 1, each with four input features (F1-F4) and an output feature.

Object No.	Features				Output
	F1	F2	F3	F4	O
1	1	0	1	2	2
2	1	1	0	3	1
3	1	0	0	0	0
4	0	2	2	1	0
5	1	1	1	0	2
0: Not Applicable, 1: Low, 2: Medium, 3: High					

Table 1. Example Data Set

To derive reducts, consider the first feature F_1. The set of objects corresponding to the feature value $F_1 = 1$ is {1, 2, 3, 5}. This set {1, 2, 3, 5} cannot be further classified solely using the relation $F_1 = 1$. It is discernible over the constraint $F_1 = 1$, which is expressed as $[x][F_1 = 1] = \{1, 2, 3, 5\}$. For the objects in the set {1, 5}, the output feature is O = 2, for the object 3, the output feature is O = 0 and for the object 2, the output feature is O = 1. Therefore, additional features are needed to differentiate between O = 0, 1, or 2. Applying this concept, the classification power of each feature can be evaluated. For instance, the feature value $F_1 = 0$ is specific to O = 0. This discernible relation can be extended to multiple features, e.g., $[x][F_1 = 1] \wedge [F_2 = 0] = \{1, 3\}$ and $[x] [F_1 = 1] \vee [F_2 = 0] = \{1, 2, 3, 5\}$, where $\wedge$ and $\vee$ refers to "or" and "and", respectively.

Reduct Generation

Most rough set based approaches generate more than one reduct for an object. This paper adapts the reduct generation procedure proposed by Pawlak (1991) and presents it in the form of "*reduct generation procedure*," as illustrated in Figure 1. The reduct generation procedure enumerates all possible reducts with one, two and three features that are presented in Table 2.

Step 1: Set object number i = 1.
Step 2: Select object i and find a set of reducts with 1 to (m – 1) features.
Step 3: Set i = i +1. If all objects have been considered, go to Step 3; otherwise go to Step 1.
Step 4: Terminate the algorithm and output the result.

Figure 1. Reduct Generation Procedure

Object No.	F1	F2	F3	F4	O	Reduc No.	U	F1	F2	F3	F4	O
1	1	0	1	2	2	1	1	x	x	1	x	2
						2		x	x	x	2	2
						3	2	1	x	1	x	2
						4		1	x	x	2	2
						5		x	0	1	x	2
						6		x	0	x	2	2
						7		x	x	1	2	2
						8	3	1	0	1	x	2
						9		1	0	x	2	2
						10		x	0	1	2	2

Table 2. Partial Reducts for Data in Table 1

2.2 Support Vector Machines

SVMs based on the statistical theory have been developed as the tools for classification, regression, and density estimation in noisy data (Vapnik, 1998). There are three significant features in SVMs. The first is the generalization theory, which leads to a structure risk minimization (SRM) model. The generalization error is from either the model or hypothesis space. The SRM model improves the generalization ability through extending the margins in the feature space. The second is the kernel functions, which maps non-linear system into a linear feature space without explicitly requiring an exact map function. SVMs computational problem is connected with the size of the feature space. This makes SVMs perform efficiently over neural networks. The last feature is that the parameters are found by solving a quadratic programming problem with linear equality and inequality constraints, which return the global optimal solution. By doing so, the estimation errors can be minimized.

SVMs are designed for a binary classification. Generally, there are two types of approaches for a multi-class classification. One is that multi-class problems have been tackled by indirectly by combining a series of binary problems. Another is considering all data in one optimization formulation. Several methods based on the combining approach are one-versus-rest, one-versus-one, and DAG (Directed Acyclic Graph) SVMs methods (Platt et al., 2000). Using the SVMs in the one-versus-rest fashion is very common, but it has potential drawbacks when classes overlap considerably. It constructs k SVMs, where k is

the number of classes. The i^{th} SVM is trained by the i^{th} class associated with a positive label and all other examples with negative labels. The predication of an example x is determined by the maximum margin of k SVMs.

One-versus-one method is combining the binary SVMs for all pairs of classes. The DAG SVMs algorithm is a pair wise approach that exhibits large variability since each binary classifier is estimated from a small subset of the training data. It allows only a simple cost structure when different misclassification costs are concerned. As a generic approach to multi-class problems, treating all the classes simultaneously is considered. Although several extensions to the multi-class cases have been proposed (Vapnik, 1998; Bredensteiner & Bennett, 1999), its optimal extension was not obvious in relation to the theoretically best classification rule. The DAG SVMs and one-versus-one have good practical performance than the other methods (Hsu & Lin, 2002). In the DAG SVMs and one-versus-one, the training phase of k-classes classification problem is completed by $k(k\text{-}1)/2$ binary SVMs. In the testing phase, DAG SVMs uses a root binary acyclic graph with k leaves, where each node is a binary SVM. To test an example x, testing begins with a root node along with the DAG to reach a leaf node. The testing of one-versus-one is using a voting approach. The result of predicating is the largest vote number. The advantage of DAG SVM is to cut down the testing time as compared to the one-versus-one method.

From the review, RST application and SVMs appear to be both robust and efficient in automatic classification. Furthermore, the methods that automatically generate diagnostic rules have shown to have a significant aim in decision making of prediction. In this paper, the concept of feature extraction, cluster analysis, and SVMs model are used to develop a methodology for aiding the preferred supplier selection. Motivation for conducting combination of RST and SVMs is the hybrid approach capable of performing significant feature identification (dimension reduction), noise elimination (object reduction), and learning from negative information/data to improve prediction accuracy. Moreover, the hybrid approach is the combination of the individual and population based data mining approaches that are able to overcome low accuracy of prediction and other limitation. The methodology development is introduced next.

3. Methodology Development

This section illustrates the development procedure of methodology. The overall concept has been represented in Figure 2.

3.1 Rule Identification Algorithm

The rule identification algorithm is incorporated with the weights. Measurement of the weight is based on domain experts' judgment and external assessment. Basically, each feature has been considered and the "ratio estimates" method assigns the weight of each feature without any bias. Moreover, adjustment of the weights assigned through pair wise comparisons is also required. Frequency of each object is derived from the original database during a certain period. All of the weights, which include feature and frequency domains, are subjected to normalization.

In Table 3, the weight value of the feature is taken into consideration and subject to a positive normalized value, which is between 0 and 1. It can be obtained from domain experts' judgment. With each column of incidence matrix $A = [a_{ij}]_{mxn,}$ frequency f_i for reduct i, $i = 1, \ldots, m$ and weight w_j for input feature j, $j = 1, \ldots, n$ are associated. It is possible to assign frequency f_i, $i = 1, \ldots, m$ as a different kind of weight since the frequency of each object can be derived from the original database. Furthermore, it is also possible to assign weights w_j, $j = 1, \ldots, n$ since the weight of each feature can be determined from domain experts. Table 3 is with a column indicating the number of objects and a row containing weights associated with the features. Using the weight coefficients w_j and f_i, an auxiliary matrix $[e_{ij}]$ will be generated from the original reduct – input feature matrix.

The weight coefficient assigned to each feature is denoted as w_j and each object (reduct) as f_i. Using the weight coefficients w_j and f_i, an auxiliary matrix $[e_{ij}]$ will be generated from the original reduct – input feature matrix. The entries of the transformed matrix are defined as follows:

$$e_{ij} = f_{i\,x}\,(w_{j\,x}\,v_j) \qquad (2)$$

where e_{ij} = entry of the transformed reduct-input feature matrix, f_i = weight of reduct I, w_j = weight of feature j, and $v_j = 1$ if feature $j \neq$ "x"; 0 otherwise.

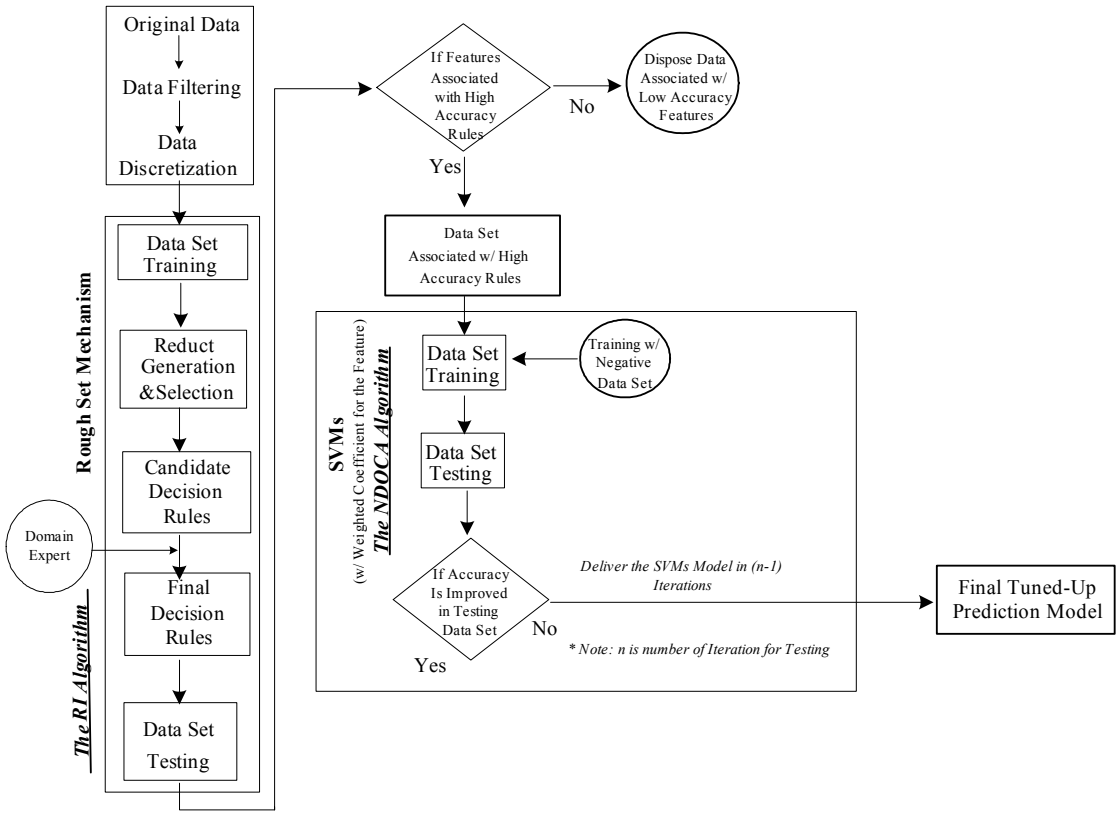


Figure 2. Conceptual framework of the hybrid data mining approach to prediction problem

Object No.	F1	F2	F3	F4	F5	F6	F7	O	**Weight f_i**
1	3	1	0	1	2	0	2	2	100%
2	3	0	1	2	1	0	3	0	100%
3	0	1	2	2	1	0	1	2	62%
4	0	1	1	1	2	0	1	1	38%
5	1	2	2	0	2	1	0	1	92%
6	2	2	0	0	2	1	1	1	54%
7	1	0	0	1	3	0	1	2	54%
8	3	2	1	1	2	1	1	1	69%
Weight w_j	80%	100%	90%	60%	70%	60%	90%		

Table 3. Data set with un-equal weight for object and feature

The RI Algorithm

Step 0.

(i) List the auxiliary matrix.

(ii) Compare the reducts (rows of matrix $[a_{ij}]$). Select the features used in only single feature reducts of the object(s).

(iii) List the number of known value for each column in $[a_{ij}]$. Select the potential features used, base on the higher number of known value (refer the results from (ii)).

(iv) Set iteration number $k = 1$.

Step 1. Compare those reducts (rows of matrix $[a_{ij}]^{(k)}$) for one specific case at a time. Select the reducts from the potential features used and based on the auxiliary matrix. If more than one solution for the reduct selection, then select the reduct which can be merged by most of objects; otherwise, select the reducts which are most frequently selected from previous iterations. Draw a horizontal line h_i through each row of matrix $[a_{ij}]^{(k)}$ corresponding to these reducts.

Step 2. For each column in $[a_{ij}]^{(k)}$ corresponding to an entry of feature, which is not "x", single crossed by any of the horizontal lines h_i , draw a vertical line v_j.

Step 3. Repeat steps 1 and 2 until one reduct has been selected for each object in the current outcome. All double-crossed entries of features of the matrix form the rules.

Step 4. If all objects have been concerned in the current outcome, transform the incidence matrix $[a_{ij}]^{(k)}$ into $[a_{ij}]^{(k+1)}$ by removing all the rows and corresponding to an entry of feature, which is not "x", included in the current outcome.

Step 5. If matrix $[a_{ij}]^{(k+1)}$ = " " (where " " denotes a matrix with all elements equal to blank, stop and output the results; otherwise set $k = k + 1$ and go to step 1.

Note that the difference between the equal and un-equal cases for the use of the RI algorithm is “Step 0 (i) is not required by equal weight case.” Consider the data set in Table 3. Determine the desired reducts (rules) in Table 4 using the RI algorithm. Repeating Steps 1-5, the final results are shown in Table 4, indicating four features 2, 3, 5, and 7 have been selected.The proposed RS based approach aims to incorporate a weight factor into each feature, process qualitative data, generate decision rules, and identify significant features. This entails that the feature (dimension) domain can be reduced tremendously.

Note that the key contribution of weight in the reduct induction is that the assigned weights help determine the preferred reducts whenever the alternative reducts are produced.

Object No.	F1	**F2**	**F3**	F4	**F5**	F6	**F7**	O
1	x	**x**	**x**	x	**x**	x	**2**	2
4	x	**1**	**1**	x	**x**	x	**x**	1
5, 6 and 8	x	**2**	**x**	x	**x**	x	**x**	1
2	x	**x**	**x**	x	**x**	x	**3**	0
7	x	**x**	**x**	x	**3**	x	**x**	2
3	x	**1**	**2**	x	**x**	x	**x**	2

Table 4. The desired reducts for Table 3

At this point, it is discerned that the weight assignment approach supports to generate the preference-based rule. Furthermore, the preferred decision rules (normally with a high accuracy) derived from the RST based approach (an individual based data mining approach) are not capable of predicting upcoming testing data sets, except when the condition part from test sets matches the preferred decision rules. Therefore, a population based data mining approach (e.g., SVMs based approach) with the consideration of negative data sub-set is introduced next.

3.2. Learning Problem Description through SVMs

The training data set is partitioned into three disjointed subsets: misclassified, not well-separated, and well-separated examples. The misclassified and not well-separated examples together are in the negative data subset whereas the well-separated examples are called in the positive data subset. For example, in the surface roughness prediction, misclassified, non-conformation part is an example of the negative data sub-set. To illustrate the structure of the data set, there is an *instance* vector $\boldsymbol{x}$ from an input space $\boldsymbol{X}$, a response or *label* $\boldsymbol{y}$ from an output space $\boldsymbol{Y}$ and a hypothesis h that forms a hypotheses space H for a learner L. For example, $\boldsymbol{X}$ represents all input features (F1 - F7) in Table 2, while $\boldsymbol{Y}$ represents one output feature (O). Assume we have

$$\mathbf{x} = (x^{(1)}, \ldots, x^{(n)})', \mathbf{X} \circledcirc R^n, \mathbf{x} \circledcirc \mathbf{X}, x^{(i)} \circledcirc R \qquad (3)$$

where R = a set of real numbers, integer $n>0$ = the size of vector x, for *multi-category classification,* $\boldsymbol{Y} = \{1,2,\ldots, m\}$. A *training set* or *training data S* is a collection of *training examples or observations* given by $z_i=(\boldsymbol{x}_i,\boldsymbol{y}_i)$. It is denoted by

$$S = (z_1.., z_l) = ((x_1, y_1), (x_2, y_2),..(x_l, y_l)) \subseteq Z^l.\ z_i \in Z = (\mathrm{X}, \mathrm{Y}), i = 1..l \qquad (4)$$

where $\ell = |S|$ is the size of the training set. There exists a true functional relationship or underlying function $f: \boldsymbol{X} \rightsquigarrow R^n \rightsquigarrow \boldsymbol{Y}$, which is often based on the knowledge of the essential mechanism. These types of model are called *mechanistic models*. A hypothesis h is an approximation to the underlying functional relationship f between variables of interest. The problem for the learner L is to learn an unknown target function $h: \boldsymbol{X} \rightsquigarrow \boldsymbol{Y}$ drawn from H and output a maximum likelihood hypothesis.

3.3 Negative Data Oriented Compensation Algorithm

It is not likely to select a perfect model for a practical problem without approximation errors in a learning algorithm. To select a perfect model, imagining that underlying function $f(x)$ is a fluctuant terrain, it is hard to fit the terrain by using a huge size of carpet $h(x)$. The reason is that only the training set and limited prior knowledge is available. The main idea of reducing the approximation error is to compensate the parts of an oversized carpet by a sequence of small sized carpets $h(i)(x)$ which is driven by the negative data subset of training data. The procedure of the Negative Data Oriented Compensation Algorithm (NDOCA) has three parameters, S0 is the training data set; T0 is the testing data set; and δ is a degree of vector similarity. For example, δ is difference between two suppliers (objects) in the preferred supplier selection. The return value of the algorithm is the predictive labels of the testing data set. Six subroutines are invoked,

1. $h^{(i)}(x)=LEARN(S_i)$
2. $P_i=PREDICT(T_i, h^{(i)}(x))$
3. $S^{\#}_{i+1} \cup S_{i+1}= DIVIDER(S_i, h^{(i)}(x))$
4. $T_i = VS(S_i, T_{i-1},\delta)$
5. $P^{\#}_i = OV(P^{\#}_{i-1},P_i)$
6. $TC(k,S)$

LEARN is for training to get the model or hypothesis; PREDICT is to predict the labels of given data set and model. These two procedures are from classical leaning algorithms such as SVMs and artificial neural networks. DIVIDER is to divide training data set into positive and negative data subsets by given the hypothesis and the function partitioner *d(h,x,y)*. DIVIDER will call PREDICT routine. In each pass, the function VS and DIVDER could be different. The following is an algorithm described as pseudo-code (Figure 3).

```
NDOCA (S0, T0, δ)
> Learning phase
1. S[0] ← S0
2. h[0] ← LEARN(S[0])
3. i ← 0
4. repeat
5. i ←i+1
6. (S#[i], S[i]) ← DIVIDER(S[i-1], h[i-1])
7. h[i] ← LEARN(S[i])
8. until TC(i,S)
9. k ← i > the number of iteration in repeat loop
> Testing phase
10. T[0] ← T0
11. P[0] ← PREDICT (T, h[0])
12. P#[0] ← P[0]
13. for i←1 to k do
14. T[i] ← VS(S[i],T[i-1], δ)
15. if T[i] ≠ Φ  *T[i] is not empty set
16. then P[i] ← PREDICT(T[i], h[i])
17. P#[i] ← OV(P#[i-1], P[i])
18. return P#[k]
```

```
DIVIDER(S[i-1], h[i-1])
1. X ← ΔY←Φ *initialize to empty set
2. foreach (x,y) in S[i-1] do *let (X,ΔY)
   be S[i-1]
3. X ← X ∪ {x}
4. ΔY← ΔY ∪ {y}
5. S[i] ← Φ
6. foreach (x, Δy[i-1]) in (X,ΔY) do
7. Δy[i] ← PREDICT(x, h[i-1])
8. if d(h[i-1], x, Δy[i-1])
9. then S[i] ← S[i]∪ {(x, Δy[i])}
10. Δy[i-1] ← Δy[i] *update ΔY
11. S#← S[i-1] - S[i]
12. return (S#[i], S[i])

VS(S[i],T[i-1], δ)
1. T[i] ← Φ
2. foreach x1 in T[i-1] do
3. foreach x2 in S[i] do
4. if vs(x1,x2) ≥ δ
5. then T[i] ← T[i] ∪ {x1}
6. break
7. return T[i]
```

Figure 3. Pseudo-code of the NDOCA

To prepare for the NDOCA learning algorithm, partitioner function *d(h,x,y)*, terminate criteria function *TC(k,S)*, and vector similarity *vs(x1,x2)* need to be provided. The performance of NDOCA very depends on the selecting of partitioner and vector-similarity function, which needs priori knowledge of learning problems. Note that the NDOCA algorithm is taken as weighted data based on weight coefficients, given by the domain experts.

4. An Empirical Study

4.1 Problem Structure and Data Set Description

Over the years, A-Metal Inc. (a pseudonym for the company) has collected over 1,000 records (objects) of machining data and wishes to investigate the machining features which have a significant impact on the quality of surface finish. Figure 4 illustrates the intelligent CNC control scheme that A-Metal is planning to implement, as opposed to the conventional CNC control that has no response capability as machining process changes.

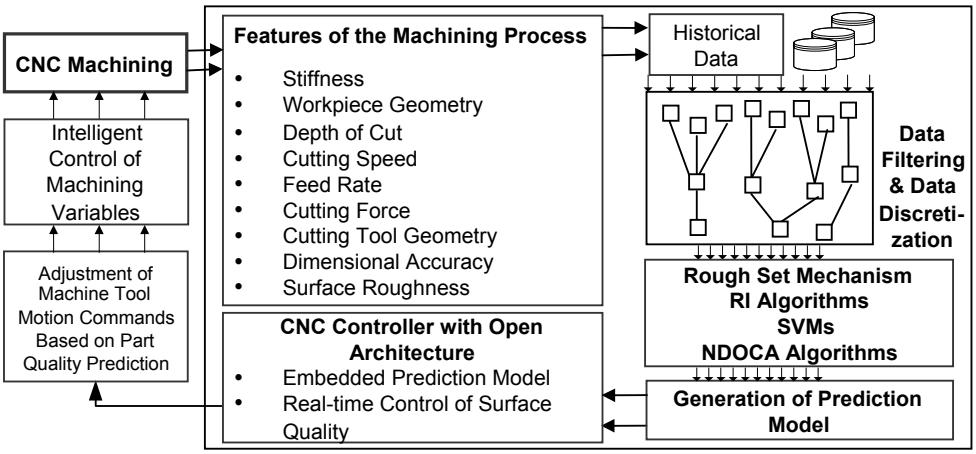


Figure 4. Structure of the closed loop machining operation process

In order to derive the rules and algorithm, conditions of the variables, which could meet the required surface roughness, were identified. Those specific variables will be used to develop the intelligent control system, and in addition can be used by industry to optimize the surface roughness of machined metal (e.g., aluminum, steel) components. Each information object was described with the eight features, F1 through F8, and one outcome, O (Table 5). The work-piece materials include three different types, including 6061-T6 aluminum, 7075-T6 aluminum, and 4140 medium carbon steel (Figure 5).
The surface roughness of the machined bores was measured along a machine Z-axis (parallel to the height of the bore). The machining has been performed on the Cincinnati Hawk CNC Turning Center. The effects of cutting speed, depth of cut, machine set up-modal stiffness, feed rate, cutting tool, tool nose

radius and resultant cutting force on the performance of surface roughness estimation were studied. After roughing and semi-finishing operations, the surface roughness was measured by means of a Taylor Hobson® surface profilometer.

(a) (b)

Figure 5. A snapshot of CNC machining (a) and a mixed array of parts consisted of 6061 Al (top), 7075 Al (middle), and 4140 medium carbon steel (the bottom two rows) (b).

	Factor	Weight
F1	Types of work piece material	.9
F2	Cutting speed	.8
F3	Depth of cut	.8
F4	Machine set up-modal stiffness	.8
F5	Feed rate	.7
F6	Cutting tool	.9
F7	Tool nose radius	.85
F8	Resultant cutting force	.75
Outcome	Surface roughness (Ra)	

Table 5. Feature set of the machining operation process

The contents of the outcome are recorded in a binary format. "ONE" means surface roughness is acceptable, while "ZERO" means unacceptable. The significant variables, which have impact on the quality of surface roughness, were determined through the rule identification algorithms. The decision pro-

duced by the algorithm became decision rules stored in the process control system.

4.2 Computational Results

To show the superiority of the proposed approach, the computational results from the RST part and the hybrid approach part are illustrated. Section 4.2.1 describes the final decision rules with significant features derived from RST. The summary of accuracy results from the test set is presented to show performance of the proposed RI algorithm. Section 4.2.2 represents solutions through the hybrid approach. Comparison among RST, SVMs, and the hybrid approach is also depicted to demonstrate accuracy of each approach in this section.

4.2.1 Rough Set Theory Part

The "Rough Set Based Decision Support System" software (Figure 6) was developed by the authors and implemented in the Advanced Manufacturing Laboratory at the University of Texas at El Paso. It was installed using an Apache 1.3 web server to enable the remote use. The system was developed with C^{++} language and the Common Gateway Interface (CGI) is used as a communication protocol between the server and client ends. The historical data were split into two data sets. One is the training data set to derive the decision rules; the other is the testing data set to verify the decision rules. Kusiak (2001) suggested the split of the data set using the bootstrapping method according to the following ratio: 0.632 for the training set and 0.368 for the test set. In this study, training data set was collected for 667 parts and testing data set was collected for 333 parts. 41 out of 667 parts in the training set were unacceptable for surface roughness, while 19 out of 333 parts in the testing set were rejected.

All decision rules derived by the RI algorithm were expressed in the form of IF-THEN rules, as illustrated in Table 6. Number of support (see the 3rd column) was recorded from the training set. The selection criteria were based on the threshold value, indicating the ratio of the number of objects supported by that individual rule to the number of total objects. In this case study, a 15% threshold value is selected based on the quality engineer's expertise. All selected decision rules should be equal or greater than this selected threshold value. For example, the first rule in Category I shows 102 acceptable parts

based on a surface roughness leading to 16% non-defective population. Category I describes the relationship between the features and the acceptable parts. The third rule in Category I is strongly supported because it represents 20% of the acceptable population. In Category II, 17% and 20% of the unacceptable parts are identified by the two rules. Overall, more simple rules (less features as conditional features) are shown in Table 6. The simple rule is treated as the desirable rule because if only two conditions are matched then the rule is fired. Based on the 15% threshold value, significant features F1, F2, F3, F5, and F8 are identified. One can observe that all rules include Feature 1 (types of work piece materials). Therefore, Feature 1 is significant in this set of rule induction. F2, F3, F5, and F8 are significant as well since they are included in the final decision rules. It can be seen that the type of work piece materials, cutting speed, depth of cut, feed rate, and resultant cutting force are important factors for the quality characteristic.

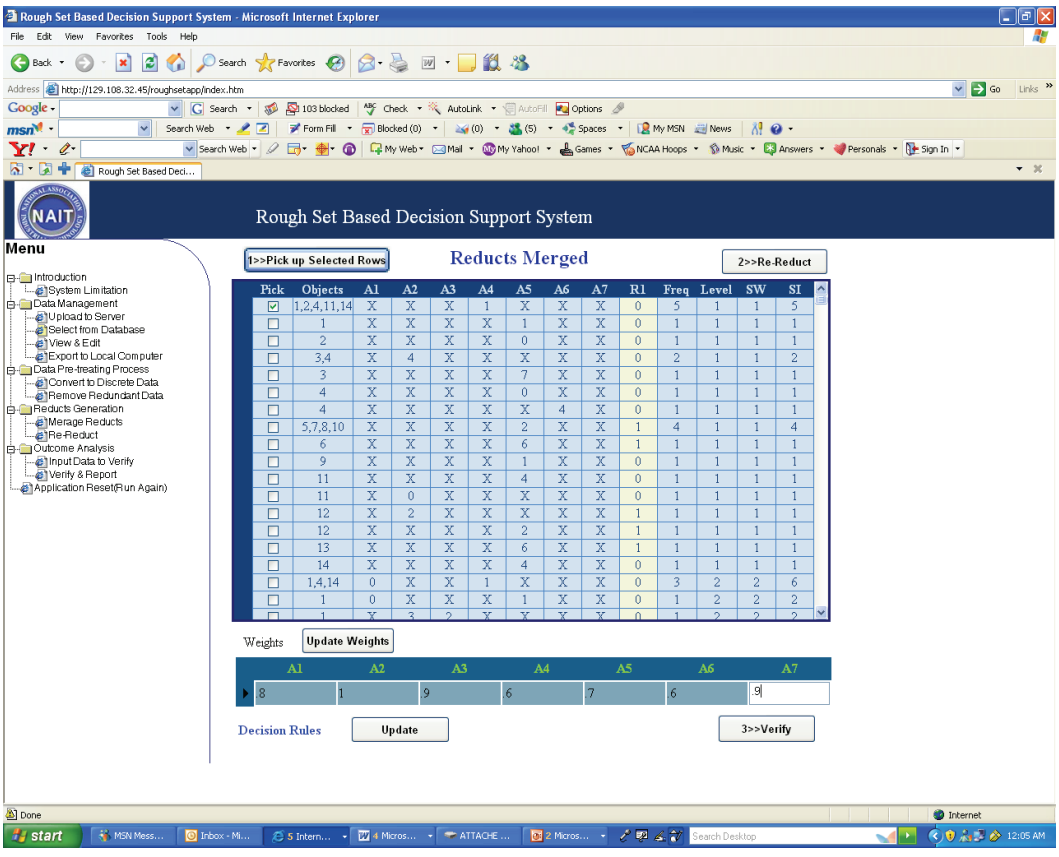


Figure 6. Screen shot of rough set application software

Rule No.	Rule expression	No. of support	% of the part population by the rule (from training set)
1	IF (F1 = Al 6061) AND (F3 = .2) THEN (D = 1).	102	16
2	IF (F1 = Al 6061) AND (F5 = .017) THEN (D = 1).	91	15
3	IF (F1 = Al 7075) AND (F8 = 600) THEN (D = 1).	125	20
4	IF (F1 = Al 7075) AND (F5 = .005) THEN (D = 1).	75	12
5	IF (F1 = Steel 4140) AND (F2 = 1200) AND (F8 = 300) THEN (D = 0).	7	17
6	IF (F1 = Al 6061) AND (F8 = 600) THEN (D = 0).	8	20

Table 6. Examples of decision rules. Note: (1) F3: depth of cut, F5: feed rate, F8: resultant cutting force, F2: cutting speed, (2) Category I includes Rule 1– 4 and Category II includes Rule 5–6.

Testing on the validity of the rules, which extracted from a data set, was carried out by the rule-validation procedure, which includes a comparison between each decision rule and each new object from the test set. One set of 314 parts with 19 defectives is used as the test set. The accuracy of results for 314 test set parts is shown in Table 7. As Pawlak (1991) explains, the "classification quality" of a feature set is the percentage of all objects in the training data set that can be unambiguously associated with the decision values based on the features in this set. At the same time, the "Diagnostic Accuracy" or so called "Classification Accuracy" for a rule set is the number of correctly classified objects from the test set to all objects in the test set. These results are animate since all of selected rules with a 15% threshold value denote close to 90% accuracy except the third rule in the first category. Four out of six rules (the 1st and 2nd rules in category I, the 1st and 2nd rules in category II) are shown over 90% accuracy. However, the good quality of rule depends on its diagnostic accuracy (Kusiak, 2001). In Table 7, the significant features are identified as F1, F2, F3, F5 and F8. Since the significant features in this case study are fathom, the dimension of interest can be reduced from 8 features to 5 features.

Test Set	Category	I				II	
333 Parts (314 acceptable vs. 19 unacceptable)	Rule ID	1	2	3	4	1	2
	Feature Set	**F1, F3**	**F1, F5**	F1, F8	F1, F5	**F1, F2, F8**	**F1, F8**
	% of the part population by the rule (from training set)	**16%**	**15%**	20%	12%	**17%**	**20%**
	Classification quality	**94.9%**	**100%**	70.1 %	88.9 %	**94.6%**	**100%**
	Diagnostic accuracy	**95.4%**	**100%**	71%	89.3 %	**95.3%**	**100%**

Table 7. Summary of accuracy results from the test set. Note: Bold text represents the threshold values of 15% case with acceptable Classification quality and Diagnostic accuracy

4.2.2 Hybrid Approach Part

The NDOCA algorithm is implemented by Perl and uses a modified SVMlight (Joachims, 2002; Joachims, 1999) as a base learning algorithm, including learning and classifying modules. Before the case is studied, the three functions-partitioner function *d(h,x,y)*, terminate criteria function *TC(k,S)*, and vector similarity *vs(x1,x2)*-need to be defined. To simplify the complexity of computation, the partitioner is defined on the feature space by *d(h, x, y)* = **iff**(h(x) < ε,**true**, **false**), ε ◎ ¸[0,0.5]. And *TC(k,S)* is defined by *TC(i, S[i])*= ***iff***(|*S[i]*|≤|*x*|, ***true***, ***false***). Basically, the Vector Similarity Euclidean method is used for training and testing data sets. The vector-similarity is a metric to describe the similar degree of two vertices. The vector-similarity plays an extremely important role in the NDOCA learning algorithm. By applying for repairing hypersurface, the first thing is to find which vertices in the testing data set need to be compensated. The vector-similarity is used to find the relationship of vertices in the negative data subset S_i and testing data subset T_{i-1}. Only those vertices in T_{i-1} with high similarity to the ones in S_i need to be compensated.

Since A-Metal Inc. would like to observe the impact of weights and negative data training, the performance measurement includes the following four different cases: 1) equal weight without non-negative data training, 2) un-equal weight without non-negative data training, 3) equal weight with non-negative data training, and 4) un-equal weight with non-negative data training. The *n*-cross validation is performed in each case. The average result of *n*-fold is the final accuracy while the minimum and maximum values of accuracy are given

as shown in Table 8. Here, the training data set (with 667 objects) used for rule induction from the previous stage is used for five-fold cross validation. Note that the data set only contains significant features (e.g., F1, F2, F3, F5 and F8).
In Table 8, one can observe that the case of equal weight without negative data training contains the lowest diagnostic accuracy with 94.8%. The case of un-equal weight with negative data training comprises the highest diagnostic accuracy with 97.3%. The case of un-equal weight without training is not exactly prevailing over the case of equal weight since the accuracy of some individual groups (e.g., group 3) in the equal weight with training case are pretty high (e.g., 97%). Therefore, it is difficult to conclude that the weight effect is predominating over the negative data training effect. In this case study, comparison of accuracy of RST, SVMs, and the hybrid approach is also investigated in order to demonstrate the advantages of applying RS rules and SVMs to prediction. The original 667 objects are applied in this case. The results are shown in Table 9. Note that the accuracy of RST is based on objects that meet the condition of the decision rules. In conclusion, most of the hybrid approaches performed better than the others.

No.	T^+	T^-	F^+	F^-	C	M	DA%	T+	T-	F+	F-	C	M	DA%
1	120	6	5	2	126	7	94.7%	122	6	3	3	128	6	95.5%
2	121	7	5	1	128	6	95.5%	122	6	2	3	128	5	96.2%
3	120	6	5	2	126	7	94.7%	120	7	6	0	127	6	95.5%
4	119	6	6	3	125	9	93.3%	119	7	8	0	126	8	94.0%
5	121	6	4	2	127	6	95.5%	121	6	3	3	127	6	95.5%
Avg.	120.2	6.2	5	2	126.4	7	94.8%	120.8	6.4	4.4	1.8	127.2	6.2	95.4%
1	118	7	9	0	125	9	93.3%	122	8	3	0	130	3	97.7%
2	121	6	4	3	127	7	94.8%	121	8	4	0	129	4	97.0%
3	122	7	2	2	129	4	97.0%	122	9	2	0	131	2	98.5%
4	120	7	5	1	127	6	95.5%	122	8	4	0	130	4	97.0%
5	120	7	5	1	127	6	95.5%	121	8	5	0	129	5	96.3%
Avg.	120.2	6.8	5	1.4	127	6.4	95.2%	121.6	8.2	3.6	0	129.8	3.6	97.3%

Table 8. Comparison of four different cases (5-fold cross validation).Note: (1) T^+ = true positive (good part), T^- = true negative (defective), F^+ = false positive, F^- = false negative, C = correct classified and M = misclassified = $F^+ + F^-$ and DA% = diagnostic accuracy = C/(C+M) * 100%; (2) Upper left: equal weight w/o negative data training; upper right: un-equal weight w/o training; lower left: equal weight with training; and lower right: un-equal weight with training.

Approach	No. of features used/No of data used	Weight included	Negative data training	Correct (%)	Incorrect (%)
[1] RST	8/677	Yes	No	95.04	4.96
[2] SVMs-1	8/677	Yes	Yes	87.90	12.10
[3] SVMs-2	8/677	No	Yes	89.30	10.70
[4] Hybrid-1	5/677	Yes	Yes	94.80	5.20
[5] Hybrid-2	5/677	No	Yes	95.20	4.80
[6] Hybrid-3	5/677	Yes	No	95.40	4.60
[7] Hybrid-4	5/677	No	No	97.30	2.70

Table 9. Comparison of proposed hybrid approach with RST and SVMs approaches. Note: (1) The accuracy of RST is based on objects meet the condition of the decision rules; (2) 5-fold cross validation is applied in all cases.

5. Conclusions

Based on the historical data, this study employed a hybrid method that connects with the causal relationships between the features of the machining process and acceptance of surface roughness. This methodology is applied to the case of surface roughness prediction. Several features that significantly impact surface roughness were identified and considered in the case study. Several experiments with the RST, SVMs, and hybrid approach (included equal and unequal weights and with or without negative data training, and different data sets) were also conducted and the results are compared. Two main algorithms are proposed in this study. One is called the RI algorithm, while the other is named the NDOCA algorithm. The RI is used to derive high accuracy decision rules and identify significant features. The NDOCA is used to improve the learning algorithm performance through compensating the base hypothesis by using the negative data set. According to the hybrid approach, combination of RI and NDOCA provides a high accuracy prediction tool for investigating features that contribute to surface roughness. The hybrid approach provides important information for acceptance of surface roughness in the machining operations. The results showed practical viability of this approach for quality control. Future research can focus on the derived rules constitute the basis for developing a rule-based intelligent control system for surface roughness in the machining operation process.

6. References

Berry, M. & Linoff, G. (1997). *Data Mining Techniques: For Marketing, Sales, and Customer Support*, John Wiley & Sons, New York, NY, USA

Bredensteiner, E.J. & K.P. Bennett (1999). Multicategory classification by support vector machines. *Computational Optimizations and Applications,* Vol. 12, pp. 53-79

Chen, C.M.; Lee, H.M. & Kao, M.T., 2004 (b). (2004). Multi-class SVMs with negative data selection for Web page classification. *Proceedings of 2004 IEEE International Joint Conference on Neural Networks*, Vol. 3, pp. 2047-2052

Cheung, D.W.; Ng, V.T.; Fu, A.W. & Fu, Y. (1996). Efficient Mining of Association Rules in Distributed Databases. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8, No. 6, pp. 911-922

Dhar, V. & Stein, R. (1997). *Seven Methods for Transforming Corporate Data into Business Intelligence,* Upper Saddle River, Prentice-Hall, New Jersey, USA

Fayyad, U.M. & Uthurusamy, R. (2002). Evolving data mining into solutions for insights. *Communications of the ACM*, Vol. 45, No. 8, pp.28-31

Hsu, C.-W. & Lin, C.J. (2002). A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, Vol. 13, No. 2, pp.415-425

Joachims, T. (1999). *Making large-Scale SVM Learning Practical*, MIT Press, MA, USA

Joachims, T. (2002). *Learning to Classify Text Using Support Vector Machines*, Kluwer Academic Publishers, Dordrecht, The Netherlands

Kusiak A., 2001(a). Feature transformation methods in data mining. *IEEE Transaction on Electronics Packaging Manufacturing,* Vol. 24, No. 3, pp.214-221

Kusiak, A., 2001(b). Rough Set Theory: A Data Mining Tool for Semiconductor Manufacturing. *IEEE Transactions on Electronics Packaging Manufacturing*, Vol. 24, No. 1, pp. 44-50

Kusiak, A. & Kurasek, C. (2001) Data mining of Printed Circuit Board Defects. *IEEE Transactions on Robotics and Automation*, Vol. 17, No. 2, pp. 191-196

Kurosu, M. & Ookawa, Y. (2002) Effects of negative information on acquiring procedural knowledge. *Proceedings of International Conferenc*e on C*omputers in Education,* 3-6 December, Vol.2, pp.1371-1372

Pawlak, Z. (1991) *Rough Sets: Theoretical Aspects of Reasoning About Data*, Boston: Kluwer Academic Publishers, MA, USA

Platt, J.; N. Cristianini & J. Shawe-Taylor. (2000) Large margin DAGs for multiclass classification. *Advances in Neural Information Processing Systems*, Vol. 12, pp.547-553, MIT Press, MA, USA

Vapnik, V. N. (1998). *Statistical Learning Theory*, John Wiley & Sons, New York, NY, USA

Ziarko, W.P. (1994) *Rough Sets, Fuzzy Sets and Knowledge Discovery,* Springer-Verlag, New York, NY, USA

21

Sequential Design of Optimum Sized and Geometric Tolerances

M. F. Huang and Y. R. Zhong

1. Introduction

Tolerancing has great impact on the cost and quality of a product. Dimensional and geometric tolerancing are designed to ensure that products meet both designed functionality and minimum cost. The task of dimensioning and tolerancing in process planning stage is to determine the working dimensions and tolerances of the machined parts by given blueprint (B/P) specifications.
A lot of research work has been carried out in dimensioning and tolerancing. In earlier studies, optimal solutions to tolerance charts have been developed to meet B/P specifications. Most researches concentrated on dimensioning and tolerancing with optimal objectives to maximize the total working tolerances based on the constraints of tolerance accumulation and machining accuracy. Linear or nonlinear programming models have been applied to obtain the optimal tolerances (Ngoi, 1992; Ngoi & Ong, 1993; Ji, 1993a; Ji, 1993b; Wei & Lee, 1995; Lee & Wei, 1998; Ngoi & Cheong, 1998a; Lee et al., 1999; Chang et al., 2000; Huang et al., 2002; Chen et al., 2003; Gao & Huang, 2003; Huang et al., 2005). Optimal methods have also been presented to allocate B/P tolerances in product design using tolerance chart in process planning (Ngoi & Cheong, 1998; Ngoi & Ong, 1999; Swift et al., 1999) but the generation of dimensional and tolerance chains being one of the most important problems. In one-dimensional (1D) cases, the apparent path tracing and tree approach were commonly used to tolerance chart for manual treatment (Ngoi & Ong, 1993; Ji, 1993a; Ji, 1993b; Wang & Ozsoy, 1993; Ngoi & Cheong, 1998b). Automatic generation of dimensional chains in assembly based on the data structure has been presented (Treacy et al., 1991; Wang & Ozsoy, 1993). Using an Expert System, assembly tolerances analysis and allocation have been implemented by appropriate algorithm in CAD system (Ramani et al., 1998). An intelligent dimensioning method for mechanical parts based on feature extraction was also introduced (Chen et al., 2001). This method could generate the dimensions of

mechanical parts for two-dimensional (2D) drawing from three-dimensional (3D) models. Recently, more valuable and attractive approaches to deal with dimensional and geometric tolerances have been developed (He & Gibson, 1992; Ngoi & Tan, 1995; Ngoi & Seow, 1996). He and Gibbon in 1992 made a significant development in geometric tolerance charting and they presented useful concepts to treat geometric dimensions and tolerances simultaneously. A computerized trace method has been extended to determine the relationships between geometrical tolerances and related manufacturing dimensions and tolerances. A new method for treating geometrical tolerances in tolerance chart has been presented (Ngoi & Tan, 1995; Ngoi & Seow, 1996; Tseng & Kung, 1999). The method identified the geometrics that exhibited characteristics similar to linear dimensions. These geometrics were first treated as equivalent dimensions and tolerances and then applied to tolerance chart directly. Tolerance zones have been utilized to analyze tolerance accumulation including geometric tolerances. The formulae for bonus and shift tolerances due to position callout have been presented (Ngoi, et al., 1999; Ngoi et al., 2000). In complex 2D cases when both angular and geometric tolerances are concerned, graphic method has been used to implement tolerances allocation (Huang et al., 2002; Zhao, 1987). In conventional tolerancing, fixed working dimensions and tolerances were designed in process planning phrase. Though this method was suitable for mass production in automatic lines, it had limitations to produce low-volume and high-value-added parts such as those found in aircraft, nuclear, or precision instrument manufacturing industry (Fraticelli et al., 1997). To increase the acceptable rate of a machined part, a method named sequential tolerance control (STC) for design and manufacturing has been presented (Fraticelli et al., 1997; Fraticelli et al., 1999; Wheeler et al., 1999; Cavalier & Lehtihet, 2000; Mcgarvey et al., 2001). This method essentially used real-time measurement information of the complete operations to dynamically recalculate the working dimensions and feasible tolerances for remaining operations. Using acquired measurement information, tool-wear effect compensation under STC has been realized (Fraticelli et al., 1999). An implicit enumeration approach to select an optimum subset of technological processes to execute a process planning under STC strategy has been presented (Wheeler et al., 1999). When measurements and working dimension adjustments would be taken to facilitate machining process and reduce manufacturing cost has also been investigated (Mcgarvey et al., 2001).

In spite of the achievement mentioned above, some issues still need further research. The previous researches focused on 1D dimensioning and tolerancing.

Though simple 2D drawings were concerned, they could be converted into 1D dimensioning and tolerancing in two different directions, i.e. in axial and diametrical directions or in axis OX and OY directions (He & Gibson, 1992; Ngoi & Tan, 1995; Ngoi & Seow, 1996; Tseng & Kung, 1999). When incline features of 3D parts are machined, complicated dimensioning and tolerancing will occur since angular tolerance will be included in tolerance chains. In addition, the relationships between orientational and angular tolerances need further investigation. Though STC strategy is able to enhance the working tolerances and acceptance rate of manufactured parts (Fraticelli et al., 1997; Cavalier & Lehtihet, 2000), how to extend this method to complex 3D manufacturing is still a new problem when sized, angular, and orientational tolerances are included simultaneously.

Based on the basic principle of STC introduced by Fraticelli et al (Fraticelli et al., 1997), the purpose of this paper is to extend the new methodology to deal with 2D sized, angular, and orientational tolerances of 3D parts. The proposed approach essentially utilizes STC strategies to dynamically recalculate the working dimensions and tolerances for remaining operations. This approach ensures that the working tolerances of a processed part are optimal while satisfying all the functional requirements and constraints of process capabilities. A special relevant graphic (SRG) and vector equation are utilized to formulate the dimensional chains. Tolerance zones are used to express the composite tolerance chains that include sized and angular tolerances to perform tolerances design. With orientational tolerances converted into equivalent sized or angular tolerances, the composite tolerance chains are formulated. Sequential optimal models are presented to obtain optimal working dimensions and tolerances for remaining operations. The working tolerances are amplified gradually and manufacturing capabilities are enhanced.

This paper is structured as follows. A new method for presenting the dimensional chains from given process planning is discussed in section 2. In section 3, a method for presenting the composite tolerance chains is discussed. In section 4, the optimal mathematical models for sequential tolerances design of 3D processed tolerances are discussed. Section 5 gives a practical example. Finally, section 6 concludes this study.

2. Automatic generation of process tolerance chains with SRG

When a n-operation part is processed by m machine tools in a particular direction, such as axial direction, the apparent path tracing or tree approach methods are usually used to generate the dimensional and tolerance chains for manual treatment (Ngoi & Ong, 1993; Ji, 1993a; Ji, 1993b; Ngoi & Cheong, 1998). If geometric tolerances are involved, only four out of total fourteen geometric tolerance specifications, which exhibit the characteristics similar to linear dimensions, are treated as equivalent dimensions and tolerances and then applied directly to tolerance chart. These four specifications are position, symmetry, concentricity, and profile of a line (surface) (He & Gibson, 1992; Ngoi & Tan, 1995; Ngoi & Seow, 1996; Tseng & Kung, 1999). In 1D case, the following dimensional and tolerance chains must be satisfied (Ji, 1993b):

$$\begin{aligned} [A]\{X\} &= \{C\} \\ [B]\{T_X\} &\le \{T_D\} \end{aligned} \qquad (1)$$

Where $\boldsymbol{A} = [a_{ij}]$ is a $m \times n$ coefficient matrix, $a_{ij} = 1$ and -1 for an increasing and decreasing constituent link of u_{di}, respectively. $a_{ij} = 0$ for otherwise. $\boldsymbol{X} = [u_1, u_2, \ldots, u_n]^T$ is a $n \times 1$ vector of the mean working dimensions. $\boldsymbol{C} = [u_{d1}, u_{d2}, \ldots, u_{dm}]^T$ is a $m \times 1$vector of mean values of B/P dimensions. $\boldsymbol{B} = [b_{ij}]$ is a $m \times n$ coefficient matrix. $b_{ij} = 1$ for an increasing and decreasing constituent link of u_{di}. $b_{ij} = 0$ for otherwise. $\boldsymbol{T}_X = [T_{u1}, T_{u2}, \ldots, T_{un}]^T$ is a $n \times 1$ vector of the working tolerances. $\boldsymbol{T}_D = [T_{d1}, T_{d2}, \ldots, T_{dm}]^T$ is a $m \times 1$vector of B/P tolerances.

When a complex part is machined, typically a number of operations are involved. Each B/P tolerance is usually expressed as a number of pertinent process tolerances. In previous researches, tremendous efforts have been contributed to 1D dimensional tolerances. Geometric tolerances as well as the interactions between them have not been investigated extensively when complex 3D parts are manufactured. When we machine a complex 3D part, two dimensions components are included to determine the position of a processed feature in 2D drawing in the given view plane. For example, for the part shown in Figure 1 (Zhao, 1987), the position of pin-hole Φ15.009±0.009 in the plane XOY is determined by coordinate dimensions and tolerances $-25 \pm \frac{1}{2} T_{N'x}$ and $28 \pm \frac{1}{2} T_{N'y}$. Similarly the position of incline plane B is determined by $L_{N'E} \pm \frac{1}{2} T_{N'E}$ and $60° \pm \frac{1}{2} T_{\alpha 1}$, where $L_{N'E}$ and $T_{N'E}$ be nominal distance and its tolerance from the axis of pin-hole to incline plane, respectively. $\alpha = 60°$ and $T_{\alpha 1}$ be nomi-

nal angle and its tolerance formed by axis OX and the normal line of incline plane, respectively.

The series of orderly processing operations of a part is generalized as the set $A_p = \{O_{p1}, O_{p2}, \ldots, O_{pn}\}$, $i = 1, 2, \ldots, n$ is the number of machining operations including turning, milling, boring, and grinding etc. The set of working dimensions and tolerances in the view plane is denoted as $\Psi = \{u_1 \pm \frac{1}{2}T_1, u_2 \pm \frac{1}{2}T_2, \ldots, u_{2n} \pm \frac{1}{2}T_{2n}\}$, where $u_i \pm \frac{1}{2}T_i$, $i = 1, 2, \ldots, 2n$ are the working dimension and tolerance components assigned to the part. Since the working dimensions include sized and angular dimensions, the corresponding tolerance can be sized or angular ones. The constraint set of B/P dimensions and tolerances is denoted as $D_{st} = \{u_{d1} \pm \frac{1}{2}T_{d1}, u_{d2} \pm \frac{1}{2}T_{d2}, \ldots, u_{d2m} \pm \frac{1}{2}T_{d2m}\}$, $i = 1, 2, \ldots, 2m$ denotes 2m B/P sized and angular dimensions and tolerances of the part. The set of B/P orientational tolerances is denoted as $T_G = \{T_{G1}, T_{G2}, \ldots, T_{Gk}\}$, $i = 1, 2, \ldots, k$ are B/P geometric tolerances. In order to establish the required tolerance equations between B/P and pertinent working tolerances, dimensional chains must be derived from process planning to represent the relations between B/P and working dimensions.

In order to discuss further this issue, we introduce a practical example shown in Figure 1(Zhao, 1987). For simplicity, only the finishing operations are taken into account. The inclined hole (Φ25.0105 ± 0.0105) and inclined plane (B) of the example have high positional precision requirements. Thus the finish operations on incline hole and incline plane are executed with jig boring and grinding machine, respectively. Point D denotes the intersection of the axis of cylinder Φ89.974±0.011 with horizontal plane W. Point C is the intersection of the axis of incline hole with plane W. Let coordinates origin O lie at the intersection point of the axis of cylinder Φ89.974±0.011 with plane A. Axis OX lies in plane A and is parallel with plane S. Axis OY is perpendicular to plane A. Axis OZ is perpendicular to plane S. The functional requirements of this part are as such: The distance from point C to D is $x_{Cd} = 8 \pm 0.07$. The functional distance between plane A and W is $y_{Cd} = 25.075 \pm 0.075$. The functional distance between incline plane B and point C is $L_{CFd} = 54 \pm 0.12$. The other requirements are shown in Figure 1. Because functional dimension x_{Cd} and L_{CFd} cannot be measured directly, the finish machining processes involved are assigned as bellows:

1. Set plane A to vertical position to guarantee that plane S is parallel with horizontal plane. Choose plane A and axial line of the shaft Φ89.974±0.011 as references. Move the table of jig boring machine to due position and process the pin hole Φ15.009 ± 0.009 and ensure that the coordinates and

tolerances of axial line of the pin hole as $x_{N'} \pm T_{N'x}/2 = -25 \pm T_{N'x}/2$, $y_{N'} \pm T_{N'y}/2 = 28 \pm T_{N'y}/2$.

2. When a measurement pin is plugged into the pin hole, it is desire that parallelism between axial line of the pin to plane A be not more than TN◎y and perpendicularity of axial line of the pin to plane S along OX axis be not more than TN⊥x.
3. Take a measurement of the related complete sized dimensions xN′, yN′, and yC.
4. Turn plane A to horizontal direction in the table of jig boring machine. Then plane A is rotated an angle of 30°. Ensure that the distance between axial line of the pin to that of incline hole is LNB± TLNB/2. Where LNB is nominal dimension of the distance form axial line of the pin to that of incline hole. TNB is the tolerance of LNB. Bore incline hole Φ25.0105 ± 0.0105 and ensure that its axial line and that of Φ89.974±0.011 is in the same plane. The angle of axial line of incline hole is α1 = 60° and its tolerance Tα1 is directly controlled.
5. Take a measurement of the related complete sized dimension LNB.

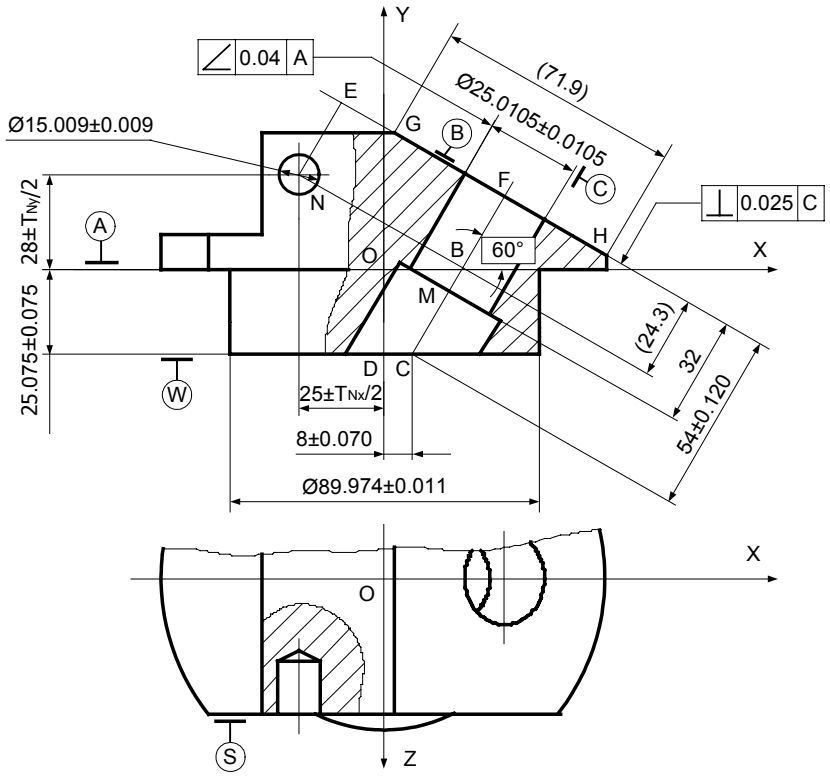


Figure 1. The 2D mechanical drawing of a 3D machined part

6. Grind incline plane B in grinding machine and guarantee that the distance between axial line of the pin to incline plane B with following dimensions and tolerances: $L_{NE} \pm T_{NE}/2$ and $30° \pm T\alpha_2/2$. Where L_{NE} and T_{NE} is nominal dimension and tolerance of the distance from axial line of the pin to incline plane B, respectively. $\alpha_2 = 30°$ and $T\alpha_2$ are nominal angle value and tolerance of inline plane B to OX axis, respectively.

In term of the above process processing, it is necessary that incline hole and incline plane of the example work piece are thus be processed economically within their dimension and tolerance ranges. The problem needs to be solved is: Establish pertinent dimensional chains in terms of the above manufacturing procedures, give the optimal model to the tolerance allocation problem, and find the optimal solutions. The finish machining process plan is generalized in table 1.

No	Operation	Reference(s)	Processing feature	Coordinates/ dimensions	tolerance
05	Boring	Plane A and axis of Φ89.974±0.011	Hole N′ Φ15.009±0.009	$x_{N'} = -25$ $y_{N'} = 28$	$T_{N'x}$ $T_{N'y}$
10	Pinning	No	Hole N′	$x_N = -25$ $y_N = 28$	$T_{N\perp x}$ $T_{N//y}$
15	Measure the complete sized dimensions x_N, y_N, and y_C				
20	Boring	Plane A and axis of Φ89.974±0.011	Incline hole Φ25.0105±0.0105	L_{NB} $\alpha_1 = 60°$	T_{NB} $T_{\alpha 1}$
25	Measure the complete sized dimensions L_{NB}				
30	Grinding	Plane A and axis of Φ89.974±0.011	Incline plane B	L_{NE} $\alpha_2 = 60°$	T_{NE} $T_{\alpha 2}$

Table 1. Finishing process plan of the part (Huang et al., 2002)

Unlike previous 1D case in conventional tolerance chart, the methods for generating dimensional chains are two-dimensional related. In other words, because every feature in the view plane has two dimension components, each link of a dimensional chain should contain two dimension components. Therefore we can use vector equation to present dimensional chains in the given 2D view plane.

In Figure 2, when incline hole is bored, the position of point C is indirectly obtained by controlling the position of pin, the distance from pin axis to that of incline hole, and angle α formed by axis OX and the axis of incline hole. Line segment NE is perpendicular to incline plane and point E is the intersection. Point F is the intersection of the axis of incline hole with incline plane. The line segment NB is perpendicular to the axis of incline hole and point B is the intersection.

To generate process tolerance chains correctly, we make use of a special relevant graph (SRG), which can be constructed directly from the process planning of the component, to express the interconnection and interdependence of the processed elements in their dimensions and tolerances in a more comprehensive way. In SRG, there are two kinds of nodes, one for the relevant elements of the component and another for their dimensions and tolerances. By searching through the SRG and coupled with the unique algorithm, dimension and tolerance chains needed relevant to the sequences of the processing plans are generated automatically.

Consider the pertinent point O, N, B, C, E, and F shown in Figure 2, the SRG model is constructed directly form the processing plan as show in Figure 3, where the dimension nodes and the element nodes are used. Dimension nodes are used to describe the dimensions relative to two pertinent elements of the work piece. Element nodes, however, are used to present the geometric elements of the work piece. The geometric elements refer to a point, a center line, or a plane of the work piece. In the graphical representation of the work piece under consideration, a block represents a dimensional node, while a circle corresponds to an element. The block drawn by slender lines is a component dimension node and the block drawn by dotted lines is a resultant one. Because two pertinent dimensions and tolerances must be included to determine the position and variation ranges of an element to origin O or the relative position to its pertinent reference(s), it is reasonable to introduce two dimension nodes to represent its two relative dimensions and tolerance components for an element. The link lines between dimension and element node indicate the interconnection and interdependence among them.

The process tolerance chains can be automatic generated through searching of the SRG coupled with the unique algorithm. The procedure is generalized as follows.

1. For each two selected resultant dimensions, choose any one of the elements relevant to them as the starting element node. Find two correspon-

ding pertinent component dimension nodes linked to it and get to another element node(s). Verify if these two component dimension nodes are linked to the same element node. If this is true, the ending element node obtained is used again as the starting element node and repeat the above process. Otherwise get two different element nodes. The two different element nodes obtained are used respectively again as the starting element node and repeat the above process until intersection element node is acquired. The searching direction is chosen to go along the SRG in a loop with the ending element node coming back to the starting element node, while the searching routes without duplicating the same element and dimension node more than once.

2. Every dimension chain can only contain two resultant dimensions and the minimum numbers of relative dimensions, otherwise, give up this loop and go to step (1).
3. Every resultant dimension is placed on the left side of equation and the other relative dimensions are placed on the right side. With these steps, it is easily to find that the four points O, N, B, and C and the five points O, N, E, F and C shown in Figure 1 and Figure 2 compose respectively a planar dimensional chain.

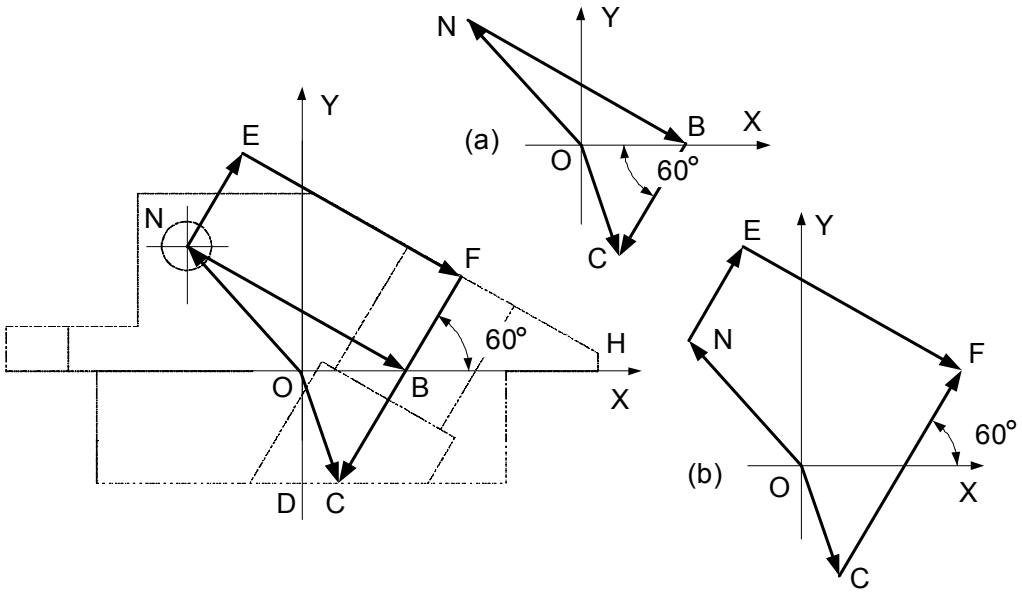


Figure 2. The vector relations between pertinent features

When incline hole is machined, the vector equation of the position of point C is:

$$\overrightarrow{OC} = \overrightarrow{ON} + \overrightarrow{NB} + \overrightarrow{BC} \tag{2}$$

Where $\overrightarrow{OC}$ is position vector of point C, $\overrightarrow{ON}$ is position vector of point N, $\overrightarrow{NB}$ and $\overrightarrow{BC}$ are relative vector from point N to point B and from point B to point C, respectively. When Equation (2) is expressed as algebraic equations, we have

$$\begin{aligned} x_N + L_{NB}\cos 30^\circ - L_{BC}\sin 30^\circ &= x_{Cd} \\ y_N - L_{NB}\sin 30^\circ - L_{BC}\cos 30^\circ &= y_{Cd} \end{aligned} \tag{3}$$

Where x_N and y_N are coordinate component of the axis of the pin. L_{NB} and L_{BC} are nominal length between point N and B, point B and C, respectively. x_{Cd} and y_{Cd} are the B/P coordinates of point C.

Similarly, when incline plane is machined, the distance from point F to point C is indirectly obtained by controlling the position of pin, the distance from pin axis to incline plane, and the angle α formed by axis OX and the normal line of incline plane. The vector equation is:

$$\overrightarrow{CF} = \overrightarrow{ON} + \overrightarrow{NE} + \overrightarrow{EF} - \overrightarrow{OC} \tag{4}$$

Where $\overrightarrow{CF}$ is relative vector from point C to point F, $\overrightarrow{NE}$ and $\overrightarrow{EF}$ are relative vector from point N to point E, and from point E to point F, respectively. It is easy to find in Figure 2 that the length of line segment L_{EF} is equal to the length of line segment L_{NB}, i.e. $L_{EF} = L_{NB}$. Also, when we represent Equation (4) into algebraic equations, we get

$$\begin{aligned} &y_N + L_{NE}\cos 30^\circ - L_{EF}\sin 30^\circ - y_C \\ &= L_{CFd}\cos 30^\circ, \text{ where } L_{EF} = L_{NB} \end{aligned} \tag{5}$$

Where L_{NE} is nominal length between point N and E. x_C and y_C are the coordinates of point C. L_{CFd} is the B/p length between point C and point F.

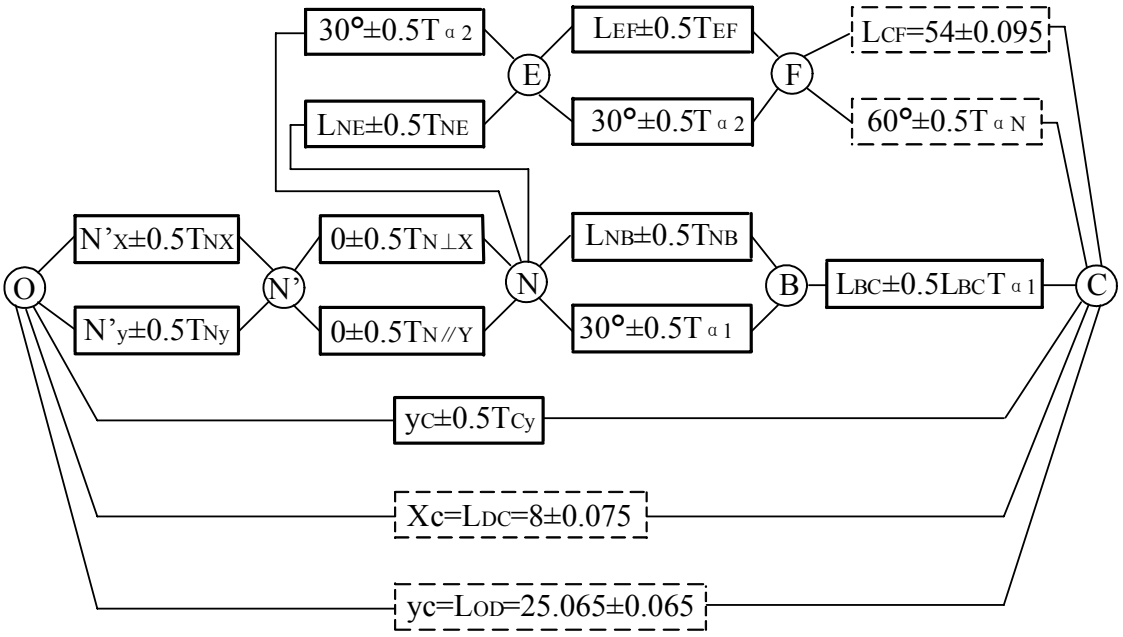


Figure 3. The SRG model of the work piece relevant to the processing plan

4. With resultant dimension chains established, the relative tolerance chain is generated in the graphic way that the resultant tolerance zone should envelope all of the pertinent component tolerance zones and it is also enveloped by design tolerance zone.

The algebraic dimensional chains related to Equation 2 and 4 are:

$$\begin{bmatrix} 1 & 0 & \cos 30^\circ & -\sin 30^\circ & 0 & 0 \\ 0 & 1 & -\sin 30^\circ & -\cos 30^\circ & 0 & 0 \\ 0 & \dfrac{1}{\cos 30^\circ} & -tg30^\circ & 0 & 1 & \dfrac{-1}{\cos 30^\circ} \end{bmatrix} \begin{bmatrix} x_N \\ y_N \\ L_{NB} \\ L_{BC} \\ L_{NE} \\ y_C \end{bmatrix} = \begin{bmatrix} x_{Cd} \\ y_{Cd} \\ L_{CFd} \end{bmatrix} \qquad (6)$$

3. Tolerance zones and tolerances accumulation

The shapes of tolerance zones in the view plane vary with the dimensions and tolerances specified to the feature. Several cases are given in Figure 4 to illustrate this issue in the view plane XOY. The different shape of parallelogram shown in Figure 4(a)-(c) corresponds to a particular tolerance zone of point A which is controlled by two different dimensions and tolerances. The tolerance zone is center at point A and its position is controlled by $L_{OA}\pm\frac{1}{2}T_{OA}$ and $Y\pm\frac{1}{2}T_Y$, $X\pm\frac{1}{2}T_X$ and $Y\pm\frac{1}{2}T_Y$, and $L_{OA}\pm\frac{1}{2}T_{OA}$ and $60°\pm\frac{1}{2}T_\alpha$, respectively. Figure 4 (d)-(e) corresponds to two different cases of tolerance accumulations.

Figure 4(d) shows the tolerance accumulation case when one-base-point is related. This case is defined when the two dimension and tolerance components of a feature are related to only one reference feature (base point). Assume that parallelogram 1 is tolerance zone of base point A and parallelogram 2 is tolerance zone of point B relative to base point A. Resultant tolerance zone of point B is obtained by adding up the above two tolerance zones geometrically. So we can move parallelogram 2 parallelly along the outline of parallelogram 1 and the zone enveloped by outmost contour of parallelogram 2 forms the resultant tolerance zone of point B. If B/P dimensions and tolerances of point B are specified as $X_B\pm\frac{1}{2}T_{BX}$ and $Y_B\pm\frac{1}{2}T_{BY}$, for acceptable point B, B/P tolerance zone (drawn by dotted lines and measured by Tx and Ty) must envelop resultant tolerance (see right hand side in Figure 4 (d)).

Figure 4(e) shows another case of tolerance accumulation when two-base-point is related. This case is defined when two dimension and tolerance components of a feature are related respectively to two different reference features (two base points). If the smaller parallelogram centers at point C is tolerance zone of point C relative to its two base points i.e. point A and B. The parallelogram center at point A and B are tolerance zone of base point A and B, respectively. Resultant tolerance zone of point C is obtained as such. First, extract tolerance zone of point C which is resultant tolerance zone of point A and B (denote as 1-2-3-4). So draw two parallel lines perpendicular to line segment AC and let the distance between them be the tolerance magnitude of base point A in the direction of line segment AC. Similarly, draw another two parallel lines perpendicular to line segment BC and let the distance between them be the tolerance magnitude of base point B in the direction of ling segment BC. The zone formed by these four lines will construct a bigger parallelogram 1-2-3-4 which centers at point C. It is the resultant tolerance zone of point C resulting from its two-base-point tolerance zones. Then, move the smaller parallelogram center

at point C parallelly along the outline of parallelogram 1-2-3-4. The zone enveloped by outmost contour of the smaller parallelogram forms resultant tolerance zone of point C. If B/P tolerance zone of point C is a rectangle (drawn by dotted lines), the rectangle must envelop resultant tolerance zone when point C is acceptable.

Tolerance accumulation and the relationships between different sorts of tolerance specifications must be solved in presenting composite tolerance chains. For given orientational tolerances shown in Figure 1, tolerance accumulation process is dependent upon the characteristic attributes they exhibit.

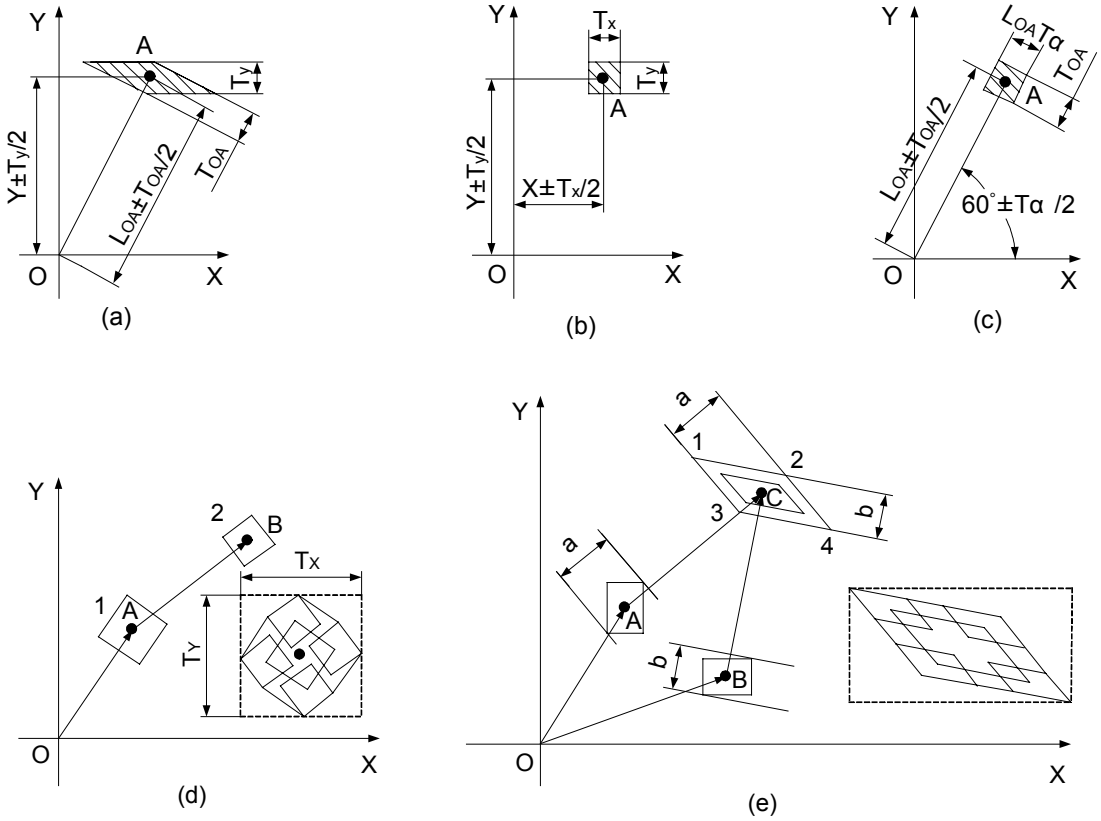


Figure 4. Tolerance zone and its stack-up in view plane XOY

In Figure 1, the tolerance zone of angularity of incline hole relative to plane A (datum A) is represented by rectangle area with shadow lines shown in Figure 5(a). Assume that the axis of incline hole has ideal geometric shape and angularity tolerance can be assured by controlling the tolerance of angle α formed by axis OX and the axis of incline hole. The expression is:

$$T_{\alpha d1} = \frac{T_{\angle}}{L_{FM}} \times 2 \tag{7}$$

Where $T_{\alpha d1}$ is equivalent design angular tolerance determined by tolerance $T_{\angle}$, which is angularity tolerance of the axis of incline hole relative to plane A. L_{FM} is nominal length of incline hole.

Similarly, in Figure 1 the tolerance zone of perpendicularity of incline plane relative to the axis of incline hole (datum C) is represent by rectangle area with shadow lines shown in Figure 5(b). Assume that incline plane has ideal geometric shape, angularity tolerance can be assured by controlling the tolerance of angle α formed by axis OX and normal line of incline plane. The expression is:

$$T_{\alpha d2} = \frac{T_{\perp}}{L_{GH}} \times 2 \tag{8}$$

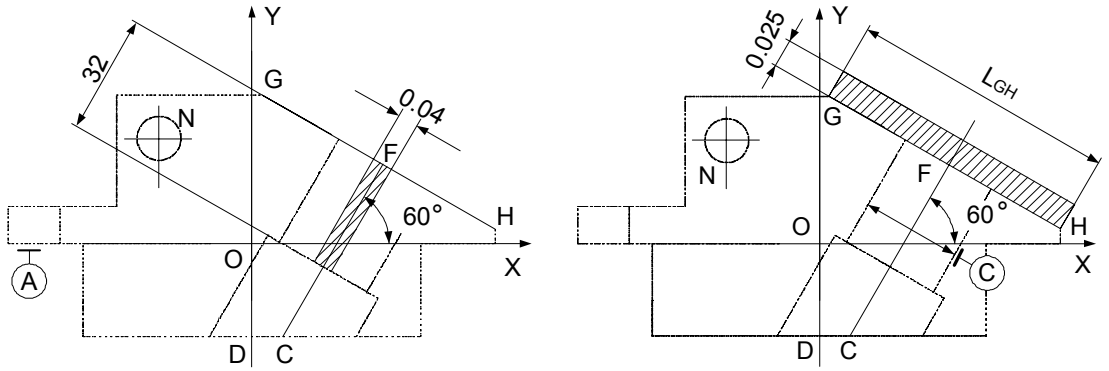
Where $T_{\alpha d2}$ is equivalent design angular tolerance determined by $T_{\perp}$, which is perpendicularity tolerance of the incline plane relative to the axis of incline hole. L_{GH} is nominal length of incline plane.

Furthermore, according to the functional role of pin, when it is plugged into pin-hole, the following equations should be satisfied:

$$\begin{cases} T_{Nx} = T_{N'x} + T_{N\perp x} \\ T_{Ny} = T_{N'y} + T_{N\parallel y} \end{cases} \tag{7}$$

Where T_{Nx} and T_{Ny} is composite tolerance component of pin axis, respectively. $T_{N'x}$ and $T_{N'y}$ is tolerance component of the axis of pin-hole, respectively. $T_{N\perp x}$ and $T_{N//y}$ is perpendicularity of pin axis to plane S in the direction of axis OX and parallelism between pin axis and plane A, respectively.

With above discussion, for Equation 2, the tolerance zone of each vector and their accumulation is shown in Figure 6. Where zone 1-2-3-4 is the tolerance zone of pin axis. Zone 5-6-7-8 is the tolerance zone of point B relative to pin axis. Zone 9-10-11-12 is the tolerance zone of point C relative to its two base points i.e. origin O and point B. Where $L_{BC}T_{\alpha 1}$ is the tolerance component perpendicular to line segment BC and T_{Cy} is another tolerance component in the direction of axis OY.



(a) (b)

Figure 5. Relationships between angular tolerance and orientational tolerances

Using the above tolerance accumulation principle discussed in Figure 4(e), the final resultant tolerance zone of point C is obtained through following steps. First, find tolerance zone of point C resulting from its two-base-point tolerance zones and denote as £, which is acquired by adding up its two base point tolerance zones in two due directions. For base point B, its tolerance zone is obtained by adding up the zone 1-2-3-4 and 5-6-7-8 geometrically. Because the direction of tolerance component of point C relative to point B is perpendicular to line segment BC (also along line segment *NB*) and its magnitude is $L_{BC}T_{\alpha 1}$, the tolerance magnitude of point B in this direction is expressed as:

$$T_{B\perp} = T_{N\perp} + T_{NB} \tag{10}$$

Where $T_{B\perp}$ is tolerance component of point B in the direction perpendicular to line segment BC. $T_{N\perp}$ is the tolerance of pin axis in the direction perpendicular to line segment BC. T_{NB} is the tolerance of L_{NB}, which is mean dimension of the distance form pin axis to that of incline hole.

On the other hand, another component of £ is in the direction of axial OY. Because the origin is fixed, the distance of £ in the direction of axial OY is nil. So £ is finally obtained as a horizontal line segment *mn* shown in the right down side in Figure 6. The length of line segment *mn* is:

$$L_{mn} = T_{Nx} + T_{Ny} tg30^\circ + \frac{T_{NB}}{\cos 30^\circ} \tag{11}$$

Zone 13-14-15-16 is resultant tolerance of point C and finally acquired by mov-

ing parallelly zone 9-10-11-12 along line segment *mn*. B/P tolerance zone *I - II- III- IV* (drawn by dotted lines) should envelop zone 13-14-15-16 (right up side in Figure 6). The algebraic equations are:

$$\begin{cases} T_{Nx} + T_{Ny} tg30^\circ + \dfrac{T_{NB} + L_{BC} T_{\alpha 1}}{\cos 30^\circ} + T_{Cy} tg30^\circ \le T_{Cxd} \\ T_{Cy} \le T_{Cyd} \end{cases} \qquad (12)$$

Where T_{Cxd} = 0.140mm and T_{Cyd} = 0.150mm are two components of B/P tolerance of point C. For Equation 4, tolerance zone of each vector and their tolerance accumulation is shown in Figure 7. Zone 1-2-3-4 and 13-14-15-16 have been discussed above. Zone 17-18-19-20 is tolerance zone of point E relative to pin axis, zone 21-22-23-24 is tolerance zone of point F relative to point E, and zone *V- VI- VII- VIII* (drawn by dotted lines) is B/P tolerance zone of point F relative to point C. Resultant tolerance zone is the one of point F relative to point C. It includes four components: zone 1-2-3-4, 17-18-19-20, 21-22-23-24, and 13-14-15-16. It is necessary that B/P tolerance zone contain its resultant tolerance zone for an acceptable part.

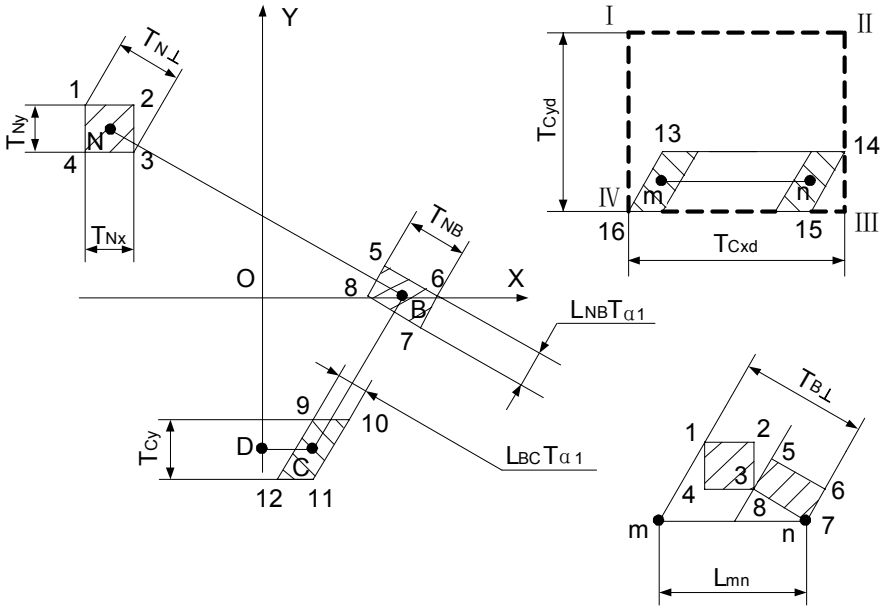


Figure 6. Original tolerance accumulation between point O, N, B, and C

When we change graphic representation into algebraic form, we can only considerer the tolerance component in the direction of line segment FC. For the

acceptable parts, the resultant tolerance component should be less than or equal to its B/P tolerance component. The algebraic equation is:

$$2T_{Nx}\sin 30^\circ + \frac{T_{Ny}}{\cos 30^\circ} + T_{NB}tg30^\circ + L_{BC}T_{\alpha 1}tg30^\circ + T_{NE} + L_{EF}T_{\alpha 2} + \frac{T_{Cy}}{\cos 30^\circ} \le T_{CFd} \tag{13}$$

Where T_{CFd} = 0.240mm is B/P tolerance of L_{CFd}.

4. 3 D sequential tolerance design

In process planning, each machining operation is specified with an appropriate tolerance based on the constraints of B/P specification and process capability. In conventional dimensioning and tolerancing, all working dimensions and process tolerances are fixed. This method, however, is suitable for mass, batch, and automated production. A new method termed STC for production of complex, low-volume, and high-value-added parts was introduced (Fraticelli et al., 1997; Fraticelli et al., 1999; Wheeler et al., 1999; Cavalier & Lehtihet, 2000; Mcgarvey et al., 2001; Huang & Zhong, (in press)). The method essentially used real-time measurement information at any completion stage of operations to exploit available space inside the dynamic feasible zone and recalculate the working dimensions and tolerances for remaining operations. It has been proved that this method can enhance the process tolerances for remaining operations and increase the acceptable rate of manufacturing.

The above researches, however, did not include geometric tolerances and were confined to 1D problem. This paper aims to extend STC method to 3D space when angular and orientational tolerances are also involved. The method essentially utilizes the measurement data of sized dimensions at appropriate completion stage of operations to evaluate the working dimensions and tolerances for remaining operations based on the process capabilities. Let actual working dimension and deviation be set $M = \{u_i^*, \Delta_{ui}, j=1, \ldots, 2n\}$, where u_i^* is acquired measurement value of working dimension u_i, $\Delta_{ui} = u_i^* - u_i$ is actual deviation of working dimension u_i^*. The original dimensional and process tolerance chains are respectively expressed in the following matrix form.

$$\begin{aligned}[A]\{X\}&=\{C\}\\ [B]\{T_X\}&\le\{T_D\}\end{aligned} \tag{14}$$

Where $\boldsymbol{A} = [a_{ij}]$ is a $2m\times2n$ coefficient matrix, $\boldsymbol{X} = [u_1, u_2,\ldots, u_{2n}]^T$ is a $2n\times1$ vector of mean working dimensions, $\boldsymbol{C} = [u_{d1}, u_{d2},\ldots, u_{d2m}]^T$ is a $2m\times1$vector of mean values of B/P dimensions, $\boldsymbol{B} = [b_{ij}]$ is a $2m\times2n$ coefficient matrix, $\boldsymbol{T}_X = [T_{u1}, T_{u2},\ldots, T_{u2n}]^T$ is a $2n\times1$ vector of working tolerances, and $\boldsymbol{T}_D = [T_{d1}, T_{d2},\ldots, T_{d2m}]^T$ is a $2m\times1$vector of B/P tolerances.

When incline features are included, a_{ij} is the function of a number of pertinent working dimensions. While $b_{ij} = \partial u_{di} / \partial u_j$ is determined by the way tolerance accumulates. For generalized description, assume that each operation associates with two components of different dimensions and tolerances in given view plane.

The generalized algorithm of 3D sequential tolerance design is expressed as following steps.

Step 1:

The original optimal tolerance design is implemented at this step. Sized, angular, and orientational tolerances are included in composite tolerance chains. Orientational tolerances are first converted into equivalent sized or angular tolerance in terms of their characteristic attributes. The composite tolerance chains are established using the methods discussed in section 3. The original optimal model is:

$$\max \sum_{i=1}^{n} \lambda_{ui} k_{ui} T_{ui} \tag{15}$$

Subject to:

$$\begin{bmatrix} b_{11}^{(1)} & b_{12}^{(1)} & \cdots & b_{12n}^{(1)} \\ b_{21}^{(1)} & b_{22}^{(1)} & \cdots & b_{22n}^{(1)} \\ \vdots & \cdots & \ddots & \vdots \\ b_{2m1}^{(1)} & b_{2m2}^{(1)} & \cdots & b_{2m2n}^{(1)} \end{bmatrix} \begin{bmatrix} T_{u1}^{(1)} \\ T_{u2}^{(1)} \\ \cdots \\ T_{u2n}^{(1)} \end{bmatrix} \le \begin{bmatrix} T_{d1} \\ T_{d2} \\ \cdots \\ T_{d2m} \end{bmatrix} \tag{16}$$

$$T_{ui\min}^{(1)} \le T_{ui}^{(1)} \le T_{ui\max}^{(1)}, i = 1,2,\ldots,2n \tag{17}$$

Where

$$\begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{12n}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & \cdots & a_{22n}^{(1)} \\ \vdots & \cdots & \ddots & \vdots \\ a_{2m1}^{(1)} & a_{2m2}^{(1)} & \cdots & a_{2m2n}^{(1)} \end{bmatrix} \begin{bmatrix} u_1^{(1)} \\ u_2^{(1)} \\ \cdots \\ u_{2n}^{(1)} \end{bmatrix} = \begin{bmatrix} u_{d1} \\ u_{d2} \\ \cdots \\ u_{d2m} \end{bmatrix} \tag{18}$$

$T^{(1)}{}_{ui}$: Original working tolerance component of dimension u_i

k_{ui}: Weight factor of tolerance T_{ui}, which is dependent upon the capacity of machining and the manufacturing cost of operation. k_{ui} is determined by the experience of a process planner.

λ_{ui}: Path selected coefficient of dimension u_i. When u_i is selected, $\lambda_{ui} = 1$, otherwise $\lambda_{ui} = 0$.

$T^{(1)}{}_{uimin}$,

$T^{(1)}{}_{uimax}$: Lower and upper bound of original tolerance T_{ui}, respectively.

When above original optimal model is solved, original working dimensions and optimal working tolerances are obtained. The operations of first stage are performed based on above original working dimensions and tolerances. Then pertinent sized dimensions are measured. Assume that u_1, u_2, …, and u_k are k measured sized dimensions. u_1^*, u_2^*, …, and u_k^* are corresponding actual measured values. The actual deviations are obtained as $\Delta_{uk}^* = u_k^* - u_k$, i=1, 2, …, k. If actual dimensions u_1^*, u_2^*, …, and $u_k^*$are within their permissible ranges, they are substituted into Equation 18. The working dimensions and tolerances for next operation step can be determined.

For dimensions u_1, u_2, …, and u_k, with their actual values measured, their tolerances T_{u1}, T_{u2}, …, and T_{uk} do not include in the tolerance chains for remaining operations. Thus the numbers of constituent tolerance links are reduced by k components, the working tolerances reassigned to remaining operations increase. The bounds of working tolerance of remaining operations can be readjusted for the purpose of ease machining.

The working dimensions and tolerances for operations of the next stage are determined by following sequential optimal model.

Step 2:

$$\max \sum_{i=k+1}^{n} \lambda_{ui} k_{ui} T_{ui} \tag{19}$$

Subject to:

$$\begin{bmatrix} b^{(2)}_{1\,k+1} & b^{(2)}_{1k+2} & \cdots & b^{(2)}_{12n} \\ b^{(2)}_{2\,k+1} & b^{(2)}_{2k+2} & \cdots & b^{(2)}_{22n} \\ \vdots & \cdots & \ddots & \vdots \\ b^{(2)}_{2m\,k+1} & b^{(2)}_{2m\,k+2} & \cdots & b^{(2)}_{2m\,2n} \end{bmatrix} \begin{bmatrix} T^{(2)}_{u\,k+1} \\ T^{(2)}_{u\,k+2} \\ \cdots \\ T^{(2)}_{u\,2n} \end{bmatrix} \le \begin{bmatrix} T_{d1} \\ T_{d2} \\ \cdots \\ T_{d2m} \end{bmatrix} \tag{20}$$

$$T^{(2)}_{ui\,\min} \le T^{(2)}_{ui} \le T^{(2)}_{ui\,\max}, i = k+1, k+2, \ldots, 2n \tag{21}$$

Where

$$\begin{bmatrix} a^{(2)}_{1\,k+1} & a^{(2)}_{1k+2} & \cdots & a^{(2)}_{12n} \\ a^{(2)}_{2\,k+1} & a^{(2)}_{2k+2} & \cdots & a^{(2)}_{22n} \\ \vdots & \cdots & \ddots & \vdots \\ a^{(2)}_{2m\,k+1} & a^{(2)}_{2m\,k+2} & \cdots & a^{(2)}_{2m\,2n} \end{bmatrix} \begin{bmatrix} u^{(2)}_{k+1} \\ u^{(2)}_{k+2} \\ \cdots \\ u^{(2)}_{2n} \end{bmatrix} = \begin{bmatrix} u_{d1} \\ u_{d2} \\ \cdots \\ u_{d2m} \end{bmatrix} - \begin{bmatrix} a^{(1)}_{11} & a^{(1)}_{12} & \cdots & a^{(1)}_{1k} \\ a^{(1)}_{21} & a^{(1)}_{22} & \cdots & a^{(1)}_{2k} \\ \vdots & \cdots & \ddots & \vdots \\ a^{(1)}_{2m1} & a^{(1)}_{2m2} & \cdots & a^{(1)}_{2mk} \end{bmatrix} \begin{bmatrix} u^{*}_{1} \\ u^{*}_{2} \\ \cdots \\ u^{*}_{k} \end{bmatrix} \tag{22}$$

$T^{(2)}_{ui}$: Second step working tolerance component of dimension u_i

$T^{(2)}_{uimin}$,

$T^{(2)}_{uimax}$: Lower and upper bound of tolerance T_{ui} for second step operations, respectively.

When above optimal model is solved, the working dimensions and tolerances for operations of the second stage are obtained. Similarly, after operations of the second stage have been performed, their actual pertinent sized dimensions are measured and substituted into mean dimension chains to determine the working dimensions and tolerances for operations of the third stage.

Step 3:

Assume that q operations have been successfully performed. The actual pertinent sized dimensions have been measured and substituted into their mean dimensional chains. The third order optimal model is expressed as follows:

$$\max \sum_{i=q+1}^{n} \lambda_{ui} k_{ui} T_{ui} \tag{23}$$

Subject to:

$$\begin{bmatrix} b^{(3)}_{1\,q+1} & b^{(3)}_{1\,q+2} & \cdots & b^{(3)}_{1\,2n} \\ b^{(3)}_{2\,q+1} & b^{(3)}_{2\,q+2} & \cdots & b^{(3)}_{2\,2n} \\ \vdots & \cdots & \ddots & \vdots \\ b^{(3)}_{2m\,q+1} & b^{(3)}_{2m\,q+2} & \cdots & b^{(3)}_{2m\,2\mathrm{n}} \end{bmatrix} \begin{bmatrix} T^{(2)}_{u\,q+1} \\ T^{(2)}_{u\,q+2} \\ \cdots \\ T^{(2)}_{u\,2n} \end{bmatrix} \le \begin{bmatrix} T_{d1} \\ T_{d2} \\ \cdots \\ T_{d\,2m} \end{bmatrix} \tag{24}$$

$$T^{(3)}_{ui\min} \le T^{(3)}_{ui} \le T^{(3)}_{ui\max}, i = q+1, q+2, \ldots, 2n \tag{25}$$

Where

$$\begin{bmatrix} a^{(3)}_{1\,q+1} & a^{(3)}_{1q+2} & \cdots & a^{(3)}_{12n} \\ a^{(3)}_{2\,q+1} & a^{(3)}_{2q+2} & \cdots & a^{(3)}_{2\,2n} \\ \vdots & \cdots & \ddots & \vdots \\ a^{(3)}_{2m\,q+1} & a^{(3)}_{2m\,q+2} & \cdots & a^{(3)}_{2m\,2\mathrm{n}} \end{bmatrix} \begin{bmatrix} u^{(3)}_{q+1} \\ u^{(3)}_{q+2} \\ \cdots \\ u^{(3)}_{2n} \end{bmatrix} = \begin{bmatrix} u_{d1} \\ u_{d2} \\ \cdots \\ u_{d\,2m} \end{bmatrix} - \begin{bmatrix} a^{(1)}_{11} & a^{(1)}_{12} & \cdots & a^{(1)}_{1k} \\ a^{(1)}_{21} & a^{(1)}_{22} & \cdots & a^{(1)}_{2k} \\ \vdots & \cdots & \ddots & \vdots \\ a^{(1)}_{2m1} & a^{(1)}_{2m\,2} & \cdots & a^{(1)}_{2m\,k} \end{bmatrix} \begin{bmatrix} u^{*}_{1} \\ u^{*}_{2} \\ \cdots \\ u^{*}_{k} \end{bmatrix}$$

$$- \begin{bmatrix} a^{(2)}_{1\,k+1} & a^{(2)}_{1\,k+2} & \cdots & a^{(2)}_{1q} \\ a^{(2)}_{2\,k+1} & a^{(2)}_{2\,k+2} & \cdots & a^{(2)}_{2q} \\ \vdots & \cdots & \ddots & \vdots \\ a^{(2)}_{2m\,k+1} & a^{(2)}_{2m\,k+2} & \cdots & a^{(2)}_{2m\,q} \end{bmatrix} \begin{bmatrix} u^{*}_{k+1} \\ u^{*}_{k+2} \\ \cdots \\ u^{*}_{q} \end{bmatrix} \tag{26}$$

$T^{(3)}{}_{ui}$: Third step working tolerance component of dimension u_i

$T^{(3)}{}_{uimin}$,

$T^{(3)}{}_{uimax}$: Lower and upper bound of tolerance T_{ui} for third step operations, respectively.

The above procedure is repeated until the last operation has been performed. Because constituent tolerance component of obtained actual sized dimensions are excluded from remaining tolerance chains while B/P tolerances remains unchanged, the proposed approach gradually enhances working tolerances of remaining operations. The final solutions of sequential tolerances are given as:

$$T = \left[T_{u1}^{(1)} \quad T_{u2}^{(1)} \qquad T_{uk}^{(1)} \quad T_{uk+1}^{(2)} \quad T_{u\,k+2}^{(2)} \quad \cdots \quad T_{u\,q}^{(2)} \quad T_{u\,q+1}^{(3)} \quad \cdots\right]^T \qquad (27)$$

5. A case study

A practical example (Zhao, 1987) but with modifications shown in Figure 1 is introduced to illustrate the proposed method. The process plan of the finish operations is given in Table 1.

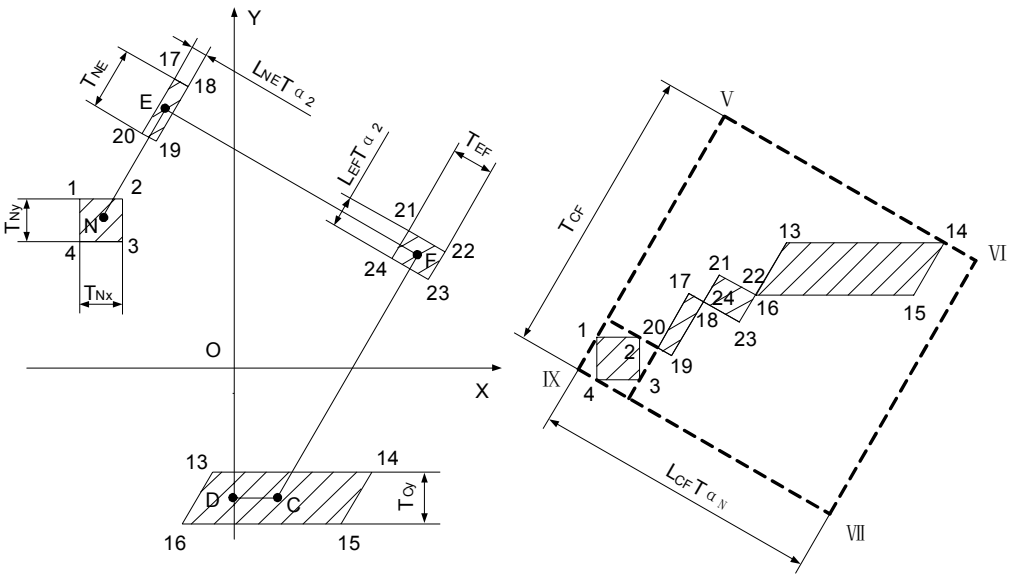


Figure 7. Original tolerance accumulation between point O, N, E, F, and C

5.1. Establishment of dimensional and tolerance chains

With the procedures discussed in section 2 and 3, dimensional and tolerance chains of this example part are given by formulation (6), (12), and (13), respectively.

5.2. Additional angular tolerance chains

According to the given process planning, the finish operations are executed with different machine tools. The accuracy of rotation working tables of machine tools provides the assurance of orientational tolerance express by perpendicularity and angularity tolerance. For jig boring machine, it is required that:

$$T_{\alpha 1} \le T_{\alpha d1} = \frac{T_{\angle}}{L_{FM}} \times 2 = \frac{0.04}{32} \times 2 = 0.0025 \text{ (rad)} \qquad (28)$$

Where $T_{\alpha 1}$ is angular tolerance of rotation working tables of jig boring machine. $T_{\alpha d1}$ is equivalent design angular tolerance determined by $T_{\angle}$. L_{FM} = 32 is nominal length of incline hole. $T_{\angle}$ = 0.04 is angularity tolerance of the axis of incline hole relative to plane A.

We can control the rotation error of rotation working table of grinding machine to ensure the perpendicularity tolerance. It is expressed as:

$$T_{\alpha N} = T_{\alpha 1} + T_{\alpha 2} \le T_{\alpha d2} = \frac{T_{\perp}}{L_{GH}} \times 2 = \frac{0.025}{71.9} \times 2 = 0.0007 \text{(rad)} \qquad (29)$$

Where $T_{\alpha N}$ is resultant angular tolerance of $T_{\alpha 1}$ and $T_{\alpha 2}$. $T_{\alpha 2}$ is angular tolerance of rotation working tables of grinding machine. L_{GH} = 71.9 is nominal length of incline plane. $T_{\perp}$ = 0.025 is perpendicularity tolerance of incline plane B relative to the axis of incline hole. $T_{\alpha d2}$ is the equivalent design angular tolerance determined by $T_{\perp}$. It is obvious that if formulation (29) is satisfied, formulation (28) is also satisfied.

5.3. Additional process capability constraints

Assume that jig boring and grinding machine have the same accuracy for their rotation working tables. The accuracy in axis OX and OY is the same. $T_{N\perp x}$ and $T_{N//y}$ have the same functional role. Thus we have

$$\begin{cases} T_{\alpha 1} = T_{\alpha 2} \\ T_{Nx} = T_{Ny} \\ T_{N\perp x} = T_{N\parallel y} \end{cases} \tag{30}$$

To ensure that the machined parts meet its designed functionality and minimum manufacturing cost, the original constraints of finishing processes are formulated in Table 2.

Operation	Tolerance	lower bound	upper bound	Weight
Boring	$T_{N'x}$	10	25	$k_1 = 1$
Boring	$T_{N'y}$	10	25	$k_2 = 1$
Pinning	$T_{N\perp x}$	7	10	$k_3 = 1$
Pinning	$T_{N//y}$	7	10	$k_4 = 1$
Boring	T_{NB}	30	75	$k_5 = 1.4$
Boring	$T_{\alpha 1}$	70″	90″	$k_6 = 1.4$
Grinding	T_{NE}	30	75	$k_7 = 1.4$
Grinding	$T_{\alpha 2}$	70″	90″	$k_8 = 1.4$
Turning	T_{Cy}	15	40	$k_9 = 1$

Table 2. Original working tolerance bounds (μm) and weights

5.4. Optimized sequential tolerance design procedure

In terms of the process planning developed for finish operations, related tolerances must be specified before any machining operation was executed. The optimization model is:

$$\max\left(k_1 T^{(1)}_{N'x} + k_2 T^{(1)}_{N'y} + k_3 T^{(1)}_{N\perp x} + k_4 T^{(1)}_{N\parallel y} + k_5 T^{(1)}_{NB} + k_6 T^{(1)}_{NE} + k_7 T^{(1)}_{\alpha 1} + k_8 T^{(1)}_{\alpha 2} + k_9 T^{(1)}_{Cy}\right)$$

s.t.

$$\begin{bmatrix} 1 & tg30^\circ & 1 & tg30^\circ & \dfrac{1}{\cos 30^\circ} & \dfrac{L^{(1)}_{BC}}{\cos 30^\circ} & tg30^\circ & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & \dfrac{1}{\cos 30^\circ} & 1 & \dfrac{1}{\cos 30^\circ} & tg30^\circ & L^{(1)}_{BC} tg30^\circ & \dfrac{1}{\cos 30^\circ} & 1 & L^{(1)}_{EF} \end{bmatrix} \begin{bmatrix} T^{(1)}_{N'x} \\ T^{(1)}_{N'y} \\ T^{(1)}_{N\perp x} \\ T^{(1)}_{N\parallel y} \\ T^{(1)}_{NB} \\ T^{(1)}_{\alpha 1} \\ T^{(1)}_{Cy} \\ T^{(1)}_{NE} \\ T^{(1)}_{\alpha 2} \end{bmatrix} \le \begin{bmatrix} T_{Cxd} \\ T_{Cyd} \\ T_{CFd} \end{bmatrix}$$

$T_{\alpha 1}^{(1)} + T_{\alpha 2}^{(1)} \le 0.0007$

$0.010 \le T_{N'x}^{(1)} \le 0.025$, $0.010 \le T_{N'y}^{(1)} \le 0.025$, $0.007 \le T_{N\perp x}^{(1)} \le 0.010$, $0.007 \le T_{N\parallel y}^{(1)} \le 0.010$, $0.030 \le T_{NB}^{(1)} \le 0.075$, $0.00034 \le T_{\alpha 1}^{(1)} \le 0.0044$, $0.030 \le T_{NE}^{(1)} \le 0.075$, $0.00034 \le T_{\alpha 2}^{(1)} \le 0.00044$, $0.015 \le T_{Cy}^{(1)} \le 0.040$.

Where

$$\begin{bmatrix} 1 & 0 & \cos 30^\circ & -\sin 30^\circ & 0 & 0 \\ 0 & 1 & -\sin 30^\circ & -\cos 30^\circ & 0 & 0 \\ 0 & \dfrac{1}{\cos 30^\circ} & -tg30^\circ & 0 & 1 & \dfrac{-1}{\cos 30^\circ} \end{bmatrix} \begin{bmatrix} x_{N'}^{(1)} \\ y_{N'}^{(1)} \\ L_{NB}^{(1)} \\ L_{BC}^{(1)} \\ L_{NE}^{(1)} \\ y_C^{(1)} \end{bmatrix} = \begin{bmatrix} x_{Cd} \\ y_{Cd} \\ L_{CFd} \end{bmatrix},$$

$L_{EF}^{(1)} = L_{NB}^{(1)}$, $k_1 = k_2 = k_3 = k_4 = k_9 = 1$, $k_5 = k_6 = k_7 = k_8 = 1.4$, $x_{N'} = -25$, $y_{N'} = 28$, $x_{Cd} = 8$, $y_{Cd} = -25$, $L_{CFd} = 54$, $T_{Cxd} = 0.140$, $T_{Cyd} = 0.150$, and $T_{CFd} = 0.240$.

The solution of the model is:

$$\left[L_{NB}^{(1)} \quad L_{BC}^{(1)} \quad L_{NE}^{(1)}\right]^T = \left[55.078 \quad 29.400 \quad 24.600\right]^T$$

$$\left[T_{N'x}^{(1)} \quad T_{N'y}^{(1)} \quad T_{N\perp x}^{(1)} \quad T_{N\parallel y}^{(1)} \quad T_{NB}^{(1)} \quad T_{NE}^{(1)} \quad T_{\alpha 1}^{(1)} \quad T_{\alpha 2}^{(1)} \quad T_{Cy}^{(1)}\right]^T =$$

$$\left[0.021 \quad 0.021 \quad 0.01 \quad 0.01 \quad 0.050 \quad 0.075 \quad 0.00034 \quad 0.00034 \quad 0.04\right]^T$$

According to the specified process planning, the pin-hole is bored in terms of dimensions and tolerances $x_{N'} = -25 \pm 0.021$ and $y_{N'} = 28 \pm 0.021$. After the first operation is executed, the pin is plugged into the pin-hole. Provided that its actual geometric deviation be within their tolerance range, that is $T_{N\perp x} = 0.010$, and $T_{N//y} = 0.010$. When coordinates of the pin and actual distance y_C are measured, assume that the acquired values are $x_N^* = -25.020$, $y_N^* = 28.020$, and $y_C^* = -25.140$. Thus the actual deviation is $\Delta_{N'x} = x_N^* - x_{N'} = -25.020 - 25 = -0.020$, $\Delta_{N'y} = y_N^* - y_{N'} = 28.020 - 28 = -0.020$, and $\Delta_{Cy} = y_C^* - y_C = -25.140 - 25.075 = -0.650$. It is obvious that $\Delta_{N'x}$, $\Delta_{N'y}$, and Δ_{Cy} are within their tolerance ranges so the operations of next stage can be carried out. Because x_N, y_N, and T_{Cy} are measured, T_{Nx}, T_{Ny}, and T_{Cy} will not be included in the tolerance chains for remaining operations. Since $T_{N\perp x}$ and $T_{N//y}$ are included in values x_N^* and y_N^*, they will not be included in the tolerance chains for remaining operations either. The optimal tolerances for next operation will be determined by following optimization model.

$$\max\left(k_5 T_{NB}^{(2)} + k_6 T_{NE}^{(2)} + k_7 T_{\alpha 1}^{(2)} + k_8 T_{\alpha 2}^{(2)}\right)$$

s.t.

$$\begin{bmatrix} \dfrac{1}{\cos 30^\circ} & \dfrac{L_{BC}^{(1)}}{\cos 30^\circ} & 0 & 0 \\ tg30^\circ & L_{BC}^{(1)} tg30^\circ & 1 & L_{EF}^{(1)} \end{bmatrix} \begin{bmatrix} T_{NB}^{(2)} \\ T_{\alpha 1}^{(2)} \\ T_{NE}^{(2)} \\ T\alpha_{\alpha 2}^{(2)} \end{bmatrix} \le \begin{bmatrix} 0.140 \\ 0.240 \end{bmatrix}$$

$$T_{\alpha 1}^{(2)} + T_{\alpha 2}^{(2)} \le 0.0007$$

$0.030 \le T_{NB}^{(2)} \le 0.160$, $0.00034 \le T_{\alpha 1}^{(2)} \le 0.00044$, $0.030 \le T_{NE}^{(2)} \le 0.160$, $0.00034 \le T_{\alpha 2}^{(2)} \le 0.00044$,

Where

$$\begin{bmatrix} \cos 30^\circ & -\sin 30^\circ & 0 \\ -\sin 30^\circ & -\cos 30^\circ & 0 \\ -tg30^\circ & 0 & 1 \end{bmatrix} \begin{bmatrix} L_{NB}^{(2)} \\ L_{BC}^{(2)} \\ L_{NE}^{(2)} \end{bmatrix} = \begin{bmatrix} x_{Cd} \\ y_{Cd} \\ L_{CFd} \end{bmatrix} - \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \dfrac{1}{\cos 30^\circ} & \dfrac{-1}{\cos 30^\circ} \end{bmatrix} \begin{bmatrix} x_N^* \\ y_N^* \\ y_C^* \end{bmatrix},$$

$k_5 = k_6 = k_7 = k_8 = 1.4$, $x_N^* = -25.020$, $y_N^* = 28.020$, $y_C^* = -25.140$, $L_{BC}^{(1)} = 29.400$, and $L_{EF}^{(1)} = L_{NB}^{(1)} = 55.078$.

The optimal solution of this problem is:

$$\left[L_{NB}^{(2)} \quad L_{BC}^{(2)} \quad L_{NE}^{(2)}\right]^T = \left[55.106 \quad 29.407 \quad 24.432\right]^T$$

$$\left[T_{NB}^{(2)} \quad T_{BC}^{(2)} \quad T_{\alpha 1}^{(2)} \quad T_{\alpha 2}^{(2)}\right]^T = \left[0.111 \quad 0.151 \quad 0.00034 \quad 0.00034\right]^T$$

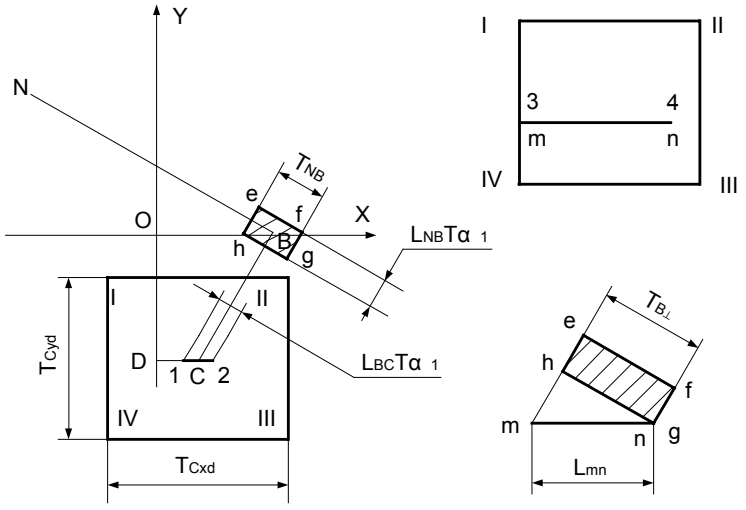


Figure 8. Tolerance accumulation for machining the inclined hole

The incline hole will be processed according to the dimensions and tolerances 55.106 ± 0.056 and 60° ± 0.00017. The tolerance accumulation for machining the inclined hole is shown in Figure 8. After incline hole has been processed, L_{NB} is measured. Let acquired value be $L^*_{NB} = 55.150$. The corresponding actual deviation $\Delta_{NB} = L^*_{NB} - L_{NB} = 0.044$ is within its tolerance range and next operations can be performed. The next step is establishment of the following optimal model:

$$\max\left(k_6 T_{NE}^{(3)} + k_7 T_{\alpha 1}^{(3)} + k_8 T_{\alpha 2}^{(3)}\right)$$

s.t.

$$\frac{L_{BC}^{(2)}}{\cos 30^\circ} T_{\alpha 1}^{(3)} \le 0.14$$

$$L_{BC}^{(2)} tg30^\circ T_{\alpha 1}^{(3)} + T_{NE}^{(3)} + L_{EF}^{(2)} T_{\alpha 2}^{(3)} \le 0.24$$

$$T_{\alpha 1}^{(3)} + T_{\alpha 2}^{(3)} \le 0.0007$$

$0.030 \le T_{NE}^{(3)} \le 0.220$, $0.00034 \le T_{\alpha 1}^{(3)} \le 0.00044$, $0.00034 \le T_{\alpha 2}^{(3)} \le 0.00044$. Where

$$L_{NE}^{(3)} = L_{CFd} + L_{NB}^* + \frac{y_C^* - y_N^*}{\cos 30^\circ},\ k_6 = k_7 = k_8 = 1.4,\ L_{BC}^{(2)} = 29.407,\ L_{EF}^{(2)} = L_{NB}^{(2)} = 55.106,$$

$$L_{NB}^* = 55.150,\ y_C^* = -25.140,\ y_N^* = 28.020.$$

The solution of the above optimal model is:

$$L_{NE}^{(3)} = 24.457,\ \left[T_{NE}^{(3)} \quad T_{\alpha 1}^{(3)} \quad T_{\alpha 2}^{(3)}\right]^T = \left[0.215 \quad 0.00034 \quad 0.00034\right]^T$$

This solution is for final operation i.e. machining the inclined plane. The tolerance accumulation for machining the inclined plane is shown in Figure 9.

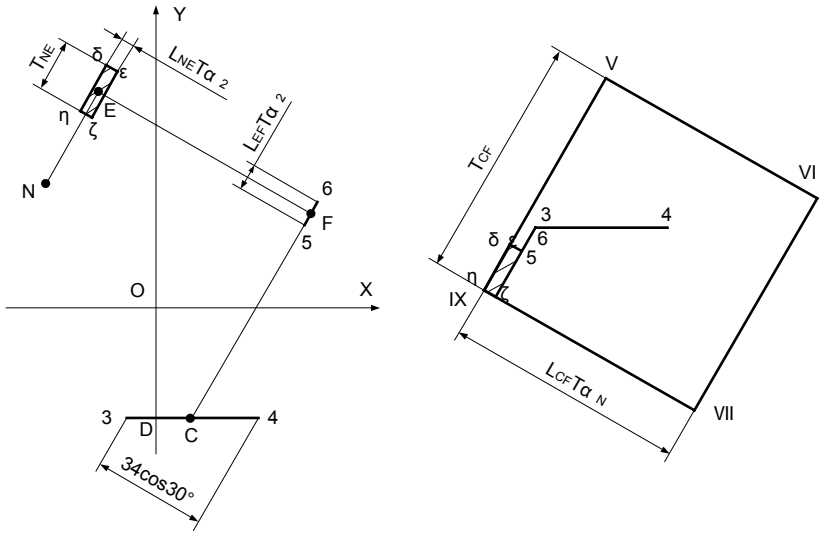


Figure 9. Tolerance accumulation for machining the inclined plane

It is not difficult to find that the working tolerances have been gradually amplified. Table 3 shows the comparison results between conventional tolerance control (CTC) and proposed sequential tolerance control (STC). Table 4 shows the variations in their pertinent process working dimensions.

	$T_{N'x}^{(1)}$	$T_{N'y}^{(1)}$	$T_{N\perp x}^{(1)}$	$T_{N\parallel y}^{(1)}$	$T_{NB}^{(2)}$	$T_{NE}^{(3)}$	$T_{\alpha 1}^{(3)}$ (rad)	$T_{\alpha 2}^{(3)}$ (rad)	$T_{Cy}^{(1)}$
STD	21	21	10	10	111	215	0.00034	0.00034	40
CTD	20*	20*	10*	10*	46	72	0.00035	0.00035	40
Ratio	1.05	1.05	1.00	1.00	2.41	2.99	0.97	0.97	1.00
In-crease	+5%	+5%	0	0	+141%	+199%	-3%	-3%	0

Table 3. Tolerance of the CTC and the proposed STC (μm). Note: The values with the sign"*"were directly given by experience in terms of the process capacities (Zhao, 1987).

Method	N'_x	N'_y	L_{NB}	L_{NE}	α_1	α_2	y_C
STC	-25	+28	55.106	24.457	60°	60°	-25.075
CTC	-25	+28	55.078	24.600	60°	60°	-25.075

Table 4. Working dimension of the CTC and the proposed STC (mm)

5.5. Comparative analysis of the proposed method

In order to analyze the effects of proposed method, comparative study is also given. The impact the weight factors have on optimal working tolerance is identified with the same model. Let $k_1 = k_2 = k_3 = k_4 = k_9 = 1$, $k_5 = k_6 = k_7 = k_8 = \omega$. The solutions of original model when $\omega = 1, 1.4, 1.5, 2, 4$ are shown in table 5. It can be seen from table 5 that when weight ω increases, $T_{N'x}^{(1)}$, $T_{N'y}^{(1)}$, $T_{N\perp x}^{(1)}$, $T_{N\parallel y}^{(1)}$, and $T_{Cy}^{(1)}$ decrease, $T_{NB}^{(1)}$ and $T_{NE}^{(1)}$ increase, while $T_{\alpha 1}^{(1)}$ and $T_{\alpha 2}^{(1)}$ remain unchanged.

Case	$T_{N'x}^{(1)}$	$T_{N'y}^{(1)}$	$T_{N\perp x}^{(1)}$	$T_{N\parallel y}^{(1)}$	$T_{NB}^{(1)}$	$T_{NE}^{(1)}$	$T_{\alpha 1}^{(1)}$ (rad)	$T_{\alpha 2}^{(1)}$ (rad)	$T_{Cy}^{(1)}$
$\omega = 1$	21	21	10	10	50	75	0.00034	0.00034	40
$\omega = 1.4$	21	21	10	10	50	75	0.00034	0.00034	40
$\omega = 1.5$	10	10	7	7	68	75	0.00034	0.00034	40
$\omega = 2$	10	10	7	7	70	75	0.00034	0.00034	35
$\omega = 4$	10	10	7	7	75	75	0.00034	0.00034	26

Table 5. Tolerance of the original model with different weights (μm)

6. Closing remarks and Conclusions

This paper presents a new graphic representation methodology for generating dimensional and tolerance chains in complex 2D drawing from 3D parts used in sequential optimal tolerance design when sized, angular, and geometric specifications are included simultaneously. This was overlooked and did not give due attention in previous literatures due to its complexity. Since geometric tolerances are also of vital importance to the functional requirements and manufacturing cost, they are necessary to be included in tolerance chains. The proposed approach copes with 3D sequential dimensioning and tolerancing by dynamic design of the working dimensions and tolerances at any completion stage of operations. The practical example shows that the proposed method can gradually amplify the working tolerances for remaining operations and raise the acceptance rate of the processed parts.

Acknowledgement
This research project is sponsored by the National Natural Science Foundation of China (grant No. 50465001) to Huang Meifa. The authors would like to thank the anonymous reviewer for their constructive comments on the earlier version of this paper.

7. References

Cavalier, T. M. and Lehtihet, E. A. (2000). A comparative evaluation of sequential set point adjustment procedures for tolerance control. *International Journal of Production Research*, Vol. 38, No. 8, May 2000, ISSN 0020-7543, 1769-1777.

Chang, C. L., Wei, C. C. and Chen, C. B. (2000). Concurrent maximization of process tolerances using grey theory. *Robotics and Computer-Integrated Manufacturing*, Vol.16, No. 2-3, April-June 2000, ISSN 0736-5845, 103-107.

Chen, K. Z., Feng, X. A., and Lu, Q. S. (2001). Intelligent dimensioning for mechanical parts based on feature extraction. *Computer-Aided Design*, Vol. 33, No. 13, Nov. 2001, ISSN 0010-4485, 949-965.

Chen, Y. B., Huang, M. F., Yao, J. C., Zhong, Y. F. (2003). Optimal concurrent tolerance based on the grey optimal approach. *The International Journal of Advanced Manufacturing Technology*, Vol. 22, No. 1-2, 2003, ISSN 0268-3768, 112-117.

Fraticelli, B. P., Lehtihet, E. A., and Cavalier, T. M. (1997). Sequential tolerance control in discrete parts manufacturing. *International Journal of Production Research*, Vol. 35, No. 5, May 1997, ISSN 0020-7543, 1305-1319.

Fraticelli, B. M. P., Lehtihet, E. A., and Cavalier, T. M. (1999). Tool-wear effect compensation under sequential tolerance control. *International Journal of Production Research*, Vol. 37, No. 3, Feb. 1999, ISSN 0020-7543, 639-651.

Gao, Y. and Huang, M. (2003). Optimal process tolerance balancing based on process capabilities. *The International Journal of Advanced Manufacturing Technology*, Vol. 21, No. 7, ISSN 0268-3768, 501-507.

He, J. R. and Gibson, P. R. (1992). Computer-Aided Geometrical Dimensioning and tolerancing for Process-Operation Planning and Quality Control. *The International Journal of Advanced Manufacturing Technology*, Vol. 7, No. 1, 1992, ISSN 0268-3768, 11-20.

Huang, M., Gao, Y., Xu, Z. and Li, Z. (2002). Composite planar tolerance allocation with dimensional and geometric specifications. *The International Journal of Advanced Manufacturing Technology*, Vol. 20, No. 5, ISSN 0268-3768, 341-347.

Huang, M. F., Zhong, Y. R. and Xu, Z. G. (2005). Concurrent process tolerance design based on minimum product manufacturing cost and quality loss. *The International Journal of Advanced Manufacturing Technology*, Vol. 25, No. 7-8, ISSN 0268-3768, 714-722.

Huang, M. F. and Zhong, Y. R. (in press). Optimized sequential design of two dimensional tolerances. The International Journal of Advanced Manufacturing Technology, (in press), ISSN 0268-3768.

Ji, P. (1993a). A linear programming model for tolerance assignment in a tolerance chart. International Journal of Production Research, Vol. 31, No. 3, March 1993, ISSN 0020-7543, 739-751.

Ji, P. (1993b). A tree approach for tolerance charting. International Journal of Production Research, Vol. 31, No. 5, May 1993, ISSN 0020-7543, 1023-1033.

Lee, Y. C. and Wei, C. C. (1998). Process capability-based tolerance design to minimize manufacturing loss. The International Journal of Advanced Manufacturing Technology, Vol. 14, No. 1, ISSN 0268-3768, 33-37.

Lee, Y. H., Wei, C. C. and Chang, C. L. (1999). Fuzzy design of process tolerance to minimize process capability. The International Journal of Advanced Manufacturing Technology, Vol. 15, No. 9, ISSN 0268-3768, 655-659.

Mcgarvey, R. G., Lehtihet, E. A., Castillo, E. D., and Cavalier, T. M. (2001). On

the frequency and locations of set point adjustments in sequential tolerance control. International Journal of Production Research, Vol. 39, No. 12, ISSN 0020-7543, 2659-2674.

Ngoi, K. B. A. (1992). Applying linear programming to tolerance chart balancing. The International Journal of Advanced Manufacturing Technology, Vol. 7, No. 4, 1992, ISSN 0268-3768, 187-192.

Ngoi, B. K. A. and Cheong, K. C. (1998a). An alternative approach to assembly tolerance analysis. International Journal of Production Research, Vol. 36, No. 11, Nov. 1998, ISSN 0020-7543, 3067-3083.

Ngoi, B. K. A. and Cheong, K. C. (1998b). The apparent path tracing approach to tolerance charting. The International Journal of Advanced Manufacturing Technology, Vol. 14, No. 8, 1998, ISSN 0268-3768, 580-587.

Ngoi, B. K. A., Lim, L. E. N., Ong, A. S., and Lim, B. H. (1999). Applying the coordinate tolerance system to tolerance stack involving position tolerance. The International Journal of Advanced Manufacturing Technology, Vol. 15, No. 6, 1999, ISSN 0268-3768, 404-408.

Ngoi, B. K. A., Lim, B. H., and Ang, P. S. (2000). Nexus method for stack analysis of geometric dimensions and tolerancing (GDT) problems. International Journal of Production Research, Vol. 38, No. 1, Jan. 2000, ISSN 0020-7543, 21-37.

Ngoi, B. K. A. and Ong, C. T. (1993). A complete tolerance charting system. International Journal of Production Research, Vol. 31, No. 11, Feb. 1993, ISSN 0020-7543, 453- 469.

Ngoi, B. K. A. and Ong, J. M. (1999). A complete tolerance charting system in assembly. *International Journal of Production Research*, Vol. 37, No. 11, July 1999, ISSN 0020-7543, 2477-2498.

Ngoi, K. B. A. and Tan, C. K. (1995). Geometrics in computer-aided tolerancing charting. *International Journal of Production Research*, Vol. 33, No. 3, March 1995, ISSN 0020-7543, 835-868.

Ngoi, K. B. A. and Seow, M. S. (1996). Tolerance control for dimensional and geometrical specifications. *The International Journal of Advanced Manufacturing Technology*, Vol. 11, No. 1, 1996, ISSN 0268-3768, 34-42.

Ramani, B., Cheraghi, S. H., and Twomey, J. M. (1998). CAD-based integrated tolerancing system. *International Journal of Production Research*, Vol. 36, No. 10,Oct. 1998, ISSN 0020-7543, 2891-2910.

Swift, K. G., Raines, M. and Booker, J. D. (1999). Tolerance optimization in assembly stacks based on capable design, *Proceedings of the Institution of Mechanical Engineers*, part B (Journal of Engineering Manufacture), Vol. 213,

No. B7, 1999, ISSN 0954-4054, 677- 693.

Treacy, P., Ochs, J. B., Ozsoy, T. M., and Wang, N. X. (1991). Automated tolerance analysis for mechanical assemblies modeled with geometric features and relational data structure. *Computer Aided Design*, Vol. 23, No. 6, July-Aug. 1991, ISSN 0010-4485, 444-453.

Tseng, Y. J. and Kung, H. W. (1999). Evaluation of alternative tolerance allocation for multiple machining sequences with geometric tolerances. *International Journal of Production Research*, Vol. 37, No. 17, Nov. 1999, ISSN 0020-7543, 3883-3900.

Wang, N. X., and Ozsoy, T. M. (1993). Automatic generation of tolerance chains from mating relations represented in assembly models. Journal of Mechanical Design, Transactions of the ASME, Vol. 115, No. 4, ISSN 0738-0666, 757-761.

Wei, C. C. and Lee, Y. C. (1995). Determining the process tolerances based on the manufacturing process capability. *The International Journal of Advanced Manufacturing Technology*, Vol.10, No. 6, 1995, ISSN 0268-3768, 416-421.

Wheeler, D. L., Cavalier, T. M., and Lehtihet, E. A. (1999). An implicit enumeration approach to tolerance allocation in sequential tolerance control. IIE Transactions, Vol. 31, No. 1, Jan. 1999, ISSN 0740-817X, 75-84.

Zhao, C. G. (1987). Graphical theory of dimensional chains and its applications. National defense industry press, China.

22

A New Rapid Tooling Process

Xiaoping Jiang and Chao Zhang

1. Introduction

Due to the globalization of the world economy and the consequent increase in competition, it is more important than ever for manufacturers to shorten their product development and manufacturing cycles. The current competitive market not only requires faster product development and reduced production time, but also demands higher quality, greater efficiencies, lower cost, and the ability to meet environmental and recycling objectives.

Tooling is a very important phase in the development and manufacturing of new products and is usually one of the most time-consuming and costly phases. Therefore, shortening the tooling lead-time plays a key role in the reduction of the overall product development and manufacturing time. Great effort has been made to develop new rapid tooling (RT) technologies that combine the recently emerged rapid prototyping (RP) processes with one or more subsequent processes.

RP normally refers to fabricating prototypes directly from computer aided design (CAD) data using a layered, additive method. Almost all products developed in the manufacturing industry arise from the creation of a three-dimensional computer model using a CAD system. Converting the CAD model into a prototype by using a RP process can be easily realized (Jacobs, 1992). A RP system can quickly generate physical objects and prototypes using liquid, powder, or sheet materials. RP parts allow designers to verify their product design at an early stage and to use three-dimensional representations of the design for sales, marketing and production.

Along with the evolution and improvements of various RP technologies, great research and development efforts have been made in recent years to develop RT technologies based on RP processes (Pham, 1998, Willis, 1997, Hejmadi and McAlea, 1996, Nelson, 1999, and Phelan, 1997). Whether the application is prototype, bridge, short-run, or production tooling, RT represents an opportunity to reduce both production time and cost. Therefore, researchers continue to

explore ways to improve RT technologies.
In this study, a new RT process using a metal shell backfilled with metal powder to provide mechanical support to the metal shell is presented and the feasibility of this new RT process is evaluated. In particular, the study is focused on (1) the packing behavior of the metal powder used to backfill the metal shell, (2) the deformation behavior of the compacted metal powder under compression, (3) the deformation behavior of the metal shell, and (4) the deformation behavior of the metal shell and compacted metal powder assembly.

2. The New Rapid Tooling Process

The proposed new RT process is shown schematically in Fig. 1. For convenience, the figure only illustrates the fabrication of half mould. The proposed new RT process involves the following major steps:

Step 1: A three-dimensional computer model of the mould is designed on a computer.

Step 2: A plastic pattern with complementary shape to the mould is fabricated using a RP process, such as Stereolithography.

Step 3: A thin metal layer is deposited onto the cavity side of the plastic pattern using an electro-chemical process to form a metal shell. Then, the metal shell is separated from the plastic pattern.

Step 4: Metal ribs are added to the back of the metal shell to increase the strength of the metal shell.

Step 5: The metal powder is packed into the metal shell to provide mechanical support to the metal shell.

Step 6: The backside of the mould is sealed to prevent leakage of the metal powder. The tool is finished and ready for moulding operations.

In addition to anticipated short lead-time and low cost, it is also expected that this RT process is environmentally friendly because the metal powder used in backing is fully reusable/recyclable.

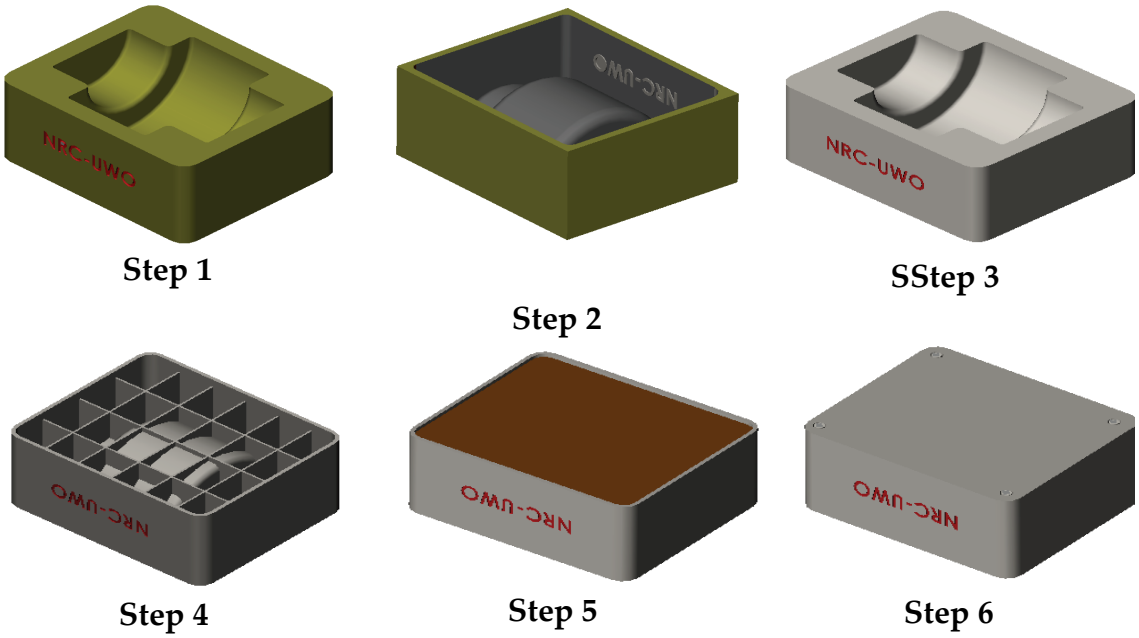


Figure 1. New rapid tooling process

3. Powder Packing Behavior

3.1 Powder Packing Models

Powder packing can be divided into single component packing, binary packing as well as multiple component packing. In single component packing, all packing particles used are of the same size and geometry. In contrast, binary or multiple components packing involves two or more packing components. The powder packing behavior is measured by the packing density, which is defined as the ratio of the volume occupied by the powder to the total volume without compression, i.e.

$$Packing\ Density = \frac{V_{POWDER}}{V_{TOTAL}} = \frac{\sum W_i / \rho_i}{V_{TOTAL}} \qquad (1)$$

where V_{TOTAL} is the total volume of the powder for single component packing or the total volume of the powder mixture for multiple component packing,

W_i is the weight of the powder component i, and ρ_i is the theoretical density of the powder material for component i. In single component packing, packing density of the powder depends on only the packing structure or the arrangement of the particles. In binary and multiple component packing, the packing density is affected by the size and proportion of each packing component. The voids formed by large particles can be filled by small particles. The voids created by small particles can be filled by even smaller particles.

3.2 Procedures for Powder Packing Experiments

For single component powder packing, the powder is first weighed and placed into a scaled cylinder or beaker. The powder in the container is vibrated for 15 minutes on a vibration machine with the topside of the powder being lightly pressed by a flat plate. After vibration, the volume of the powder is measured. The packing density is determined based on the measured data for the weight and volume of the powder.

For binary or multiple component powder packing, the coarse metal powder is weighed and placed into a scaled cylinder. The cylinder is vibrated on a vibrating machine for 15 minutes while the topside of the powder is being lightly pressed by a flat plate. After the volume of the coarse powder is measured, a weighed amount of fine powder is added to the container while the cylinder is vibrated. The volume of the mixed powders is measured and the packing density of the mixed powders is calculated. For a three-component mixture, the weighted finer powder is added to the previous mixture while the cylinder is vibrated. The final volume is measured and the packing density of the mixed powders is calculated.

For binary component powder packing involving low fluidity powders, the weighed coarse and fine powders are placed into a scaled cylinder or a beaker. The mixture is stirred so that the powders are mixed evenly. It is then vibrated. The volume of the mixed powders is measured after the vibration. The packing density is calculated.

In this study, eleven different metal powders are selected for the powder packing experiment. The particles of the selected powders have different shapes and sizes, and they are made of different materials. The characteristics of the selected powders are listed in Table 1.

Powder number	Powder name	Material	Material density (g/ml)	Geometry	Average particle size (μm)
1	Carbon Steel Ball	Carbon Steel	7.85	Spherical	3175
2	12 HP Copper Shot	Copper	8.91	Round	850
3	34 HP Bronze	Bronze	8.65	Round	450
4	Fe	Iron	7.85	Spherical	22~53
5	T-15	Tool Steel	8.19	Spherical	>150
6	T-15	Tool Steel	8.19	Spherical	80~150
7	T-15	Tool Steel	8.19	Spherical	<22
8	ATOMET 1001	Low Carbon Steel	7.85	Irregular	>150
9	ATOMET 1001	Low Carbon Steel	7.85	Irregular	<22
10	DISTALOY 4600A	Low Carbon Steel	7.9	Irregular	>150
11	DISTALOY 4600A	Low Carbon Steel	7.9	Irregular	<22

Table 1. Characteristics of selected powders

3.3 Results of the Single Component Powder Packing Experiments

The packing density depends on the characteristics of the particles. Generally, for powder packing, the density of the powder material has no significant influence on its packing density. Particles of the same size and shape will have the same packing density despite of the difference in their theoretical densities (Leva and Grummer, 1947). The main factors affecting the packing density for single component powder packing are particle size, particle shape, and the ratio of the diameters of the container to the particle.

(a) The effect of the ratio of the diameters of the container to the particle

McGeary (1962) studied the effect of the ratio of the diameters of the container to the particle D/d (D is the container diameter and d is the particle diameter) and concluded that if the ratio D/d is greater than 50, the packing density tends to reach the maximum value. Experiments are carried out here for ratios D/d

from 3.5 to 39.4 using 3175 μm (1/8-inch) diameter carbon steel balls (Powder #1) and for the ratio *D*/*d* of 57.6 using 12 HP copper shorts of diameter of 850 μm (Powder #2). In the carbon steel ball packing tests, different diameter containers are used to create different *D*/*d* ratios. The experimental results presented in Table 2 show the effect of the ratio *D*/*d* on the packing density. The lowest packing density, 0.55, occurs at the lowest ratio *D*/*d*, which is 3.5. The highest packing density is 0.65 when the ratio *D*/*d* is 57.6. It can be observed from Table 2 that the packing density increases with the increase of the ratio *D*/*d*. However, the packing density does not change much when the ratio *D*/*d* is greater than 7.66.

Powder number	2	1	1	1	1	1	1
D/d	57.6	39.4	15.1	10.9	7.66	4.93	3.50
Packing density	0.65	0.63	0.61	0.62	0.62	0.58	0.55

Table 2. Single component packing density for different D/d

(b) The effect of the particle shape

The particle shape varies significantly depending on the manufacturing process used and influences the particle packing, flow, and compression properties. The greater the particle surface roughness or the more irregular the particle shapes, the lower the packing density (Shinohara, 1984). For a gas atomized metal powder, the shape is almost spherical and for water atomized metal powder, the shape is more irregular (German, 1998). Some particle shapes of the selected powders used in this study are shown in Fig. 2.

Table 3 gives the comparison of the packing densities for powders with different particle shapes. The powders with irregular particle shapes, DISTALOY 4600A (Powders #10 and #11) and ATOMET 1001 (Powders #8 and #9) powders, have a lower packing density, which is 0.49, as compared with the packing density of the powders of the spherical shape with the same size (Powders #5 and #7), which is 0.63. Therefore, the packing density of the powders with irregular shapes is 22% lower than that of the powders with the spherical shape.

(c) The effect of the particle size

The results shown in Table 3 also indicate the effect of the particle size on the packing density of the powder. It can be seen that the packing densities for the

powders with spherical shape and round shape are between 0.60 and 0.63, and it is 0.49 for the powders with irregular shapes, despite of the difference in the particle size. Thus, particle size has no significant effect on the packing density. However, a test for the particle fluidity by pouring the powders onto a plate with smooth surface that is at a 45° angle to the horizontal plane reveals that the particles demonstrate a low fluidity if the particle size is less than 22 μm.

Powder number	1	2	3	4	5	6
Shape	Spherical	Round	Round	Spherical	Spherical	Spherical
Size (μm)	3175	850	450	22~53	>150	80~150
Packing density	0.63	0.65	0.63	0.63	0.63	0.60
Powder number	7	8	9	10	11	
Shape	Spherical	Irregular	Irregular	Irregular	Irregular	
Size	<22	>150	<22	>150	<22	
Packing density	0.63	0.49	0.49	0.49	0.49	

Table 3. Single component packing density for different particle shapes and sizes

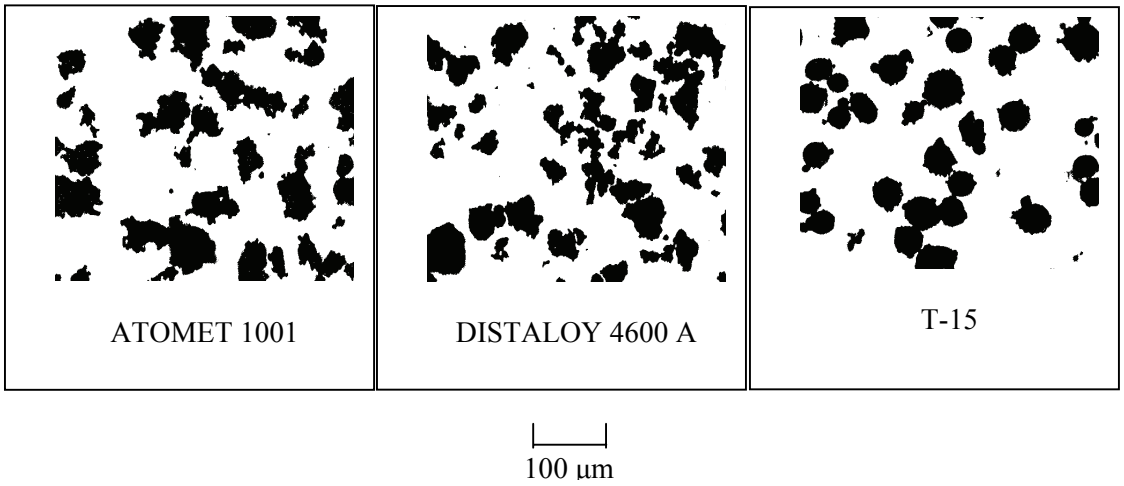


Figure 2. Optical micrographs of powders with different shapes

3.4 Results of the Binary and Tertiary Powder Packing Experiments

The results of the single component powder packing experiments indicate that the maximum packing density is about 0.65. For the new RT process considered in the current study, a higher packing density is required to achieve sufficient load transfer ability. Adding certain amount of smaller particles into a

packing structure consisted of large particles can greatly improve the packing density. Small particles are used to fit into the interstices between large particles, and smaller particles can be used to fit into the next level of pores. Thus, the packing density can be improved. This is the basic principle for the binary or multiple component packing. The factors that affect the binary or tertiary packing density, such as the size ratio and the mixing ratio of the packing components, are considered in this study. The mixing ratio is defined as the ratio of the weight of the large particle to the total weight of the powder mixture and the particle size ratio is defined as the ratio of the size of the large particle to the size of the small particle.

(a) The effect of the particle size ratio

To exam the effect of the particle size ratio of the packing components on the packing behavior of binary and tertiary mixtures, the experiments are conducted for different particle size ratios at the mixing ratio of 0.74 for binary mixtures, and 0.63 for the large size particles in the tertiary mixture and 0.23 for the middle size particles in the tertiary mixture. Table 4 gives the packing densities of binary and tertiary mixtures at different particle size ratios. The results show that adding small particles into a packing structure of large particles can greatly increase the packing density. The packing density of the binary or tertiary mixture increases between 9% and 44% as compared with the single component packing density. The increase in the packing density for the binary mixture with a low particle size ratio (Cases 4-6) is in the range of 9% ~ 14% and it is 32% ~ 33% for the binary mixture with a high particle size ratio (Cases 2 and 3).

Case	Powder mixture	Particle size ratio	Packing density			Packing density increase (%)
			Large particle	Small particle	Mixture	
1	#1+#3+#7	144: 20.5: 1	0.63	0.63	0.91	44
2	#1+#4	(59.9~144): 1	0.63	0.63	0.84	33
3	#2+#4	(16.0~38.6): 1	0.65	0.63	0.86	32
4	#5+#7	6.82: 1	0.63	0.63	0.71	13
5	#2+#6	(5.67~10.6): 1	0.65	0.60	0.71	9
6	#1+#2	3.74:1	0.63	0.63	0.72	14

Table 4. Binary and tertiary packing density for different particle size ratios

The increase in the packing density for the tertiary mixture is 44%. The basic requirement of good multiple component packing is that small particles can freely pass through the voids between large particles. For spherical component packing, the minimum size ratio that satisfies this requirement can be determine using the packing models shown in Fig. 3.

There are two extreme packing conditions in the ordered single component packing. The simple cubic packing, as shown in Fig. 3 (a), produces the largest interstice between particles. The face-centered cubic packing shown in Fig. 3 (b), on the other hand, produces the smallest interstice between particles. The size of the fine particles should be smaller than the throat gate dimension of large particles so that the fine particles can freely pass through the throat gate between large particles. In Fig. 3, R is the radius of the large sphere, and r is the radius of the small sphere. For the face-centered packing model, the relation between R and r can be expressed as:

$$\frac{R}{R+r} = \cos 30^{\circ} \qquad (2)$$

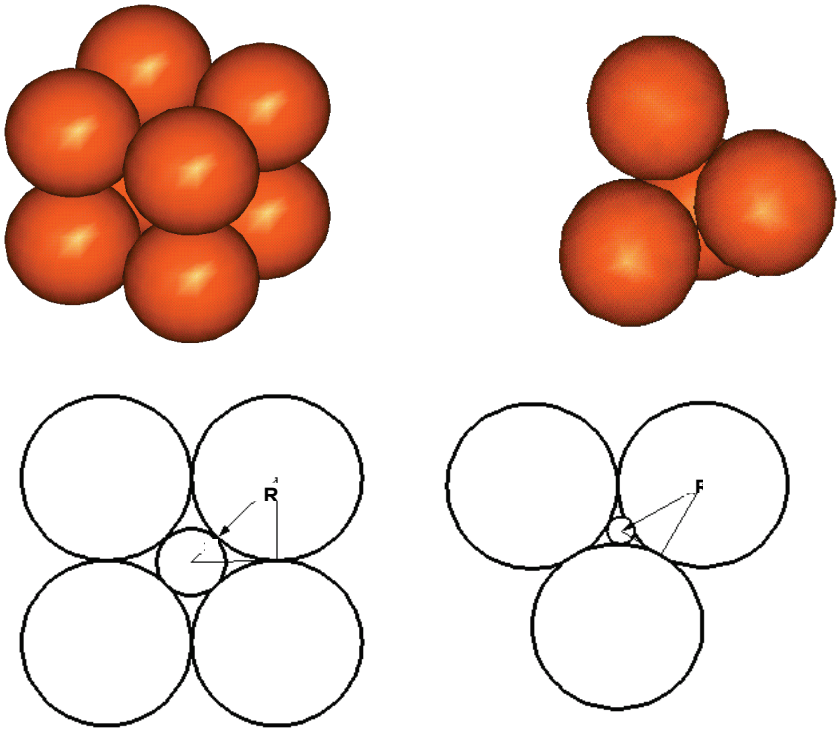


Figure 3. Throat gate structures between particles. (a) Simple cubic packing; (b) Face-centered cubic packing

From Eq. (2), we have $R/r = 6.46$. For the simple cubic packing, the relation becomes

$$\frac{R}{R+r} = \cos 45^{\circ} \qquad (3)$$

Therefore, $R/r = 2.41$

It can be concluded that the minimum particle size ratio, R/r, for small particles to fill the voids between large particles without pushing them apart is 2.41. When the ratio *R*/*r* is greater than 6.46, all of the small particles can pass the throat gates and enter the interstices between large particles. In order to obtain a higher packing density, the particle size ratio should be greater than 6.46.

The experimental results shown in Table 4 reflect the effect of particle size ratio. The particle size ratios in Cases 1 to 3 are much higher than 6.46. Thus, the packing densities in these cases are higher than those in Cases 4 to 6. In Case 6, the particle size ratio is lower than 6.46, but higher than 2.41. So, the small particles can only partially fill the voids between the large particles. The packing density increases compared with the single component packing density. However, it is lower than that with high particle size ratio. In Case 5, the size ratio varies from 5.67 to 10.6 and it does not totally satisfy the particle size ratio requirement for good binary packing, which leads to a lower packing density. The particle size ratio in Case 4 is 6.82 and it is greater than the minimum particle size ratio requirement for good binary packing, which is 6.46 based on ordered packing. However, the packing density is also low. This is due to the fact that the actual powder packing is not ordered packing. The result suggests that the minimum particle size ratio for actual powder packing to achieve a good binary packing should be higher than 6.82. As expected, the highest packing density is obtained from tertiary powder packing, Case 1, which is 0.91.

It is observed that the binary packing density for the mixture of Powder #2 and Powder #4 (Case 3) is slightly higher than that for the mixture of Powder #1 and Powder #4 (Case 2). This may attribute to the fact that the single component packing density for Powder #1 is lower than that for Powder #2 as shown in Table 3. It is also noticed that the binary packing density is between 0.71 and 0.72 when the particle size ratio is lower than the minimum particle size ratio requirement for good binary packing and it is 0.84 to 0.86 when the particle size ratio is higher than the minimum particle size ratio requirement. Therefore, the particle size ratio has little effect on the binary packing density

once the size ratio is lower or higher than the minimum particle size ratio requirement for good binary packing.

(b) The effect of the mixing ratio

The experiments are conducted for binary mixtures at different mixing ratios to investigate the effect of the mixing ratio on the packing density of binary powder mixtures. Table 5 shows the experimental results of packing densities for four different binary mixtures at different mixing ratios. The packing density varies from 0.67 to 0.86. It can be seen from the results that there is an optimal mixing ratio for each binary mixture at which the packing density of the binary mixture is maximal.

When small particles are added to fill the voids between the large particles, the porosity of the binary powder mixture decreases. Therefore, the packing density of the binary mixture increases. When the small particles fill all of the voids without forcing the large particles apart, the packing density of the binary mixture is at its maximum value. Further addition of small particles will force the large particles apart and the packing density will decrease. The optimal mixing ratio falls in the range of 0.71 - 0.77.

Mixture	#2+#6	#1+#4	#2+#4	#5+#7
Particle size ratio	5.67~10.6	59.9~144	16.0~38.6	6.82
Mixing ratio	Binary packing density			
0.65	0.70	0.82	0.83	0.68
0.68	0.71	0.82	0.84	0.69
0.71	0.72	0.83	0.85	0.70
0.74	0.71	0.84	0.86	0.71
0.77	0.70	0.82	0.86	0.72
0.80	0.69	0.81	0.85	0.70
0.83	0.68	0.80	0.83	0.68
0.86	0.67	0.77	0.80	0.67

Tabele 5. Binary packing density at different mixing ratios

4. Deformation Behaviour of Compacted Metal Powder under Compression

The effects of various parameters on the deformation behavior of compacted metal powder under compressive loading are investigated experimentally in

order to examine the feasibility of the proposed new RT process. The experimental results are used to obtain the elastic properties of the compacted metal powder under various loading conditions. These are important parameters for the deformation analysis of the metal shell and powder assembly used in the new RT process.

4.1 Compression Experiments

The metal powders used for the compression experiments are given in Table 6. Three different kinds of powders are selected to provide different particle shapes and hardness. As shown in Table 6, T-15 tool steel powder has much higher hardness than that for ATOMET 1001 and DISTALOY 4600A. For T-15, both coarse and fine size particles are used to examine the compression behaviour of powder mixtures. For ATOMET 1001 and DISTALOY 4600A, only coarse size particles are used. The sizes of the powders are chosen so that the size ratio of coarse powder and the fine powder is greater than 7. The mixing ratio of the coarse and fine powders is varied between 0.70 and 0.80, which gives a higher packing density as shown in Table 5. The compression tests are carried out using an Instron Mechanical Testing System according to ASTM standard B331-95 (ASTM B331-95, 2002) in an axial compression die shown schematically in Fig. 4. The powder is dried in an oven at 105oC for 30 minutes before the compression test to remove any absorbed moisture. The powder is vibrated for 15 minutes in the single component powder compression test after being loaded into the compression die.

Powder	Particle size		Material properties				Shape
	Coarse (μm)	Fine (μm)	ϱ (g/ml)	E (GPa)	HRB	ν	
T-15	150-350	6- 22	8.19	190~210	220	0.27~0.3	Spherical
ATOMET 1001	45-150	-	7.85	190~210	48	0.27~0.3	Irregular
DISTALOY 4600A	45-150	-	7.85	190~210	79	0.27~0.3	Irregular

Table 6. Characteristics of selected powders ϱ – Density; E - Young's modulus; HRB – Hardness; ν -Poisson's ratio

The coarse and fine powders are carefully mixed in the die and are then vibrated for 15 minutes before the compression test of the mixed powders. The loading and unloading rate is 10 kN/min, and the maximum compressive stress used is 138 MPa, corresponding to the maximum injection moulding pressure used for forming most engineering plastics.

4.2 Results

(a) The effect of powder material properties on powder compressive properties

Table 7 shows the results of the single loading-unloading compression experiments for the three coarse powders listed in Table 6. The powder compact density is defined as the ratio of the volume occupied by the powder to the total volume after the compression. It can be seen that the powder material properties have a significant effect on the compressive characteristics of the powder. The total strain under the same loading condition for the T-15 tool steel powder is 0.157, which is the smallest among all powders considered.

Powder (coarse)	Total strain	Compact density	Packing density	Powder condition after compression
T-15 Tool Steel	0.157	0.627	0.63	Loose
DISTALOY 4600A	0.375	0.692	0.49	Block
ATOMET 1001	0.462	0.766	0.49	Block

Table 7. Effect of powder material properties on powder compressive deformation behavior

In contrast, the total strain for the ATOMET 1001 powder is largest, 0.462, three times that of the T-15 tool steel powder. For the purposes of comparison, the packing density obtained in Section 3 is also listed in Table 7. It can be seen that the change between the powder compact density and packing density is smallest for the T-15 tool steel powder that corresponds to the smallest total strain. Therefore, as expected, powders with higher hardness produce smaller compressive strain and density change under the same compressive load.

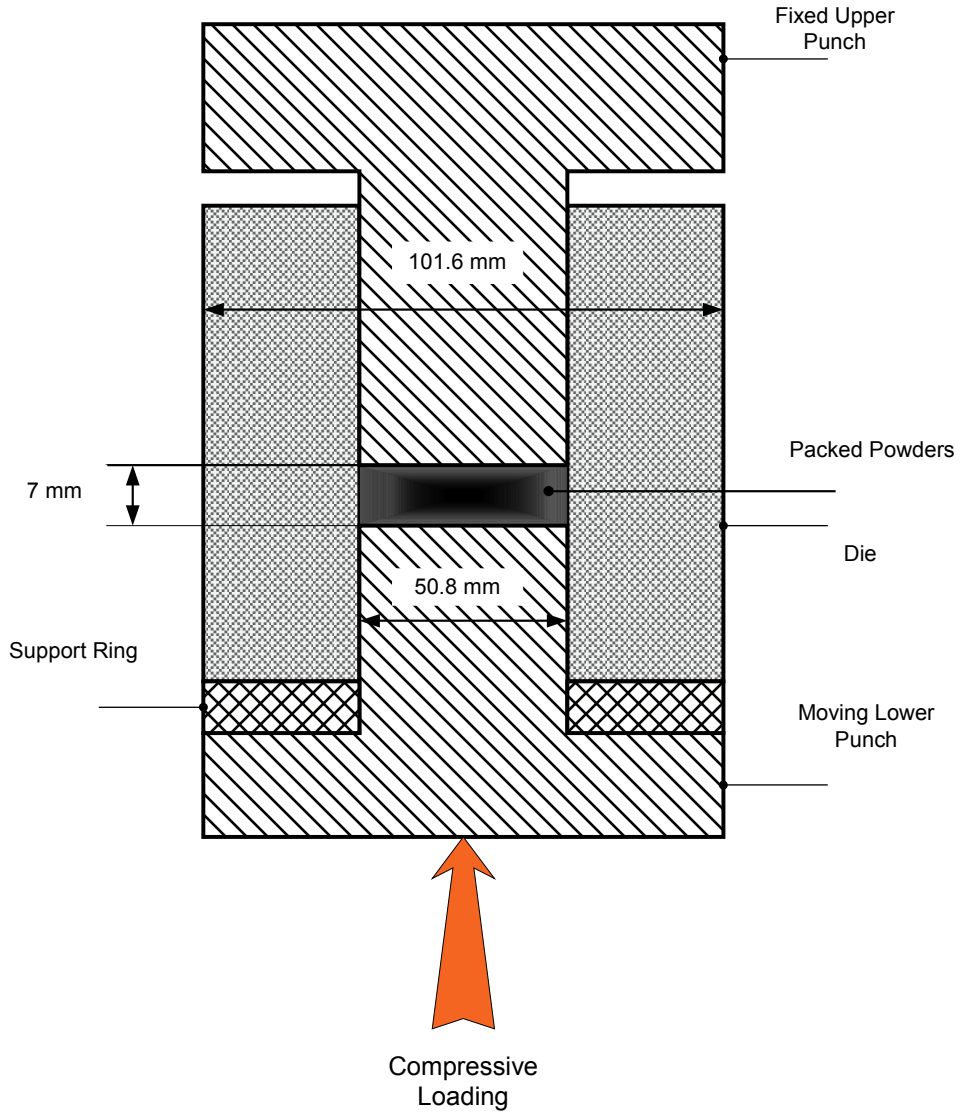


Figure 4. Die and punches for the compression test

It is also observed that the T-15 tool steel powder has the lowest compact density after compression although it has the highest packing density before compression. Therefore, for the same powder compact density, harder materials can support bigger loads. This suggests that powders with high hardness are preferred for the backing application in the proposed new RT process. In addition, the test indicates that soft powders such as DISTALOY 4600A and ATOMET 1001 tend to form blocks after compression. Such powder blocks cannot be reused for tooling applications because they lose the filling capabil-

ity that powders possess. In contrast, the T-15 tool steel powder remains in loose condition after compression at a compression stress of up to 138 MPa. Such powders are better choices for the application in the proposed new RT process from a reusable point of view. Therefore, the T-15 tool steel powder is used in the experiments conducted in subsequent sections.

(b) The effect of the mixing ratio on the compressive properties of binary powder mixtures

The RT process considered in the current study requires a higher packing density to achieve sufficient load transfer ability. The addition of smaller particles into a packing structure consisting of large particles can greatly improve the packing density. Experiments that involve the binary powder mixture of the coarse T-15 powder and fine T-15 powder at different mixing ratios using a single loading-unloading compression cycle are also carried out. The mixing ratio is defined as the ratio of the weight of the coarse powder to the total weight of the powder mixture. Table 8 shows the total compressive strain and corresponding powder compact density of the binary powder mixture at different mixing ratios. It can be seen from the results that at the mixing ratio of 0.77, the total strain is minimal and the compact density is maximum. This is also the optimal mixing ratio for T-15 tool steel powder mixture at which the powder packing density is maximal as shown in Table 5. It is clear that the optimal mixing ratio corresponding to a maximum powder packing density produces the least compressive deformation.

(c) Compressive behavior of powders in multiple loading-unloading cycles

To investigate the effect of loading history on the deformation behavior of the compacted metal powder, multiple loading-unloading experiments for the T-15 binary powder mixture with a mixing ratio of 0.77 are carried out using a five-cycle loading pattern shown in Fig. 5.

Powder Mixture	Mixing Ratio	Total Strain	Compact Density
T-15 Tool Steel	0.80	0.183	0.783
T-15 Tool Steel	0.77	0.162	0.822
T-15 Tool Steel	0.74	0.179	0.810

Table 8. Effect of the mixing ratio on binary powder compressive deformation behavior

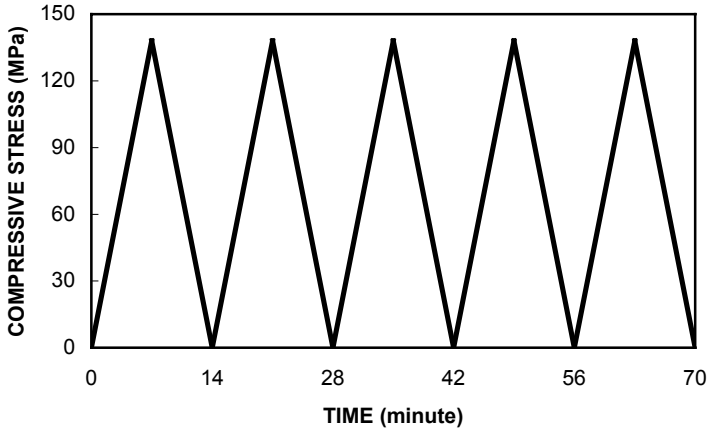


Figure 5. Loading pattern of the five-cycle compression test

(i) Loading history and critical point

The loading-unloading curves for the five-cycle compression test are shown in Fig. 6. The first loading curve is significantly different from the succeeding unloading and reloading curves. For the same load, it showed that the total deformation is twice as much as the other curves. Upon unloading during the first cycle, approximately 50% of the deformation is recovered, indicating a large amount of irreversible deformation during the first load cycle. After the unloading, the next reloading curve crosses the previous unloading curve at a certain stress level.

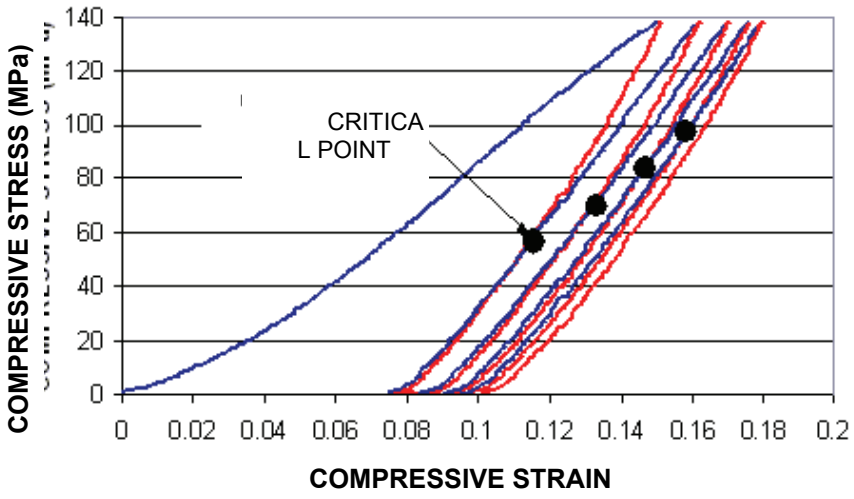


Figure 6. Typical loading-unloading curves of the five-cycle compression test

A pair of unloading and subsequent reloading curves form a cross point, as shown in Fig. 6. This cross point is referred to as the critical point. The unloading and reloading curves become parallel and closer to each other as the reloading and unloading cycles proceed. The tangent of the unloading or the reloading curve increases over cycles and approaches a constant value.

The critical point has two features. First, when the load is below the critical point, the reloading curve lies on the left side of the unloading curve of the previous cycle and the two curves essentially overlap with each other, indicating that the deformation below the critical point is mostly elastic in nature. On the other hand, when the reloading load goes beyond the critical point, the strain of reloading exceeds that in the previous unloading process, and the curves shows a hysteresis. Secondly, the stress corresponding to the critical point moves higher with an increased number of cycles, as shown in Fig. 6.
The deformation behavior and the critical point phenomenon can be understood from the deformation mechanisms of the powder compact. During the first loading cycle, the vibration packed powder particles only have point contacts with each other and will go through a large amount of irreversible deformation through such mechanisms as relative particle movement, plastic deformation at contacting points, and perhaps particle fracture for brittle particles (Carnavas, 1998). Elastic deformation will increase with the increase in the load and the decrease in the irreversible deformation. Upon unloading in the first load cycle, only the elastic component of the deformation is recovered, leaving a significant amount of irreversible deformation. During the succeeding loading cycles, the irreversible deformation mechanisms have largely been exhausted, and therefore a major portion of the deformation is elastic in nature. In particular, when the load is below the critical point, the deformation is essentially elastic and completely reversible. However, when load is high enough, i.e., beyond the critical points, some of the irreversible deformation mechanisms, such as local plastic deformation and particle relative movements, can further contribute to the unrecoverable deformation. It can be expected that with the proceeding of repeated loading-unloading cycles, the available sites and the amount of irreversible deformation will be gradually reduced, and therefore resulting in increased critical point and tangent of the loading-unloading curves.

These features indicate that the elastic properties of compacted powders can be controlled with properly designed loading-unloading cycles.

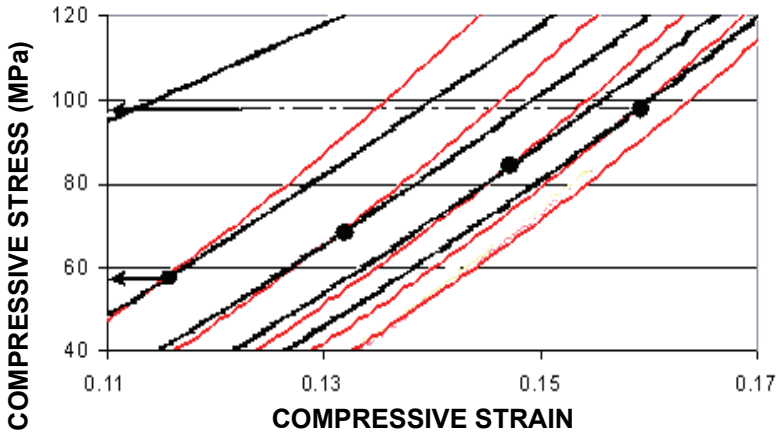
(ii) The effect of the mixing ratio on the critical point

Figure 7 shows the effect of the binary mixing ratio on the position of the critical point. The compressive stresses corresponding to the first and the fourth critical points are shown in Table 9. It can be seen that the binary powder mixture with a mixing ratio of 0.77 has higher critical points compared to the powder mixtures with mixing ratios of 0.74 and 0.80, respectively. The mixing ratio of 0.77 corresponds to the highest powder packing density in the binary packing system. A higher critical point means a higher deformation resistance in the subsequent reloading. High deformation resistance is beneficial for maintaining the integrity of tooling under working conditions for the intended RT application.

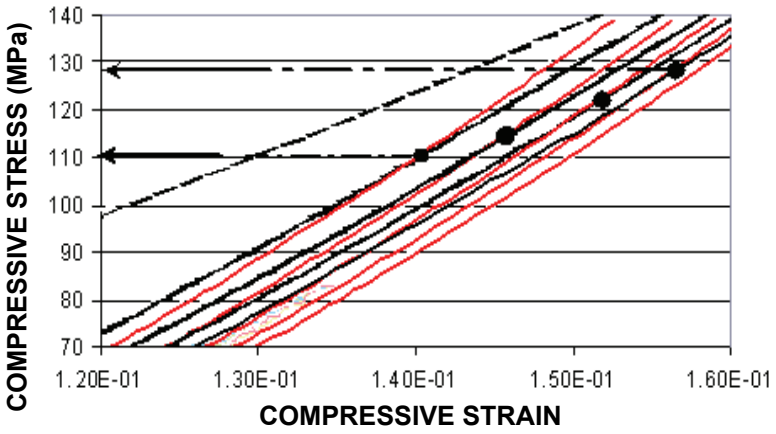
(d) The effect of loading history on the compressive properties of binary powder mixtures

Excessive deformation of compacted powders in the proposed RT application is undesirable and must be avoided. To minimize the deformation of the power compacts under repeated compressive load cycles, two new compression tests are designed and carried out to examine the deformation behavior of compacted powders under cyclic loading below the critical point using the binary T-15 tool steel powder mixture. The loading patterns for the two tests are shown in Figs. 8 and 9, respectively.

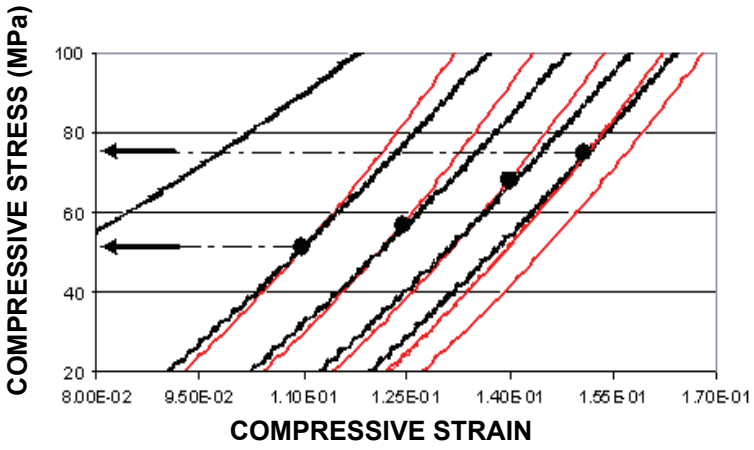
In the first test, after two cycles of loading and unloading at the maximum load of 138 MPa, three more cycles of loading-unloading are added with a maximum load set at 60% of the previous maximum load, as shown in Fig. 8.



(a) Mixing ratio: 0.74



(b) Mixing ratio: 0.77



(c) Mixing ratio: 0.80

Figure 7. Locations of critical points

The new maximum load after initial loading was chosen as 60% of the initial maximum load so that the it is safely below the critical points on the initial loading curves (The stresses at critical points are at least 80% of the maximum load for the mixture with a mixing ratio of 0.77 as shown in Fig. 7b) and still practical for injection moulding of engineering plastics.

In the second test, the powder first undergoes five loading and unloading cycles with a maximum load of 138 MPa, and three more cycles are followed with the maximum load set at 60% of that used in previous cycles, as shown in Fig. 9. Figure 10 shows the loading-unloading curves of the compression test using the loading pattern shown in Fig. 8 and powder mixtures with mixing

ratios of 0.74, 0.77, and 0.80, respectively. For all mixing ratios studied, there is no further strain increase observed in the three loading-unloading cycles with reduced maximum load. Details of the stress-strain curves for the mixing ratio of 0.77 are given in Fig. 11. Figure 12 shows the experimental results using the second loading pattern shown in Fig. 9. Again, there is no further increase in the total strain observed in the three loading-unloading cycles with reduced maximum load.

Powder mixture	Mixing ratio	At the first critical point (MPa)	At the fourth critical point (MPa)
T-15 tool steel	0.80	52	75
T-15 tool steel	0.77	110	127
T-15 tool steel	0.74	58	98

Table 9. Compressive stress at the first and fourth critical points

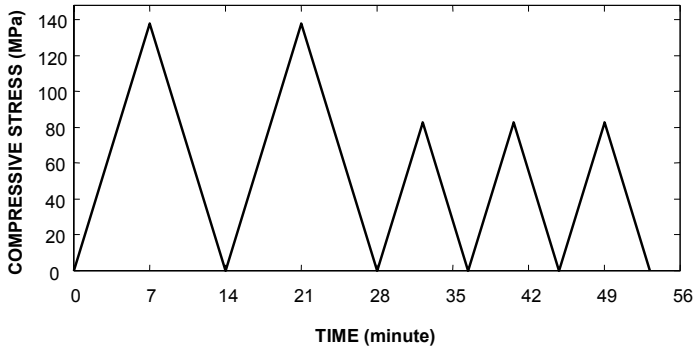


Figure 8. First loading pattern

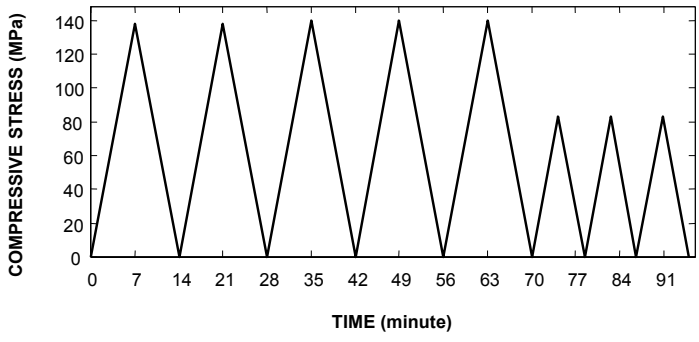
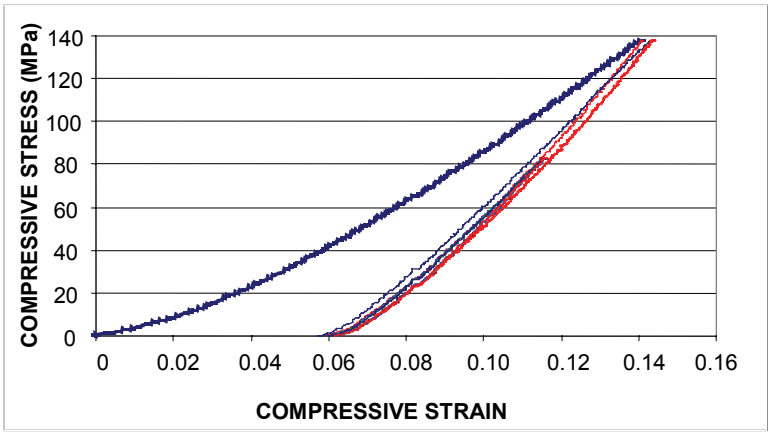
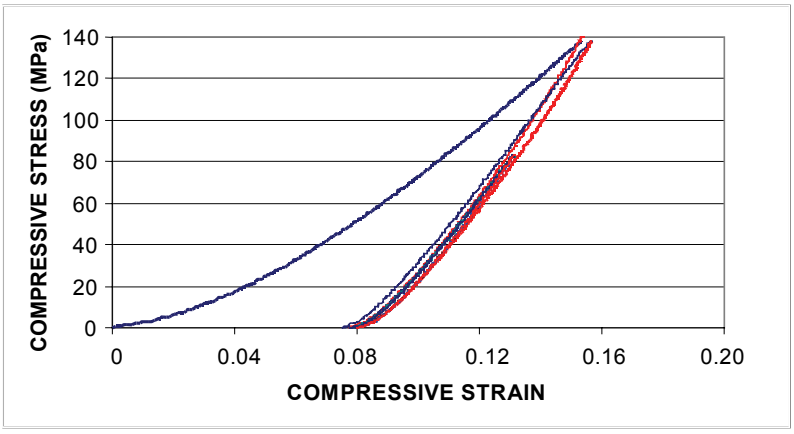


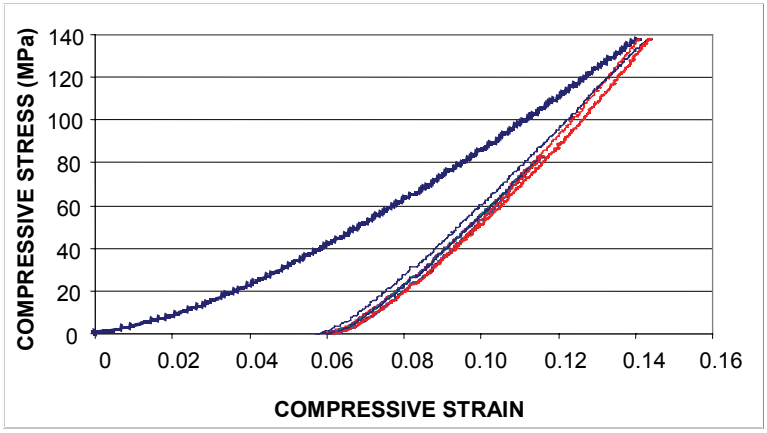
Figure 9. Second loading pattern



(a) Mixing ratio 0.74



(b) Mixing ratio 0.77



(c) Mixing ratio 0.80

Figure 10. Loading-unloading curves using the first loading pattern

These results indicate that by properly pre-compressing the metal powders, plastic deformation of the backing metal powders can be eliminated in the proposed RT process. It is also seen that the increase in the total strain in the first five loading-unloading cycles is smallest for the mixing ratio of 0.77 among all powder mixtures examined. This is consistent with the results from the compression test using a single loading-unloading cycle.

Table 10 shows the change in total strain and powder compact density between the second loading and the last loading for the loading pattern shown in Fig. 9. It is seen that the powder mixture with a mixing ratio of 0.77 has the least increase in both the total strain and powder compact density between the second loading and the last loading.

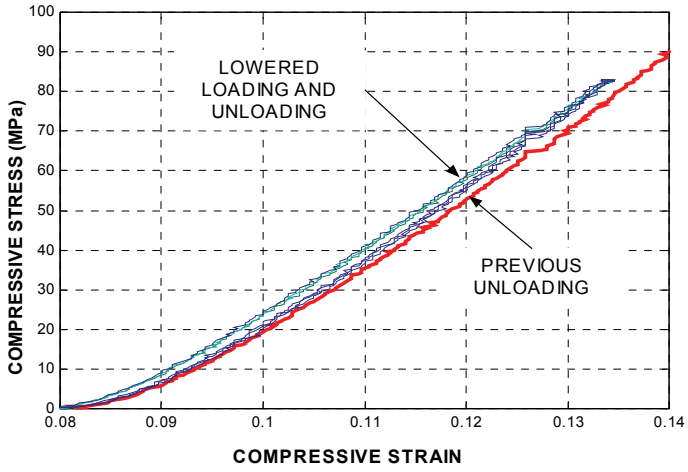
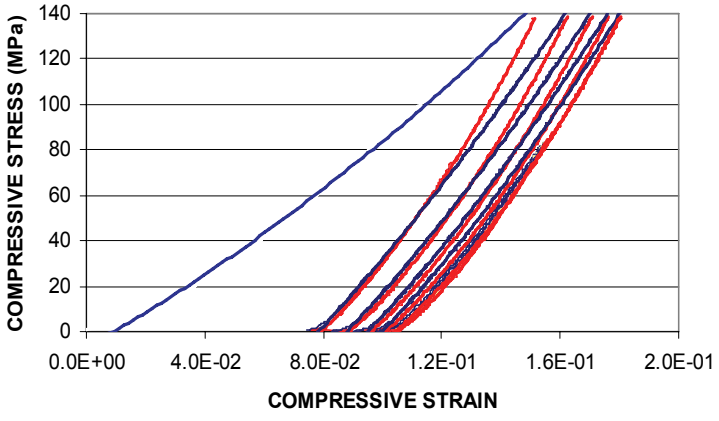
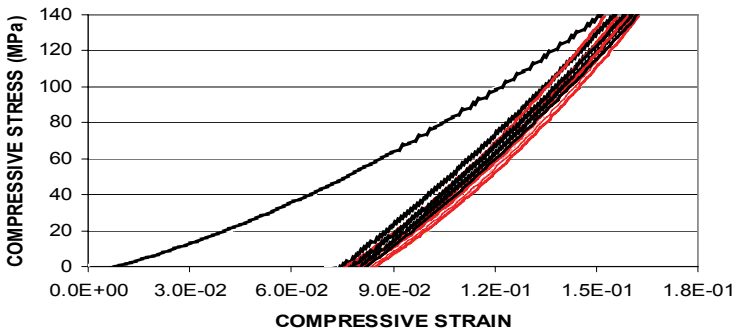


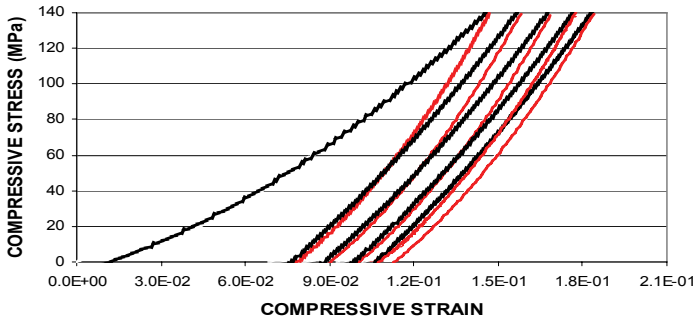
Figure 11. Detailed loading-unloading curves of the test using the first loading pattern (Mixing ratio: 0.77)



(a) Mixing ratio: 0.74



(b) Mixing ratio: 0.77



(c) Mixing ratio: 0.80

Figure 12. Loading-unloading curves using the second loading pattern

The increase in the total strain is less than 5% for the powder mixture with a mixing ratio of 0.77, indicating that the plastic deformation after the second loading cycle is very small.

4.5 Elastic Properties of Compacted Powders

Elastic properties, such as the Young's modulus and Poisson's ratio of compacted powders are important parameters for the deformation analysis of the metal shell and powder assembly used in the new RT process. Linear elasticity (Young's modulus and Poisson's ratio) is often sufficient to describe the behavior of compacted powders (Cambou, 1998). However, few studies have reported the elastic properties of unsintered compacted powders (Carnavas, 1998). Hehenberger *et al.* (1982) interpreted the unloading curves of uniaxial compression of compacted powders in terms of the Young's modulus and Poisson's ratio.

Powder/ Mixing ratio	Total strain			Compact density increase (%) between the second loading and the last loading
	Second loading (ε_S)	Last loading (ε_L)	$\Delta\varepsilon$= (ε_L- ε_S)	
T-15/0.80	0.157	0.183	0.026	8.1
T-15/0.77	0.155	0.162	0.007	0.6
T-15/0.74	0.162	0.179	0.017	1.8

Table 10. Changes of the total strain and compact density between the second loading and last loading

The Young's modulus and Poisson's ratios for the compacted binary T-15 powder mixture are obtained experimentally in this study using the data from the uniaxial compression test. The compaction of powders involves a number of different mechanisms. These mechanisms are interrelated through particle interactions that in turn depend on the distributed and individual particle properties. In contrast, unloading of the compacted powder is largely a linear elastic process. Therefore, the elastic properties of compacted powders can be more accurately determined from the elastic unloading process.
Typical unloading and reloading curves used for the calculation of the elastic properties of the compacted powder are shown in Fig. 13. Although the curves are not entirely linear, both the unloading and reloading curves have an apparently linear portion that is consistent with elastic deformation. In this linear portion, the reloading compressive strain does not exceed the previous unloading strain and the curve lies on the left of the previous unloading curve in the stress-strain diagram. In the current study, the elastic properties of compacted powders are calculated from the linear portion of the unloading curve. The linear portion is taken from the point with 20% of the maximum loading stress to the critical point. The elastic parameters of compacted powders are calculated using linear elastic theory. Figure 14 shows the unloading curve obtained from the compression test in the fifth unloading phase for the T-15 powder mixture with a mixing ratio of 0.77.

The Young's modulus of a compacted powder is calculated as follows:

$$E = \frac{\Delta\sigma}{\Delta\varepsilon} = \frac{\sigma_B - \sigma_A}{\varepsilon_B - \varepsilon_A} \tag{4}$$

where σ is compressive stress [MPa] and ε is compressive strain. 'A' and 'B' could be any two points on the straight line as shown in Fig. 14. A linear regression analysis shows that the R-squared value for the curve shown in Fig. 14 is 0.9982. It indicates that the trend line of the selected segment matches well with the actual experimental data. The tangent of the trend line is 1679.8 MPa, which is the Young's modulus of the compacted T-15 powder mixture at a mixing ratio of 0.77.

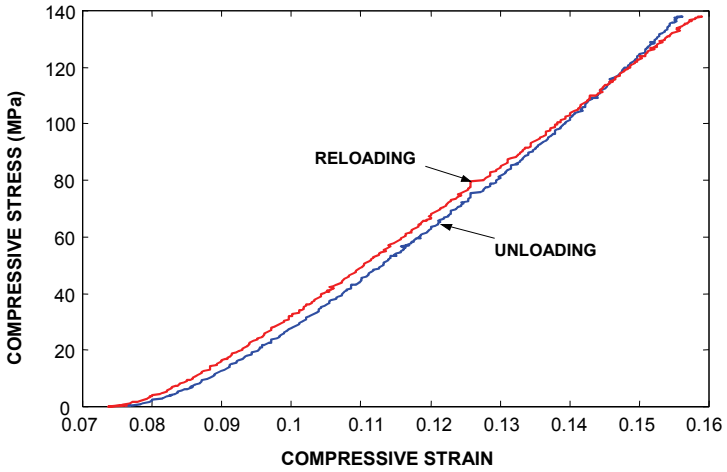


Figure 13. Typical unloading and reloading process

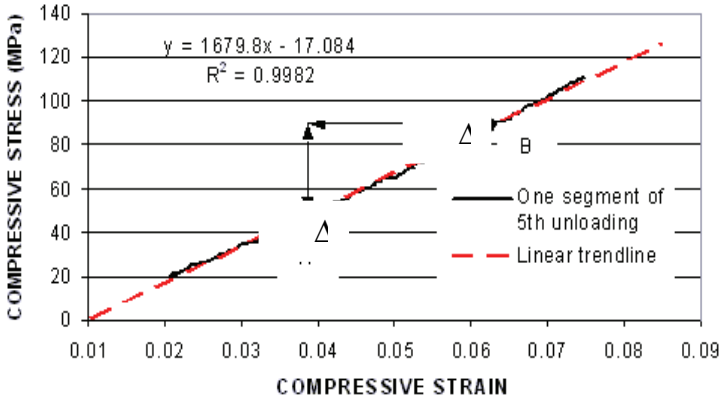


Figure 14. Selected unloading curve (Mixing ratio: 0.77)

For the calculation of the Poisson's ratio, it is assumed that the compression die is rigid, and the deformation on the inner surface of the die in the radial and circumferential directions is zero. From Hooke's Law, the triaxial stress can be written in cylindrical coordinates as (Gere, 2001):

$$\sigma_z = \frac{E}{(1+\nu)(1-2\nu)}\left[(1-\nu)\varepsilon_z + \nu(\varepsilon_r + \varepsilon_\theta)\right] \qquad (5)$$

where E is Young's modulus and v is Poisson's ratio. When ε_r and ε_θ are equal to zero, Eq. (5) can be rewritten as:

$$\sigma_z = \frac{E}{(1+\nu)(1-2\nu)}(1-\nu)\varepsilon_z \qquad (6)$$

If the Young's modulus E is assumed to be a constant in Eq. (6), the compressive stress and strain have a linear relationship. Letting $\Omega = \frac{E\varepsilon_z}{\sigma_z}$, Eq. (6) can be rewritten as:

$$2\nu^2 - (\Omega - 1)\nu + (\Omega - 1) = 0 \qquad (7)$$

The values of the Poisson's ratio for the T-15 powder mixture at a mixing ratio of 0.77 are calculated based on the experimental data using Eq. (7). They are tabulated in Table 11 and shown in Fig. 15. The results show that the Poisson's ratio decreases with an increase in compressive stress. This agrees with the results from Hehenberger's study (1982). Since the Poisson's ratio is the ratio of lateral strain to axial strain during elastic deformation, the decrease in the Poisson's ratio with increasing stress means that the increments of lateral strain will become smaller with each increment of compressive stress (strain).

Compressive Stress σ_z (MPa)	Ω	Poisson's Ratio ν
110	0.8454	0.242
95	0.8210	0.258
80	0.7875	0.277
70	0.7570	0.293
55	0.6909	0.323
40	0.5750	0.367
25	0.3200	0.437

Table 11. Poisson's ratios for T-15 tool steel powder mixture at the mixing ratio of 0.77

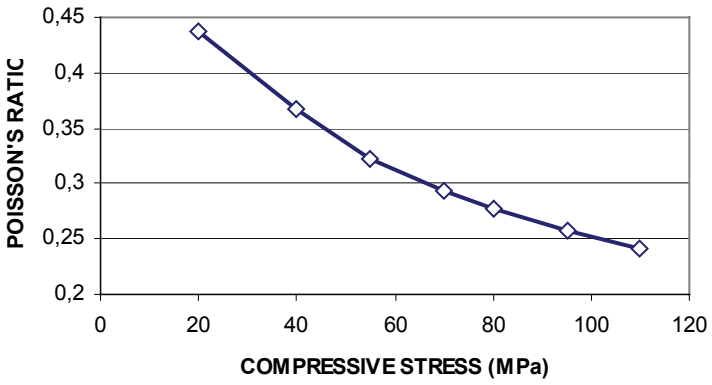


Figure 15. Variation of Poisson's ratio with compressive stress

5. Deformation of the Metal Shell

The front side of the metal shell used in the proposed new RT process has the shape complementary to the mould to be fabricated and the backside of the metal shell is hollow with a number of reinforcing ribs as shown in Fig 1. These ribs divide the space inside the metal shell into a number of square cells. The deformation analysis is conducted for a single cell formed by four adjacent ribs in the metal shell.

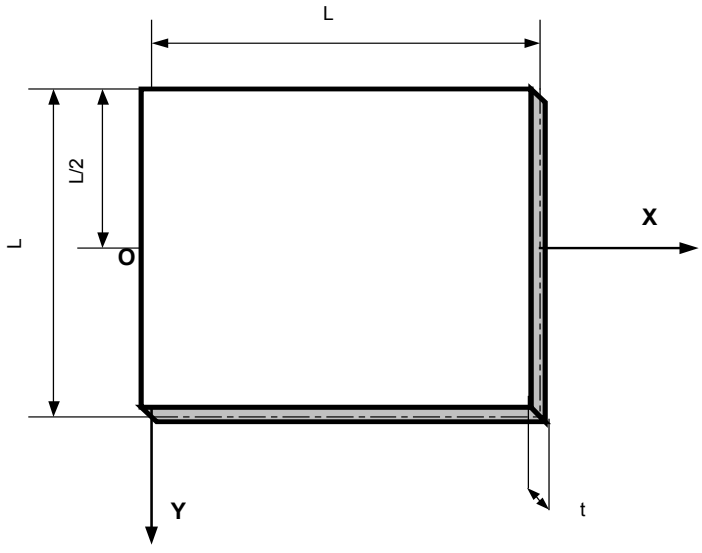


Figure 16. Dimensions of the thin plate

The ribs are assumed rigid and their deformation can be neglected during loading. Thus, the top surface of each cell can be considered as a thin plate with its four edges being clamped (fixed). Therefore, the model used for the deformation analysis of the metal shell is a square metal plate, which is equal to the size of a single cell, fixed at its four edges with a distributed force applied on it. This model is called the thin plate model. The dimensions of the thin plate are shown in Fig. 16, where *t* is the thickness of the plate and *L* is the edge dimension of the plate. The deformations of the thin plate can be analyzed using both the traditional elastic theory for a thin plate (Gould, 1998) and the finite element analysis (FEA) (Gould, 1998). Both methods are used in this study for the analysis of the deformation of the metal shell. The software I-DEAS (Lawry, 2000) is used for the FEA simulations. The comparison of the results from the two methods is presented.

5.1 Analysis Based on Traditional Elastic Theory

The assumptions used in traditional elastic theory for a thin plate are (Gould, 1998):

a. The material of the plate is elastic, homogeneous, and isotropic.
b. The plate is initially flat.
c. The thickness "*t*" of the plate is small in comparison to its lateral dimension *L*. The smallest lateral dimension of the plate is at least ten times larger than its thickness, i.e. $t / L \leq 0.1$.
d. Deformations are small in comparison to the thickness.

The maximum deformation occurs at the center of the plate. For a square plate with clamped edges, the maximum deformation based on the traditional elastic theory is (Gould, 1998):

$$\omega_{\max} = 1.26 \times 10^{-3} \frac{p_0 L^4}{D} \tag{8}$$

where

$$D = \frac{Et^3}{12(1-v^2)}$$

ω_{max} is the maximum deformation of the plate, p_0 is the uniformly distributed load per unit area on the plate, E is the plate material elastic modulus, v is the plate material Poisson's ratio, and t is the thickness of the plate.

5.2 Analysis Based on FEA

The simulation function is used to conduct the finite element analysis. The three-dimensional solid element and free mesh method are used in the simulation so that there is no restriction on the thickness of the plate.

5.3 Results

Since the stress and deformation of the thin plate varies with the plate thickness and loading, the deformation analysis is carried out for a thin plate with different thickness as well as different loading using both the FEA and the traditional elastic theory. The dimensions of the plate are 25.4×25.4 mm^2 and the materials of the plate is nickel (E=210 GPa and ν=0.31). Table 12 and Fig. 17 present the results on the maximum deformation of the metal shell at different relative shell thickness, t/L, with a uniformly distributed load of p_0 =138 MPa on the plate surface.

t (mm)	t/L	ω_{Max} (mm)		Difference (%)
		Elastic Theory	FEM Method	
0.2	0.0079	466.92	234	99.5
0.4	0.0157	58.37	48.9	16.2
0.6	0.0236	17.29	16.0	7.5
0.8	0.0315	7.30	6.98	4.4
1.0	0.0394	3.74	3.66	2.1
1.2	0.0472	2.16	2.15	0.5
1.4	0.0551	1.36	1.42	-0.04
1.6	0.0629	0.91	0.938	-3.1
1.8	0.0709	0.64	0.671	-4.8
2.0	0.0787	0.46	0.498	-8.3
2.4	0.0945	0.27	0.30	-11
2.8	0.1102	Not applicable	0.197	Not applicable
3.0	0.1181	"	0.165	"
3.4	0.1339	"	0.119	"
3.8	0.1496	"	0.091	"
4.0	0.1575	"	0.081	"
4.5	0.1772	"	0.062	"
5.0	0.1969	"	0.049	"

Table 12. Maximum metal shell deformation at different shell thickness

It can be seen that the thickness of the metal shell has a strong influence on the shell deformation. As shown in Fig. 17, when the relative shell thickness, t/L, is greater than 0.08, the deformation is not sensitive to the shell thickness. Table 12 shows that the maximum deformation calculated by both methods agree well in the range of t/L = 0.03 - 0.07, for which the difference between the two methods is less than 5%.

The smallest difference in the maximum deformation predicted by the two methods occurs at the relative thickness t/L = 0.055. In addition, when the relative thickness is greater than 0.1, the traditional elastic theory is not applicable for the prediction of the deformation. However, there is no such limitation for the FEA method using a three-dimensional element.

Table 13 shows the variation of the maximum deformation of the metal shell with loading at the relative shell thickness t/L = 0.0394. The results from both methods agree very well and the maximum difference is less than 2.3%.

6. Analysis of the Deformation of the Shell-Powder Assembly

The mechanical structure of the metal shell backfilled with the metal powder can be treated as a thin metal plate on an elastic block of compacted powders.

Loading (MPa)	ω_{max} (mm)		Difference (%)
	Elastic Theory	FEM Method	
137.9	3.74	3.66	2
124.1	3.36	3.29	2.1
110.3	2.98	2.93	1.8
96.5	2.61	2.56	2.1
82.7	2.24	2.19	2.3
68.9	1.87	1.83	2.0
55.2	1.49	1.46	2.3
41.4	1.12	1.10	1.8
27.6	0.747	0.731	2.2
13.9	0.374	0.366	2.0
6.9	0.187	0.183	2.0
3.4	0.0934	0.0914	2.1

Table 13. Maximum metal shell deformation at different loading

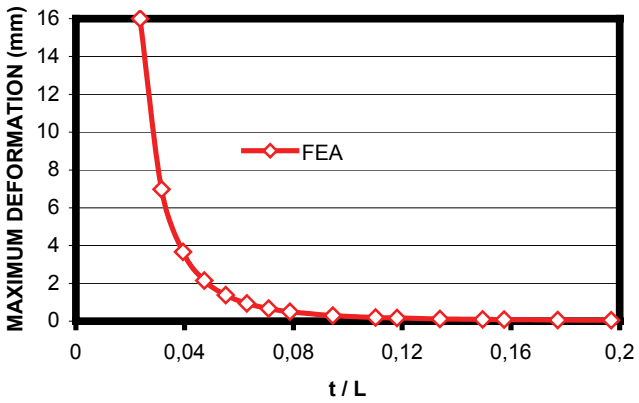


Figure 17. Variation of the maximum deformation of the shell with the shell thickness

Using the traditional elastic theory to analyze this type of structure is difficult although this is a common type structure in engineering applications. Therefore, the FEA method is used for the deformation analysis of the metal shell and powder assembly. The deformation behavior of the metal shell and powder assembly at different metal shell thickness, cell dimensions and compressive loading is investigated. The results can be used to optimize the structure of the shell-powder assembly.

6.1 Model and Boundary Conditions for the Shell-Powder Assembly

The analysis is carried out for a single cell of the metal shell and metal powder assembly. The model and the boundary conditions used in the analysis are shown in Fig. 18, where t is the thickness of the metal shell, H is the high of the powder, and L is the edge dimension of the cell. The material of the metal shell is nickel. The binary mixture of T-15 powders, Powders #5 and #7, at the mixing ratio of 0.77 is used to backfill the metal shell. The elastic properties of the binary powder mixture are from those given in Table 11. The boundary conditions of the model are also shown in Fig. 18. A uniform distributed stress, the loading, is applied on the top surface of metal shell. The four edge surfaces and the bottom of the shell-powder assembly are fixed.

6.2 Results

The simulations are performed for two different cell sizes of the shell-powder assembly, L=25.4 mm and L=12.7 mm, at different shell thickness, t, and differ-

ent loading, *P*. *H* is taken as 50 mm. The variations of the maximum deformation of the shell-powder assembly with the shell thickness and loading are shown in Fig. 19 for the cell size of *L*=25.4 mm and *L*=12.7 mm. It is observed that the cell size, *L*, and shell thickness, *t*, especially the cell size, significantly affect the deformation of the shell-powder assembly.

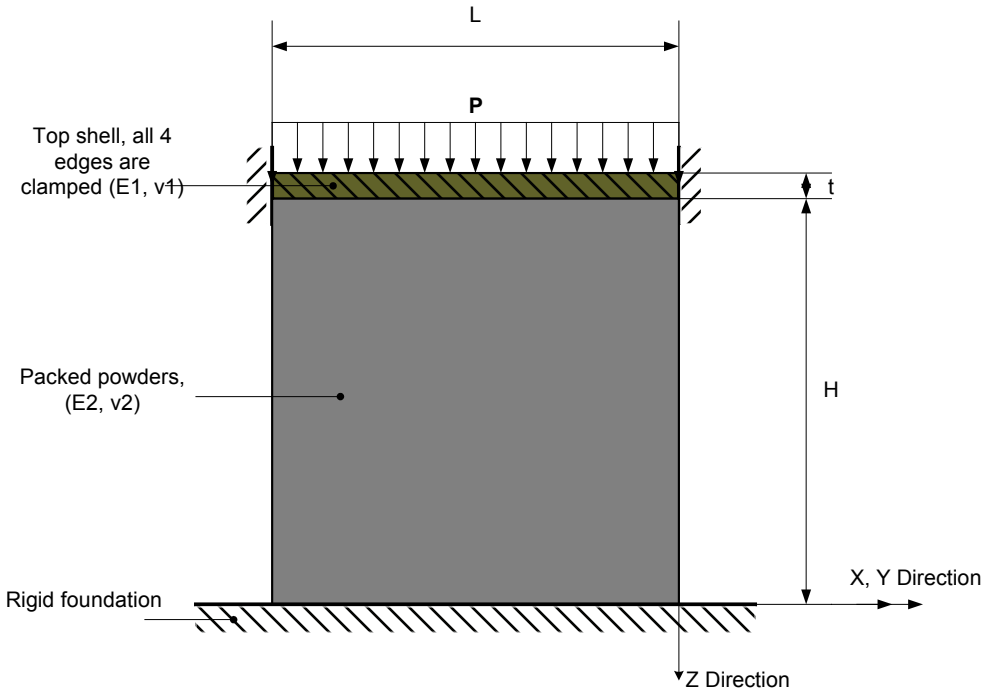
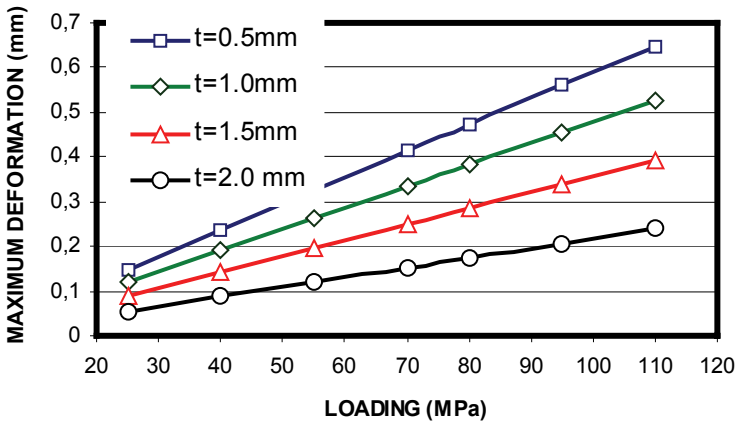
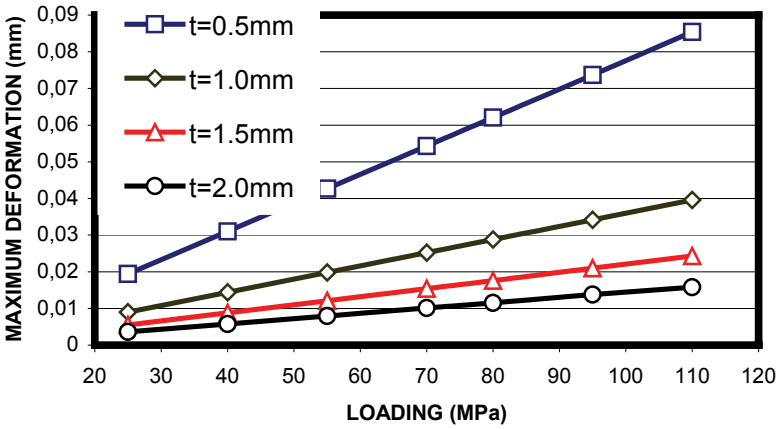


Figure 18. Shell-powder assembly and boundary conditions



(a) L = 25.4 mm



(b) L = 12.7 mm

Figure 19. Variation of the maximum deformation of the shell-powder assembly with the shell thickness and loading

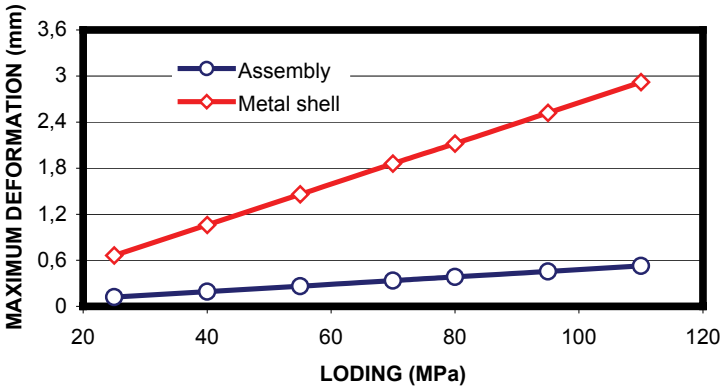


Figure 20. Comparison of the maximum deformations between the metal shell and the shell-powder assembly

At the same shell thickness, the maximum deformation can be lowered by a factor of 7-15 when the cell size is reduced from 25.4 mm to 12.7 mm.
For the same cell size, the maximum deformation can be lowered to 2.7 and 5.3 times when the shell thickness is increased from 0.5 mm to 2 mm.
The comparison of the maximum deformation of the metal shell and the shell-powder assembly is given in Fig. 20, where the parameters are H=50 mm, L=25.4 mm, and t=1 mm. It can be seen that the use of the backfilled metal powder greatly improves the structure's ability to resist deformation. The

maximum deformation of the shell-powder assembly at a compressive loading of 110 MPa is only 18% that of the metal shell without the powder backing support. Without powder support, the maximum deformation of the metal shell increases drastically with the increase in loading as shown in Fig. 20. In contrast, the increase in the maximum deformation with the increase of the loading is very slow if the back support metal powder is used.

7. Conclusions

In the single component packing, the particle shape has the most significant effect on the powder packing density. The spherical and round particles produce higher packing density, and therefore, are desirable for the intended RT application. The dimension ratio of the container and the particle (D/d) has less effect on the packing density when $D/d \geq 7.66$, but packing density started to drop significantly when the ratio D/d is less than 5. The particle size has no significant effect on the packing density. Mixing particles with different sizes can greatly increase the packing density because the voids among large particles can be filled by small particles. In the current study, the packing density of three-component packing can reach 0.91 and binary packing density can reach 0.86. The particle size ratio is a very important parameter for multiple component packing. For best packing results, the size ratio of the large particle to the small particle should be higher than at least 6.82 so that all small particles can enter the interstices between large particles. On the other hand, particle size should not be too small to avoid low fluidity. The mixing ratio is another important parameter affecting multiple component packing density. There exists an optimal mixing ratio for a binary mixture at which the packing density is maximal. The optimal mixing ratio is in the range of 0.71 - 0.77.

The deformation behavior of compacted powders under uniaxial compression depends heavily on the properties of the powder materials, the mixing ratio of powders of different sizes, as well as the loading history. It is possible to obtain the required elastic properties of the compacted powders used for the proposed RT application by adjusting these factors. The powder with the higher hardness has less compressive deformation under the same loading conditions. Higher hardness also helps to prevent compacted powders from forming blocks after compression, which is necessary for powder reuse. The T-15 tool steel powder is the best choice among the materials studied for the proposed new RT process. In addition, the mixing ratio of 0.77 is found to

give the highest deformation resistance. The critical point is an interesting and important phenomenon for powder compression. It defines the limit for the operating load that produces the smallest increase in plastic strain under the tool working condition. The loading history is important for the deformation behavior of compacted powders. A higher loading stress level generally leads to higher operating loads without further plastic deformation in subsequent loading. The load level that corresponds to the critical point increases with the number of loading cycles. The deformation resistance of the compacted powder can be improved by increasing the number of loading cycles. The unloading curves have a linear portion and are almost parallel in a certain range of the stress-strain curve. The Young's modulus and Poisson's ratios can be obtained using the experimental data from the linear portion.

The maximum deformation of the metal shell predicted by the elastic theory and FEA method are in good agreement in a certain range of the relative shell thickness (t/L). The elastic theory is applicable only for a thin plate. However, there is no such restriction for the FEA method. The relative metal shell thickness has a significant effect on the deformation of the metal shell. Furthermore, the use of the metal powder to support the metal shell can greatly improve its ability to resist deformation. The deformation of the metal shell-powder assembly depends also strongly on the dimension of the cell size and the thickness of the metal shell. Use of a proper cell dimension and shell thickness can limit the deformation of the shell-powder assembly within the tolerance, which indicates that the proposed new rapid tooling process is feasible.

8. References

ASTM B331-95 (2002). *Standard Test Method for Compressibility of Metal Powder in Uniaxial Compaction*, ASTM International.

Cambou, B. (1998). *Behavior of Granular Materials*, Springer Verlag.

Carnavas, P. C. (1998). Elastic Properties of Compacted Metal Powders, *Journal of Material Science*, Vol. 33, pp. 4647-4655.

Gere, J. M. (2001). *Mechanics of Materials*, 5th edition, Brooks/Cole, Pacific Grove, CA.

German, R. M. (1998). *Powder Metallurgy of Iron and Steel*, John Wiley & Sons Inc., New York.

Gould, P. L. (1998). *Analysis of Shells and Plates*, Prentice Hall Upper Saddle River, New Jersey.

Hehenberger, M., Samuelson, P., Alm, O., Nilsson, L. and Olofsson (1982). Experimental and Theoretical Studies of Powder Compaction, *Proceedings of the Conference on the Deformation and Failure of Granular Materials*, IUTAM, pp. 381–390, Delft, September, 1982.

Hejmadi, U., and McAlea, K. (1996). Selective Laser Sintering of Metal Molds: The Rapid Tooling Process, *Proceedings of the Solid Freeform Fabrication Symposium*, pp. 97-104, Austin, TX, August, 1996.

Jacobs, P. F. (1992). Rapid Prototyping & Manufacturing Fundamentals of Stereo Lithography, Society of Manufacturing Engineers, Dearborn, MI.

Lawry, M. H. (2000). I-DEAS Master Series™ (Mechanical CAE/CAD/CAM Software), *Student Guide*, Structural Dynamics Research Corporation, Eastman Dr. Milford, OH.

Leva, M. and Grummer, M. (1947). Pressure Drop Through Packed Tubes, Part III, Prediction of Voids in Packed Tubes, *Chemical Engineering Progress*, Vol. 43, pp. 713-718.

McGeary, R. K. (1962). Mechanical Packing of Spherical Particles, *Journal of the American Ceramic Society-McGary*, Vol. 44, No. 10 pp. 513-522.

Nelson, C. (1999). Rapid Tooling Mold Inserts for Pressure and Gravity Die Casting Using SLS Technology, *Transactions of World of Die Casting*, Nov., pp. 441-446.

Pham, D. T. (1998). Techniques for Firm Tooling Using Rapid Prototyping, *Proceeding of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, Vol. 212, No. B4, pp. 269-277.

Phelan, M. (1997). The Two-Week Tool, *Automotive Industries*, March, pp. 62-63.

Shinohara, K. (1984). Rheological Property of Particulate Solids, *Handbook of Powder Science and Technology*, New York, pp. 396-404.

Willis, S. (1997). Fast Rapid Tooling for Injection Molds, *Proceedings of the Seventh International Conference on Rapid Prototyping*, pp.1-12, San Francisco, March, 1997.

23

SCM Innovation for Business Globalization Based on Coupling Point Inventory Planning

Koshichiro Mitsukuni, Yuichi Nakamura and Tomoyuki Aoki

1. Introduction

We introduce a new SCM (Supply Chain Management) solution based on Coupling Point Inventory Planning theory for the repetitive products. This solution has been achieved three-way optimums; quick response to order, the available inventory level flatly, and the multi-items replenishment under the limited capacity.

The new SCM solution is organized to combine three methods. The first method is the Coupling Point establishment for quick response to order when the demand lead time is shorter than the supply lead time. Coupling Point (CP) is defined where the position of demand and supply lead-times are equal on the supply chain process.

The second method is the re-order calculation to control the inventory level flatly without demand forecast. It plans the theoretical necessary inventory including safety stock under the demand fluctuation, and calculates the re-order quantity. The re-order quantity is calculated by the difference of the theoretical necessary inventory and the measured actual stock.

The third method is the multi-items replenishment by the margin stock ratio under the limited capacity when the supply capacity and the demand quantity are different. The margin stock ratio is calculated by the theoretical necessary inventory and measured actual available inventory.

The organization of this paper is as follows: In chapter 2, the problem of the recent SCM is stated. In chapter 3, the concept of the coupling point inventory planning theory is stated. In chapter 4, the method of the coupling point establishment is introduced. In chapter 5, the method of the re-order quantity calculation for inventory level flatly without the demand forecast is introduced. In chapter 6, the method of the multi-items replenishment by the margin stock ratio under the limited capacity is introduced. In chapter 7, the expected effect by the coupling point inventory planning method is described. In chapter 8, the application of coupling point inventory planning method is introduced.

Finally, the paper will be concluded in chapter 9 with the brief discussion of the result.

2. Problem of recent SCM

2.1 Problem of global SCM

Increasing number of international companies have been employed SCM based on demand forecast in recent years. However, in the global SCM, because of the various processes and procedures in transportation (such as, custom clearance, container handling, ocean transportation, physical transportation) of the parts and the products could take anywhere from a week up to four weeks. Further more, various parts (or the products) are often made in different areas of the world to be assembled and sold all over the world. As the total supply lead-time is necessary 4 ~ 5 months, the demand forecast before the laps of lead time is inaccurate.

In case of the supply lead time is long, the inventory management method is used the fixed-cycle ordering system in some companies. Figure 1 shows one of the characteristics of the fixed-cycle ordering system. As this system absorbs the demand fluctuation by the inventory, the inventory level becomes to fluctuate. Then, international companies have the difficulty with excess or shortage inventory (Lagodimos & Anderson, 1993, Shacham, 1993).

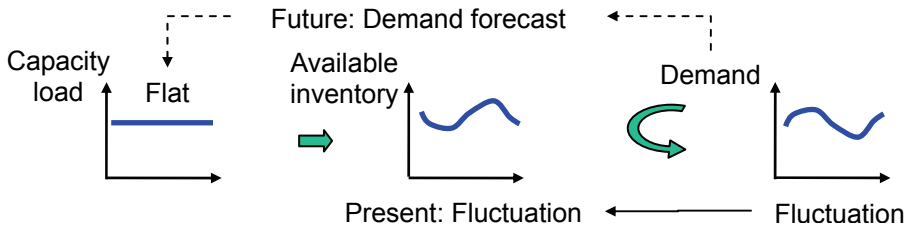


Figure 1. characteristics of the fixed-cycle ordering system

2.2 Problem of domestic SCM

On the other hand, companies of using SCM in domestic can shorten supply lead-time and decrease on hand stock by controlling the balance between demand and the supply as well as shortening the time and the process. It takes to communicate the demand information between downstream (such as, cus-

tomer, consumer) and upstream (such as, materials, parts) companies. However, the shortening planning cycle becomes small bucket size of the production capacity.

In case of the supply lead time and the ordering cycle are short, the inventory management method is used the fixed-size ordering system in some companies or Just-In-Time system. Figure 2 shows one of the characteristics of the fixed-size ordering system. As this system controls the inventory level flatly, the capacity / load level becomes to fluctuate. Then, domestic companies have also the difficulty with excess or shortage inventory (Kimura & Terada, 1981, Huang & Kusiak, 1996).

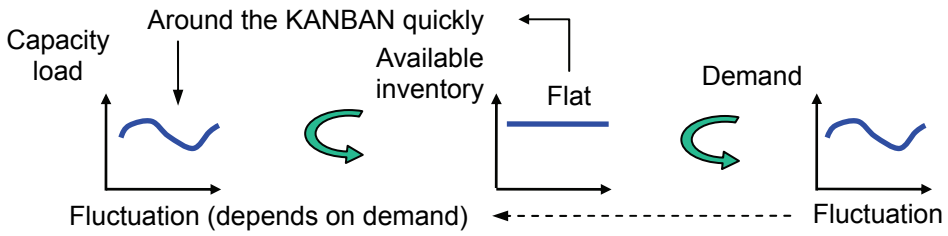


Figure 2. characteristics of the fixed-size ordering system

2.3 Reasons of inventory occurrences

As just described, we are able to recognize two reasons of the inventory occurrences. Figure 3 shows the first reason is the imbalance between the demand and the supply lead time.

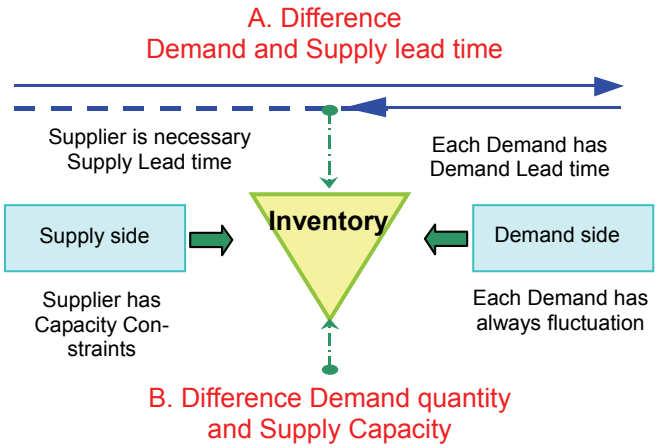


Figure 3. Elucidation of inventories occurrence factor

For example, the transportation lead time is too long in the global SCM. The second reason is the imbalance between the demand quantity and the supply quantity or capacity. For example, the bucket size or capacity is small in the domestic SCM.

3. Concept of Coupling Point Inventory Planning

3.1 Recognizing of physical phenomenon

The reasons of inventory occurrences are described by Figure 4. The expecting and the excess inventory situations are related by two reasons on the matrix.
In the first situation, excess inventories do not occur when the demand lead time is longer than the supply lead time or when they are equal, because production based on order is possible. But it can not cope with the case when the demand lead time is shorter than the supply lead time. Manufacturing and stockpiling are taken to prevent the loss of sales opportunities. This is called Production based on demand forecast.
In the second situation, surplus inventories do not occur when the demand quantity is equal to the supply quantity. Inventories occur as their difference when the supply quantity is greater than the demand quantity.

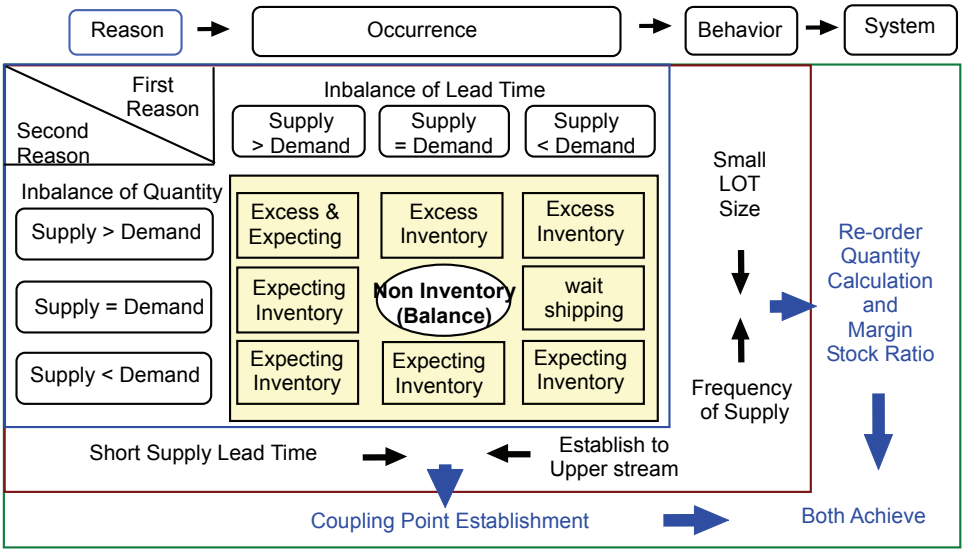


Figure 4. Concept of coupling point inventory planning.

This difference causes a shortage when the supply quantity is smaller than the demand quantity. In this case, both current production and inventories are used up to prevent the loss of sales opportunity. Seasonal products are the good examples.
In some cases, inventories are caused by the imbalance of the lead time. In the other cases, surplus inventories are caused by the imbalance of the quantity or capacity. The former is called Expecting inventories, and the latter is called Excess inventories. Consequently, inventory occurrences are recognized to occur on the physical phenomenon.
Some SCM software based demand forecast method controls the excess or the shortage inventory. Although these methods are useful for production planning and management, the situation of the excess or the shortage inventory does not clear. Because, in the explanation of the inventory is necessary the inventory theory.

3.2 Definition of Coupling Point Inventory Planning

We have been proposed to solve combine the two reasons of inventories occurrence. The bottom and the right side in Figure 4 shows, the first solution is the coupling point establishment for quick response to order when the demand lead time is shorter than the supply lead time. The second solution is the reorder quantity calculation for inventory level control flatly. The third solution is the multi-items replenishment by the margin stock ratio under the limited capacity when the supply capacity and the demand quantity are difference.
The new SCM solution is organized to combine three methods. We give a generic name 'Coupling Point Inventory Planning Theory' to three methods (Mitsukuni et al., 1997, 1999, 2002, 2003).

4. Method of Coupling Point establishment for quick response to order

4.1 Definition of Coupling Point

The view point of the supply chain lead time is shown Figure 5. CP is determined by demand and supply lead time. Supply lead time is defined as the time it takes to send products from a certain position in the process to a customer. For example, the supply lead time at the position of supplier is the sum of the time of all the sub-processes (such as parts maker, set maker, and deliv-

ering etc).

Demand lead time is defined as the time it takes for customers to receive products from ordering. Demand lead time is determined by the customer requirement.

Using these two kinds of lead time, Coupling Point (CP) is defined as follows: the stock positions where the supply and the demand lead times are equal. Then, total supply lead time is divided at CP, and we are able to quick response to order on the downstream side. Then, we are aware the new lead time from upstream to CP.

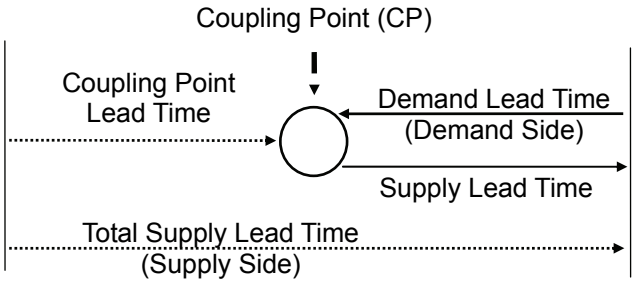


Figure 5. Definition of Coupling Point

Equation of Coupling Point definition is shown Figure 6, and the Coupling Point i^* is shown in the equation (1). In this equation, $P(i)$ denotes the processing time of a sub-process i, where $i = 1$ to n increasing from demand side to supply side.

The stock position situated at the entering position of sub-process i is also defined to be i. Ld is the demand lead time which depends on customers' requirement.

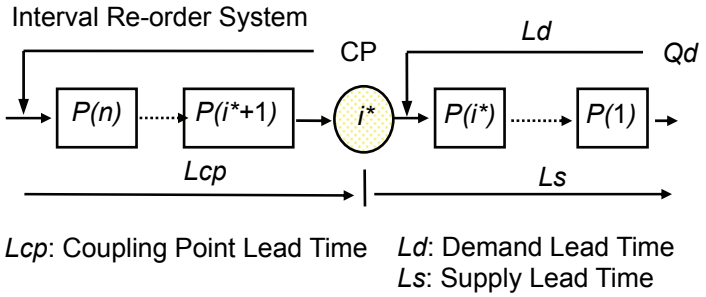


Figure 6. model of Coupling Point

$$i^* = \left\{ \max_{1 \le i \le n} i \,\middle|\, \sum_{j=1}^{i} P(j) \le Ld \right\} \tag{1}$$

The new supply lead time from sub-process i to customer Ls is denoted by equation ($\sum_{j=1}^{i} P(j)$). The coupling point lead time from sub-process $p(n)$ to CP Lcp is denoted by equation ($\sum_{j=i^*+1}^{n} P(j)$).

4.2 Necessary Inventory of Coupling Point

We assume that the replenishment system of CP to be an interval reorder system, and that the demand is under normal distribution. The necessary stock of inventories In is shown as equation (2). In this equation, Qd is the mean value of demand size per unit period, σd is the standard deviation. Lcp is the coupling point lead time, C is the re-order interval, k is the coefficient of safety stock (depends on service level).

$$In = Qd(Lcp + C) + k\sqrt{(Lcp + C)} \cdot \sigma d \tag{2}$$

Next, we assume that the process consists of n stage sub-processes, in which inventories are stocked exclusively at the stock position of CP i^*. Lcp is substituted by ($\sum_{j=i^*+1}^{n} P(j)$). Then necessary stock of inventories $In(i^*)$ is equation (3).

$$In(i^*) = Qd\Big(\sum_{i=i^*+1}^{n} P(i) + C\Big) + k \bullet \sqrt{\Big(\sum_{i=i^*+1}^{n} P(i) + C\Big) \cdot \sigma d} \tag{3}$$

If CP is established where position is near the upstream side, the expecting inventory becomes to decrease. On the opposite, if CP is established where position is near the downstream side, the expecting inventory becomes to increase. Consequently, the expecting inventory is depending on the position of CP.

4.3 Example of coupling point establishment

An example of the supply chain model process is shown Figure 7. The supply chain model process has 5 companies. The most of the upstream company is a parts manufacturing, and supply lead-time is 14days (2week).

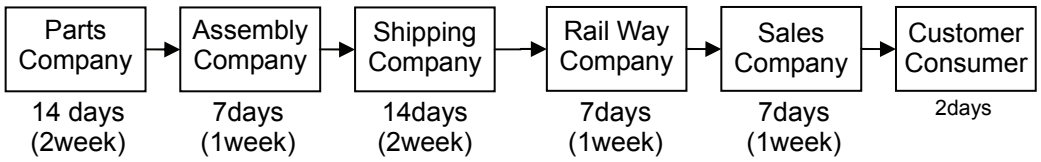


Figure 7. SCM Model Process

Similarly, each company is placed on the supply chain process from the upstream to downstream. The second company is an assembly manufacturing, and lead-time is 7days (1week). The third company is a shipping and railway transportation, and lead-time is 21days (3weeks). The forth company is a sales, and lead-time is 7days (1week) to customize and package. The most of the downstream side is customers or consumers, and required delivery lead time is two days.
An example of the establishing Coupling Point (CP) is shown Figure 8. As Customers require 2 days deliver, CP is established at the exit of the sales company for quick response to orders. Consequently, the coupling point lead time from the parts to CP is 49days (7weeks).

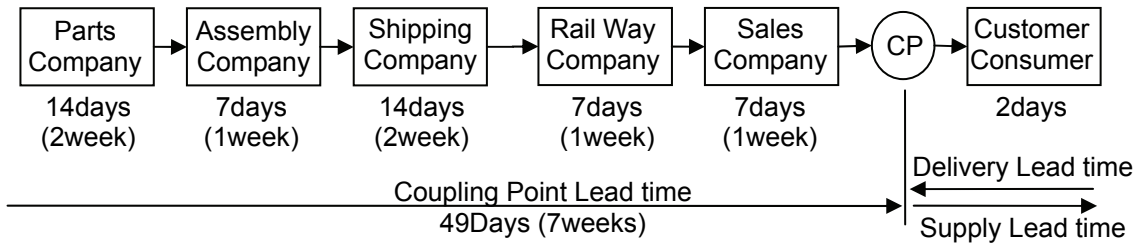


Figure 8. Example of Establishing Coupling Point

5. Method of Re-order Quantity Calculation for inventory level flatly

5.1 Definition of re-order quantity

We should consider the anticipation inventory of the upstream side of CP. The method of re-order quantity determination is shown Figure 9.
On each item, each planning period, the total inventory is defined theoretical necessary inventory *In*.

The theoretical necessary inventory is planed by the lead-time of the supply chain process, planning cycle, average demand quantity and its standard deviation. *In* includes the safety stock.

On the other hand, as the actual stock on supply chain process and the actual stock on hand are measured by the fact and are reported to inventory planning. The total of measured actual stock of each item is defined available inventory *Ia*. When the available inventory is less than the theoretical necessary inventory, the difference quantity is the shortage.

Thus, the difference quantity should be replenished. Then the difference quantity of available and theoretical necessary is defined re-order quantity. When the available inventory is greater than the theoretical necessary inventory, the difference quantity of excess is the over. In this situation, the replenishment is unnecessary.

Consequently, the total inventory becomes flat by the theoretical necessary inventory. And the out-of-stock situation becomes to decrease by the safety stock.

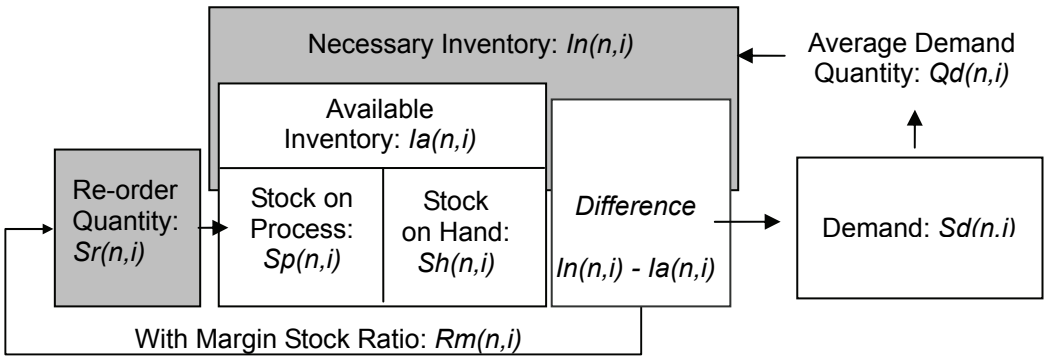


Figure 9. Definition of Re-order Quantity

5.2 Model of Re-order Calculation

An example of the supply chain logical model is shown Figure 10. In this model, between the parts manufacturing and the sales company are denoted by the abstraction of one logical supply chain process. The inventory planning is considered at CP. Each item denotes n. The coupling point Lead-time of each item denotes $Lcp(n)$ (week). The inventory planning cycle denotes C(week), and each planning period denotes i.

At CP, on each item n, each planning period i, the actual demand quantity (out) denotes $Sd(n,i)$ (unit), the actual arrival (in) quantity denotes $Sa(n,i)$ (unit), the result of on hand stock quantity after in / out denotes $Sh(n,i)$ (unit). $Sh(n,i)$ is measured by the fact, and is reported to inventory planning. The co-efficient of the safety stock denotes $k(n)$. The moving average demand quantity under the moving average term m denotes $Qd(n,i)$ (unit/period), and its standard deviation denotes $\sigma d(n,i)$. $Qd(n,i)$ and $\sigma d(n,i)$ are calculated by the actual demand quantity $Sd(n,i)$. The actual stock on the physical supply chain process denotes $Sp(n,i)$ (unit). $Sp(n,i)$ is measured by the fact, and is reported to inventory planning. The re-order quantity denotes $Sr(n,i)$ (unit) and the margin stock ratio denotes $Rm(n,i)$. Then, the inventory planning directs $Sr(n,i)$ and $Rm(n,i)$ to the physical supply chain process.

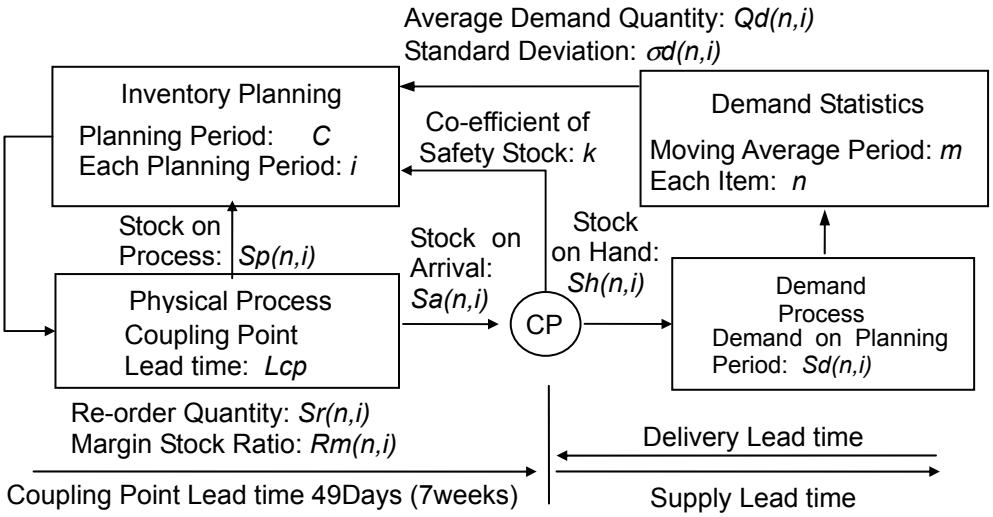


Figure 10. Model of re-order calculation

5.3 Algorithm of re-order calculation

The demand quantity $Qd(n,i)$ in moving average term m of each item n, each planning period i is calculated by equation (4). And its standard deviation is calculated by equation (5).

$$Qd(n,i) = \frac{\sum_{x=i-m}^{i-1} Sd(n,x)}{m} \tag{4}$$

$$\sigma d(n,i) = \sqrt{\frac{\sum_{x=i-m}^{i-1} (Qd(n,i) - Sd(n,x))^2}{m-1}} \tag{5}$$

The moving average term m is determined to consider by the products life cycle, seasonable demand, and trend of each item n. The moving average term m is recommended greater than 40~50 as a population of the demand data. For example, in case of actual sales term is 1 year, the planning period is divided a week, the result of moving average term m is 52. Then the seasonal variation is able to include the standard deviation, because the influenced of seasonal dispersion is calculated by the 52 weeks including maximum / minimum data in the year.

The theoretical stock on supply chain process of each item n, each planning period i denotes $Fp(n,i)$ (unit), and is calculated by equation (6). $Fp(n,i)$ is multiplied by average demand quantity Qd (n,i) and Coupling Point lead time $Lcp(n)$.

$$Fp(n,i) = Qd(n,i) \cdot Lcp(n) \tag{6}$$

The theoretical on hand stock of each item n, each planning period i denotes $Fc(n,i)$ (unit), and is calculated by equation (7). $Fc(n,i)$ is multiplied by average demand quantity Qd (n,i) and planning cycle C(week).

$$Fc(n,i) = Qd(n,i) \cdot C \tag{7}$$

The theoretical inventory of re-order interval of each item n, each planning period i denotes $Fd(n,i)$ (unit), and is calculated by equation (8). $Fd(n,i)$ is a total of $Fp(n,i)$ and $Fc(n,i)$.

$$Fd(n,i) = Fp(n,i) + Fc(n,i) = Qd(n,i) \cdot (Lcp(n) + C) \tag{8}$$

The theoretical safety stock of each item n, each planning period i denotes $Fs(n,i)$ (unit), and is calculated by equation (9). $Fs(n,i)$ is calculated by the coefficient of safety stock k, standard deviation of the average demand $\sigma d(n,i)$, process lead-time $Lcp(n)$, and planning cycle C.

$$Fs(n,i) = k(n)\sqrt{(Lcp(n) + C)} \cdot \sigma d(n,i) \tag{9}$$

The theoretical necessary inventory of each item n, each planning period i de-

notes $In(n,i)$ (unit), and is calculated by equation (10). $In(n,i)$ is a total of $Fd(n,i)$ and s(n,i).

$$In(n,i) = Fd(n,i) + Fs(n,i) = Qd(n,i) \cdot (Lcp(n) + C) + k(n)\sqrt{(Lcp(n) + C)} \cdot \sigma d(n,i) \quad (10)$$

The available inventory of each item n, each planning period i denotes $Ia(n,i)$, and is calculated by equation (11). $Ia(n,i)$ is a total of actual stock on supply chain process $Sp(n,i)$ and actual stock on hand $Sh(n,i)$.

$$Ia(n,i) = Sp(n,i) + Sh(n,i) \quad (11)$$

The re-order quantity of each item n, each planning period i denotes $Sr(n,i)$ (unit), and is calculated by equation (12). $Sr(n,i)$ is subtracted from the theoretical necessary inventory $In(n,i)$ to the available inventory $Ia(n,i)$. The margin stock ratio $Rm(n,i)$ is calculated by equation (13) and will describe in Chapter 6.

$$Sr(n,i) = In(n,i) - Ia(n,i) \quad (12)$$

$$Rm(n,i) = \frac{Ia(n,i)}{Fd(n,i)} \quad (13)$$

5.4 Flow of re-order calculation

The flow of the re-order quantity calculation is shown Figure 11.
Step1~Step3 are initialize procedures of inventory planning. In Step1, the planning cycle C and the moving average term m are established. In Step2, the Coupling Point lead time Lcp, the co-efficient of safety stock k, and the past demand Sd of each item n are established. The co-efficient of safety stock k is determined by the service ratio. For example, k=1.64, 1.96, 2.25, and 3.27 are depending on the service ratio 90%, 95%, 97.5%, and 99.9%. As the past demand Sd is established between (i-m) to (i-1) of the moving average term m, current period i is established m+1.
Step4 is a statistics procedure of the demand. The average demand quantity and the standard deviation between (i-m) to (i-1) is calculated by equation (4) and (5), then the results are established at position i.
Step5 is a calculation procedure of the theoretical necessary inventory In, and In is calculated by equation (6)to(10), then the result is established at position i.
Step6~7 is a calculation procedure of available inventory Ia, and Ia is calculated

by equation (11), then the result is established at position *i*.

Step8 is a calculation procedure of re-order quantity *Sr* and margin stock ratio *Rm*. *Sr* is calculated by equation (12) and *Rm* is calculated by equation (13), then results are established at position *i*.

Step9 is a direction procedure of the re-order quantity *Sr* and the margin stock ratio *Rm* to physical supply chain process. When the planning period becomes next, the planning period is 1 up. The re-order quantity *Sr* at the prior period is moved to the actual stock on supply chain process *Sp* at the current period.

Similarly, the arrival from the actual stock on supply chain process *Sp* at prior period is moved to the arrival quantity *Sa* at current period. In the management of the work shop, the manager should be doing action for the abnormal when the manager finds the difference of the calculated and the actual quantity.

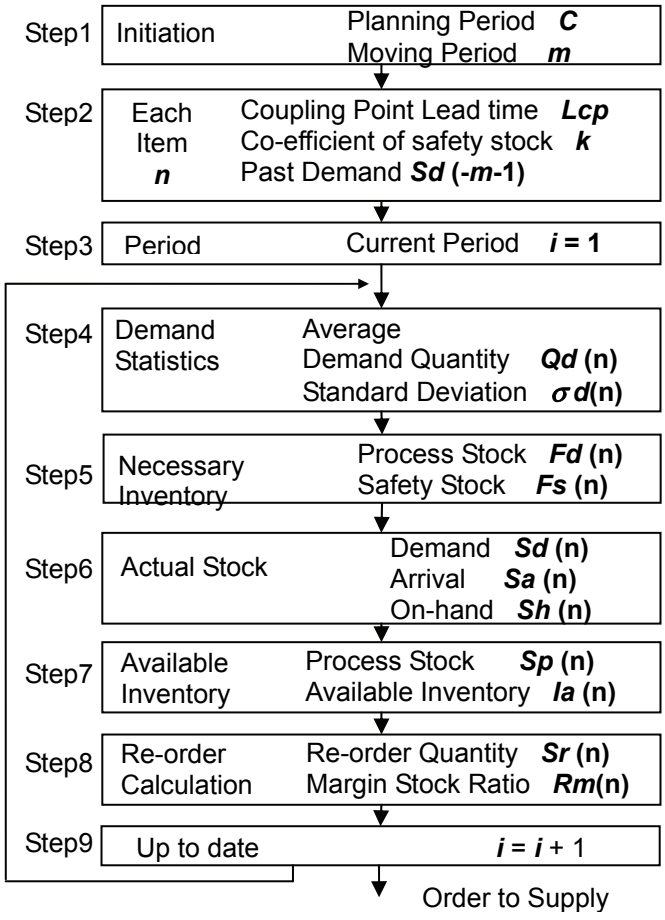


Figure 11. Flow of Re-order Quantity calculation

5.5 Example of inventory level control flatly

The effect of the inventory level control is shown Figure 12. The model process is shown Figure 8. The country of the sales company is U.K., the country of the factory is China. The average demand quantity Qd is 1702, and the standard deviation ◎d is 894.
At first, they have used the demand forecast. However, they were worried the excess inventory. Then we proposed re-order quantity calculation. We have been designed pilot items and have been tried the pilot running.
As a result, the re-order quantity is able to calculate without demand forecast, and the total inventory of supply chain process becomes flat by theoretical necessary inventory. We were also sure to decrease the total inventory.

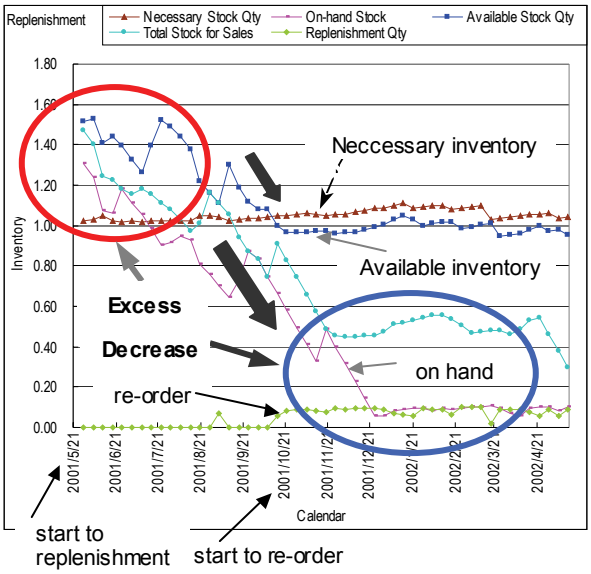


Figure 12. Example of inventory level control

6. Method of multi-items replenishment by Margin stock ratio

6.1 Definition of Margin Stock Ratio

Figure 13 shows the margin stock ratio means the estimation of the out-of-stock situation occurrence. The margin stock ratio is able to calculate by the theoretical necessary inventory except safety stock and measured actual available inventory.

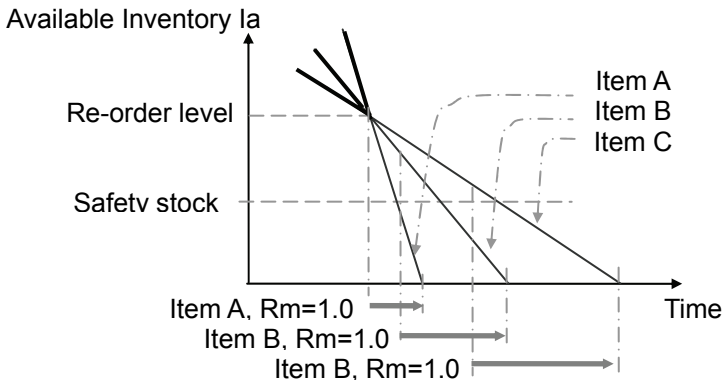


Figure 13. Definition of margin stock ratio

The margin stock ratio $Rm(n,i)$ is defined by the theoretical inventory of reorder interval period $Fd(n,i)$ and available inventory $Ia(n,i)$, denoted by equation (13).

When the available inventory is empty, the margin stock ratio Rm denotes zero, and this value means to occur the out-of-stock situation. When the available inventory is holding a lot, the margin stock ratio is bigger than one. While the margin stock ratio increases, the out-of-stock situation decreases. Thus, we are determined to assign the production capacity depending on margin stock ratio.

6.2 Determination of Replenishment Items

We are able to recognize two cases of the limited process capacity E. In the first case, the total replenishment quantity exceeds the limited process capacity E. In this case, we should determine the priority of the replenishment items. In the second case, the total replenishment quantity is below the limited process capacity E. In this case; we consider the remaining capacity for the precedence of the replenishment items.

Under the limited capacity, the choices of reorder items among multi-items is determined by the prediction of the out-of-stock situation called margin stock ratio $Rm(n,i)$.

Figure 14 shows, the replenishment items j is arranged by the margin stock ratio $Rm(n,i)$ in an ascending order.

The replenishment quantity is summarized by each item while the total replenishment quantity is equal or less than the limited capacity. It is denoted by

equation (14). Thus, we get the replenishment item's number x^*, and determine the replenishment items $j = 1$ to $x^* + 1$.

Furthermore, although the last item x^*+1 exceeds the limited quantity, it is carried over to the next planning.

$$x^* = \left\{ \max_{1 \le x \le m} x \;\middle|\; \sum_{j=1}^{x} Sr(j) < E \right\} \tag{14}$$

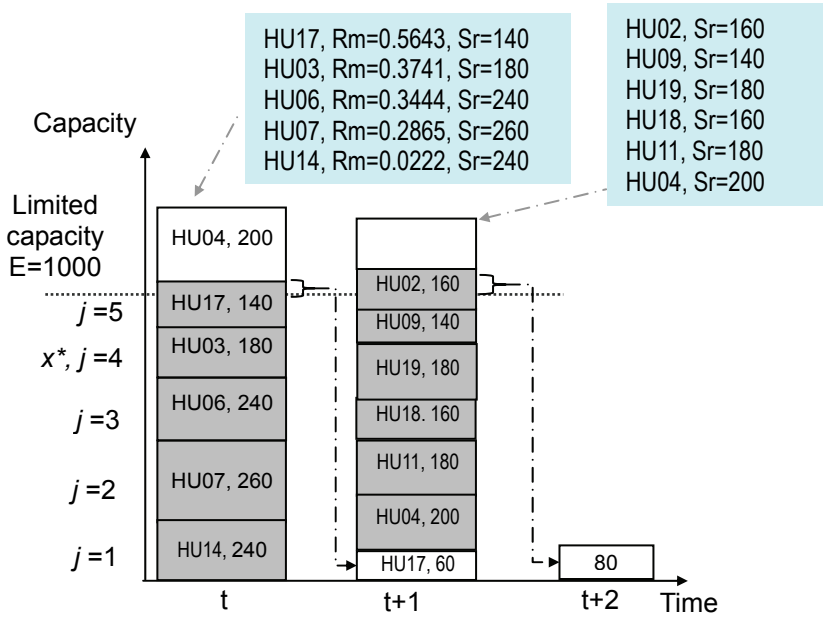


Figure 14. Determination of Replenishment Items

6.3 Calculation example of margin stock ratio

The determination of replenishment items by margin stock ratio are calculated by three steps.

Step 1 is the calculation of the margin stock ratio $Rm(j)$ with each item by equation (13) and makes its list. Table 1 shows the calculation example of the margin stock ratio $Rm(j)$.

Step 2 is the arrangement of the margin stock ratio $Rm(j)$ in an ascending order. Table 2 shows the example of the arrangement.

Step 3 is the determination of the replenishment items under the limited capacity E by equation (14). The replenishment items j are determined from 1 to

x^*+1. For example, in Table 2, when the limited capacity E is 1000 unit / day, x^* is 4, then the replenishment items are HU14, HU07, HU06, HU03 and HU17. The last item is x^*+1.

Furthermore, the exceeded limited capacity (60 units) carries forward to the next planning, and the limited capacity of next planning E is 940. Similarly, the limited capacities E are 1400 and 1800, as shown in Table 2.

Item	Qd	Lcp	C	Fd	Ia	Rm	Sr
HU01	70	1	1	140	124	0.8857	140
HU02	80	1	1	160	140	0.8750	160
HU03	90	2	1	270	101	0.3741	180
HU04	100	2	1	300	173	0.5767	200
HU05	110	2	1	330	439	1.3303	220
HU06	120	2	1	360	124	0.3444	240
HU07	130	3	1	520	149	0.2865	260
HU08	140	3	1	560	764	1.3643	280
HU09	70	1	1	140	121	0.8643	140
HU10	80	1	1	160	153	0.9563	160
HU11	90	2	1	270	171	0.6333	180
HU12	100	2	1	300	350	1.1667	200
HU13	110	2	1	330	456	1.3818	220
HU14	120	2	1	360	8	0.0222	240
HU15	130	3	1	520	539	1.0365	260
HU16	140	3	1	560	617	1.1018	280
HU17	70	1	1	140	79	0.5643	140
HU18	80	1	1	160	105	0.6563	160
HU19	90	2	1	270	211	0.7815	180
HU20	80	1	1	160	221	1.3813	160
Total	2000			6010	5045		4000

Table 1. Calculation Sample of Margin Stock Ratio

Arranged					E=		
Order	Item	Rm	Sr	Σ Sr	1000	1400	1800
1	HU14	0,0222	240	240	*	*	*
2	HU07	0,2865	260	500	*	*	*
3	HU06	0,3444	240	740	*	*	*
4	HU03	0,3741	180	920	*	*	*
5	HU17	0,5643	140	1060	*x*+1	*	*
6	HU04	0,5767	200	1260		*	*
7	HU11	0,6333	180	1440		*x*+1	*
8	HU18	0,6562	160	1600			*
9	HU19	0,7815	180	1780			*
10	HU09	0,8643	140	1920			*x*+1
11	HU02	0,8750	160	2080			
12	HU01	0,8857	140	2220			
13	HU10	0,9562	160	2380			
14	HU15	1,0365	260	2640			
15	HU16	1,1018	280	2920			
16	HU12	1,1667	200	3120			
17	HU05	1,3303	220	3340			
18	HU08	1,3643	280	3620			
19	HU20	1,3813	160	3780			
20	HU13	1,3818	220	4000			

Table 2. Determination Sample of Replenishment Item

6.4 Example of multi-items replenishment under the limited capacity

The effect of the capacity constraints replenishment by margin stock ratio is shown Figure 15. We have used the metal cutting center, the center was organized 4 machine lines, and kind of products is 6. The limited capacity of the center is 5800 (unit/day). In the upside of the figure, total re-order quantity is over the limited capacity. We make an inventory planning by the margin stock ratio every day. As a result, in the downside of the figure, the capacity / load became to flat by the margin stock ratio.

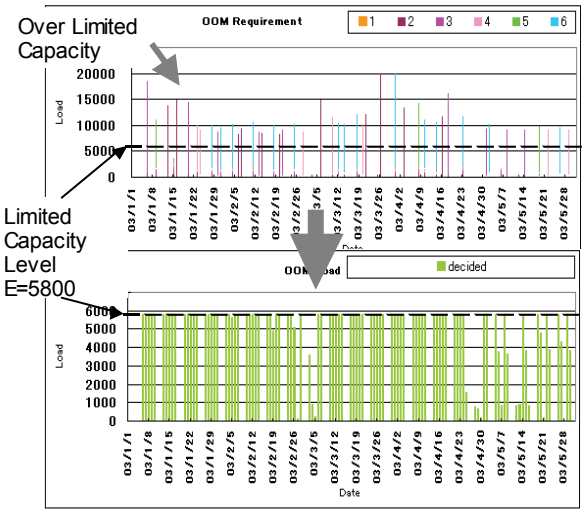


Figure 15. Example of multi-items replenishment under the limited capacity

7. Expected effect

We have been sure the expected effect of Coupling Point Inventory Planning.

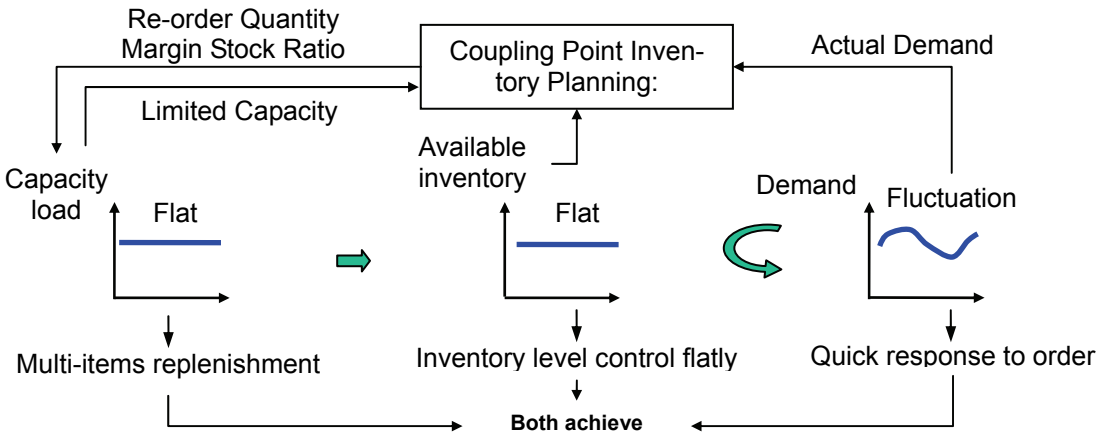


Figure 16. Expected effect of proposed method.

Figure 16 shows this method contributes to achieve three-way optimums;
(1) Quick response to order by CP establishment, as in Figure 8,
(2) To control the available inventory level flatly by the re-order calculation, as in Figure 12,

(3) The multi-items replenishment under the limited capacity by the margin stock ratio as in Figure 15.

8. Application of global SCM

We have some applications of global SCM shows Table 3. In this paper, we introduce to apply *J* Company.

J Company is the manufacture of electronics stationary to build the global SCM. The parts are made in Japan, the products are assembled in China, and the products are sold in the U.S.A, France, U.K, and Germany. Figure 17 shows they should consider the shipping logistics. We thought on boat means to equal the moving warehouse. Then we have been decided to apply Coupling Point inventory theory in 2001.

At first, we have been designed to apply on 100 main items in the number of all 1000 sales items, and have been chosen to pilot 2 devices and 3 accessories. The introduced method has been used 1 year on the pilot running. The inventory and the Shortage-ratio levels were the expectation, and the inventory became to decrease about 3 million US dollars in the global operation of 5 items. The introduced method was successfully applied.

Then, after the pilot, they have been spread to apply 40 items, and currently, they are sure to decrease inventory about 18 million US dollars in the global operation.

from 1993 to 2004

Corp	Kind of products	Effect
A	Window frame	Inventory reduce 25%, delivery 5 days
B	Computer Storage	Inventory reduce 25%, delivery 2 weeks
C	Electronics parts	Inventory reduce 85%, delivery 7 days
D	Communication device	cash reduce 28Mus$, delivery 7 days
E	Car navigator	Inventory reduce 25%, delivery 3 days
F	Personal computer	ROA up to 3.2, delivery 3 days
G	Semi conductor	cash reduce 28M.us$, delivery 7 days
H	Home electric	cash reduce 95M.us$, delivery 3 days
I	Window frame sales	cash reduce 57M.us$, delivery 7 days
J	Electronics stationary	cash reduce 18M.us$, delivery 2 days
K	Foods can	cash reduce 47M.us$, delivery 7 days
L	Chemical 5 divisions	Inventory reduce 15~25%, delivery 7~15 days
M	Chemical plastics	Loss reduce 5%, delivery 7 days
N	Motor car parts	Inventory reduce 25%, 1 day delivery

Table 3. Applications of global SCM

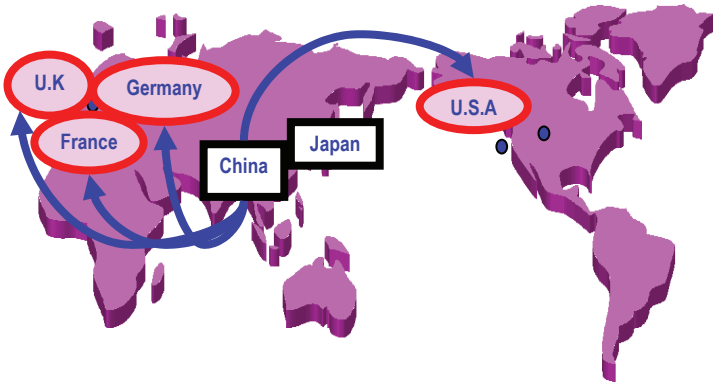


Figure 17. Application of global SCM

9. Conclusions

In this paper, we introduced a new supply chain solution based on Coupling Point Inventory Planning. The introduced method has been developed by Hitachi, Ltd., in 1993. This method has been achieved three-way optimums; quick response to order, the available inventory level flatly without demand forecast, and the multi-items replenishment under the limited capacity.
The introduced method was successfully applied to decrease inventory in some cases. We are also successfully applied to replenish without demand forecast and bull whip.

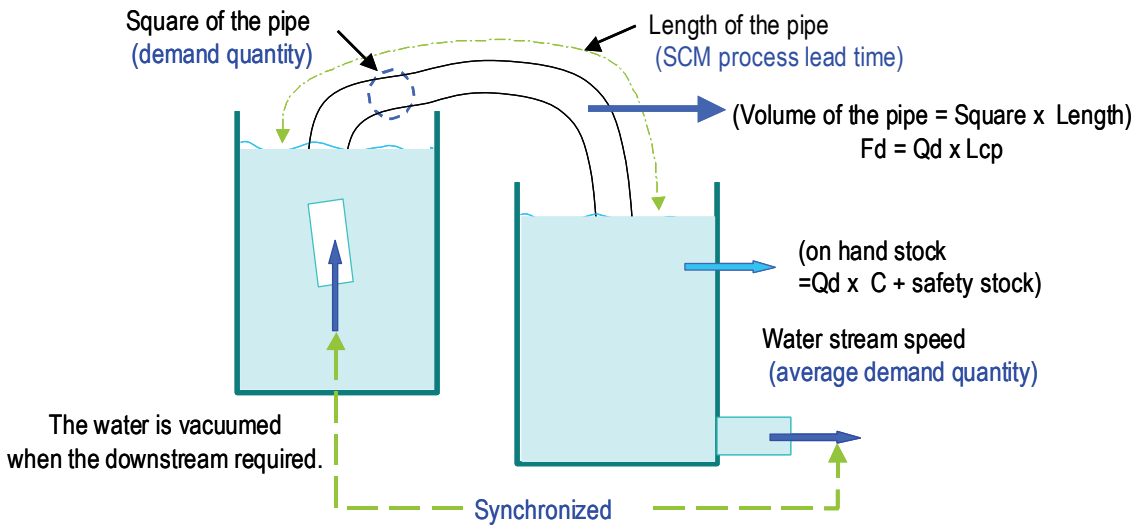


Figure 18. Understanding by siphon (analogy)

We also apply to combine the production system based on order and demand forecast in case of non-repetitive products, to experience the various situations of demand and supply, and to spread in the global SCM.
Lastly, we are able to easily understand the Coupling Point Inventory Planning theory. Figure 18 shows that the siphon is an analogy of SCM based on Coupling Point Inventory Planning. The water flows from upstream to downstream by using siphon without the pump. The volume of the pipe is able to calculate by using the length and square of the pipe. The water at the upstream is vacuumed depending on synchronized demand when the water of the downstream is used. The water stream has a lead time between from the upstream to the downstream. However, the siphon is not using of the future demand forecast after the laps of lead time.
Furthermore, we should challenge much more to use for Global Just-In-Time by Coupling Point inventory theory.

10. References

Lagodimos, A. G.; Andeson, E. J.,"*Optimal positioning of safety stocks in MRP*", Int. J. of Production Research 1993, Vol. 31, No. 8, 1797-1813.

Shacham, J. "*The Third Generation of Production Management Technology*", in Autofact of the 1993 Society of Manufacturing Engineers Conference, MS93-250.

Kimura, O.; Terada H.,"*Design and Analysis of Pull System, a method of multi-stage Production Control*", Int. J. of Production Research, Vol. 19, No. 3, pp. 241-253, 1981.

Huang, C. C. and. Kusiak, A., "*Overview of Kanban systems*", Int. J. of Computer Integrated Manufacturing, Vol. 9, No. 3, pp. 169-189, 1996.

Mitsukuni, K.; Komiya, F. Sugiyama, K. Y.Tomita, H.Maki, and N.Komoda,: "Coupling Point Production Control System for Quick Response to Orders and Minimum Inventories," in Proc. of 6th IEEE Int. Conf. on Emerging Technologies and Factory Automation, pp.154-159 (1997).

Mitsukuni, K., Tsushima, I. and Komoda, N.: "*Evaluation of Optimal Ordering Method for Coupling Point Production System*," in Proc. of 7th IEEE Int. Conf. on Emerging Technologies and Factory Automation, pp.1469-1474 (1999).

Mitsukuni, K. Koyama, M. Nakamura Y.: "*New Supply Chain Concept Based on Coupling Point Inventory Planning*," in Proc. of IEEE Int. Symposium on Industrial Electronics, pp.1358-1363 (2002).

Mitsukuni, K. Nakamura, Y. and Aoki T.: "*New Supply Chain Planning Method Based on Coupling Point Inventory Planning*," in Proc. of IEEE Int. ETFA2003, Vol. 2 pp. 13 - 18 (2003).

24

Relative Control and Management Philosophy

Che-Wei Chang

1. Introduction

Silicon wafers for the semiconductor industry are extremely complex materials with characteristics such as high purity levels, crystallographic perfection, precise mechanical tolerances, complicating efforts to effectively monitor process stability and control quality for individual product types. Material of silicon wafer can be doped with more than 12 kinds of dopants, such as B, C, N, Al, Si, Sb and others. Currently, the sizes of the firm's products are 4-, 5-, 6-, 8- and 12-inch. Considering dopants and sizes, and each kind of product has different attributes according to which, 7~12 minutes are required to slice a piece of wafer. About 2 minutes are required to inspect the quality of a piece of wafer. A wafer can be easily broken during inspection, because of its thinness and brittleness (Lin et al., 2002). Moreover, slicing is a kind of cutting technique that has difficulty in yielding the required precision. Three scenarios will incur damage on the work piece: (1) inaccurately estimating the precision of the slicing machine, (2) engineers set parameters and change the type of material and (3) inconsistently controlling the wafer quality by applying the sampling method owing to the small batch size of wafer slices in the production model.
Consequently, given unstable yields of synchronously multiple quality characteristics are unstable or drifting accuracy of wire saw machines, inspectors must consider employing machine control and monitoring measures. Five synchronously occurring precision quality characteristics, namely thickness (THK), bow, warp, total thickness variation (TTV), center thickness (CTRTHK) and total indicator reading (TIR) must be simultaneously inspected using automatic testing equipment (ASTM F534, 1995; ASTM F657, 1995; Takeshi, 1998). Those multiple quality characteristics destabilize the slicing. The case firm used quantitative methods, such as process capability indices (PCIs) and statistical process control (SPC) charts, are severely limited in monitoring slicing problems (Lin et al., 2002).

This chapter proposes relative control and management philosophy that in-

volving three stages to explore slicing problems and enhance slicing quality and process. The first stage, applies focus groups procedure that can explore an engineer's knowledge and expertise. Organizations can effectively use focus groups to create knowledge of stable processes, optimal settings and quality control. Interactive discussions indicate that the focus groups can enhance productivity and effectiveness of decision either by accelerating the decision process or by elevating the quality of the resulting decisions. Moreover, the proposed procedure allows an engineer to rapidly adjust a manufacturing system to eliminate related phenomenon and enhance slicing quality and process capability (Lin et al., 2004). The second stage, applies grey situation decision-making (GSDM) is used to screen the worst quality characteristic from the synchronously occurred multiple quality characteristics to monitor the process. Then the exponential weighted moving average (EWMA) control chart is presented to demonstrate and verify the feasibility and effectiveness of proposed discussions. The third stage, applies the Chinese philosophy of yin and yang to illustrate wafer slicing quality, and provides decision makers with philosophical thoughts for balancing the simultaneous consideration of various factors (Lin et al., 2005). Furthermore, to increase process yield and accurately forecast next wafer slice quality, grey forecasting is applied to constantly and closely monitor slicing machine drift and quality control.

2. Methodology

2.1 Focus Groups

Focus groups are discussion groups brought together to share perceptions on a defined area of interest to generate knowledge and hypotheses, opinions and attitudes to evaluate commercial ventures, ideas, or the assessment of needs is indispensable. Typically eight to twelve participants are conducted by a skilled moderator who introduces the topic and encourages the group to discuss the topic among themselves. Participants are experts on the topic, since the topic is what they think, feel, or do. A discussion guide directs the discussion through topics in an expected order. The moderator guides conversation gently through each topic until that part of the discussion has unproductive, and may return to later if reemerges in a different context. While allowing the moderator to probe and clarify implied or unclear meanings, this flexibility also allows participants to raise important issues and nuances, which research-

ers often do not foresee. Focus groups rely on the dynamics of group interaction to reveal participants' similarities and differences of opinion (Krueger and Casey, 2000; Morgan, 1997). Participants of relatively homogeneous focus groups have the opportunity to stimulate, support and build on each other's ideas on the topic. Consequently, focus groups reduce the changes of making errors in creating survey questions and, hence improve validity.

Group interaction, spontaneity and sincerity, peer support, descriptive depth, and the opportunity for unanticipated issues to arise - can effectively enable focus groups to create relevance to stable the slicing process, optimal settings and raising the slicing yield. Furthermore, this relatively non-threatening group setting is a cost-effective and efficient means of learning about and elucidating different processes unstable problems by confronting and overcoming difficulties in communication. Focus groups are used in this study to provide some insight into what experiential engineers and their professional knowledge to find slicing problems easier, particularly in terms of information and advice, and the reasons why.

2.2 Grey Situation Decision-Making and Its Algorithm

Grey situation decision-making (GSDM) provides a procedure to deal with one event that involves multiple situations in the same event and choose the best or the worst situation what they occur. The definition and algorithm of the method are as follows (Deng, 2003; Lin, et al., 2002).

Definition 1. Let $a_i,\ i=1,\ 2,\ ...,\ m$ be the sample screening events and $b_j,\ j=1,\ 2,\ ...,\ n$ be the countermeasures of the multiple quality characteristics in the process. Then, a_i and b_j, are referred to as a combined event, S_{ij}, also called a "Situation" and represented as

$$S_{ij}=\left(a_i,\ b_j\right) \tag{1}$$

Definition 2. Evaluating a criterion for the effectiveness of multiple quality characteristics is called "Target".

Definition 3. If $S_{ij}=\left(a_i,\ b_j\right)$ is a situation, then let p represent the number of

target. Using the countermeasure, b_j, which relates to the sample screening event, a_i, the effectiveness of a_i and b_j, is written as, E_{ij}^p. Let M be a mapping, $M\left(E_{ij}^p\right) = R_{ij}^p$, where R_{ij}^p is the value of the mapping between E_{ij}^p and E_{ij}^p, and is an element of E^p. Let X^+ be positive space. If M satisfied, (1) $M\left(E_{ij}^p\right) = R_{ij}^p \in R^p$ and $R_{ij}^p \in [0, 1]$ and (2) $R_{ij}^p \in X^+$, then M can be called the mapping effectiveness measurement. The properties of M are as follows.

(1) The upper-bound effective measuring target of M is "higher-the-better." That is

$$R_{ij}^p = \frac{E_{ij}^p}{\max\limits_i E_{ij}^p} \tag{2}$$

(2) The lower-bound effective measuring target of M is "lower-the-better." That is

$$R_{ij}^p = \frac{\min\limits_i E_{ij}^p}{E_{ij}^p} \tag{3}$$

(3) The moderate effective measuring target of M is "target-is-the-best." That is

$$R_{ij}^p = \frac{\min\limits_i \left\{E_{ij}^p, E_0^p\right\}}{\max\limits_i \left\{E_{ij}^p, E_0^p\right\}} \tag{4}$$

Where

$E_0^p = \frac{1}{n}\sum_{i=1}^{n} E_{ij}^p$;

i is the index of sample, and j is the index of the countermeasure for quality characteristics.

Definition 4. Let the situation, S_{ij} S, have a measuring target for n quality characteristics. If the mapping of E_{ij}^{p} is $M\left(E_{ij}^{p}\right)=R_{ij}^{p}$, then $R_{ij}^{1},\ R_{ij}^{2},\ ...,\ R_{ij}^{n}$ exist; therefore, the synthetic effective measuring of R_{ij}^{p} for one of the quality characteristics is,

$$R_{ij}^{\Sigma}=\frac{1}{n}\sum_{p=1}^{n}R_{ij}^{p} \tag{5}$$

Consider n countermeasures, $b_1,\ b_2,\ ...,\ b_n$, to deal with for the sample screening event, a_i. Associated mapping synthetic effective measuring vectors, R_i^{Σ}, exist and can be expressed as,

$$R_i^{\Sigma}=\left[R_{i1}^{\Sigma},\ R_{i1}^{\Sigma},\ ...,\ R_{in}^{\Sigma}\right] \tag{6}$$

Definition 5. Let R_i^{Σ} be the synthetic effective measuring vector of a_i, expressed as, $R_i^{\Sigma}=\left[R_{i1}^{\Sigma},\ R_{i2}^{\Sigma},\ ...,\ R_{in}^{\Sigma}\right]$. If $R_{ij}^{\Sigma^*}$ satisfies the following condition,

$${}_kR_{ij}^{\Sigma^*}=\max_j\ {}_kR_{ij}^{\Sigma},\quad j\in J=\{1,\ 2,\ ...,\ n\} \tag{7}$$

then $S_{ij}^{*}=\left(a_i,\ b_j^{*}\right)$ is "satisfied situations"; b_j^{*} is the satisfied countermeasure of the quality characteristic of sample screening event, a_i, and $R_{ij}^{\Sigma^*}$ is the best situation of the satisfied situation.

2.3 Chinese Philosophy – Relative Management and Control

Einstein (1920) and Laozi state that the world contains no absolutes. Laozi is one of the most influential philosophers during the past 2500 years of Chinese civilization, and in the US the New York Times once chose Laozi as one of the greatest authors of all time (Laozi and Roberts, 2004). Laozi's book, the Dao De Jing, which describes around 5000 Chinese characteristics, described all things

as originating from the "Way," which is present within all things in the universe. Laozi saw all things as relative. Einstein is one of the most influential physicists in the 20th Century's greatest minds. In 1929, TIME noted in a cover story that "Albert Einstein's theories have altered human existence not at all." In the relativity propounded by Einstein, everything is relative. Specifically, speed, mass, space and time are all subjective. Nor are age, motion or the movements of the planets capable of being objectively measured rather they are judged according to the whim of the observer.

Laozi's book (Laozi and Roberts, 2004), the Dao De Jing, based on the idea that the world contains no absolutes. Laozi saw all things as relative. Notably, management issues are also relative rather than absolute. Figure 1 displays a yin and yang symbol that has been modified to apply to main factors and noise factors. This chapter applies the concept of yin and yang to quality management. The main blocks of color on the yin and yang symbols represent the major effects of decision factors influencing slicing quality. Meanwhile, the small circles of opposite color represent the noise factors affecting decision factors.

The curve symbolizes the constant change in the balance between yin and yang. The above demonstrates the belief that there are no absolutes: nothing is ever entirely yin or yang, but rather a balance always exists between these two forces, just as with time cycles; that is, the process moves in a never-ending cycle characterized by "departing, becoming distant and returning." Consequently, the law of yin and yang involves a balance between yin and yang and an integration of the positive and the negative, light and dark, hot and cold drives all change in the world and provides the life force in the universe. From a directional perspective, Laozi's philosophical thought is focused on balance and continual change. Yin and yang are dependent opposites that must maintain a constant balance.

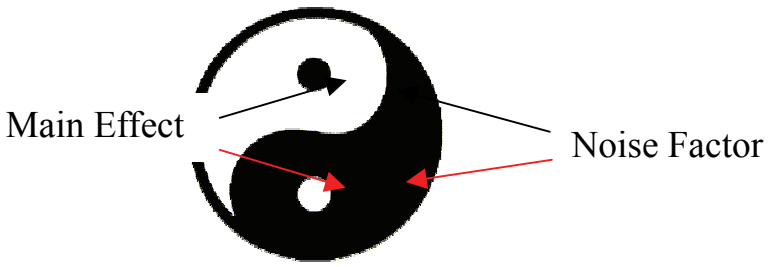


Figure 1. Main Effect Factor and Noise Factor

Decision makers are constantly eager to reselect quality characteristics for accurately monitoring process quality without incurring additional costs. The question then arises of how to make appropriate decisions. Generally, based on cost, operability, productivity and benefit considerations, most decision makers ignore uncontrolled noise factors. However, identifying which control factor is the main effect factor and which are noise factors has always confused decision makers seeking to make appropriate decisions. Thus, this chapter applies the Chinese philosophy of yin and yang, to illustrate relative management issues, and provide decision makers with philosophical thoughts for balancing the simultaneous consideration of various factors.

3. Case Implementation

The case firm divides departments according to by class of clean room. Process one is growing crystals. Moreover, process two includes slicing, edge grinding, lapping, backside treatment and etching. Process three includes wafer polishing, cleaning, packing and inspection. This chapter focuses on slicing, and thus more participants come from the process two departments than from other departments. In relative control and management philosophy procedure, three stages and steps of process are suggested. Figure 2 illustrates implement procedure of relative control and management philosophy.

3.1 Focus Groups Processes

Six steps are proposed for the focus groups processes. Figure 2 of stage 1 illustrates the focus groups implementation procedure.

Step 1.1: Specify the research and define the purpose

Focus groups require a clear, specific statement of purpose to develop appropriate questions and obtain the useful information. Statements of purpose that are broad and general create difficulty in identifying potential participants, developing questions, and obtaining useful results.

Step 1.2: Determine focus group participants
Generally, focus group participants must meet two criteria:

1. they must have the process experience required by the research goals, and
2. they must be able to communicate this experience to the group.

This study required all participants to have process management experience and process quality control experience. The practical need to provide adequate communication among participants, the moderator, and project consultants made this requirement crucial. Four focus groups of 15 persons were conducted to identify the factors that influenced silicon quality and the process capability.

Focus group 1:
administrative department, four managers participating that include the general manager, department one manager, department two manager and department three manager.

Focus group 2:
engineer group, four persons participating that include department two of the section chief and three engineers.

Focus group 3:
quality control department, with the manager and section chief participating.

Focus group 4:
consultant group, involving a process consultant, whose main responsibility was to solve process problems at the firm; and four project consultants participated, whose main responsibilities were to increase yield and process capability.

Step 1.3: Decide the moderator

The moderator can control the influences on the success of the focus group. Therefore, the main role of the moderator is to facilitate open, uninhibited dialogue. Thus, the moderator should play several roles that depend on sensitive balancing and an objective and detached perspective. Especially, a focus group moderator should deal tactfully with outspoken group members, maintain the focus of the discussions, and ensure that every participant gets the opportunity to contribute.

Step 1.4: Conduct the focus group and determine the method of data collection

The participants and moderator sat around a circle table. The discussions were recorded on tape and an assistant moderator took notes. The participants were asked to speak one at a time to ensure that all comments could be clearly heard on the tape. According, focus groups create knowledge of stable processes, optimal settings and quality control, all of which influence quality and process capability factors related to slicing problems. Auxiliary data sources include ISO documents, computer records and observations of behavior designed to help focus groups to make precise decisions.

Step 1.5: Discuss the topic of the slicing problems in the focus group

Slicing is an increasingly complex process. Effectively monitoring the individual product process stability and quality is difficult. However, when the yield of high quality wafers is unstable, or when the wire saw machine drifts, the inspector must carefully control and monitor the machine. Slicing is a kind of sawing that cannot easily yield the knife drift required of a wire knife. The work-piece can be damaged in three ways.

(1) Frequency changes may adjust the precision of the slicing machine.

(2) Controlling whole process quality by sampling is difficult, since production is in small batches. The crystal grown using the raw wafer material, such as a silicon ingot can be sliced into 250~300 pieces. Standard sampling requires only one or two wafers to be sampled to monitor and control slicing. Such a small number of samples cannot provide sufficient information to determine process quality.

(3) Engineers set parameters and change the type of material, thus destabilizing the slicing.

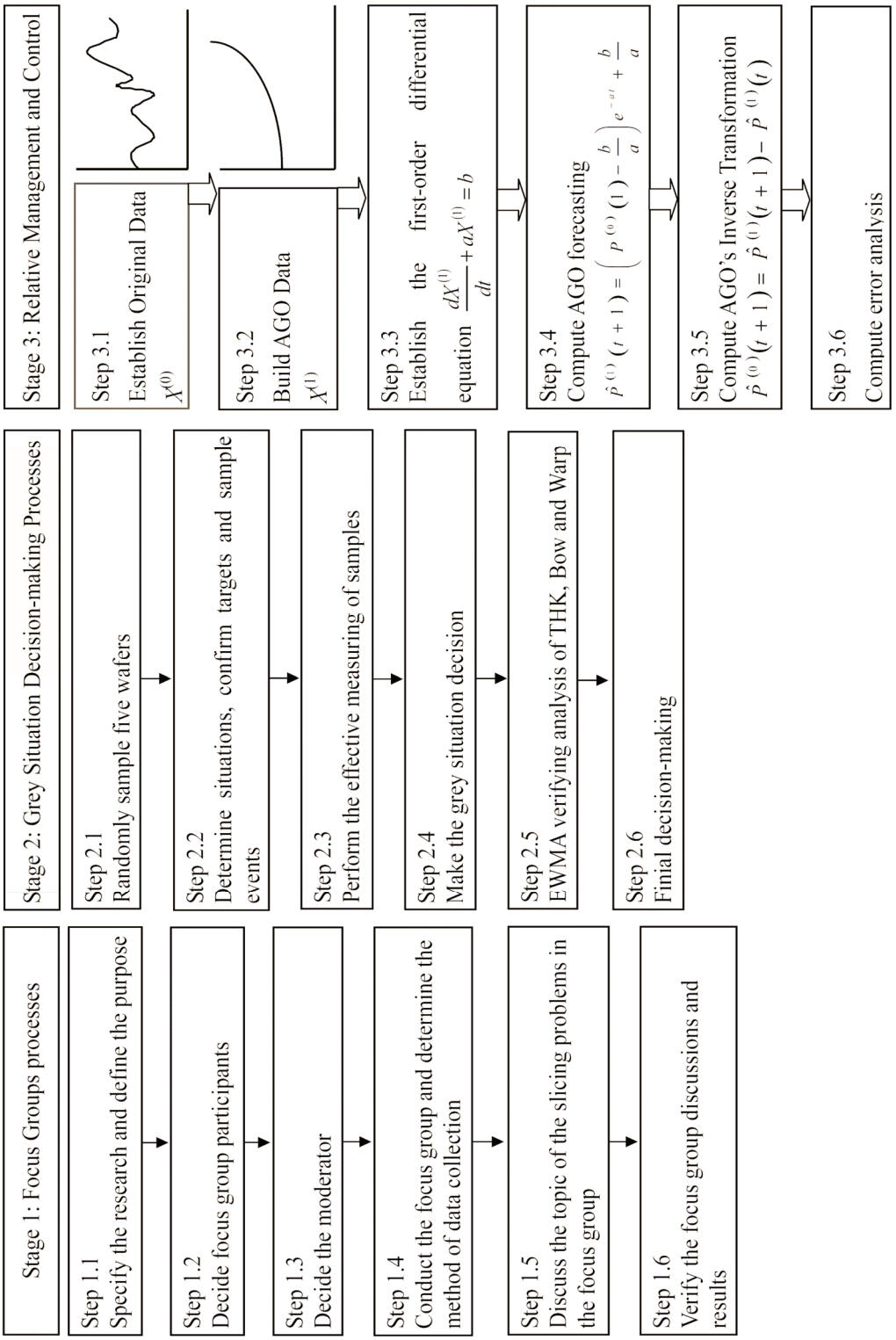


Figure 2. Relative Management and Control Implementation Procedure

Two topics are identified based on the above, including wire knife drift, abnormal work-piece impact yield following slicing, and the influence of parameters on PCIs:

(1) Adjusting the wire knife drift influences sliced wafer quality
This topic relates to the adjusting wire knife time and engineering procedures. Managing knowledge concerning the slicing knife can increase yield, quality and process capability.
(2) Influence of parameters on process capability
This topic concerns slicing machine precision and wire knife drift. The focus groups explore how parameter settings influence slicing stability.

Step 1.6: Verify the focus group discussions and results
To understand slicing problems of wire knife, defective yield, and parameters settings, all focus groups discussed how inspecting an entire ingot can require examining 284 wafer slices.

Adjusting wire knife drift influences slicing quality

(A) Analyzing Wire Knife

THK parameter was set to 750±30(μ) and the bow was □15(μ). Figures 3 and 4 illustrate the processing of 284 wafers, including THK and various bow values, and wire knife times. Figure 3 illustrates the thickness variation and the wire knife adjusted time, the average THK is 760.222(μ), the yield is 0.87 and the wire knife is adjusted 44 times. The wife knife was adjusted once, and adjusted time appearance at the 14th, 18th and 29th sliced wafers, and so on. The wire knife was repeatedly adjusted, and adjust time appear at the sliced wafers of 72nd~75th, 126th ~132nd, 141st~143rd and 223rd~228th. Figure 4 reveals that the average bow is 7.549(μ), while the standard deviation is 12.78(μ) and the yield is 0.83. Figure 4 shows that when executing the procedure of adjusted wire knife, the bow becomes extremely unstable and this unstable variation influences the following wafers.

Statistical data from Figure 4 reveals that engineers identified the following problems with adjusting the wire knife: (1) Adjusting the wire knife is appropriate, but the engineers do not perform the adjustment procedure, with a probability of 0.43; or (2) it is not yet time to adjust the wire knife but engineers

perform the adjustment procedure, with a probability of 0.20. The post-adjustment wafer yield is just 0.57, and the inspected machine gradually increases the process capability. Consequently, engineers should separate individual wafers from the preceding and following ones when adjusting the wire knife. Such separation reduces the likelihood of problems occurring in subsequent edge grinding and lapping processes. The focus groups concluded that two general causes exist for unstable slicing. When the machine appears unstable, engineers can stabilize the slicing by adjusting the wire knife. Moreover, if the wire knife undergoes considerable abrasion and engineers have adjusted the wire knife, the slicing becomes extremely unstable.

(B) Concluding the Results

The focus groups concluded that unstable wire knife causes poor bowing, and adjusting the wire knife cannot reduce the defect rate to below that of bow. Therefore, firm should focus on controlling bow quality. Slicing problems can be classified as either machine or engineer-related. These problems can be further subdivided into another two groups, namely machine-related problems and human-related problems.

Machine-related problems arise after long periods of continuous saw operation, and involve reduced machine precision impacting quality and yield. The saw machine must be periodically adjusted. Meanwhile, human-related problems typically relate to the wire knife procedures used, which are crucial to stabilizing slicing.

The wire knife is adjusted based on engineer experience, and no specific rules are followed. Inexperienced engineers are likely to make the THK too thick or too thin, meaning the anticipated results will not be obtained.

Engineer adjustments to the wire knife involve adjusting the pressure, angle and force. Engineers can use a wafer box with a different color to the color box used online to distinguish abnormal wafers and thus facilitate inspection. Hence, firms must standardize their procedures. Timely application of engineer expertise is critical in the slicing process. Every engineer must be educated and be knowledgeable regarding methods of increasing yield.

(2) Parameters that influence process capability

(A) Process Capability Analyze

Cpk is conventionally used to assess process capability and process variation

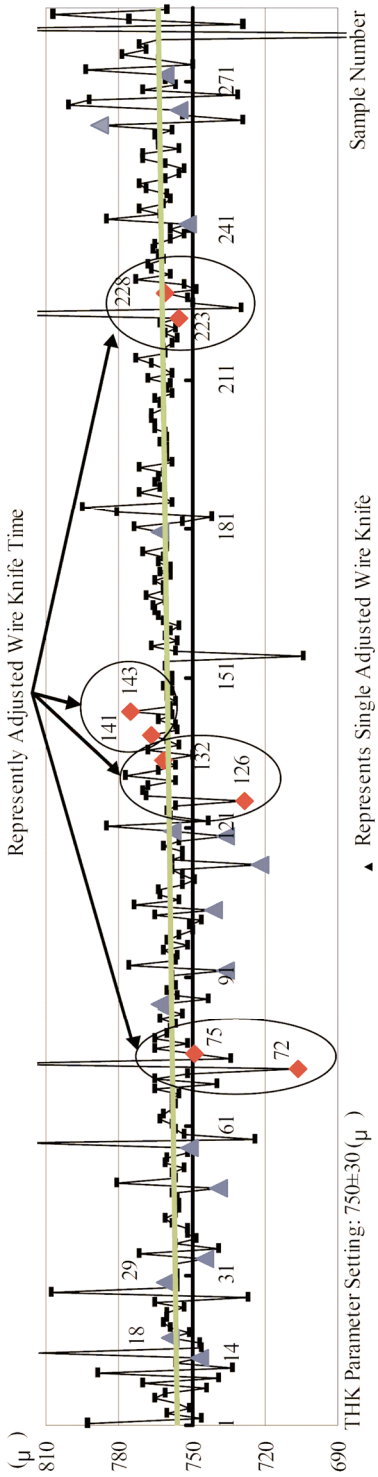


Figure 3. Inspecting an Entire Ingot of THK

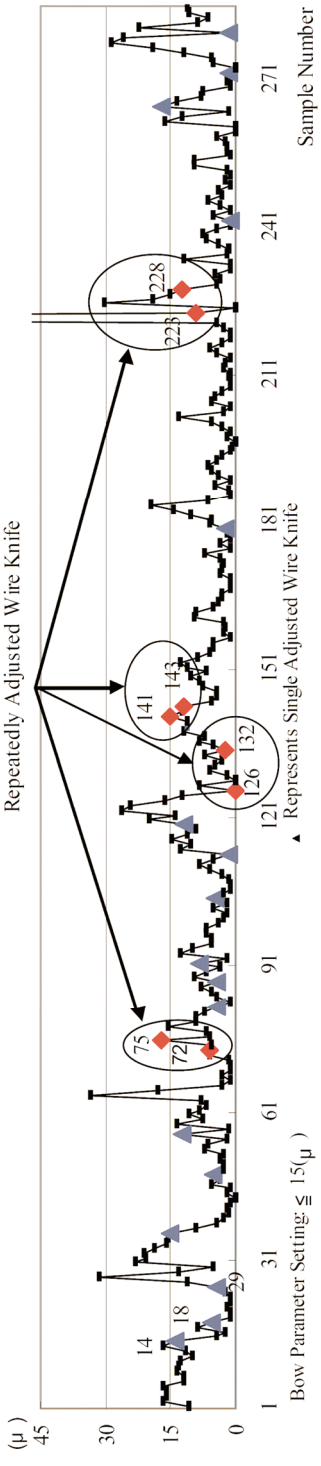


Figure 4. Inspecting an Entire Ingot of Bow

In the semiconductor industry, and is defined as,

$$C_{pk} = \min\left\{C_{pu}(U),\ C_{pl}(L)\right\} \tag{8}$$

where $C_{pu}(U)=\frac{USL-\mu}{3\sigma},\ C_{pl}(L)=\frac{\mu-LSL}{3\sigma}$; USL denotes the upper specification limit and LSL is the lower specification limit; μ represents the process mean, and σ is the standard deviation under constant controlled conditions.

In the subject firm, the parameter settings include THK, bow, knife rotational speed and slicing speed. The Cpk value of THK and bow is set at 1.0. Table 1 lists the THK quality characteristic, using Eq. (8) to calculate the process capabilities Cpu, Cpl and Cpk,. Column 2 illustrates the parameter of THK, 750±30(μ). Cpl = 1.128 is significantly higher than Cpu = 0.555 and Cpk = 0.555, demonstrating that slicing process capability is unstable. However, the average THK for slicing is 760.222(μ); that is the central line is increased by 10.222(μ). Column 3 illustrates the parameter for THK assumption shifting 10(μ) to 760(μ), and the Cpk index is recalculated as 0.738. Notably, CpU and CpL are stable. The right column reveals the defective yield after adjusting the wire knife, and the Cpk index is recalculated as 1.084, consistent with the standard process capability, Cpk = 1.0. Consequently, the groups suggest that engineers check the THK parameter setting and the wire knife vibration. Following any adjustment of the wire knife, engineers should perform wafer slicing before proceeding to the next process. Table 2 presents the bow quality characteristic process capabilities, Cpu, Cpl and Cpk,. The Cpk index for the slicing process capability is always unstable; the Cpk index for the slicing process capability is unstable. This finding relates to slicing quality yield and also influences edge grinding, lapping and polishing. The bow of PCI is lower than that of THK. The bow instability reduces lapping and polishing process yield.

(B) Reaching Conclusions

The focus groups concluded that the causes of machine instability can be classified into three problem types, all machine related:

(a) First, engineers can use the Taguchi method to modify parameter settings to raise process capability.

Parameter Setting / Process Capability	Customer Target 750(μ)	Target Shift 10μ 760(μ)	Take of Defective Wafer
Upper PCI (CpU)	0.555	0.756	1.122
Lower PCI (CpL)	1.128	0.738	1.084
PCI (Cpk)	0.555	0.738	1.084

Table 1. PCIs Analysis THK Quality Characteristic

Parameter Setting / Process Capability	Customer Target □15(μ)	Take of Defective Wafer
Upper PCI (CpU)	0.539	0.606
Lower PCI (CpL)	0.544	0.589
PCI (Cpk)	0.539	0.689

Table 2. PCIs Analysis Bow Quality Characteristic

(b) Second, the defect ratio increases with wafer size. Therefore, 8-inch wafers have a lower yield than 4-, 5-, and 6-inch wafers. Moreover, when slicing an 8-inch ingot, engineers must inspect and tightly control the wire knife for the bow to reduce abnormalities.

(c) Third, the wire knife life cycle severely influences process capability. Generally, a wire knife can slice 1000-1200 wafers. However, wire knives break easily. To prevent knife breakage, engineers can use a model to forecast knife life cycle. Extremely unstable bow value indicates that it is time to replace the old wire knife with a new one.

(C) PCIs Verifying Analysis of 4-, 5-, 6-, and 8-inch Wafers

To verify the effectiveness of the new parameter settings, engineers monitor the slicing of 4-inch, 5-inch, 6-inch and 8-inch ingots. Table 3 lists the PCIs of the THK and bow quality characteristics. The 4-inch ingot has THK parameter setting of 525±15(μ) and bow parameter setting of □10(μ). The ingot can be sliced into 346 pieces. The PCI of the THK quality characteristic is 2.60 and the bow quality characteristic is 1.84. Table 3 confirms that the 5-inch, 6-inch and

8-inch ingots PCIs of the bow are lower than those of THK. Notably, the process capability reduces when slicing large wafers, namely, 8-inch wafers have lower PCIs than 4-, 5-, and 6-inch wafers. Especially the bow PCI = 0.83 for slicing an 8-inch ingot, which is below the customer value (PCIs = 1.0). Consequently, 8-inch ingot slicing should be carefully monitored and the bow tightly controlled to reduce abnormalities.

Wafer Size / Quality Characteristic / PCIs Analysis	4-Inch		5-Inch		6-Inch		8-Inch	
	THK	Bow	THK	Bow	THK	Bow	THK	Bow
Customer Specification (Unit: μ)	525±15 □10		625±30 □100		525±20 □40		853±30 □100	
Sample Size	346	346	334	334	181	181	334	334
Cpk	2.60	1.84	1.64	1.46	1.40	1.01	1.07	0.83

Table 3. PCIs Analysis THK and Bow Process Capability of 4-, 5-, 6-, and 8-inch Wafers

Consequently, decision makers must own enough wisdom to judge, supervise and evaluating trade-offs multiple decision problems. Typically, slicing problems can be divided into machine and engineer-related. A fishbone diagram derived from the preceding is illustrated in Figure 5. These problems can be further subdivided as follows.

(A) Four machine-related problem types.

(1) The saw operates for a long time, and machine precision impacts quality and yield. The saw machine must be periodically adjusted.

(2) Engineers can use quantity method, such as the parameter design of Ta guchi methods to modify parameters setting to increase process capabil ity.

(3) When producing large size of wafer, the risk of defective yield increases. Therefore, 12-inch wafers have a lower yield than 4-, 5-, 6- and 8-inch wa fers. Moreover, when slicing an 8-inch ingot, engineers must inspect and tightly control bow to reduce abnormalities.

(4) The wire knife life cycle severely affects process capability. Generally, a knife can slice 1000-1200 wafers. The wire knife breaks easily. To prevent the knife's break, engineers can use a forecast model to predict the knife life cycle. If the bow value successive extremely unstable, then it is time to change a new wire knife.

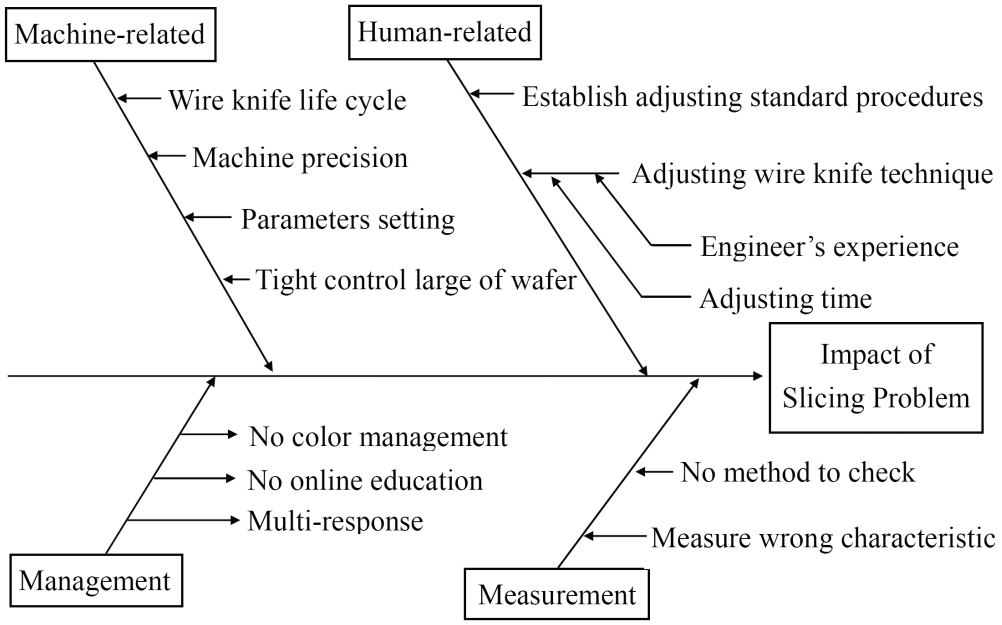


Figure 5. Slicing Problems of Root Causes

(B) Three human-related problem types.

(1) Usually, adjusting the wire knife procedures is the most important technique for stabilizing the slicing. The wire knife is adjusted according to an engineer's experience, and no rules are followed. If engineers have insufficient experiences, they are likely to make the THK too thick or too thin. The anticipated results will not be obtained. When an engineer adjusts the wire knife, he or she must adjust the pressure, angle and force. Engineers can use a different color wafer box from the online used color box to distinguish abnormal wafers for convenient inspection. Taguchi methods can solve problems of multi-response quality characteristics.

(2) The setting up of the wire knife affects the slicing yield. Hence, the firm must establish standard procedures. Timely application of an engineer's expertise is critical in the slicing process.
(3) Every engineer must be educated and have relevant knowledge increase yield.

Above analysis allow engineers to clearly understand how quality and process capability affect silicon wafer slicing. Interactive discussions indicate that the decision-making groups can enhance the productivity and effectiveness of decision-making, either by accelerating the decision-making process or by increasing the quality of the resulting decisions. Moreover, the proposed procedure allows an engineer to adjust rapidly a manufacturing system to eliminate problematic phenomena and increase slicing quality and process capability.

3.2 Grey Situation Decision-Making Processes

Six steps are proposed for the GSDM processes. Figure 2 of stage 2 illustrates the GSDM implementation procedure.

Step 2.1: Randomly sample five wafers

Sample five work-pieces whose samples have been completely confirmed at random, and measure the multiple quality characteristics at five points on each work-piece, using an ADE6300, measuring instrument. Then average the measured data of these points (See Table 4.)

Quality Characteristics	Work-pieces of Sample				
	1	2	3	4	5
THK	789.00	744.00	759.00	753.00	752.00
Warp	26.60	20.80	27.00	22.80	21.90
Bow	10.90	16.80	15.95	16.05	16.70

Table 4. Measured Multiple Quality Characteristics of Wafer (Unit:μ)

Step 2.2: Decide upon the situations, confirm the targets and sample events.

(1) Event: decide the screening samples of the quality characteristics, and can be defined as a_1.

(2) Countermeasure quality characteristic 1, THK (defined as b_1); quality characteristic 2, warp (defined as b_2); quality characteristics 3, bow (defined as b_3).

(3) Situation:

$S_{11} = (a_1,\ b_1)_{=}$ (sample screening of the quality characteristics; countermeasure for quality characteristic 1). In S_{ij}, i is the index of the sample; j is the index of the quality characteristics.

$S_{12} = (a_1,\ b_2)_{=}$ (sample screening of the quality characteristics; countermeasure for quality characteristic 2);

$S_{13} = (a_1,\ b_3)_{=}$ (sample screening of the quality characteristics; countermeasure for characteristic 3);

(4) Target:

Target 1: THK is the target-is-the-best effective measured value, and $E_{11}^1 = 789,\ E_{21}^1 = 744,\ E_{31}^1 = 759,\ E_{41}^1 = 753,\ E_{51}^1 = 752$.

Target 2: Warp is the lower-the-better effective measured value, and $E_{12}^2 = 26.6,\ E_{22}^2 = 20.8,\ E_{32}^2 = 27.0,\ E_{42}^2 = 22.8,\ E_{52}^2 = 21.9$.

Target 3: Bow is the lower-the-better effective measured value, and $E_{13}^3 = 10.9,\ E_{23}^3 = 16.8,\ E_{33}^3 = 15.95,\ E_{43}^3 = 16.05,\ E_{53}^3 = 16.7$.

Step 2.3: Measuring the samples

According Step 2.2, THK, warp and bow are the target-is-the-best, lower-the-better, and lower-the-better quality characteristics, respectively. The dimensionless linear normalization is simplified as,

Target 1: Use Eq. (4) to compute the effective measured value of THK,

$$E_0^1 = \frac{1}{5}(789 + 744 + \ldots + 752) = 759.4,$$

and

$$R_{11}^{1} = \frac{\min\left\{E_{11}^{1}, 759.4\right\}}{\max\left\{E_{11}^{1}, 759.4\right\}} = \frac{759.4}{789} = 0.9625$$

Similarly, $R_{21}^{1} = 0.9797$, $R_{31}^{1} = 0.9995$, $R_{41}^{1} = 0.9916$ and $R_{51}^{1} = 0.9903$.

Target 2: Use Eq. (3) to compute the effective measured value of warp,

$$R_{12}^{2} = \frac{\min\limits_{i} E_{ij}^{2}}{E_{12}^{2}} = \frac{20.8}{26.6} = 0.7820$$

Similarly, $R_{22}^{2} = 1$, $R_{32}^{2} = 0.7704$, $R_{42}^{2} = 0.9123$ and $R_{52}^{2} = 0.9498$.

Target 3: Use Eq. (3) to compute the effective measured value of bow,

$$R_{13}^{3} = \frac{\min\limits_{i} E_{ij}^{3}}{E_{13}^{3}} = \frac{10.9}{10.9} = 1$$

$R_{23}^{3} = 0.6488$, $R_{33}^{3} = 0.6834$, $R_{43}^{3} = 0.6791$ and $R_{53}^{3} = 0.6527$.

Step 2.4: Make the grey situation decision

Eq. (5) yields the synthetic effective measured value as:

$$R_{11}^{\Sigma} = \frac{1}{5}\left(R_{11}^{1} + R_{21}^{1} + \ldots + R_{51}^{1}\right) = \frac{1}{5}\left(0.9625 + 0.9797 + \ldots + 0.9903\right) = 0.9847$$

Similarly, $R_{12}^{\Sigma} = 0.883$ and $R_{13}^{\Sigma} = 0.7328$.

Thus, the worst quality characteristic in the wafer is bow. Bow is therefore monitored. However, the firm currently monitors the THK quality characteristic. The synthesized effective measured values of THK are the highest. Therefore, the process capability of THK is very stable and the manufacturer need not spend much money or time to inspect and monitor this characteristic.

Step 2.5: EWMA (Robert, 1959; Lucas and Saccussi, 1992) verifying analysis of THK and bow

Based on the adjusting the drift of the wire knife impacts the quality of slicing discussions, the exponential weighted moving average (EWMA) control chart is presented to demonstrate and verify the feasibility and effectiveness of proposed discussions.

In this step, an EWMA control chart detects and quickly sets off the alarm in the case of an abnormality quality quickly. Therefore, the method can effectively monitor a little drift in the process. The effective measured value of bow and THK are plotted on an EWMA chart. A univariate EWMA chart is modeled as

$$Z_t = \lambda \bar{X}_t + (1-\lambda) Z_{t-1}, \quad t = 1,\ 2,\ \ldots,\ n \tag{9}$$

Whereλ is the weighting factor (defined by the decision maker) and typical values for λ are between 0.05 and 0.3 in SPC applications; $\bar{X}_t$ is the subgroup average for the current subgroup at time t (or the current observation if the subgroup size is one (n = 1); the value of Z at time zero, Z0, is either a target value or the overall average of the selected subgroups (also defined by the decision maker).

The upper and lower control limits for the EWMA statistics are as follows.

$$UCL = Z_0 + \frac{3\sigma}{\sqrt{n}} \sqrt{\left(\frac{\lambda}{1-\lambda}\right)\left(1-(1-\lambda)^{2i}\right)} \tag{10}$$

and

$$LCL = Z_0 - \frac{3\sigma}{\sqrt{n}} \sqrt{\left(\frac{\lambda}{1-\lambda}\right)\left(1-(1-\lambda)^{2i}\right)} \tag{11}$$

where Z0 is the starting value (defined by the decision maker as either the target value or the process mean value), and n is the size of the subgroup.

The process standard deviation, σ, is estimated using the $\bar{X}$ chart and setting $\lambda = 0.3$ and n = 2 to monitor and inspect bow and THK. In this chart, 124 samples are generated while the process is controlled.

The measured data for bow are obtained from the machine sensors, and the THK is inspected at five points on each work-piece by the measuring instrument, ADE6300 and averaging these points as a point in Figure 6.
In Figure 6, two machine sensors, S1 and S2, monitor the slicing process. S1 is near the slicing plane of the work-piece and monitors bow characteristic. For example, in Figure 7, the bow value is the mean of the max and min values of the drifts, (2+7)/2=4.5. S2 is near the slicing knife and monitors the vibration of the knife. In this paper, S1 is only used for the bow, but S2, which helps to prevent S1 from measuring in a biased way, is not directly related to the Bow.

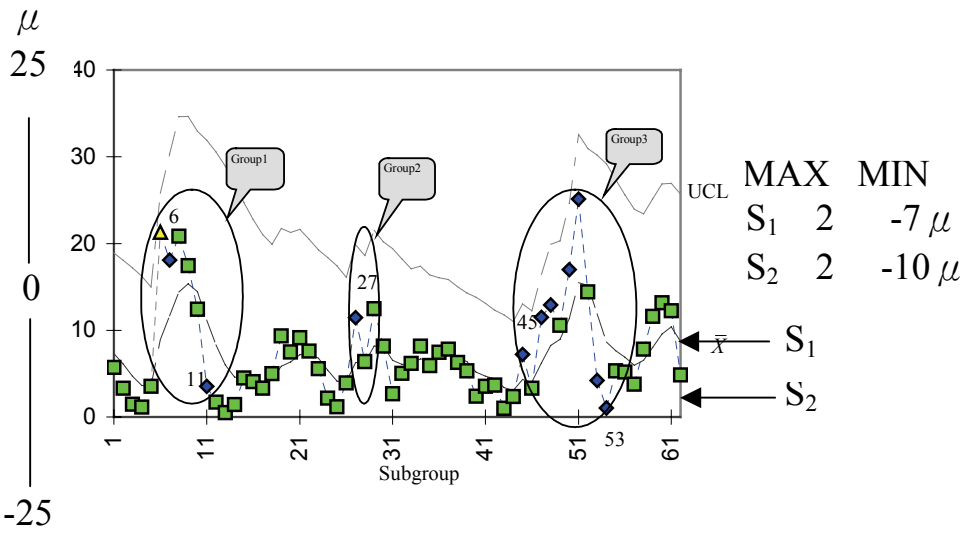


Figure 6. The Machine Sensors of Bow Information

Figure 7 shows an upper bound of the bow's EWMA chart by $\overline{X}$ counts and Figure 8 shows the THK. In Figure 7, the out-of-control conditions appear at the 6th, 27th and 45th signals. In Figure 8, the process is out-of-control at the 26th and 46th signals. The abnormal quality alarms of bow in Groups 2 and 3 are the same as those of THK in Groups 1 and 2. In Figure 7, the out-of-control signals start from the 6th $\overline{X}$ count, but in Figure 8 the $\overline{X}$ count of THK is under control.

Consequently, the quality of bow should be monitored more frequently than the quality of THK. Therefore, the effectiveness of monitoring the worst char-

acteristic, bow, using an EWMA control chart is the same as that of using the GSDM; that is, bow dominates other characteristics.

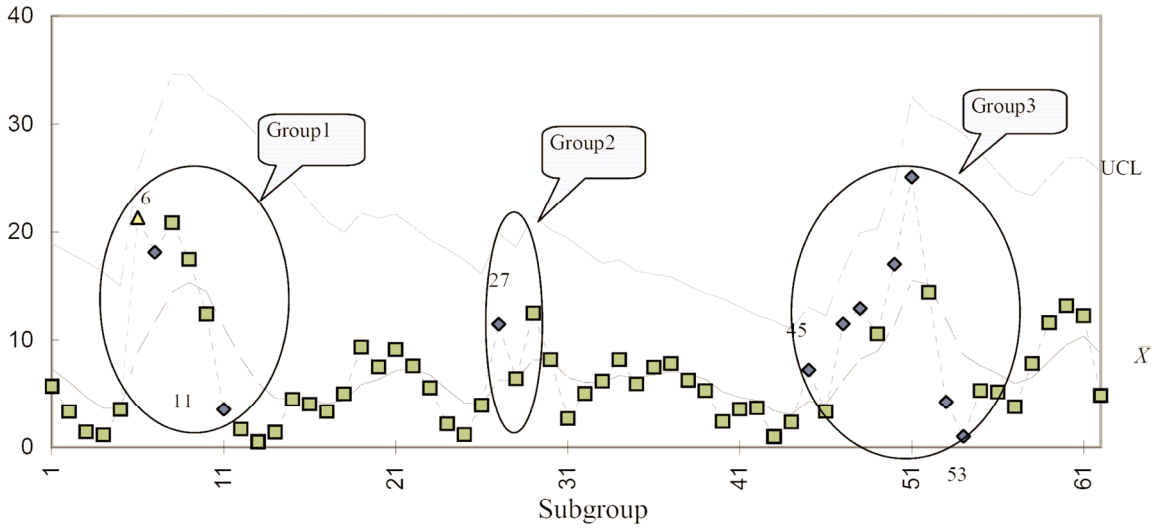


Figure 7. Upper Side of Bow's EWMA Chart by $\overline{X}$ Counts

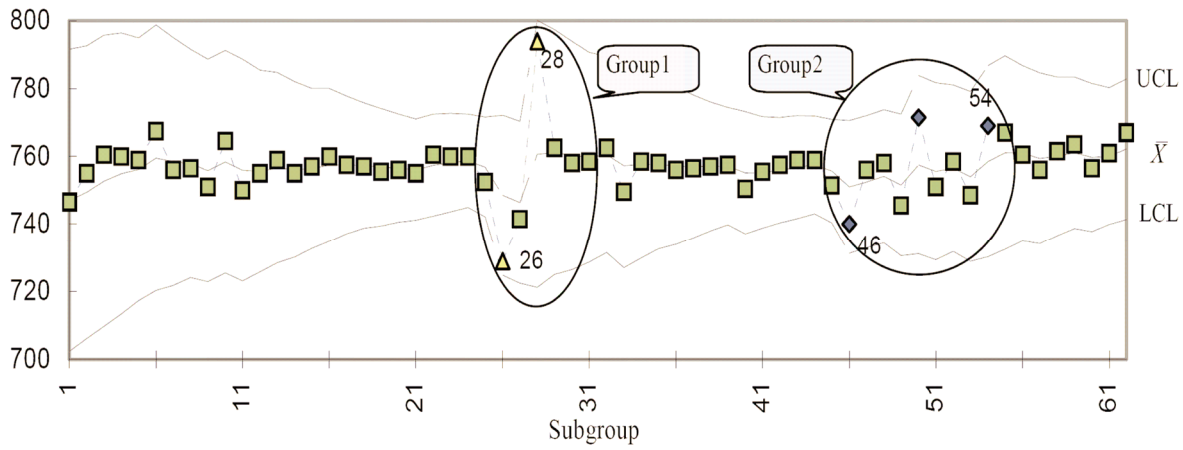


Figure 8. THK's EWMA Chart by $\overline{X}$ Counts

Step 2.6: Final decision-making

(1) By the GSDM, or the EWMA control chart, bow is the worst quality characteristic. Bow is thus unstable in the slicing process.

(2) According to the definition of THK (ASTM F534, 1995; ASTM F657, 1995), the distance between *a* and *b* can be altered by subsequent lapping and polishing. However, bow and warp, cannot be changed. Hence, bow is the critical characteristic, but THK influences the quality of the wafer.

3.3 Relative Management and Control

Simultaneously monitoring and controlling multiple quality characteristics is costly. Above the result of GSDM processes, bow is the most difficult quality characteristic to monitor and control. Bow characteristic can be considered the main effect factor, and the other characteristics, such as THK and warp characteristics can be considered noise factors. Decision makers can apply the slicing machine sensors to costlessly balance the main effect factor and the noise factors. By using the concept of yin and yang variation and collocating grey forecasting, GM (1, 1) model (Deng, 1989 and 2003; Lin and Lin, 2001; Hsieh, et al., 2002; Lin and Yang, 2003), precise drift and quality control from the slicing machine can be constantly monitored and forecast. If the GM (1, 1) model can accurately predict Bow value, then the main effect factor, noise factors and parameter settings are balanced and the slicing process is controlled. Otherwise, the slicing process and parameter settings were not controlled.

Table 5 lists and compares four forecasting models, namely the GM (1, 1) model, regression model, time series and neural network. Deng (1989, 2003) demonstrated that four-data is used to construct the GM (1, 1) model and forecast. In the slicing process, a small batch size production model cannot provide sufficient information to clarify the quality of the entire process. That is, available data are insufficient for designing a predictive model, but Grey forecasting does not require much data. In conventional forecasting methods, decision makers deal with variables by considering numerous samples, and then analyze and compare the relationships among these samples by assuming that populations must obey some identified distribution. However, Grey forecasting can be used with very few data and can use an arbitrary distribution of data to forecast output value. The main feature of “Grey forecasting” is that it can be used with limited data, including just three or four observations, and can make an objective forecast based on an arbitrary data distribution (Deng, 1989, 2003). The Grey forecasting model thus is extremely appropriate for forecasting and controlling the slicing process of the small batch size production model.

The Grey system theory treats all variables as a grey quantity within a certain range. Grey forecasting model then accumulates available data to derive the internal regularity. The model examines the nature of internal regularity in managing the disorganized primitive data. The model was established by transferring the arranged sequence into a differential equation.

	Methodology			
	GM(1, 1)	Regression Model	Time Series	Neural Network
Data	$\geqq 4$	$\geqq 30$ (suggested)	$\geqq 100$ (suggested and as much as better)	as much as better
Model Complexity	Low	Low	High	Low
Self Learning Ability	Yes	No	No	Yes
Model Style	Dynamic	Statistic	Statistic	Dynamic

Table 5. Compare Forecasting Methodologies

The following illustration thoroughly describes the method used to construct the model adopted herein by creating a sequence of one order linear moving GM (1,1). Figure 2 of stage 3 presents the first order differential equation of GM (1,1) model and the algorithms of the method are as follows (Deng, 2003; Lin and Yang, 2003; Lin et al., 2005):

Step 3.1: Establish original data series

$$X^{(0)} = \left(P^{(0)}(1), P^{(0)}(2), \cdots, P^{(0)}(n) \right) \tag{12}$$

The variables, including $P(1), P(2), \cdots,$ and $P(n)$, are used to construct the Grey forecasting model.

Step 3.2: Build accumulated generating operation (AGO) series

When a model is constructed, the Grey system must apply one order accumulated generating operation (AGO) to the primitive sequence in order to provide the middle message of building a model and to weaken the variation ten-

dency. In the slicing process, the production system can be regarded as a closed system. The main effect factor and noise factors simultaneously exist in a wafer during slicing. According to AGO, noise factors can thus be eliminated (Deng, 1999). Herein, $X^{(1)}$ is defined as $X^{(0)}$'s one order AGO sequence. That is,

$$\begin{aligned} X^{(1)} &= \left(P^{(1)}(1),\ P^{(1)}(2),\ \cdots,\ P^{(1)}(n)\right) \\ &= \left(\sum_{t=1}^{1} P^{(0)}(t),\ \sum_{t=1}^{2} P^{(0)}(t),\ \cdots,\ \sum_{t=1}^{n} P^{(0)}(t)\right) \end{aligned} \tag{13}$$

Step 3.3: Establish the first-order differential equation

$$\frac{dX^{(1)}}{dt} + aX^{(1)} = b \tag{14}$$

Where t denotes the independent variables in the system, a represents the developed coefficient, b is the Grey controlled variable, and a and b denote the parameters requiring determination in the model.

Step 3.4: Compute AGO forecasting

The approximate relationship can be obtained as follows by substituting $\hat{a}$ obtained in the differential equation, and solving Eq. (14):

$$\hat{P}^{(1)}(t+1) = \left(P^{(0)}(1) - \frac{b}{a}\right)e^{-at} + \frac{b}{a} \tag{15}$$

where

$$\hat{a} = \begin{bmatrix} a \\ b \end{bmatrix} = \left(B^T B\right)^{-1} B^T Y_N$$

$$B=\begin{bmatrix} -\frac{1}{2}\left[P^{(1)}(1)+P^{(1)}(2)\right] & 1 \\ -\frac{1}{2}\left[P^{(1)}(2)+P^{(1)}(3)\right] & 1 \\ \vdots & \vdots \\ -\frac{1}{2}\left[P^{(1)}(n-1)+P^{(1)}(n)\right] & 1 \end{bmatrix}$$

$$Y_N=\left[P^{(0)}(2),\ P^{(0)}(3),\ \cdots,\ P^{(0)}(n)\right]^T$$

Step 3.5: Compute AGO's inverse transformation

When $\hat{P}^{(1)}(1)=\hat{P}^{(0)}(1)$, the acquired sequence one order Inverse-Accumulated Generating Operation (IAGO) is acquired and the sequence that must be reduced as Eq. (16) can be obtained.

$$\hat{P}^{(0)}(t+1)=\hat{P}^{(1)}(t+1)-\hat{P}^{(1)}(t) \qquad (16)$$

where $\hat{P}^{(1)}(t+1)$ is the predicted value of $\hat{P}^{(1)}(t+1)$ at time $k+1$.
Given $t=1,\ 2,\ \cdots,\ n$, the sequence of reduction is obtained as follows:

$$\hat{X}^{(0)}=\left(\hat{P}^{(0)}(1),\hat{P}^{(0)}(2),\cdots,\hat{P}^{(0)}(n+1)\right)$$

Where $\hat{P}^{(0)}(n+1)$ is the Grey elementary predicting value of $P(n+1)$.

Step 3.6: Compute residual error

After the above model is generated and developed, further tests are necessary to understand the error of forecasted value and actual value. To demonstrate the efficiency of the proposed forecasting model, this paper adopts the residual error test method to compare the actual value and forecasted value. Herein, Eq. (17) and Eq. (18) are used to compute the residual error and the average re-

sidual error of Grey forecasting.

$$\text{Error} = \left| \frac{P(t) - \hat{P}(t)}{P(t)} \right| \qquad j = 1,\ 2,\ \cdots,\ n. \tag{17}$$

$$\text{Average Error} = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{P(t) - \hat{P}(t)}{P(t)} \right| \qquad j = 1,\ 2,\ \cdots,\ n. \tag{18}$$

Currently, the firm applies the bow quality characteristic to monitor slicing quality, and thus this view can be accurately verified. The Bow parameter is set to 0±15(μ) (customer object value), and the GM (1, 1) model is applied to forecast the Bow variation and wire knife. Compared with the online and forecasting values, if the forecasting values consistently exceed twice the 10% residual error, then an alarm in the slicing process system will sound. The alarm information indicates that the slicing system is unbalanced, and slicing quality may be out-of-control at the next wafer to be sliced. Therefore, engineers must adjust the wire knife and check the slicing process parameter settings. According to GM (1,1) Grey forecasting model, this study estimates 40 samples as online verifying analysis and Table 6 lists the actual output values, forecasted output values and residual error.

The verifying and forecasting model comprises three parts. The first part is monitoring and forecasting slicing wafer from No. 1 to 8. When the forecasting values exceed the 10% residual error by two times, the engineer does not stop the slicing machine to check and adjust the parameter settings of the correlative slicing process until the bow is out-of-control. The second part is monitoring and forecasting from No. 9 to 33. When the forecasting values exceed twice the 10% residual error, the engineer stops the slicing machine to check and adjust the correlative slicing process parameters setting. The third part is monitored and forecast slicing wafer from No. 34 to 40. This part does not consider forecasting values until the bow is out-of-control.

This chapter uses part one samples to illustrate GM (1, 1) algorithms and the results are listed in columns 3 to 5 of Table 6. The explanation follows:

From Eq. (12) and Table 6, the primitive sequence $X^{(0)}$ is

$$X^{(0)} = (14, 7, 9, 9)$$

No.	Actual output value	Forecasted output value	Residual error (%)	Average residual error (%)	No.	Actual output value	Forecasted output value	Residual error (%)	Average residual error (%)
1	14	14.000	-		21	13	13.387	0.030	
2	7	7.387	0.055		22	11	10.592	0.037	
3	9	8.290	0.079	0.055	23	5	5.374	0.075	
4	9	9.302	0.034		24	3	2.726	0.091	
5	11	10.438	0.051		25	9	9.160	0.018	
6	10	11.712	0.171		26	9	8.656	0.038	
7	10	13.142	0.314		27	8	8.180	0.022	0.041
8	29	-	-		28	8	7.730	0.034	
9	7	7.000	-		29	7	7.304	0.043	
10	8	7.826	0.022		30	7	6.903	0.014	
11	8	8.320	0.040	0.026	31	9	8.845	0.017	
12	9	8.845	0.017		32	8	8.320	0.040	
13	6	9.403	0.567		33	8	7.826	0.022	
14	5	9.996	0.100		34	6	7.362	0.227	
15	14	10.627	0.241		35	6	6.925	0.154	
16	10	10.000	-		36	15	6.516	0.566	
17	6	6.239	0.040		37	14	-	-	-
18	8	7.551	0.079		38	14	-	-	-
19	9	9.139	0.015		39	18	-	-	-
20	12	11.061	0.078		40	27	-	-	-

Table 6. Verifying Analysis

From Eq. (13), the one order AGO sequence of $X^{(1)}$ is obtained as follows:

$$X^{(1)} = (14, 21, 30, 39)$$

Additionally, matrix B and constant vector Y_N are accumulated as follows:

$$B = \begin{bmatrix} -17.5 & 1 \\ -25.5 & 1 \\ -34.5 & 1 \end{bmatrix} \quad Y_N = \begin{bmatrix} 7 \\ 9 \\ 9 \end{bmatrix}$$

From Eq. (14), $\hat{a}$ is obtained as

$$\hat{a} = \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} -0.115 \\ 5.357 \end{bmatrix}$$

This forecasting model is identified by incorporating *a* and *b* into Eq. (15)

$$\hat{P}^{(1)}(t+1) = \left(P^{(0)}(1) + 46.5\right)e^{(0.115)t} - 46.5 \qquad (19)$$

The forecast values 7.387, 8.290, 9.302, 10.438, 11.712 and 13.142 of the output value of slicing wafers from Nos. 2 to 7, respectively, are obtained by substituting $t = 1,\ 2,\ \cdots,\ 7$ into Eqs. (16) and (19), and the reduction sequence is

$$\hat{X}^{(0)} = (\ 7.387, 8.290, 9.302, 10.438, 11.712, 13.142\ )$$

From Table 6, which lists the actual output value of slicing wafer, the forecast output value of slicing wafer, residual error, and average residual error can be determined by substituting $\hat{P}^{(0)}(t+1)$ and $P^{(0)}(t+1), t = 1,\ 2,\ \cdots,\ 7$ into Eq. (17) and Eq. (18).

The first part of Table 6 involves two sections. The first section involves slicing wafer values from No. 1 to No. 5, while the second section involves slicing from No. 6 to No. 7. In the first section, the average residual error is 5.5%. This section illustrates that the bow characteristic forecast output value could accurately forecast the actual output value. That is, the slicing process presents the normal situation and the parameter settings are balanced. However, in the second section, the residual errors are 17.1% and 31.4%, respectively. Despite this, the two values are still under control and the engineer does not stop the slicing

machine to check the parameter settings for the correlative slicing process. Thus, in the eighth slicing wafer value is out-of-control. Therefore, the unbalance forecasted values provide warning information that slicing process is unbalance. The second part involves three sections. The first section is slicing wafer values from No. 9 to No. 12; the second section involves slicing wafers from No. 13 to No. 15, and the third section involves slicing wafers from 16 to No. 33. In the first section, the average residual error is 2.6%, indicating that the slicing process is under control. Although, the actual output values and the forecasted output values are under control in the second section, the residual errors are 56.7%, 10% and 24.1%, and thus continue to exceed 10%. The engineer stops the slicing machine and checks the correlative slicing process parameters setting. After checking the process parameter setting, the slicing process produces wafers effectively until No.33. In the third part, the residual errors from Nos. 34 to 36 are 22.7%, 15.4% and 56.6%, respectively. These residual errors continue to exceed 10%, but the actual output values are under control. The engineer does not consider forecasting values until the bow becomes out-of-control. The slicing process is unstable from No. 34 to 40. The noise factors from No. 34 to 36 are unbalanced and cause unstable yields.

4. Conclusion

Relative control and management philosophy includes three stages; they are focus groups, GSDM and relative management and control.

The first stage, focus groups processes allow engineers to clearly understand how quality and process capability influence silicon wafer slicing. Interactive discussions indicate that the focus groups can improve the productivity and effectiveness of decisions, either by accelerating the decision process or by increasing decision quality. Moreover, the proposed procedure allows engineers to rapidly adjust a manufacturing system to eliminate problematic phenomena and increase slicing quality and process capability.

The second stage, GSDM provides relative concept to screen the synchronously occurring abnormalities multiple quality characteristics. The main contribution of this paper is that is uses only five historical samples to screen out the worst quality characteristic from existing quality characteristics in the slicing process.

The third stage, relative management and control applies the Chinese philoso-

phy of yin and yang to illustrate relative management of main effect factor and noise factor to control and monitor slicing process quality, and provides decision makers with philosophical thoughts for balancing the simultaneous consideration of various factors. Yin and yang refers to the Chinese idea of balance combined with continual change, an idea that can provide a continuous method of measuring and balancing slicing process. Furthermore, Grey forecasting detects and rapidly triggers the alarm in the case of an abnormality quality. Thus, the matching grey forecasting method is applied to increase process yield and accuracy of forecasting the quality of slicing the next wafer and constantly monitor accurate drift and quality control from the slicing machine.

Continuing to control and monitor slicing quality is not absolute, but rather requires balancing the two forces of "under control" or "out-of-control". Specifically, decision makers continue to monitor the slicing balance between the main effect factor and noise factors. Yin and yang thus is not a theory, thought or method. Instead yin and yang is a practical management philosophy that allows decision makers to use their hearts and minds to manage in a balanced way. Furthermore, it is essential for decision makers to learn the managerial art and wisdom of yin and yang philosophy.

Acknowledgement
The author would like to thank the author's master Dr. Hong Tao-Tze teaches the principle of yin and yang philosophy (www.taijimen.org) and the National Science Council of the Republic of China for financially supporting this research under Contract No. NSC 94-2416-H-264-007.

5. References

ASTM F534. (1995). *Annual Book of ASTM Standards,* ISBN: 0803122268.
ASTM F657. (1995). *Annual Book of ASTM Standards,* ISBN: 0803122268.
Deng, J. (1989). Introduction to Grey System Theory, *The Journal of Grey System*,

Vol., 1 No., 1, pp.1-24, ISSN: 0957-3720.

Deng, J. (2003). *Grey System Theory and Applications*, Kao-Li, Taiwan, ISBN: 9575-847288.

Einstein, A. (1920). *Relativity: The Special and General Theory*, New York: Henry Holt and Company, ISBN: 1-58734-092-5.

Hsieh, C. H.; Chou, J. H. & Wu, Y. J. (2002). Optimal Predicted Fuzzy PI Gain Scheduling Controller of Constant Turning Force Systems with Fixed Metal Removal Rate, *The International Journal of Advanced Manufacturing Technology*, Vol., 19, No., 10, pp. 714-721, ISSN: 0268-3768 (Paper) ISSN: 1433-3015 (Online).

Krueger, R. A. & Casey, M. A. (2000). *Focus Groups: A Practical Guide for Applied Research*, 3rd Spiral edition, Sage Publications, Thousand Oaks, CA, ISBN: 0761920714.

Laozi & Roberts, M. (Translator) (2004). *Dao De Jing: The Book of the Way*, University of California Press, ISBN: 0520242211.

Lin, C. T. & Yang, S. Y. (2003). Forecast of the Output Value of Taiwan's Optoelectronics Industry Using the Grey Forecasting Model, *Technological Forecasting & Social Change*, Vol., 70, No., 2, pp. 177–186, ISSN: 0040-1625.

Lin, C. T.; Chang, C. W. & Chen, C. B. (2004). Focus Groups: Impact of Quality and Process Capability Factors on Silicon Wafer Slicing Process, *International Journal of Manufacturing Technology and Management*, Vol., 2, No., 2, pp. 171-184, ISSN: 1368-2148 (Print) ISSN: 1741-5195 (Online).

Lin, C. T.; Chang, C. W. & Chen, C. B. (2005). Relative Control Philosophy-Balance and Continual Change for Forecasting Abnormal Quality Characteristics in a Silicon Wafer Slicing Process, *The International Journal of Advanced Manufacturing Technology*, Vol., 26, No., 9-10, pp. 1109 – 1114, ISSN: 0268-3768 (Paper) 1433-3015 (Online).

Lin, C. T.; Chen, C. B. & Chang, C. W. (2002). Screening Synchronously Occurred Multiple Abnormal Quality Characteristics Screening in a Silicon Wafer Slicing Process, *The Asian Journal on Quality*, Vol., 3, No., 1, pp.48-60, ISSN: 1598-2688.

Lin, Z. C. & Lin, W. S. (2001). The Application of Grey Theory to the Prediction of Measurement Points for Circularity Geometric Tolerance, *The International Journal of Advanced Manufacturing Technology*, Vol., 17, No., 5, pp. 348-360, ISSN: 0268-3768 (Paper) 1433-3015 (Online).

Lucas, J. M. & Saccussi, M. S. (1992) Exponentially Weighted Moving Average Control Charts: Properties and Enhancements, *Technometrics*, Vol., 24, pp.

216-231, ISSN: 0040-1706.

Morgan, D. L. (1997). *The Focus Groups Guidebook, Focus Group Kit 1*, Sage Publications, Thousand Oaks, ISBN: 0761908188.

Robert, S. (1959). Control Chart Tests Based on Geometric Moving Averages, *Technometrics*, Vol., 1, pp. 239-250, ISSN: 0040-1706.

Takeshi, H. (1998). *Ultraclean Surface Processing of Silicon Wafers: Secrets of VLSI Manufacturing*, Springer, ISBN: 3540616721.

25

Multidimensional of Manufacturing Technology, Organizational Characteristics, and Performance

Tritos Laosirihongthong

1. Introduction

Over the last ten years manufacturing technology use has been studied in several countries and a stream of findings has been coming in. The purpose of this study is to investigate manufacturing technology use in the Thai automotive industry, and to (1) examine findings concerning certain manufacturing technology dimensions, (2) investigate the relationships among manufacturing technology use, organizational characteristics (i.e. size, ownership and unionization), and performance, and (4) use the findings to shape a concept of multidimensional view of manufacturing technology. In the past, many studies have used data from the US, Australia, and other developed countries (Boyer *et, al*., 1997; Sohal, 1999; Dean *et, al*, 2000: Park, 2000). The findings from this study using data of the Thai automotive industry are a useful contribution to international applicability of manufacturing technology.

This chapter is organized into five sections. The next section summarizes the literature and theoretical background. Research methodology and data analysis incorporating sample selection, questionnaire design, and reliability and validity of measurement instruments is described in Section 3. Research findings and conclusion is presented in Section 4 and 5 respectively.

2. Literature Review and Theoretical Background

2.1 Technology dimensions

Certain classes of manufacturing technology are appropriate for particular competitive manufacturing strategy. For example, computer numerical control (CNC), computer-aided design (CAD), computer-aided manufacturing (CAM) or computer-aided engineering (CAE) are appropriate for a strategy seeking

flexibility. Manufacturing technologies have been grouped and classified in several different ways, some based on the level of integration, or the nature of the technology. (Rosenthal, 1984; Warner, 1987; Adler, 1988; Paul and Suresh, 1991; Small and Chen, 1997).
Swamidass and Kotha (1998), in an empirical study, found that nineteen technologies used in manufacturing could be classified into four groups based on the volume and variety considerations of the production process. Their empirical results indicate that manufacturing technology could be classified into four groups:

1) *Information exchange and planning technology*
2) *Product design technology*
3) *High-volume automation technology* and
4) *Low-volume flexible automation technology*.

A notable conclusion of their study being that *High-volume automation technology* could be used to serve the low variety and high volume production strategy, while *Product design technology* and *Low-volume flexible automation technology* could be used to serve the high variety and low volume production strategy. The implication is that technology dimensions have far reaching consequences for the manner in which companies use them. This study decides to use the empirically-established dimensions of manufacturing technology reported by some previous studies, as described in section 3, to guide this study.

2.2 Manufacturing technology use and organizational characteristics

A number of previous studies have indicated that organizational characteristics (i.e., firm size, ownership, year in operation, sales volume, and labor union membership) have an influence on the adoption and implementation of manufacturing technology (Ettlie, 1984; Chen *et al*, 1996; Millen and Sohal, 1998; Schroder and Sohal, 1999; Swamidass and Winch, 2002). Summary of these findings are explained as follow:

2.2.1 Size

Manufacturing and operations management researchers have found that large companies show a higher degree of manufacturing technology implementation than small and medium companies (Paul and Suresh, 1991; Mansfield, 1993;

Sohal, 1999; Swamidass and Kotha 1998). This is attributed in the literature to the fact that large companies have superior technological know-how because of their access to more human, financial and information resources compared to small to medium companies. Researchers have come to agree that size is an important variable when it comes to manufacturing technology use. For example, Small and Yasin (1997) recommend that future research in management of manufacturing technology should adopt a contingency approach to find out how organizational variables such as firm size, industry structure, and planning approach influence the relationship between adoption of manufacturing technology and overall plant performance.

2.2.2 The nationality of plant ownership

Although a number of studies to investigate the relationship between organizational variables and technology use have been conducted in developed countries, such studies are not common in developing countries. Peter *et al*, (1999) state that the nationality of ownership of companies reflects the differences in management practice in manufacturing technology implementation due to differences in national culture. Sohal (1994) reports a number of significant differences in manufacturing technology use (e.g. computer hardware, computer software, plant and equipment) and management effort (e.g. source of manufacturing technology information, financial appraisal techniques, training, and benefits) between Australia and the United Kingdom. Lefley and Sarkis (1997) studied appraisal/assessment of manufacturing technology capital projects in the USA and UK and found different degrees of success in manufacturing technology implementation. Kotha and Swamidass (1998) report a significant effect of the nationality of a company (Japan vs. USA) on manufacturing technology use.

Further, Schroder and Sohal (1999) found that Australian-owned companies rate the anticipated benefits of increased throughput, sales, and investment in manufacturing technology more highly than foreign-owned companies from South Korea, Taiwan, Japan, USA, and New Zealand operating in Australia.

2.2.3 Unions

It has been widely suggested that effective implementing of manufacturing technology depends on the human factor or employees and their flexibility (Goldhar and Lei, 1994; Upton, 1995; Lefebvre *et al*, 1996). This often means that labor unions have to set aside their traditional work rules and job control

strategies to allow team work and consultation (Osterman, 1994). Successful adoption of manufacturing technology also requires worker to attain new levels of operational skills and a higher level of commitment to improve product quality (Osterman, 1994). This can often be achieved through agreement with the union and management as in the case of Harley-Davidson Motor Company.

Chen *et al*, (1996) note that a company equipped with all the computerized or automated manufacturing technologies may be surprised to find that ultimate success is largely determined by the human factor. They also give the example of a plant, operated with the help of 300 robots, which had higher productivity and poorer quality performance than a more labor-intensive plant with a labor union.

Other major issue related to the adoption and implementation of manufacturing technology is employee commitment and cooperation (Krajewski and Ritzman, 1993; Chen and Gupta, 1993). Tchijov (1989) reports that plants with labor union membership exhibit the resistance to the adoption of manufacturing technologies. On the contrary, Dimnik and Richardson (1989) found that there was no relationship between union membership and adoption of manufacturing technology in a sample of auto-parts manufacturers in Canada.

Small and Yasin (2000) investigated human factors in the adoption and performance of manufacturing technology in unionized organizations. They found a union effect on the adoption of just-in-time production system only. For all other technologies investigated in their study, there was no significant union effect. Thus, given the above, there is no clear evidence of union effect on manufacturing technology use; it deserves more investigation.

2.3 Performance measures

Performance measures are multidimensional. Several researchers have investigated the relationship between manufacturing technology implementation and performance (Paul and Suresh, 1991; Chen and Small, 1994; Small and Yasin, 1997; Small, 1999; Swamidass and Kotha, 1998). This study classifies the wide range of performance measures in the literature into three groups:

(1) strategic measures
(2) organizational measures and
(3) business and market performance measures.

2.3.1 Strategic measures

Researchers suggest that the performance measures of manufacturing technology implementation should be strategically focused (Millen and Sohal, 1998; Sohal, 1999; Efstathiades *et al*, 2000; Sun, 2000). These measures include many dimensions including quality and flexibility.

Quality has surfaced in many performance measures. For example, Dimnik and Richardson (1989) note that the key performance measures in evaluating manufacturing technology in the automotive industry in Canada are cost, quality and flexibility. Other researchers recommend other two dimensions while investigating the auto industry; product quality, and service quality comprising both pre- and after-sale service (Curkovic *et al*, 2000). In the literature this study find that quality performance measure may incorporate percent defective, rejection rate, customer complaints, and product accuracy (Paul and Suresh, 1991; Laosirihongthong and Paul, 2000).

Flexibility is an important component of performance especially in the automotive industry (Zairi, 1992; Zammuto & O'Connor, 1992; Sohal, 1994; Boyer, 1996). Small and Chen (1997) define flexibility as the ability to respond quickly to changing customer needs. They also classify manufacturing flexibility into two dimensions, "time-based flexibility" which focuses on the speed of response to customer needs, and "range-based flexibility" which is concerned with the ability to meet varying customization and volume requirements in a cost-effective manner. In addition, time-based performance of automotive suppliers is critical, and manufacturing lead-time is especially critical in this industry (Jayaram et al, 1999).

2.3.2 Organizational performance

The specific measures of organizational performance include the degree to which manufacturing technology have improved work standard, skills of employees, image of the company, and coordination and communication within the company (Millen and Sohal, 1998; Sun, 2000; Efstahiades *et al*, 2000). Organizational measures are related to workflow, work standardization, communication, and management control (Dean et al, 2000).

2.3.3 Business and market performance

A third set of measures is reported by Small and Yasin (1997), who suggest that business and market performance measures could be tied to revenue from manufacturing operation, return on investment, overhead cost, time-to-market

for a new product, and market share of existing/new products. Some of these measures are financial performance measures. Swamidass and Kotha (1998) investigated the relationship between manufacturing technology use and financial performance. They found that the relationship is not significant, and conclude that perhaps strategic rather than financial benefits might have been the primary reason for investing in manufacturing technology. Therefore, this study did not use financial performance measure.

In summary, performance measures used in manufacturing and operations management researches while investigating manufacturing technology use are varied. However, there is a common understanding that there are three important but broad dimensions of performance measures -- quality, flexibility, and organizational measures. This study uses these three dimensions for performance measurement reflecting the successful for manufacturing technology implementation.

2.4 Guiding Research Question

The discussion of key variables and their relationships above provide the basis for the guiding research question of the study based on the three technology types and three performance dimensions discussed above:

Whether High-volume automation technologies, data-interchange technologies, and low-volume automation technologies, either individually or collectively affect one or more of the performance measures, which are quality performance, flexibility and organizational performance.

3. Research Methodology and Data Analysis

3.1 Sample and data collection

This study selected only companies who are listed with Thailand Industrial Standard Institute and Thai Automotive Institute. The companies surveyed in this study all produce products classified in the automobile and parts/components industry sector. Questionnaire used in this study consists of three parts: the degree of manufacturing technology use, perceived manufacturing technology benefits/performances, and organizational characteristics. It includes fifteen manufacturing technology (Boyer et al., 1997; Burgess and

Gules, 1998; Efstahiades et al., 2000; Boyer and Pagell, 2000; Efstathiades et al., 2002), thirteen perceived performance measures (Small and Yasin, 1997; Park, 2000), and four organizational characteristics including size of the company (measured by a number of employees), type of ownership, and existence of labor union.

Characteristics	Description	%
Respondents	MD/VP/P	10.20
	Factory/Production Mgr.	37.80
	General Manager	14.50
	Engineering Mgr.	22.70
	QA/QC Mgr.	18.80
Company size (number of employees) [1]	Small to medium <= 200	58.40
	Large > 200	41.60
Ownership	Thai-owned	30.40
	Foreign-owned	14.30
	Joint-venture	55.30
Labor union	Labor union present	30.45
	No labor union	69.55
Main product classifications	Body parts	21.42
	Chassis parts	25.58
	Suspensions parts	12.25
	Electrical parts	8.20
	Accessories	11.45
	Trim parts	21.10
Existing quality management system	ISO/QS9000 certified	94.58
	None	5.42

Table 1. Characteristics of respondents [1] Size classification according to Ministry of Industry, Thailand

Totals of 480 questionnaires this study distributed to factory, general, engineering, and quality assurance managers who have a responsibility for manufacturing technology implementation in their own companies. Questionnaires were sent to the respondents by given directly (for return by mail) at the suppliers' monthly meeting of one Japanese assembler and one American assembler. One respondent per company was asked to indicate the degree of implementation for fifteen manufacturing technology and perceived performance after the implementation. The usage attributed to these technologies and performances was measured using Likert's five-point scale where 1 = not used or

not satisfied and 5= extensively used or very satisfied. A total of 124 questionnaires were returned giving a response rate of 25.83 percent, comparable to the rates in previous such research (Sohal, 1996; Small and Chen, 1997). Table I exhibits the characteristics of respondents.

3.2 Non-Respondent Bias

A random sample of 30 companies from the 356 non-respondents was selected to compare the respondents with non-respondents. The following classificatory data this study are collected from the 30 non-respondents through the phone: (1) size (employment), (2) ownership, (3) ISO 9000 certification, and (4) unionization. All 30 non-respondents contacted by phone provided classificatory information requested by phone. In Table II, this study indicates the result of the comparison between responding and non-responding sample. The Chi-square values indicate that the two samples are statistically different. Major differences between respondents and non-respondents being that the sample of respondents have larger firms, foreign-owned firms, more ISO-certified firms, and more unionized forms.

If this study assume that the sample of 30 non-respondents is representative of all non-respondents, the findings of this study are pertinent to the 124 manufacturers who participated in this study.

Organizational characteristics	**Respondents**	**Non-respondents**	***Chi-sq.***	**Chi-Sq. table (.05 significance, 2-tail)***
Size =< 200 employees	72(58%)	12(40%)		4.89
Size > 200 employees	52(42%)	18(60%)	17.1	
Thai owned	37(30%)	6(20%)		7.57
Foreign owned	17(13.7%)	15(50%)		
Joint venture	70(56.3%)	9(30%)	52.8	
ISO/QS9000 certified	117(94.4%)	19(63.4%)		4.97
None	7(5.6%)	11(36.6%)	8.2	
Labor union present	37(30%)	10(33.4%)		5.14
No labor union	87(70%)	20(66.6%)	10.4	

Table 2. Comparison of Respondents with a Random Sample of Non-Respondents. * The Chi-squared values for size, ownership, ISO certification and union are all larger than the Chi-square table values for .05 significance (2-tail). Thus, the respondents are not similar to the random sample of non-respondents

Generalization of the findings to non-respondents must be done with care. Given that the sample of 124 firms participating in this study is substantial, the findings are valuable even if they are not representative of the entire Thai auto industry.

3.3 Data analysis

3.3.1 The reliability and validity of empirical measures

The internal consistency of our measures was verified using Cronbach's alpha (Cronbach, 1951); a value greater than 0.6 was treated acceptable (Chen and Small, 1994). Content validity was established from literature review, expert and practitioner opinions, and pre-testing with a small number of managers. Construct validity was ensured by factor identification through principal component factor analysis (Nunnally, 1967). Factors are selected using these three rules: (a) minimum Eigenvalue of 1, or cumulative factor variance explained in excess of 70 percent; (b) minimum factor loading of 0.4 for each item; and (c) the simplicity of factor structure. Factor analysis was used to find factors to explain dependent variables (performance measures) and independent variables (technology use). SPSS software was used to perform principal component analysis including an orthogonal transformation with Varimax rotation. The results are shown in Tables III (for technology factors) and VII (for performance factors).

In order to test the validity of perceptual performance measures, this study conducted a correlation analysis between selected objective external measures with self-reported perceptual data on performance for 20 per cent of the companies randomly selected (n = 30) from our sample of 124 respondents. Selected objective external measures were obtained from the Monthly Suppliers Evaluation Reports--MSER (Sriwatana, 2000; Vibulsilapa, 2000) concerning delivery, quality, cost, and organizational reputation. Correlation analysis between MSER data and survey data was conducted, specifically, the correlation analysis between MSER data and survey-based composite values of flexibility, quality performance, and organizational performance for a random sample of 30 companies. The resulting correlation coefficients are 0.77, 0.81, and 0.73 respectively. Therefore, this study considers the perceptual performance measures acceptable (Swamidass and Kotha, 1998; Lewis and Boyer, 2002).

4. Research Findings

4.1 Technology use (factors) confirm prior studies

Multi-item scales are developed for each construct (technology and performance) in this study. Before creating the final scales, the data are checked for normality and outliners. As shown in Table III and VII, the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy is 0.887 (for technology factors) and 0.894 (for performance). A minimum Kaiser-Meyer-Olkin score of 0.5 is considered necessary to reliably use factor analysis for data analysis (Small, 1999). Score over 0.80 are considered very strong. Similarly, the Bartlett test of sphericity (the higher, the better) was 987.32 (technology factor) and 1322.t (performance) with significance value (Small, 1999).
The results of rotated principal component factor analysis show that three factors explain 63.25 per cent of the total variance (Table III). These technology factors are used in subsequent analysis to examine the relationships between technology use and organizational characteristics, as well as technology use and performance.
In Table III, the result indicates that seven technologies load on the first factor. This factor consists of technologies that can be used to reduce direct labor costs in repetitive operations and high-volume production with low variety of products. Therefore, the study names this factor as "*High-volume automation technologies."*
The second factor consists of five technologies that relate to planning and data interchange. Therefore, the study names this factor as "*Data-interchange technologies*," which parallels the "information exchange and planning technologies" reported by Swamidass and Kotha (1998) using US data. The third factor includes technologies that provide low-volume manufacturing flexibility that permits low-volume high variety production. This study, therefore, calls this factor, *"Low-volume flexible automation technologies.*"

The three factors that emerged from data of the Thai automotive industry are similar to technology factors that determined from factor analysis of some previous studies.

Thus, it is important to note that manufacturing technology factors that were identified in this study are robust and are stable across time and national boundaries.

Technology Factors	**Mean**	**S.D.**	**orsExtracted fact**		
			1	**2**	**3**
High-volume automation technologies					
Automated material handling	2.15	1.21	0.774		
Automated assembly system	2.38	1.35	0.732		
Automated storage/retrieval system	1.74	1.02	0.715		
Automated inspection system	2.42	1.01	0.701		
Computer-aided manufacturing	3.15	1.04	0.554		
Barcode system	2.88	1.33	0.568		
Pick and place robots	2.04	1.41	0.520		
Average mean score	*2.39*				
Data interchange technologies					
Material resources planning	2.54	1.11		0.726	
cturing centerFlexible manufa	1.98	1.06		0.711	
Computer-aided process planning	2.22	1.21		0.702	
Computerized statistical process control	2.16	1.05		0.566	
Electronic data interchange	2.53	1.02		0.511	
Average mean score	*2.28*				
Low-volume flexible automation technologies					
Computer numerical control	3.88	1.44			0.818
Pneumatic and hydraulic equipment	3.72	1.32			0.735
Computer-aided design	3.25	1.51			0.598
Average mean score	*3.62*				
Kaiser-Meyer Olkin adequacy(KMO)			0.887		
ericityBartlett's test of sph			987.32		
Significance			0.00000		
Cronbach's Alpha			0.875	0.902	0.821
Eigenvalues			3.488	2.876	2.034
Varience explained			24.45	22.62	16.18
Total variance explained			24.45	47.04	63.25

Table 3. Technology Facts (Rotated Comonent Matrix). Note: 1 – Lowest, 5 - Highest

4.2 Technology factors and organizational characteristics

4.2.1 Size

In Table IV, this study compares the use of three different technology dimensions (factors) in large versus small/medium firms. The table shows that there is a significant difference between large and small-to-medium companies in the use of *High-volume automation technologies* (p=.025) and *Low-volume flexible automation technologies* (p=.002). There is no significant difference in the use of *Data-interchange technologies* (p=.103). *Data-interchange technologies* form the backbone of manufacturing systems now and these technologies have been around longer the other technologies. *The implication is that all manufacturers, regardless of size, equally depend on Data-interchange technologies.* One reason being, these technologies are easily implementable on PCs, which are affordable by even small manufacturers. For example, MRP and Electronic Data Interchange (EDI) (see Table III) that are included in this dimension could be implemented using ordinary PCs. *The findings reveal that plant size has differential effect on the various technology factors*.

Technology Factors	**Sig.**	***Small-to-medium*** **Composite mean**	***Large*** **Composite mean**
High-volume automation technologies	0.025*	2.87	2.74
Data-interchange technologies	0.103	2.01	2.23
Low-volume flexible automation technologies	0.002**	2.66	3.37

* Significant at 0.10 level. ** Significant at 0.05 level. *** Significant at 0.01 level.

Table 4. Technology Factors and Size of Company. * (Employees <= 200 = small-to-medium; employees > 200 = large.)

4.2.2 Ownership

Table V reports the use of the three different dimensions of manufacturing technologies in Thai-owned, foreign-owned and jointly-owned firms. According to the table, the following is revealed:

- In foreign-owned plants, *High-volume automation technology* use is significantly higher than its use in either Thai-owned (p=.001) or joint-venture plants (p=.001).
- In Thai-owned plants, *Low-volume flexible automation technology* use is higher than the use of this technology in either joint ventures (p=.001) or foreign-owned (p=.001) plants. Apparently, Thai plants produce more low volume components.
- Plant ownership has no effect on *Data-interchange technologies.* In an earlier section, this study reported that plant size has no effect on *Data-interchange technology* use. Taken together with this finding, it is important to note that *Data-interchange technologies* are relatively more mature technologies, easily implementable without much capital or resources, and is immune to size and ownership.

Technology Factors		**Thai-owned**	**Joint-venture**	**Foreign-owned**
High-volume automation technologies	Mean score →	2.35	2.22	2.61
Significance of Joint venture and column	Joint venture	p=0.182 (ns)		
Significance of Foreign-owned and column	Foreign-owned	p=0.001***	p=0.001***	
Data-interchange technologies	Mean score →	2.53	2.72	2.45
Significance of Joint venture and column	Joint venture	p=0.225 (ns)		
Significance of Foreign-owned and column	Foreign-owned	p=0.743 (ns)	p=0.351 (ns)	
Low-volume flexible automation technologies	Mean score →	3.47	3.18	3.01
Significance of Joint venture and column	Joint venture	p=0.001***		
Significance of Foreign-owned and column	Foreign-owned	p=0.001***	p=0.423 (ns)	

Table 5. Technology Factors and Ownership. * Significant at 0.10 level. ** Significant at 0.05 level. *** Significant at 0.01 level. ns = not significant

4.2.3 Unionization

Very few studies have investigated the effect of unionization on manufacturing technology use. Tchijov (1989)'s found that plants with labor union membership exhibit the resistance to adoption of new technologies. This study does not measure union membership of employees, if measures if the plant is unionized or not. As shown in Table VI, the use of *Data interchange technologies* is significantly higher (p=.013) in plants with labor unions, and the use of *High-volume automation technologies* is higher in non-union plants (p=.011). It is a notable finding that unionization does have an effect in the use of at least a certain technology.

Technology Factors	**Sig.**	**Labor union**	**Non-union**
		Composite mean	**Composite mean**
High-volume automation technologies	p=0.011*	2.53	2.62
Data-interchange technologies	p=.013**	2.77	2.32
Low-volume flexible automation technologies	p=0.644	3.32	3.15

Table 6. Technology Factors and Labor Unionization. * Significant at 0.10 level. ** Significant at 0.05 level. *** Significant at 0.01 level.

4.2.4 Performance measures

A principal component factor analysis is used to reduce and group the thirteen individual performance items in the survey into three performance factors, "*Flexibility performance*", "*Quality performance*", and "*Organizational performance*". The three performance factors together explain 71.55 percent of the total variance (Table VII).

4.3 Technology factors and performance

As a rule, this study finds that there is little association between technology use and performance factors (Table VIII), the one exception being *High-volume automation technology*, which is associated with *Quality Performance* (Pearson r =

0.236; p = 0.000). Three multiple regression models to estimate performance using technology use dimensions are reported in Table IV. According to the table, only quality performance is explained by one of the technology dimensions (*High-volume automation technologies).*

An inference from this study is that, for the auto industry, high-volume automation is an essential ingredient for quality. This inference may be limited to the auto industry because of the sample.

Performance measures	***Mean**	**S.D.**	**Extracted Factors**		
			1	**2**	**3**
Flexibility performance					
Delivery lead time	3.87	0.84	0.720		
Responsiveness to customer needs	3.65	0.78	0.815		
Production change overtime	3.42	0.92	0.736		
Set-up time	3.33	0.76	0.884		
Average mean score	*3.57*				
Quality performance					
Defective ratio along the process	3.66	0.88		0.833	
Rejection ratio within the process	3.47	0.91		0.784	
Customer complain	4.22	1.02		0.746	
Frequency of inspection	3.85	0.77		0.626	
Accuracy of product	4.01	0.98		0.689	
Average mean score	*3.84*				
Organizational performance					
Upgrading human skills	3.72	0.74			0.843
Company's image	3.88	0.83			0.744
Work standardization	4.21	0,98			0.832
Reducing bargaining of skilled labor	3.18	0.86			0.675
Average mean score	*3.75*				
Kaiser-Meyer Olkin adequacy(KMO)			0.894		
Bartlett's test of sphericity			1322.7		
Significance			0.00000		
Cronbach's Alpha			0.922	0.916	0.842
Eigenvalues			2.133	3.411	2.756
Varience explained			24.22	28.72	18.61
Total varience explained			24.22	52.94	71.55

Table 7. Performance Factors (Rotated Component Matrix
Note: 1 – Lowest, 5 – Highest

Technology Factors	***Flexibility***	***Quality***	***Organizational***
High-volume automation technologies	0.005 p = 0.843	0.236 p = 0.000***	0.054 p = 0.331
Data-interchange technologies	0.054 p = 0.466	0.082 p = 0.342	0.037 p = 0.693
High-flexible automation technologies	0.993 p = 0.215	0.051 p = 0.442	0.027 p = 0.578

Table 8. Correlation Analysis between Technology Factors and Performance Factors. * Significant at 0.10 level. ** Significant at 0.05 level. *** Significant at 0.01 level

Technology Factors		***Small-to-medium***	***Large***
	Sig.	**Composite mean**	**Composite mean**
High-volume automation technologies	0.025*	2.87	2.74
Data-interchange technologies	0.103	2.01	2.23
Low-volume flexible automation technologies	0.002*	2.66	3.37

Table 9. Technology Factors and Size of Company* * Employees <= 200 = small-to-medium; employees > 200 = large. * Significant at 0.10 level. ** Significant at 0.05 level. *** Significant at 0.01 level.

5. Conclusions and Future Studies

The most notable theme here is that findings from this study confirm several findings reported in the literature based on data from other nations. First, the study concurs with previous studies that show the size of companies influences the use of manufacturing technology. The reasoning is now this study known; large companies can afford the higher cost of adopting these technologies. Also, managerial resources necessary in planning and implementing such technologies are available in larger companies (Ariss *et al*, 2000).

Second, this study found that technology use is a function of the nationality of the plant ownership. For example, finding indicates that *High-volume automation technologies* such as automated material handling, automated assembly

system and robots are more likely to be adopted in foreign-owned companies than in Thai-owned and joint-venture companies. Foreign-owned companies perhaps tend to adopt more technologies because of their superior financial, technical and managerial resources, technological capabilities, and abilities to transfer those technologies. Further, foreign-owned plants may replicate the use of technology in plants back home, which is invariably a more developed nation compared to Thailand. The findings concerning the effect of the nationality of ownership on technology use concurs with studies on technology implementation in Australia (Sohal *et al*, 1991), in the UK (Sohal, 1994), and the USA (Kotha and Swamidass, 1998).

Third, The multidimensional view of technology reported by Swamidass and Kotha (1998) using a US sample holds up this study in the sample of firms from Thai auto industry; further, the two samples are several years apart.

5.1 Some Directions for Future Studies

5.1.1 The need for more investigations of the unionization-technology link

A notable finding of this study is that the use of Data interchange technologies, at least, is significantly higher in plants with labor unions. Could it be that these technologies reduce the influence or soften the effect of unionization? Do they reduce the need for employees in functions affected by unionization? Is it possible that unions do not resist the adoption of Data-interchange technologies? The search for answering to these questions is a worthy line of investigation for the future.

5.1.2 A proposed concept of manufacturing technology use

This study, confirms the emerging multi-dimensional view of technology use with collected data in Thailand with a specific industry. Further, the multiple technology factors that this study found in Thailand are similar to those found in the USA. This is a testimony to the robustness of the technology factors, which transcend national borders. Additionally, in an earlier study by Swamidass and Kotha (1998), which reported the multiple dimensions of technology, the data came from a survey nearly 10 years earlier than the Thai survey reported here. Therefore, it appears that the technology dimensions/factors are stable across time.

In addition, this study confirms findings concerning the effect of plant size, and the nationality of ownership. Taken together, empirical research to this point encourages the following Theory of manufacturing technology use for testing and retesting in the future for its confirmation and establishment: " *In the complex manufacturing environment made of people, technology and procedures, manufacturing technology is not homogenous but has consistently distinct dimensions. These technology dimensions are robust and exist across national boundaries and time. However, technology use is a function of plant size, and the nationality of plant ownership".*

5.2 Limitations

While this study is based on responses from nearly 150 firms, our non-response bias test shows that the responding firms are larger, more foreign-owned, more ISO-certified, and more unionized, compared to non-respondents. In the future, a more representative sample may be investigated. Boyer et al (1997) found that companies benefit from manufacturing technology investments when there is adequate and matching investments in the infrastructure. This study did not investigate this aspect of technology use in more details. Therefore, this study would encourage studies that test the above concept in order to expand it to cover the role of infrastructure investments.

Acknowledgements
The author would like to thank Paul M. Swamidass, Professor and Director of Thomas Walter Center of Technology, Auburn University, Alabama, for his valuable suggestions on the first revision of this manuscript.

6. References

Adler, P.S. (1988). *"Managing flexible automation"*. California Management Review. 20 (1), 35-56.

Ariss, S.S., Raghunathan T.T. and Kunnathar A. (2000). *"Factors affecting the adoption of advanced manufacturing technology in small firms"*. S.A.M. Advanced Management Journal, Spring.

Attasathavorn, J. (2001). *"Reports of Thai Automotive Industry"*. For Quality Magazine, March-April, 36-49. (in Thai).

Bank of Thailand. (2000). *The Economics Report during January – March 2001*, 65-80. (in Thai).

Boer, H., Hill, M. and Krabbendam, K. (1990). *"FMS implementation management: promise and performance"*. International Journal of Operations and Production Management, 10 (1), 5-20.

Board of Investment (BOI). (1995). *Report of the Investment of Automotive Industry in Thailand*. Board of Investment, Bangkok. (in Thai).

Boyer, K. (1996). *"An assessment of managerial commitment to lean production"*. International Journal of Operations and Production Management, 16 (9), 48-59.

Boyer, K, Leong, G.K., Ward, P.T., and Krajewski, L.J. (1997). *"Unlocking the potential of advanced manufacturing technologies"*. Journal of Operations Management, 15, 331-347.

Chen, I.J., Gupta A., and Chung, C.H. (1996). *"Employee commitment to the implementation of flexible manufacturing systems"*. International Journal of Operations and Production Management, 16 (7), 4-13.

Chen, I.J. and Gupta, A. (1993). .*Understanding the human aspect of flexible manufacturing system through management development*.. International Journal of Management Development, 10 (1), 32-43.

Chen, I.J. and Small M.H. (1994). *"Implementing advanced manufacturing technology: An Integrated Planning Model"*. OMEGA, 22 (1), 91-103.

Cronbach, L.J. (1951). Coefficient *Alpha and the Internal Structure of Tests: Psychometrika*, 16, 297-334.

Curkovic, S., Vickery, S.K., and Droge, C. (2000). *"An empirical of the competitive dimensions of quality performance in the Automotive supply industry"*. International Journal of Operations and Production Management, 20 (3), 386-403.

Dean, A.S., Mcdermott, C., Stock, G.N. (2000). *"Advanced manufacturing technology: Does more radicalness mean more perceived benefits?"*. The Journal of High Technology Management Research, 11(1), 19-33.

Dean, J.W. Jr. and Snell, S.A. (1991). .*Integrated manufacturing and job design*.. Academy of Management Journal, 34 (4), 776-804.

Dean, J.W. Jr., Yoon, S.J., and Susman, G.I. (1992). .*"Advance manufacturing technology and organizational structure: Empowerment or subordination?*. Organizational Science, 3 (2), 203-229.

Dimnik, T. and Richardson, R. (1989). *"Flexible automation in the auto parts industry"*. Business Quarterly, 54 (4), 46-53.

Efstathiades, A., Tassou A.S., Oxinos G., Antoniou A. (2000). *"Advanced manufacturing technology transfer and implementation in developing countries: The case of the Cypriot manufacturing industry"*. Technovation, (2), 93-102.

Ettlie, J.E. (1984). *Implementation strategy for discrete parts manufacturing innovation*. In: Warner, M. (Ed.), In Microelectronics, Bookfield, VT.

Federation of Thai Industries (FTI). (2000). *Reports of The Thai Automotive Industry. Working group of automotive industry*, Bangkok. (in Thai).

Japan International Corporation Agency (JICA). (1995). *The Study on Industrial Sector Development Supporting Industries in the Kingdom of Thailand*. Tokyo: International Corporation. (in Thai).

Jayaram, J., Vickery, S.K. and Droge, C. (1999). *"An empirical study of time-based competition in the North American automobile supplier industry"*. International Journal of Operations and Production Management, 19 (10), 1010-1033.

Kotha, S. and Swamidass, P.M. (1998). *"Advanced manufacturing technology use: exploring the effect of the nationality variable"*. International Journal of Production Research, 36 (11), 3135-3146.

Krajewski, L.J., and Ritzman, L.P. (1993). *Operation Management: Strategy and analysis. 3rd Edition*: Addison, Reading, MA.

Laosirihongthong, T. and Paul, H. (2000). *"Implementation of New Manufacturing Technology and Quality Management System in Thai Automotive Industry"*. Proceedings of IEEE International Conference in Management and Innovation of Technology, November 12-15, 2000, Singapore.

Lefley, F. and Sarkis, J. (1997). *"Short-termism and the appraisal of AMT capital projects in the USA and UK"*. International Journal of Production Research, 35 (2), 341-369.

Mansfield, E. (1993). *"The diffusion of flexible manufacturing system in Japan, Europe and the United States"*. Management Science, 39 (2), 149-159.

Meredith, J. R. (1987). *"Implementing new manufacturing technologies: managerial lessons over the FMS life cycle"*. Interface, November-December, 51-62.

Millen, R. and Sohal A.S. (1998). *"Planning processes for advanced manufacturing technology by large American manufacturers"*. Technovation, 18 (12), 741-50.

Nunnally, J.C. (1967). *Psychometric Theory*, McGraw Hill: New York, NY

Park, Y.T. (2000). *"National systems of Advanced Manufacturing Technology (AMT): Hierarchical classification scheme and policy formulation process"*. Technovation, (20), 151-159.

Paul, H. and Suresh B. (1991). *"Manufacturing strategy through planning and control techniques of advanced manufacturing technology"*. International Journal of Technology Management, 6 (3-4), 233-242.

Paul, H. and Laosirihongthong T. (1999). *ISO9000 Implementation in Thailand: Experience form Thai Autoparts Industry. Proceedings of the 14th International Conference in CAD/CAM, Robotics and Factory in the Future*, Narosa Publishing House, 527-532.

Peter, B., Lee, G., and Sohal, A.S. (1999). *Lessons for implementing "AMT: Some case experience with CNC in Australia, Britain and Canada"*. International Journal of Operation and Production Management, 19 (5/6), 515-526.

Rosenthal, S.R. (1984). *"Progress toward the 'factory of the future"*. Journal of Operation Management, 4 (3), 405-415.

Schroder, R. and Sohal A.S. (1999). *"Organizational characteristics associated with AMT adoption: Towards a contingency framework"*. International Journal of Operations and Production Management, 19 (12), 1270-1291.

Small, M.H. (1999). *"Assessing manufacturing performance: an advanced manufacturing*

technology portfolio perspective". Industrial Management & Data Systems, 99(6), 266-277.

Small, M.H. and Chen I.J. (1997). *"Organizational development and time-based flexibility: An empirical analysis of AMT adoption"*. International Journal of Production Research, 35 (11), 3005-3021.

Small, M. H. and Yasin M.M. (1997). *"Developing a framework for the effective planning and implementation of advanced manufacturing technology"*. International Journal of Operations and Production Management, 17 (5), 468-489.

Sohal, A.S. (1994). *"Investing in advanced manufacturing technology: Comparing Australia and the United Kingdom"*. Benchmarking for Quality Management & Technology, 1 (1), 24-41.

Sohal, A.S. (1999). *"Introducing New Technology into a Small Business: A Case Study of Australia Manufacturers"*. Technovation, 19 (3), 187-193.

Sohal, A.S., Samson, D. and Weill, P. (1991).,*"Manufacturing and technology strategy: a survey of planning for MANUFACTURING TECHNOLOGY"*. Computer Integrated Manufacturing System, 4, 71-79.

Sriwatana, T. (2000). *Summary of suppliers performance evaluation. The Monthly Suppliers Evaluation Reports, 1998-2000*. Toyota Motor (Thailand) Company Limited. 45-70.

Sun, H. (2000). *"Current and future patterns of using advanced manufacturing technologies"*. Technovation, (20), 631-641.

Swamidass, P.M. (2000). *Encyclopedia of Production and Manufacturing Management*, Kluthis studyr Academic Publishers. 400-405.

Swamidass, P.M. and Kotha S. (1998). *"Explaining manufacturing technology use, firm size and performance using a multidimensional view of technology"*. Journal of Operation Management, 17, 23-37.

Tchijov, I. (1989). *"CIM Introduction: Some Socioeconomic Aspects"*. Technological Forecasting and Social Change, Vol. 35 (2-3), 261-275.

Thai Automotive Institute (TAI). (2000). *Thailand Automotive Industry Directory 2000*. Bangkok. (in Thai).

Thailand Development Research Institute (TDRI). (1999). *The development of Thailand's Technological Capability in Industry*, Bangkok. (in Thai).

Vibulsilapa, S. (2000). *Suppliers evaluation results. Quality Assurance Supplier Evaluation Reports, 1998-2000*. Isuzu Motor (Thailand) Company Limited, 81-124.

Warner, T. (1987). *"Information technology as a competitive burden"*. Sloan Management Review, Fall, 55-61.

Zairi, M. (1992). *"Measuring success in AMT implementation using customer-supplier interaction criteria"*. International Journal of Operations and Production Management, 12 (10), 34-55.

Zammuto, R.F. and O'Connor, E.J. (1992*)."Gaining advanced manufacturing technology benefits: The roles of organization design and culture"*. Academy of Management Review, 17, 701-728.

26

Engineering Change Management in Distruted Environment with PDM/PLM Support

Joze Tavcar and Joze Duhovnik

1. Introduction

Globalization has dramatically changed the way in which products are produced by manufactures of all sizes. Small to medium sized organizations are now just as likely to engage in global outsourcing projects as large multinational teams (Tosse, 2005). Global distributed teams need to effectively communicate and collaborate throughout the entire product development process to produce innovative products of the highest quality in the shortest period of time.

In industry, engineering change management (ECM) is recognized as a problem that receives too little attention relative to its importance. Wright's (Wright, 1997) conclusion is that from the manufacturing perspective ECM is a disturbance obstructing smooth product manufacture, but such a perspective ignores ECM's capacity to provide the incentive for product improvement. Wright's conclusion is that a number of coordinated research programs are required to establish the ground rules for maximizing the product design benefits from EC activity. Many and especially late ECs are very costly for any development project. ECs consume one third to one half of the total engineering capacity and represent 20 to 50 % of total tool costs (Terwiesch & Loch, 1999). The key contributors to long EC lead times are: complex approval process, snowballing changes, scarce capacity and organizational issues. Loch (Loch & Terwiesch, 1999) analyzed the process of administering engineering chain orders within a large vehicle development project. Despite the tremendous time pressure in development projects, EC process lead times are in the order of several weeks, months and even over one year (Loch & Terwiesch, 1999). A detailed analysis has shown a low proportion of value-added time in the EC process – less than 8.5 %. An EC spends most of its lifetime waiting for further processing. Loch suggests the following improvement strategies in order to reduce EC lead time: flexible capacity, balanced workloads, merged tasks and sharing resources (pooling).

Huang (Huang et al., 2003) investigated the current state of ECs in current industrial practice. Huang focused on big manufacturing companies and found that it is necessary to develop methodologies and techniques to improve the ECM practices. There was no evidence that ECM software packages had been used within the surveyed companies. Current ECM practices vary between companies, from formal to ad hoc approaches. Current tools dominating at new product development and introduction process are low-cost, low-function personal productivity tools like spreadsheets, project management and word processing according to AMR Research (O'Marah, 2004).

ECM support can be implemented in commercial PDM/PLM or ERP software. There are web-based ECM systems that provide better information sharing, simultaneous data access and prompt communication (Huang et al., 2001). But even a high level of information technology for ECM is very often paper based, especially in smaller companies (Huang et al., 2001). The reasons for this are that computer aids are not well known to EC practitioners and some of existing computer aids do not reflect good EC practice. In some cases, comprehensive functionality of some systems undermines their focus and imposes intensive data requirements (Huang et al., 2001).

Rouibah (Rouibah & Caskey, 2003) focused on cases in which complex product development involves more than one company – distributed engineering change management. The concurrent design process results in a parameter network that tells us how closely different components are interrelated. The knowledge contained in this network helps manage cross-company activities during the ECM process.

A review of the references emphasizes the problem of engineering changes in companies and offers quite specific solutions for complex products. This paper establishes a general model of engineering change management and applies it to distributed manufacturing and product development teams. Distributed environment requires specific methods, organization, communication skills and information system. The reference ECM model helps engineers recognize the main problems and improve the process. This was also confirmed on examples from industrial practice.

2. Characteristic design and product levels

Product development involves four characteristic levels of design. Each of them requires certain very specific activities (Prasad, 1996). The characteristic

design levels could therefore ensure very clear definitions of the activities and thus provide the necessary software and other support for all phases of the design process (Duhovnik at al., 1993). The following four levels of the design process have become established in professional literature: original, innovative, variation and adaptive (Table 1) (Žavbi & Duhovnik, 2001). On the basis of the above design levels, design tasks can be determined and distributed among them.

- **Original design** means the designing of entirely new products, whereby a new working principle is determined for a new or known function. In the process of designing from scratch, one therefore needs to define the working principle, model of shape, functionality and technical shape.
- **Innovative design** means designing products by varying the working principles which fulfil the required function to the optimum degree. In innovative design one needs to define the model of shape, functionality and technical shape.
- **Variational design** means designing products by varying loads, therefore comparable models of shape are obtained. In variational design one needs to define the functionality and technical shape.
- **Adaptive design** means designing products by adapting their dimensions to the technical and technological possibilities for their manufacture. In adaptive design one needs to define the technical shape. This shape is conditioned both by optimization of microtechnology (special features of the manufacturing technology) and by the shape design of details (ergonomics, assembly, etc.). Adaptive design is a dominant type of design (Table 1) and typical of the engineering change process.

The characteristic design phases are: determination of design requirements, conceptual design, embodiment design and preparation of technical documentation (Horvath & Vergeest, 2000). During their work, designers will require different types of support, depending on the phase of design or abstraction of the product they are working on at the time (Rude, 1998), (Suh, 1990).

The type and number of changes that need to be made later are largely determined already in the phase of product development. If a thorough analysis is not performed taking into account all phases of the product's life cycle, from

market requirements and manufacturing technology to maintenance, the number of necessary changes will obviously be greater (Duhovnik & Tavčar, 2002).

	Product levels	Description	Design levels
I.	New product / new technology	New working principles	ORIGINAL DESIGN
II.	New product / known technology	Variation in working principles	INNOVATIVE DESIGN
III.	Variant of an established product	Variation in functionality	VARIATION DESIGN
IV.	Product improvement (change)	Variation in dimensions	ADAPTIVE DESIGN

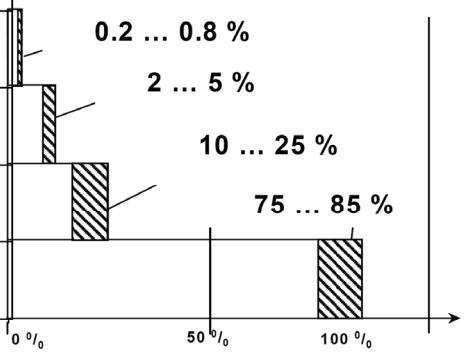


Table 1. Relationship between design and product levels (Duhovnik & Tavčar, 2002).

The entire product family and the possibility of its upgrading have to be envisaged already during the product's conceptual design. A clear presentation of the influence of concurrent engineering methods on the number of changes is given in Prasad's work (Prasad, 1996). For the sake of comprehensive analysis, it should be emphasized that change management begins already during conceptual design and later phases of product development.

3. Generalised engineering change process

Generalized model of engineering change process (figure 2) helps us understand and compare procedures in different types of production and consequently find the most appropriate methods for a specific enterprise. Each change begins with an idea. It is important to stimulate the employees to creativity as well as to ensure an easy collection of ideas and their tracking. Collecting of proposals for changes must be possible and accessible in a simple manner, throughout the company and also from the outside, servicing personnel and salesmen being the most important participants. It is necessary to ensure that proposals are collected centrally and that they are properly documented.

In the next step, the idea itself should be transformed into a proposal for a change. The information system plays and important role in arrangement and collection of the required data. Arranging also includes analyzing and testing, if applicable. It needs to be ensured that each proposal is subject to appropriate professional discussion, which, due to economic reasons, can be conducted in several stages. Each change must go through the process of approval, where the consequences of the change are calculated from all perspectives, e.g. in terms of costs and technical feasibility. Once the change has been approved, it should first be provided for changes in documents and their distribution, following which the change needs to be implemented in the production process, servicing etc.

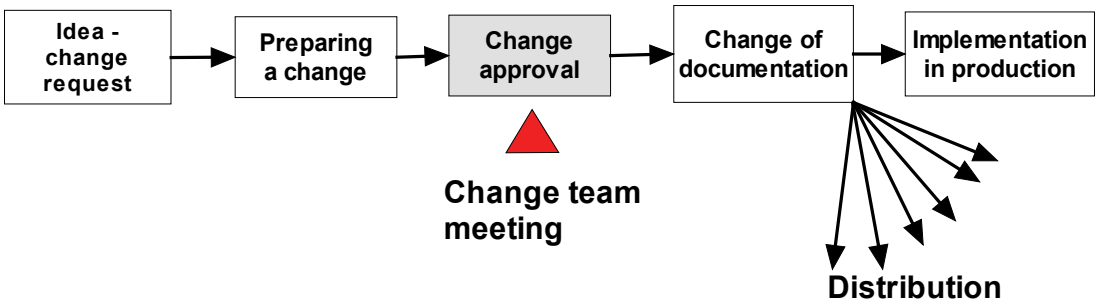


Figure 2. Generalised engineering change process of a product (Tavčar & Duhovnik, 2005)

The objective of this paper is to develop a method that will help distributed companies recognize weak points in their engineering change management system and improve it. Systematic analyses in various companies showed that the criteria presented in figure 3 have to be fulfilled for ECM to be managed well. It is very important for all of the stated considerations namely, communication, decision making, organization, process definition and information system to fulfil the minimum threshold criteria. The impact of an individual criterion depends on the type of production. The quality of communication primarily affects the first three phases of the EC process shown in figure 2. A clear definition of the process and the information system affects all phases of the EC process. Organization has the greatest influence on change preparation, which includes additional research and prototype production.

3.1 Communication

To support developmental-design activities, it is important to be able to identify the relevant communication channels, as well as the frequencies and contents of communication (Frankenberger & Badke, 1998). The predominant type of communication varies considerably with the design level. In new product development, the world outside of the core development team serves as an important source of information, and creative dialogue will predominate. At the level of variants, designers are considerably more limited and dependent on the information that has been organized within the information system; this is even truer in the case of product changes. Poor communication is the most frequent reason for problems in ECM.

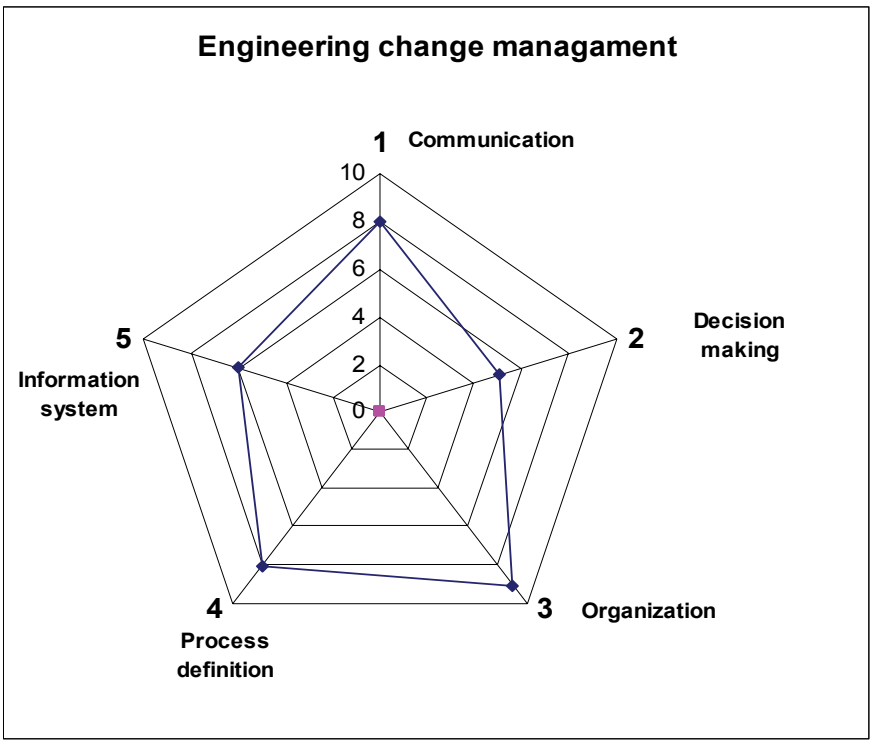


Figure 3. Criteria necessary for effective change management (ECM)

The following forms of communication in EC process were recognized: creative dialogue, review and approval, informing team members and searching for information (Tavčar & Duhovnik, 2000). The type of communication varies

with the phase of engineering changes. During initial stages of the engineering change process, there are many considerations to be taken into account and harmonized, and many decisions to be made. This part of the process cannot be formalized. Informal communication is very important, since it is the source of creativity (Prasad, 1996). Physical proximity between project team members is the best way to accomplish a creative dialogue. In the distributed teams is a creative dialogue enabled with videoconferencing. Later during ECM, especially during distribution, the range of people requiring access to product data becomes wider. For this reason, access to the electronic form of documents and communication via the information system is very important.

Regular and effective communication is the necessary prerequisite for the functioning of virtual teams. Virtual team members need specific skills to communicate and work well (Tavčar at al., 2005). In addition to technical knowledge required to use the communications equipment, special features of work in a virtual team also need to be taken into account, e.g. regular responses, which are important for building trust. Each virtual team member must be independent and must show initiative. Individual skills, such as, for example, knowledge of a foreign language in a multilingual team, cannot be mastered overnight, which should be taken into account as early as team formation. Training in the use of unified software in the entire team (e.g. 3D viewer and red-lining) is needed. The EC leader must prepare the schedules and rules for regular meetings (Kayworth, 2000). Team members must take the time to get to know each other well, because this improves communication and increases the level of effectiveness. In strong personal relationships, communication is frequent but short. Relationships in virtual teams are developed and strengthened through a proactive effort to solve problems (Hart & Mcleod, 2003). Product development and engineering change management requires intense communication; the use of a video system is therefore essential. Based on studies (Harvey & Koubek, 1998), there is no difference between personal face-to-face and video communication in product development. A large difference is seen, however, if only audio or text communication is used. In complex tasks, such as change approval, the type of communication medium employed (e-mail, audio, and video) has a strong impact on effectiveness, while in simpler tasks this has no marked effect (Kayworth, 2000). Communication becomes more effective once the team develops a common vocabulary (Kayworth, 2000).

Successful work in virtual development teams requires certain special skills for team members (Tavcar at al., 2005):

- Willingness to cooperate and work in EC team.
- Effective communication in a virtual team (trust building).
- Initiative and ability to find information and make decisions.
- Mastery of a common spoken and written technical language (similar background is an advantage in communication).
- Working with the communications software
- Ability to access and work with product data.
- Specialised knowledge (compatible with other team members)

As a rule, communication involves a feedback loop between the sender and the recipient. It is very important for effective communication that the sender immediately receives a confirmation that the recipient correctly interpreted the information. Whenever one writes a message, a reply is needed in order to know that the intent of the message was achieved. In a conversation, however, confirmation is often expressed simply through mimics. Within a familiar team, even a small hint will suffice and everyone will understand the message. However, recipients from different cultural environments or different types of expertise will require a clear, modified explanation. Effective communication in an EC team requires as many communication channels as possible: audio, video and textual.

Creativity requires an optimum level of communication (Leenders at al., 2003). Overly intense or overly limited communication reduces creativity. Communication is the driving force of development teams. Both individuals and the team as a whole require an appropriate level of autonomy to develop their creativity. These needs can be fulfilled with regularly scheduled formal and informal communication. Good dissemination of information and distributed communication (each member with all of the others) must be ensured. This is in agreement with a German study (Frankenberger & Badke, 1998) that reports that 80% of a designer's time is composed of routine work that individuals perform independently, and 20% of conflict situations, which need to be solved, and decisions that have to be made. According to studies, designers solve 88% of all problems in co-operation with others (Frankenberger & Badke, 1998), by relying on the experience and knowledge of their co-workers and the synergy effect of the team.

Trust in a virtual team

Individual studies have confirmed that well managed preparation can accelerate the building of virtual teams and thus increase their effectiveness (Huang at al., 2003). During their life cycle, virtual teams pass through various phases, which need to be taken into account during team management. Members of each virtual team are initially strangers to each other, with a considerable degree of mistrust. Effective communication and functioning of the team as a whole begins only when trust develops between the members. Team management must always take into account the phase the team is undergoing at the time. Kasper (Kasper & Ashkanasy, 2001) builds trust in virtual teams on a common business understanding and business ethics. Common business understanding includes a correct understanding of the virtual team's goals, distribution of roles, clear definition of tasks, management, and a joint identity. This needs to be clear right from the start. A virtual organisation requires clearly set rules to enable trust to be built.

3.2 Decision-making in a distributed engineering change management teams

Decision-making is the bottleneck point during the ECM process. Work is more efficient if decisions are made by one person. However, it is difficult for one person to have all of the complex knowledge that is required for such decision-making. It is common practice for decisions to be adopted by a team, an EC committee. However, in this case the danger is that responsibility could be shifted from one person to another. A good process also contains clear delimitations of competencies concerning decision-making and interventions in the case of complications.

The leader of distributed ECM team needs to be additionally trained for work in a virtual team. To ensure engineering change execution, a constant overview over the current status and activities of the individuals is necessary. Appropriate division of work is essential - the interdependence of tasks serves as a source of creativity, but it also brings about greater problems in coordination. The change approval is the main mile stone in the EC process. Representatives from all phases of the product life cycle should be involved in the EC team. At approval process EC team members should have a chance to exchange their opinions. Videoconferencing has more channels of communication and therefore has advantages compared to approval by e-mail. A clear decision-making structure helps to speed up decision-making and EC process (Vilsmeier, 2005).

Distributed EC teams change their team members often. Special attention should be put to activities at initialisation of a new EC team. The goals should be set clearly, adequately trained individuals should be selected, and the necessary infrastructure for communication and work should be provided. For good co-operation, the team members should have complementary, and partially also the same, knowledge. Virtual development work requires careful planning and monitoring. The EC leader should know how work in the teams is going, and distribute information between team members about the project as a whole. Independence between teams makes work more productive, but cross-team communication offers new potential for creativity. Inter-team communication is therefore indispensable. For good functioning of the EC team, personal contacts between the members must be well developed and should provide mutual support. Technology will make work easier, but will not be crucial for the effectiveness of virtual teams (Lurey & Raisinghani, 2001). Another role of project leaders in virtual teams is to ensure building of the team, taking into account the cultural specificities of individual members (Kayworth, 2000). Complex tasks require very intense communication, which can be ensured only in a systematic way. Virtual teams are more effective when they deal with less demanding tasks (Leenders at al., 2003).

3.3 Organization

The organizational structure should support the EC and design processes. For more effective work, it is necessary to separate changes in already products undergoing manufacture from the projects intended for developing new products (Tavčar & Duhovnik, 1999). This division of work can ensure shorter response times. One should be aware that ECs are very unpredictable, which causes variable loads and long response times (Loch & Terwiesch, 1999). Additional research can be especially time consuming. One of the possible solutions in distributed teams is flexible working hours, which are adjusted to the amount of work. An additional useful measure may be for projects to share their employees (sharing resources). This would mainly involve specialists for individual areas, e.g. surface treatment, noise and vibration, especially in the case of technically demanding products. It is recommended that team members should be prepared in advance to tackle typical problems. For good use of the capacities, it is necessary to ensure a good overview over the occupancy and flexibility when work is assigned.

The analysis of changes requires an interdisciplinary approach and excellent communication between the participants. Close connections between the sell, R&D, service, purchase and production departments are very important in the decision-making process, especially when individual variants are discussed. The organizational structure should encourage good communication and quick inclusion of individuals when necessary. An EC passes through several team members and suppliers at various locations. An appropriate approach is usually to assign several persons responsible for implementation of each change; these persons monitor the change and initiate appropriate actions if the process is stalled anywhere. The response should be quick and reliable and capacities should be flexible.

The EC management starts at initiation of a new product development process. An early inclusion of strategic suppliers in the first phases of product development opens new technological possibilities and reduces product development time. If possible, the purchased components which will be used in the serial product as well, should be incorporated already in the prototype. Similarly, tools suppliers are included in the conceptual design phase and minor corrections of the product's shape often prevent unnecessary problems. Developmental teams must be provided with the best possible communications facilities. It is important to keep records on justifications for decisions, for example, and predetermined developmental phases. There records are indispensable in the EC process in the case EC is not executed by the same people as product development. Product development yields product data; these should be documented in an appropriate format and should enable controlled electronic access through all phases of the product life cycle. Figure 6 shows extended product structure with all the necessary documentation. PDM systems (Product Data Management) were developed to enable controlled access to technical documentation (3D models, drawings, documents).

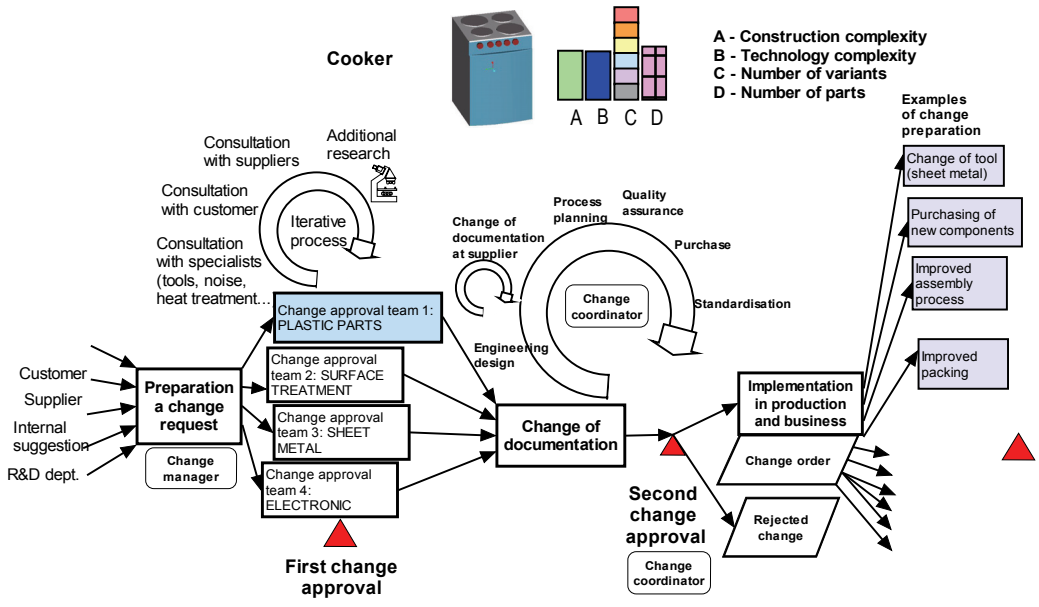
3.4 Process definition

Quick and reliable implementation of ECs requires a detailed process definition, which should be well understood by all participants. The EC process has characteristic milestones, which are presented in figure 2, but the execution should be in line with the company's special features and goals. A common mistake in practice is to use the same process for small changes and for new products. This causes a great deal of waiting and long lead times during change implementation. A clear division of processes and people who are in

charge of them has been proved to be successful. Workflow in an information system significantly contributes to tractability and transfer rates between individual workplaces. It should be taken into account that changes always involve a large degree of unpredictability. It also often turns out that additional research is necessary, as well as cooperation with external suppliers, customer approvals etc. An effective ECM system ensures reliable operation, especially in such exceptional cases.

Case study: Engineering change process in serial production

After completed product development, product data are determined in detail and can be changed only according to strictly defined procedures. There should be no hold-ups in the serial production process. Since the correlation between different fields is high, the production process must be carefully planned and the communication channels must be provided for. Product changes cause a chain of corrections and costs related to tool change, already purchased components, servicing etc. Communication via workflow increases productivity and reliability.



1. Change requests and suggestions come from different sources to change manager.	2. Collecting of product and component stock data and change pre-assessment.	**3. First change approval** by an expert team at adaptive or variation design levels.	4. Change of technical and business documentation serves as additional test of feasibility.	5. Change coordinator controls workflow, stops change process or perform **second change approval** and coordinates change implementation.	6. Change order triggers change implementation.	7. Distribution of technical and business documentation.	8. Date D - change is applied in production.

Figure 4. Engineering change process in the manufacture of household appliances

Serial production of household appliances is usually based on assembling of elements and modules produced by different suppliers. The quality and timing of delivery by several suppliers should be guaranteed. Umbrella companies should be in charge of marketing and development of end products. The product development time is reduced by the transfer of component development to strategic suppliers. The manufacture of household appliances is an example of distributed engineering change management.

In the case of less complex products, the commission for approval of changes can be always the same. In the manufacture of household appliances, however, the range becomes so huge that it is more sensible to form distributed groups for characteristic types of changes, e.g. sheet metal, plastics and surface treatment (figure 4). In this way the working process in smaller groups is more effective. Flexibility can be achieved in different ways: a group of selected specialists can be called according to the problem; virtual group is defined throughout the flexible workflow. The documentation about changes should be transparently accessible in the information system.
In the manufacture of household appliances, there are many design related changes (consumer needs). From the technical standpoint, it is more difficult to control a vast number of changes and the entire logistics than individual changes. A change of documentation simultaneously also constituted a feasibility study, in order to reduce the product development time. With PLM systems and program solutions, the two-phase approach became established: change review and approval in the first phase, and entry of the change in the documentation in the second. Based on the analysis of household appliance manufacturing, the following has been established: Approval regarding the feasibility of a change in a two-phase chain is not the best approach. The most effective way for making a definite decision on change approval is a creative dialogue between the team members. The dialogue can be conducted by means of a video conference (figure 4). Flexible workflow is vital for the process of modifying documentation.

- There is a high degree of unpredictability. Therefore, workflow must be flexible, so that the way can be defined simultaneously, according to the needs.
- An overview of each individual document's status should be provided in terms of its current location. Easy access and user-friendliness are important.

- A change should be implemented in a predefined sequence; however, those included in the process should be able to consult anybody, including external suppliers. In this way, a virtual group is formed and it can function effectively, as if it was located in the same place.

With the large number of variants and also of participants in the process, computer support becomes indispensable for communication. No individual alone can have a good overview of the numerous processes that take place simultaneously. The EC process must be determined with flexible workflow, so that each participant receives only those documents in which he or she needs to make changes. There should be also a user friendly link to related documents, for example product data of the assembly where the changed part is build in. The inclusion of external suppliers in the information system is especially important for good flow of information and effective decision-making, so that these can be independent of the location.

3.4 Information system

The information system constitutes the necessary infrastructure for effective ECM. Based on data from (Huang et al., 2003), (Huang et al., 2001), the analysis of changes in the information system appears to be more an exception than a rule. This is because during any EC, areas from sales and production to development are interwoven. An orderly information system should be upgraded so as to fulfil the specific requirements regarding adaptability of development and orderly production, and this is a very difficult task. The PLM system proved to be very suitable for this purpose, because in addition to change descriptions, all other documentation and product data is also available (figure 5). One must keep in mind that electronic communications have their limitations. This method of communicating is very convenient for informing the team members, but it cannot replace a creative dialogue within the team. Engineering Change approval require an intense communication via videoconferencing.

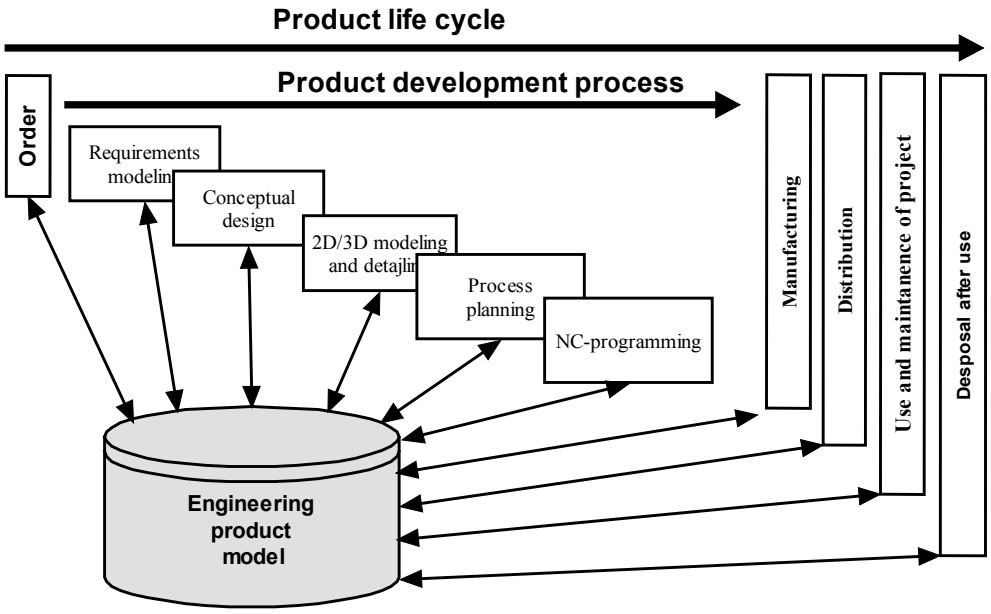


Figure 5. Ready access to product data and close connection to supply throughout the whole product life cycle is important for EC process (Grabowski, 1993)

Comparison and overlapping between PDM/PLM and ERP

A database about the building blocks and products with key information about the products is the main part of the ERP (Enterprise Resource Planning). Software usually allows standard monitoring of material management and production planning process. The use of PDM systems was initially limited to technical departments, support to storing and to accessing files, generated during computer aided design processes. The PDM system later became a tool for management of information on products throughout their lifecycle. PDM (Product Data Management) has been renamed into PLM (Product Lifecycle Management). The physical boundaries of an enterprise are not also a boundary to the information flows. The basic PDM systems user functions are: monitoring the design process and control of later changes, management of products structure, classification and project management.

The PDM and ERP systems are often overlapping (e.g. products structure) during the engineering change process. A good coordination between the two systems is a pre-requisite for a successful work. It is necessary to be able to take advantage of each of the systems and connect them into an effective system (Bartuli & Bourke, 1995).

The higher the integration the larger the volume of data transferred between the PDM and ERP systems, which cause overlapping to a certain degree. It is necessary to make a decision about the prevailing system and how is the master data transferred forward or linked between both systems.

Some ERP systems designers expanded the functionality of their systems also to the PDM sphere (example: SAP). It is necessary to check if the functionality is not limited because production systems are based on different principles.

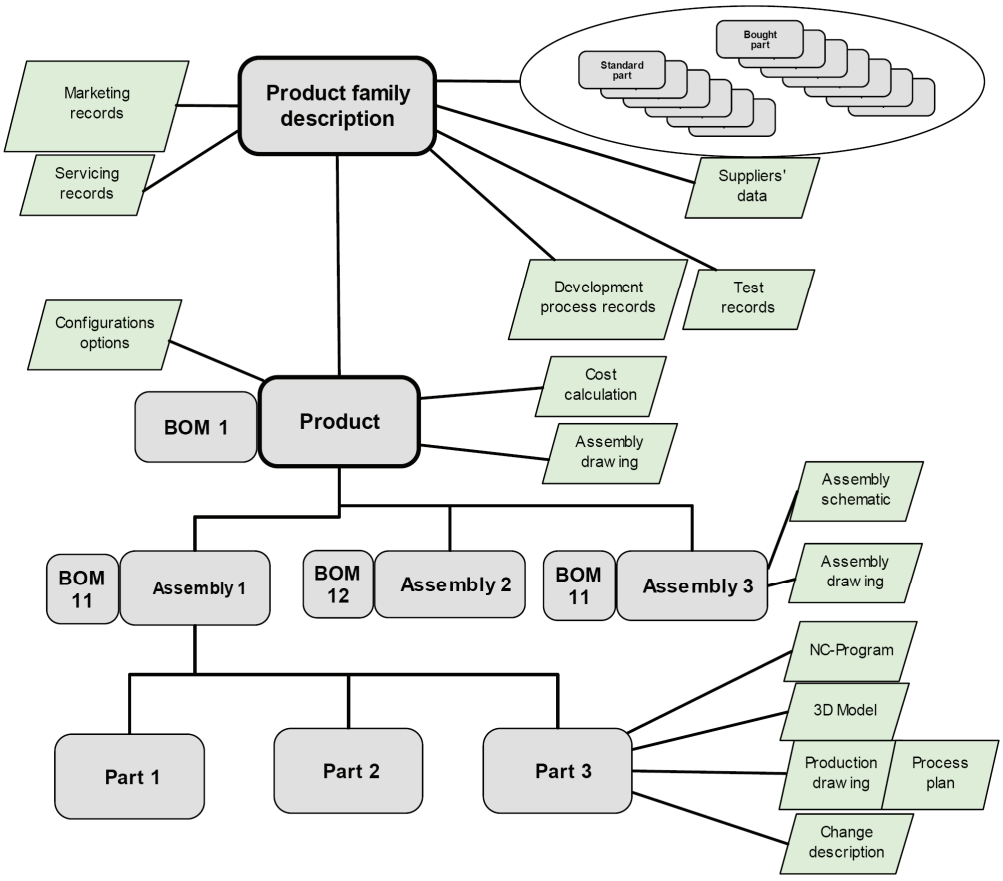


Figure 6. Expanded product structure.

Among the known cases it is not so much about the support during the development phase as it is about storing the results – the verified documentation and control of access for a wide range of users. For the purpose of changes management, a uniform information system is an advantage

Documents change in PDM/PLM system

PLM system is indispensable tool for documents change management. PLM system is with the access control and workflow useful first of all in distributed environment. Figure 7 shows an example of various document statuses of a design drawing from serial production. Access rights, for example, also vary based on the document's status. In the phase Under development, a design engineer can change a 3D model or drawing as much as he wants. Once the document is completed, a design engineer changes its status to Prototype. The document's owner can make changes only if he goes back to the Under development phase. Changes in the document status function as electronic signatures. Document status can be changed only by selected users: only standardiser can change the document status to Approval, and only head of the product development group to Issued. From this stage, the only possible status change is to Invalid or to new version (Revision). This approach ensures that all issued copies are stored. As soon as a document's status is changed to Issued, a TIFF drawing file format is created – this is intended only for viewing and printing. The history of document status changes is also recorded, i.e. their dates and names of the persons who made them. Originals are no longer stamped paper drawings, but PLM computer files with appropriate statuses.

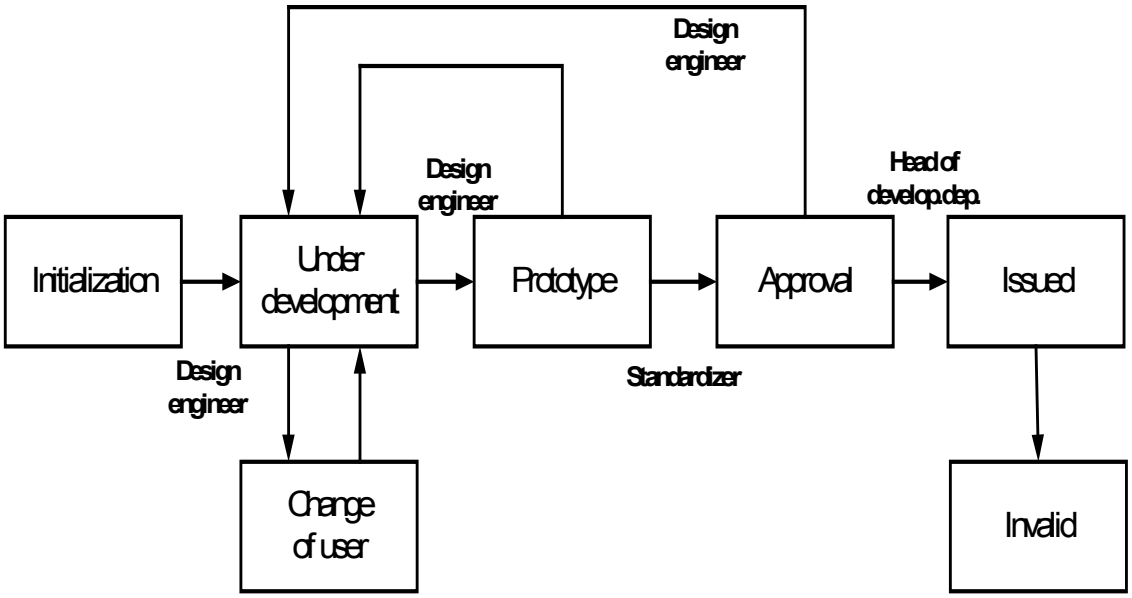


Figure 7. Document approval procedure (case from serial production)

The PLM system proved to be very suitable for this purpose, because in addition to the description of the changes, all other documentation is also available. One must keep in mind that electronic communications have their limitations. This method of communicating is very convenient for informing the team members, but it cannot replace a creative dialogue within the team.

Documentation on the entire product family and the planned variants are the results of product development. Extended product structure serves as the starting point for the preparation of variants. The process for the preparation of variants is considerably shorter, because data and knowledge used was accumulated already during product development. One must also make sure that new findings, e.g. from testing performed on the models, are also entered in the information system in a clear manner, so that data would be easily accessible later. The preparation of variants focuses on the production process, therefore communication between variant developers and other team members takes place mainly via data which is formally entered into the information system.

The PDM system allows movements along the structure and a direct access to a building block, its description and corresponding documents. The product's hierarchical structure is defined by assemblies' part lists. Individual components and assemblies should be marked or documented independently in order to be applicable for new products without any modifications. In the PDM systems with an object user interface, the related building blocks form a structure. A product's model is represented by a structure of building blocks and documents and a building block in a structure is like a card, accessible through descriptions with attributes.

Access to documents control

Technical information systems (PDM) enable several kinds of data handling via security mechanisms. Access rights to data are changing throughout the product's lifecycle. For example, the owner can freely modify his or her documents during the process of creation, however, any modifications of archive data require a special procedure. In the operational systems, there are basic levels of access to files, such as reading, writing, deleting, execution and visibility of files. They are usually different for the owner of a file, compared to other users. There are many more options in PDM/PLM systems. Rights are related to a specific phase in a document's or building block's lifecycle. Different

options allow a number of combinations and consequently different levels of protection. Some of them are presented below:

- **Promote** Shift forward in a product's or document's lifecycle. For example, once the documents have been checked, the head of design moves them from the Checking to Approved status.
- **Demote** Shifting by one or more phases backwards in a product's or document's lifecycle. For example, a designer spots an error in a drawing. In order to make the necessary corrections, it is necessary to go back to the phase In progress where he or she has the right to change the document.
- **Modify** The right to change meta data, i.e. building blocks' or documents' attributes.
- **Lock** A user exports a document to his or her local computer and has the right to lock the document in order to make sure that no-one replaces the file while the corrections are being made.
- **Check in** The right to replace a file in the PDM/PLM system.
- **Create** The right to create a new building block or document.
- **Change owner** A possibility to control the access is through the ownership of documents, which can change during the lifecycle.

Users are divided into groups by tasks, defined by the same priorities, which makes maintenance easier. The characteristic groups are as follows:

- designers
- technologist
- project managers
- PDM/PLM system administrator
- special roles, such as changes manager, responsible for classification
- other employees (manufacturing, purchase of goods, sales departments etc.)
- outside users, such as customers, suppliers

Access to data should be adapted to each user's role. Automated procedures of documents flow and their approval as well as procedures in case of a change can contribute significantly to time efficiency.

Improvements and other changes are always implemented during the regular production process. Therefore, implementation must be performed rapidly, in order to prevent delays. This is enabled primarily by a well informed chain of persons who implement changes in the company, from changes in documentation to those of production (orderly workflow). Access to documents which change with time must be provided in electronic form, along with a well structured system for revisions and distribution (a PLM function). Integration with the ERP system is indispensable (stocks need to be checked, new orders entered and the production plan must be harmonized with the new conditions).

Version control

Control of different documents versions during the engineering changes process requires a special attention. Once the product documentation has been approved, all subsequent changes in the archives are stored in the form of once valid and filed documents. When a change is introduced, a new document version, i.e. building block is released.

The user, responsible for changes, should create a new object engineering change, where a description and reasons for the change are given and a link to the corresponding building block is created (Figure 8). The change is activated only when it has been approved. In the next step, the object engineering change is linked to all documents that should be changed. Later, new versions of the building block and marked documents are created (Figure 8). When new versions are released it is possible to keep the old situation documented. New documents versions are set in the initial position. It enables the owners to make changes, however, the documents revision is required. Only when all the documents have been revised and approved it is possible to adopt the final engineering change.

PDM/PLM systems offer different possibilities regarding the visibility of building blocks and documents. It is possible to set them in such a way that only the last versions are visible. A clear access to old versions is possible via links between building blocks and documents.

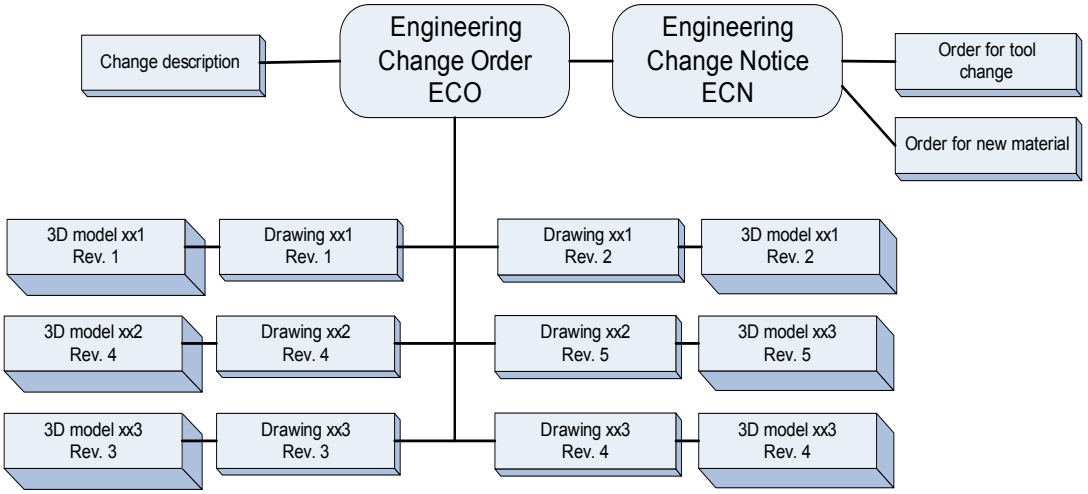


Figure 8. An engineering change object links the old and the new versions of building blocks and corresponding documents

Workflow Management

The role of the PLM system can be illustrated colourfully by comparing it to the conveyor belt in the production process. It is clear to everyone that a good transport between different machines significantly affects the productivity. In the same way, it is necessary to connect jobs within a team in order to introduce changes and provide them with the necessary information. Taking account of the fact that changes are often subject to iterative work – documents often circulate among designers, technologists and quality control – the role of support to documents transfer is even greater.

The PDM/PLM systems control the information flow among different workplaces according to a pre-determined sequence. The majority of routine activities have been automated. However, documents should be in the electronic form in order to allow their transfer within a computer network. The use of such systems makes sense, especially when it comes to absorbing the EC process, dominated by fragmented team work, be it from the location or time point of view. The electronic documents actually do not circulate (they always remain on the PDM/PLM system's WEB server). The announcements are transferred, the access privileges change and access by means of indicators is easier. Workflow management is also supported by office business systems.

Transition to electronic operations in a distributed environment requires the process re-engineering, which takes advantages of all possible technical solutions. Workflow can be divided into the following steps:

- Modelling the information flows
  (sequence of activities, volume of data, frequency of activities…).
- Setting the access rights for users and those in charge of approvals
- Prospective study of data transfer between documents and data bases, without having records in several places.
- Setting-up the necessary network connections and software in order to make the data flow work.

Case study: Workflow in Engineering change process
A proposal for a change can be made by a wide range of users (Figure 9). The proposer should describe the change, state the reasons and provide a rough assessment of costs for its introduction. If necessary, a drawing or any other documents should be attached to the proposal. Information, such as name, date and mark are taken from the system automatically. The proposer should submit the complete proposal for assessment.

Assessment of a proposal for a change
When a proposal comes to the inbox of the person in charge of changes he or she verifies if all the necessary information is there and if the change is sensible. If necessary, the changes should be discussed via a videoconference with the parties concerned. An administrator in charge of the change is appointed and those in charge of the execution are proposed. Once approved, the person in charge of changes replaces the proposal by an order and sends it to the administrator for further processing.

Preparing a change and changing the documentation
The persons in charge of changes are also e-mail recipients. The persons and corresponding workflow can be set simultaneously. Contact persons in the development, technology, control, purchase and standardisation departments are appointed. They divide the labour within the department according to the type of a change and available capacities. For typical changes, the administrator selects the relevant persons in charge of the execution, which accelerates the flow. The physical location of the participants is irrelevant, the only precondition is a high speed Internet connection and access to the PDM/PLM system.

In the PLM system, a proposal or an order for a change is presented as an icon, containing data attributes. The icon is the linking element and all documents,

related to the change, are attached to it. If necessary, drawings, building blocks and documents are added by users involved in the change. Documents can be changed only if they are related to the engineering change order (ECO). In this case, a new documents revision is created (Figure 8). Each change remains documented even after it had been introduced. Each document within the system is stored only once. Interrelations among documents provide for transparency and traceability.

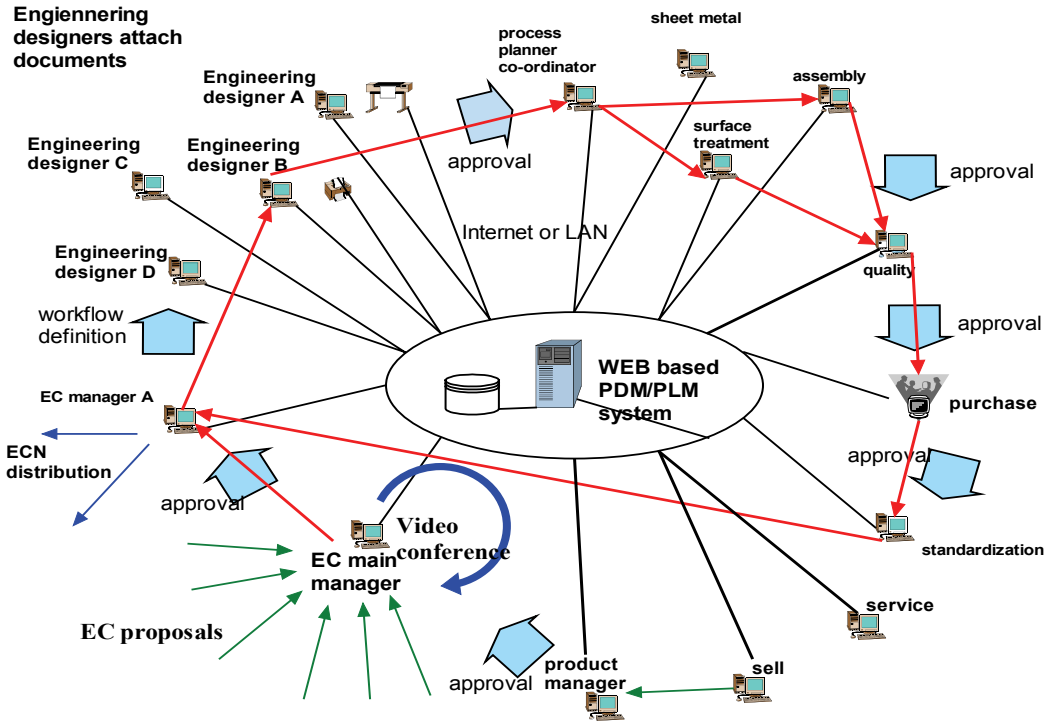


Figure 9. Introducing the changes; users, connected in a chain

Engineering Change Order (ECO) and users notification

When the appointed people have done their job – changes of technical documentation – the ECO with the attached documents comes to the administrator. He or she then makes a proposal for the Engineering Change Notice (ECN), checks the supplies and notifies the concerned. When the administrator takes care of the documents, necessary to introduce a change in the production process, he or she begins the distribution process. First, a list of the ECN recipients is prepared. The users confirm the receipt by the electronic signature.

Different documents, such as tools or manifold request, are attached to the ECN. Through the ECO connection, all modified drawings (Figure 8) are available. Only the ECN is distributed in the electronic form. Afterwards, the relevant document is only a few clicks away.

4. WEB based Information system

The recent advance in Web-based technologies has the potential to greatly enhance PDM's functionality and capability, and to overcome the obstacles that many traditional PDM systems are confronted. A multi-tiered (Figure 10), collaborative environment with advanced Web technologies is well suited for a PDM system to operate in (Xu & Liu 2003). Web based PDM system do not need any additional software installation at clients computer. Software updates need to be executed on servers only, it is indispensable advantage in big systems (Woerner, 2005).

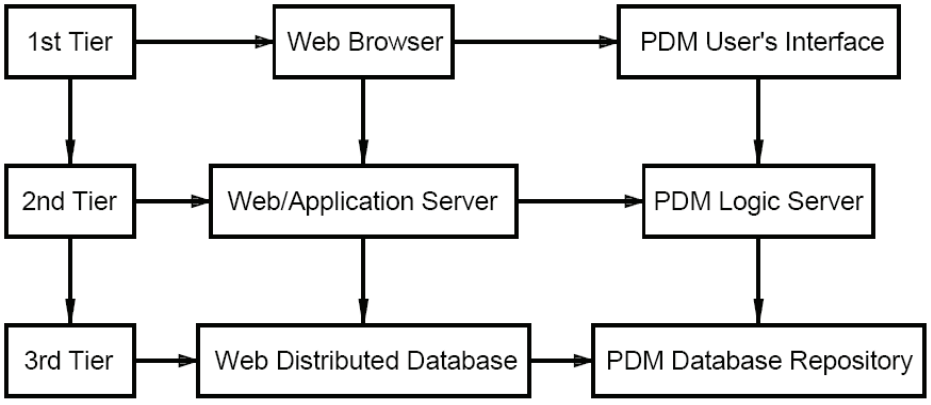


Figure 10. Three-tiered PDM and Web

Distributed engineering change management team consists of different specialists: design engineers, process planners, purchasing, sales, production personnel… Only some of them need to modify data and all team members need to access product data fast and in a user-friendly way. Viewers are a cheap workplace. For example, neither Pro/Engineer licence nor an expensive computer, enabling running the extensive software, is necessary in order to view a Pro/Engineer drawing. It is possible to add comments in the form of layers, without changing the original document. In the PDM systems, the comments

are stored in special files, linked to the corresponding documents. The most popular formats are well supported, which enables switching the layers and assemblies on and off. Integration between the PDM system and a viewer should enable reading and changing the attributes in the PDM system, movement of objects along the lifecycle and access to e-mail. During the process of changes management in the distributed environment, a clear access to all graphic information, related to the EC is of utmost importance. A fast access to a drawing or a 3D model enhances the understanding and clears up many misunderstandings.

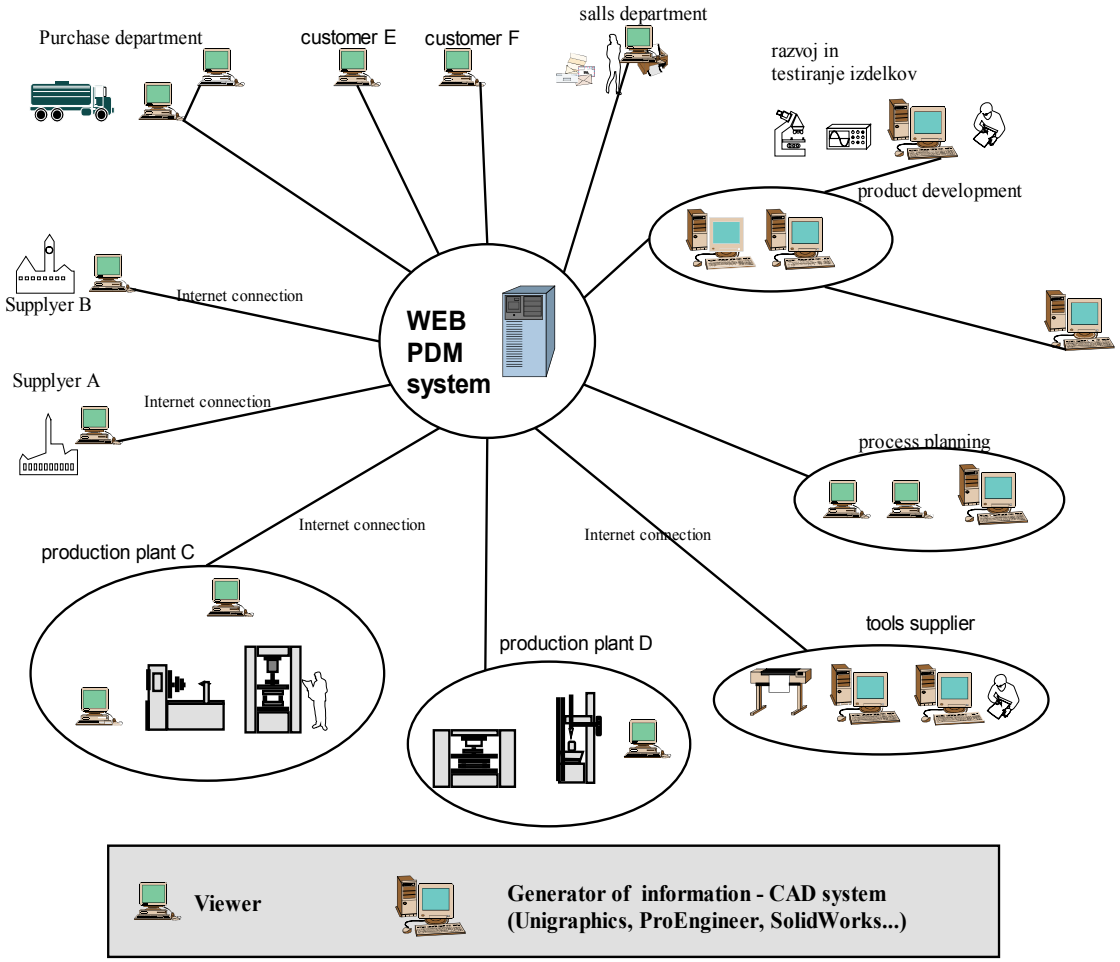


Figure 11. Generators and information users

4.1 Web and security

The added risk to the use of Web and Internet is that the information access is provided across a wide area, unprotected by the security firewalls. As more and more companies rely on the communication and data transmission through Internet, the potential risk of exposing the confidential information to outsiders increases (Leong, 2003). To address this problem, encryption / decryption is incorporated in the system for the user to encrypt a message before sending it out and others to decrypt it on receipt (Woerner, 2005). The encryption reduces the performance of transaction for about 50 %. Other possible solutions include the use of firewall and virtual private network (VPN). Effective application of network security technology in PDM environment can also help increase the confidence of Internet data transmission, and more importantly, the confidence of the PDM users.

Public Key Infrastructure (PKI) enables parties to an e-commerce transaction to identify one another by providing authentication with digital certificates, and allows reliable business communications by providing confidentiality through the use of encryption and authentication, data integrity, and a reasonable basis for nonrepudiation through the use of digital signatures (Carayannis, 2005). PKI uses public/private-key pairs. Public-key cryptography works in such a way that a message encrypted with the public key can be decrypted only with the private key, and, conversely, a message signed with a private key can only be verified with the public key.

Subject to some adjustments, devices that have been developed for the purpose of safe electronic banking can also be applied to the distributed PDM systems. Tools for a safe transfer of data via the Internet are available. First of all, each user should take care of data protection on his or her own computer.

5. Conclusions

Traditional engineering change process is based on teams that work together at the same location and on physical prototyping. Therefore, a move to a virtual environment is difficult. Communication technology is getting better, cheaper and more widely available. Effective work and communication in a virtual team require special skills (Tavčar et al., 2005). The best way to acquire such skills is via personal experience (Horvath et al., 2003). ECM processes require intense communication, which is possible via videoconferencing. New generations of developers need new features for successful work in global

teams. This process can be significantly accelerated by appropriate education, training and management. Independent individuals who know what they can expect from a global development team and are familiar with communication technologies will become involved in such teams with optimism, and full of trust. At the same time, they will be aware of what is necessary to maintain trust and work effectively: quick responses, good dissemination of information to all, and initiative.

A few special features were incorporated in the procedure for implementing engineering changes, and we would like to draw attention to them. The engineering change procedure is divided into two parts; in the first one, decisions are passed, and in the second emphasis is on rapid implementation. When assessments are made, a managed and intensive communication via electronic media is an indispensable part of the reliable and correct decision-making process. In addition to the technical aspects of prototype production, access to all data on the product, both technical and manufacturing data, is also important. The process of changes also includes component and tool suppliers, who can retrieve data directly from the central system. Engineering change is an iterative process at adaptive or variation design level (Duhovnik & Tavčar, 2002). It is important to recognize the design level of EC while it determines action plan and what support is needed for different phases of design process.
Engineering change management is an integral part of each production process, but it is frequently paid insufficient attention. From the technical standpoint, changes here are smaller in extent, but due to the interweaving of various fields, they are that much more important. It was shown on several examples from practice that appropriate communication needs to be provided between all those who are involved in the process of change implementation. This is especially important in the decision-making phase. In order to shorten change implementation times, one can use reserves found in technical implementation, accessibility of information and workflow between the users. Special attention was put on PDM/PLM system support. Communication skills, decision making, organization, process definition and information system were recognized as key factors for efficient ECM in distributed environment.

6. References

Bartuli G., Bourke R., The Best of Both Worlds, Planning for Effective Coexistence of Product Data Management and MRP II Systems, APICS - The

Performance Advantage, APICS, The Educational Society for Resource Management, USA, 1995, Feb., March

Carayannis E. G. and Turner E., Innovation diffusion and technology acceptance: The case of PKI technology, Technovation, In Press, Corrected Proof, Available online 28 July 2005

Duhovnik J., Tavčar J., Reengineering with rapid prototyping, TMCE 2002 : proceedings of the fourth international symposium on tools and methods of competitive engineering, april 22-26, 2002, Wuhan, P.R. China. Wuhan, China , 2002, Pages 117-130

Duhovnik J., Tavčar J., Concurrent engineering in real and virtual tool production. *Concurr. eng. res. appl.*, 1999, vol. 7, no. 1, str. 67-79.

Duhovnik J., Tavčar J., Koporec J., Project Management with Quality Assurance, Computer Aided Design, Butterworth Heinemann, Oxford, Vol. 25, Number 5, pp 311-320, 1993.

Frankenberger E., Badke-Schaub P., Influences on Design Productivity - Empirical Investigations of Group Design Processes in Industry, The Design Productivity Debate, Springer-Verlag London, 1998.

Grabowski H., Anderl R., Polly A. (1993), Integrirtes Produktmodell, Beuth Verlag

Hart R. K. and Mcleod P. L., Rethinking Team Building in Geographically Dispersed Teams: One Message at a Time, Organizational Dynamics, Volume 31, Issue 4, January 2003, Pages 352-361

Harvey C. M., Koubek R. J., Toward a model of distributed engineering collaboration, Computers & Industrial Engineering, Volume 35, Issues 1-2, October 1998, Pages 173-176

Horvath I., Duhovnik J., Xirouchakis P., Learning the methods and the skills of global product realization in an academic virtual enterprise. *Eur. j. eng. educ.*, 2003, vol. 28, No. 1, 83-102. http://www.tandf.co.uk/journals.

Horvath I., Vergeest J. S. M., Engineering design research: anno 2000, Proceeding of the International design conference - Design 2000, Dubrovnik, Croatia, 2000.

Huang G. Q., Yee W. Y. and Mak K. L., Current practice of engineering change management in Hong Kong manufacturing industries, Journal of Materials Processing Technology, Volume 139, Issues 1-3, 20 August 2003, Pages 481-487

Huang G. Q., Yee W. Y. and Mak K. L., Development of a web-based system for engineering change management, Robotics and Computer-Integrated Manufacturing, Volume 17, Issue 3, June 2001, Pages 255-267

Jarvenpaa S. L., Leidner D. E., Communication and Trust in Global Virtual Teams, JCMC

Kasper-Fuehrer E. C. and Ashkanasy N. M., Communicating trustworthiness and building trust in interorganizational virtual organizations, Journal of Management, Volume 27, Issue 3, 6 May 2001, Pages 235-254

Kayworth T. and Leidner D., The global virtual manager: a prescription for success, European Management Journal, Volume 18, Issue 2, April 2000, Pages 183-194

Leenders R. Th. A. J., van Engelen J. M. L. and Kratzer Jan, Virtuality, communication, and new product team creativity: a social network perspective, Journal of Engineering and Technology Management, Volume 20, Issues 1-2, June 2003, Pages 69-92

Leong K. K., Yu K. M. and Lee W. B., A security model for distributed product data management system, Computers in Industry, Volume 50, Issue 2, February 2003, Pages 179-193

Loch C. H. and Terwiesch C., Accelerating the Process of Engineering Change Orders: Capacity and Congestion Effects, Journal of Product Innovation Management, Volume 16, Issue 2, March 1999, Pages 145-159

Lurey J. S. and Raisinghani M. S., An empirical study of best practices in virtual teams, Information & Management, Volume 38, Issue 8, October 2001, Pages 523-544

O'Marah K., Trends in New Product Development and Introduction Processes, AMR Reseach, 2004, Boston, USA

Pirola-Merlo A., Härtel C., Mann L. and Hirst G., How leaders influence the impact of affective events on team climate and performance in R&D teams, The Leadership Quarterly, Volume 13, Issue 5, October 2002, Pages 561-581

Rouibah K. and Caskey K. R., Change management in concurrent engineering from a parameter perspective, Computers in Industry, Volume 50, Issue 1, January 2003, Pages 15-34

Prasad B., Concurrent Engineering Fundamentals, Vol. I Integrated product and process organization, Technomic, Lancaster, USA, 1996.

Rude S., Wissenbasiertes Konstruieren, Shaker Verlag, Aachen, 1998

Smith P. G. and Blanck E. L., From experience: leading dispersed teams, Journal of Product Innovation Management, Volume 19, Issue 4, July 2002, Pages 294-304

Suh N. P., The Principles of Design, Oxford University Press, Inc., New York,1 1990.

Tavčar J., Duhovnik J., Model of Communication in the Design Process, International conference Design 2000, Dubrovnik, Croatia, 2000.

Tavčar J., Žavbi R., Verlinden J., Duhovnik J., Skills for effective communication and work in global product development teams. *J. eng. des.)* [Print ed.], 2005, vol. 16, No.. 6, 557-576. http://www.tandf.co.uk/journals.

Tavčar J., Duhovnik J., Engineering change management in individual and mass production. Robot. comput.-integr. manuf.. [Print ed.], 2005, letn. 21, št. 3, str. 205-215. Http://www.sciencedrect.com/science/journal/07365845.

Terwiesch C. and Loch C. H., Managing the process of engineering change orders: the case of the climate control system in automobile development, Journal of Product Innovation Management, Volume 16, Issue 2, March 1999, Pages 160-172

Tosse T., Product Data Management as a Platform for Global Product Lifecycle Management, ProductData Journal, Cross-Comain Engineering, Vol. 12, No. 2, 2005, ProSTEP iViP Association, Darmstadt

Vilsmeier J., Change and Configuration Management of Software and Hardware for the Eurofighter, ProductData Journal, Cross-Comain Engineering, Vol. 12, No. 2, 2005, ProSTEP iViP Association, Darmstadt

Xu X. W. and Liu T., A web-enabled PDM system in a collaborative design environment, Robotics and Computer-Integrated Manufacturing, Volume 19, Issue 4, August 2003, Pages 315-328

Woerner J. and Woern H., A security architecture integrated co-operative engineering platform for organised model exchange in a Digital Factory environment, Computers in Industry, Volume 56, Issue 4, May 2005, Pages 347-360

Wright C., A review of research into engineering change management: implications for product design, Design Studies, Volume 18, Issue 1, January 1997, Pages 33-42

Žavbi R., Duhovnik J., Model of conceptual design phase and its applications in the design of mechanical drive units. V: LEONDES, Cornelius T. (ur.). Computer-aided design, engineering, and manufacturing : systems techniques and applications. Vol. V, The design of manufacturing systems. Boca Raton [etc.]: CRC Press, cop. 2001, Pages 7/1-38.

27

Study of Flexibility and Adaptability in Distributed Supply Chains

Felix T. S. Chan and H. K. Chan

1. Introduction

Uncertainties in demand and supply, which are two major contributions to system dynamics, are unavoidable attributes in supply chains. Agent technology has been a renowned enabler to achieve flexibility and adaptability, which are regarded as the distinctive characteristics of future supply chains, to overcome system dynamics (Chan and Chan, 2005). In order to testify the usefulness of these characteristics, a series of simulation study have been conducted by the authors to investigate the effects of these two characteristics on distributed supply chains, which are subject to uncertainty. In fact, this article aims at presenting the simulation results and drawing conclusion in relation to these two characteristics on supply chain dynamics.

The research motivation of this article originates from two reported research. Chan and Chan (2005) performed a survey on related literature, and concluded that agent technology would be a potential problem solver in modelling future supply chains. They then developed a multi-agent based simulation model for supply chains (Chan and Chan, 2004). By making use of this model, they followed the same line of research direction by introducing flexibility and adaptability through coordination mechanisms in their investigation. As a pilot study, a simulation study with the said flexibility in a single product environment has been reported (Chan and Chan, 2006). This chapter further extends their study with focus on a multi-product environment. Simulation results indicated that introduction of flexibility in due date and quantity is able to reduce the total cost of the system under study, as compared with traditional stochastic model which makes use of safety stock to counteract with system dynamics. Like flexibility, additional adaptability could be able to improve the performance of the supply chain further, with even better improvement.

The organisation of the rest of this chapter is as follows: Section 2 presents re-

lated literature. The research methodology, i.e. the simulation model, will be briefly explained in Section 3. Simulation results and key findings will be discussed in Section 4 to Section 6: results with respect to flexibility study are summarised in Section 4; effects on information sharing will be discussed in Section 5; and results in regards to adaptability will be presented in Section 6. Section 7 is the concluding section for future research direction.

2. Literature Review

2.1 Distributed Problem Solving in Supply Chains

Supply chain can be viewed as a network of participating corporations working together to achieve the system goals. It can be defined as a "connected series of activities which is concerned with planning, coordinating and controlling of materials, parts and finished goods from supplier to customer" (Stevens, 1989). Supply chain management aims at optimising all activities through the supply chain, so that products and services are supplied in the right quantity, to the right time, and at the optimal cost. In this connection, coordination among supply chain members is of vital importance. Due to the distributed nature of global supply chain, agent technology has been employed to model supply chains in some reported literature. As a matter of fact, agent technology provides channels for integrating the independent echelons of the entire supply chain as a networked system (Gjerdrum *et al*., 2001). Mutli-agent system (MAS), a branch of Distributed Artificial Intelligence, consists of more than one autonomous agent. One of the critical research challenges in a large portion of agent-based applications is coordination (Tambe et al., 1999).

MAS is a typical example of distributed problem solving technique that gains high attention in recent supply chains research. Swaminathan et al. (1998) presented a multi-agent approach to model supply chain dynamics. They developed a supply chain library of software components such that customised supply chain models can be built from the library. Sadeh et al. (2001) presented an agent-based architecture for dynamic supply chain called MASCOT (Multi-Agent Supply Chain cOordination Tool). MASCOT is a re-configurable, multi-level, agent-based architecture for coordinated supply chain. Agents in MASCOT serve as wrappers for planning and scheduling modules. Above

mentioned researches are focusing on the architectural issues and lacking of higher coordination mechanism, which is a common weakness in many agent-base research in the supply chain domain. One reason may due to the fact that coordination is more problem specific and it is not easy to generalise a theory for different supply chains. Nevertheless, agents in a MAS is loosely coupled and are not controlled by a central controller, it is easy to loss distributed functions. Coordination is an effective tool to prevent the system from such problem, i.e. chaotic behaviour in agent's terminology.

2.2 Information Sharing in Supply Chains

Information transfers among independent companies in supply chains tend to be distorted and can be misguided up-stream members regarding their inventory and production decisions, which is the well known Bullwhip Effect (Lee *et al*., 1997). It is commonly believed that information sharing may reduce the impact of demand uncertainty (Lin et al., 2002). However, information sharing among companies is not always possible because of privacy of corporate information, and trust among corporations. In addition, incompatibility among heterogeneous information systems can also hinder information sharing among them. Therefore, information sharing has been over-emphasised as a generic cure for supply chain dynamics (Raghunathan, 2001).

After Lee et al. (1997) had coined the Bullwhip Effect, Lee et al. (2000) conducted another study to analyse how information sharing can improve the supplier's order quantity decision in a two-level supply chain, with a known autoregressive demand process. In respond to their study, Raghunathan (2001) showed that the manufacturer could make use of its own information with respect to the entire order history in order to reduce the variance of its forecast. As a consequence, there is no need to make investment for sharing information. More research on information sharing with respect to supply chains can be found. Cachon and Fisher (2000) studied a supply chain subject to stationary stochastic demand. They compared a traditional information policy that does not use shared information against a policy with full information sharing. They observed that share information among supply chain members could reduce cost. They, however, argued that implementing information technology to accelerate and smooth the physical flow of goods through a supply chain, i.e. simply flowing goods through the supply chain more quickly and more evenly, may produce greater improvement than sharing information.

2.3 Coordination in Supply Chains

Quantity / price discount is a common strategy to provide coordination channel among supply chain members. Quite a lot of research could be found with respect to discount policy. For example, Viswanathan and Piplani (2001) considered an incentive policy such that a vendor offers a discount to buyers if they place orders only at the times as specified by the vendor. One common weakness of the reported research with such channel coordination is that deterministic demand is assumed. Therefore, impact of system dynamics on the proposed model has not been studied. Facing uncertain demand, for example, retailers prefer to place an order late in most case in order to gather enough time to collect more information (Chen and Xu, 2000). However, this leads to insufficient production times and hence production cost would probably be increased.

Coordination can also exist in the form of contracting. Quantity flexibility contract "provides flexibility with no explicit penalty for exercise, by adopting constraints as a way to motivate appropriate behaviour" (Tsay, 1999). By introducing quantity flexibility, the retailer can place an order earlier due to the flexibility that is introduced in the quantity range and the supplier may only need to finish the order with quantity that is within the committed range. In addition, the retailer may request less quantity of goods to be shipped if the actual demand is lower than what is expected. This philosophy, which is also the research direction of this study, can provide incentive to both supplier and retailer.

2.4 Research Direction

Effective coordination strategies will be very important for agents in next-generation of multi-agent systems (Lesser, 1998). These agents will need to be highly adaptive due to their ''open'' operating environments where the configuration and capabilities of other agents and network resources could be changed dynamically. One of the ways that such agents can be adaptive is to consider multiple ways of solving their sub-problems so that they can adjust their solution to produce the best possible result, subject to the restrictions on available processing, communication, and information resources, etc. In fact, quantity flexibility as proposed in this study is one of the possible ways to provide agents with a set of possible solutions so that the best solution could be finalised dynamically through the proposed coordination mechanism.

Agents can also be more adaptive if they are not restricted to solving one goal at a time, but are able to flexibly arrange their activities to solve multiple goals concurrently. This is exactly the idea of the proposed adaptive coordination mechanism in this study.

Based on these findings, Chan and Chan (2006) studied the effects of demand and supply uncertainties as independent variables in an agent-based supply chains with single product type, and suggested a coordination mechanism with quantity flexibility to react with such uncertainties. It was found that the performance of the supply chain under studied outperforms the same one with stochastic model, where the performance measures are total cost. This study is a natural extension of previous Chan and Chan's work (2006) whereas multi-product supply chain system will be studied here. By employing the same quantity flexibility approach to the said system, it is found that total system cost is improved when compared with the stochastic model, which is in line with their findings.

In addition, the effects of information sharing on the proposed coordination mechanism have been studied as a benchmark. Finally, adaptability nature has also been added in the proposed mechanism in order to further improve the system performance.

3. Supply Chain Model

3.1 The Agent-based Model

As mentioned before, this study makes use of the agent-based model which was developed by Chan and Chan (2004). Since the main focus of this article is not on the agent-based model, only a brief sequence diagram as quoted in Chan and Chan (2006) is included as in Fig. 1. For detail discussions on the agent-based model and associated operations, please refer to Chan and Chan (2004). The agent-based simulation program was written in JAVA.

Description of events	Agent which announces the job		Other agents
[a] Order is needed, announce job to other agents	Job Announcement Agent	→	Job Reception Agent
[b] Job Reception Agent relays the job to Bid Evaluation Agent			↓
[c] Other agents are considered to submit a bid			Bid Evaluation Agent ↓
[d] Other agents submit a bid according to their own conditions	Bid Reception Agent	←	Bid Evaluation Agent
[e] Deadline for bid submission is reached (no more bids will be accepted)	Bid Reception Agent		
[f] Notify Job Award Agent	↓		
[g] Job Award Agent sends offer to other agents according to the ranked bids	Job Award Agent	→	Bid Evaluation Agent
[h] Acknowledge of job offer	Job Award Agent	←	Bid Evaluation Agent
[i] Contract is made	↓ Incoming Contract Agent		↓ Outgoing Contract Agent

Figure 1. Simplified sequence of operations among agents (Source: Chan and Chan, 2006)

3.2 The Supply Chain

In the simulation study, the model consists of three customers and four suppliers. Total number of product types is three. Simulation study has been carried out to verify the usefulness of the proposed flexibility and adaptability concept. Length of simulation is 465 periods while the first 100 periods are ignored for calculation in order to minimise the start-up effect. The final performance measures are based on the last 365 periods (i.e. T = 365). If one period is equal to one day, then the effective length of simulation run is one year. Each simulation setting will be run with 10 different random seeds and the average is reported in order to minimise the random effect. Together with the 16 sets of independent variables (to be discussed in Section 3.3), a total of 160 simulation runs were carried out for each strategy. Since there are 3 sets of strategy, total number of simulation runs is 160 x 3 = 480 sets. In fact, more simulation runs have been conducted (e.g. against different capacity levels as discussed in Section 3.3.), however, only the two independent variables are the main focus of this study.

3.3 Dependent and Independent Variables

Total system cost is recorded as the dependent variable for comparing different coordination mechanisms.

Setting	Demand Uncertainty	Supply Uncertainty
1	1	1
2	1	2
3	1	3
4	1	4
5	2	1
6	2	2
7	2	3
8	2	4
9	3	1
10	3	2
11	3	3
12	3	4
13	4	1
14	4	2
15	4	3
16	4	4

Table 1. Different settings of the simulation study

On the other hand, there are two independent variables in this study, namely, demand uncertainty and supply uncertainty (i.e. variation of capacity of each supplier). Each of these variables is modelled by varying the variance of the corresponding Normal distribution at four levels (from 1 to 4 where 1 means the least uncertain and 4 is the most uncertain). Therefore, there are 16 sets (4 x 4) of different simulation runs for each strategy as summarised in Table 1 in regard to the demand and supply uncertainties. Different settings mean different combination of uncertain demand and supply as shown in Table 1. The higher the number, the higher is the degree of uncertainty as expressed in terms of variance (or standard deviation) of the associated Normal distribution. In addition, capacity level is also modelled as the third independent variable. However, only representative results will be presented since this parameter is relatively insensitive with respect to this study.

4. Simulation Results with Flexibility

4.1 The Coordination Mechanism

In the order-up-to stochastic model, the so-called order-up-to level in fact consists of a basic quantity plus a safety stock, as illustrated in equation (1):

$$S = \mu (T_o + L) + \upsilon \sigma \sqrt{(T_o + L)} \qquad (1)$$

where, S is the re-order level;
μ is the mean of demand;
σ is the standard deviation of demand;
υ is the service level that the retailer would like to achieve;
T_o is the review period;
L is the order lead time.

The rationale behind is to use the safety stock (the latter term in equation (1)) as a buffer to compensate the effect of uncertainties. An order is placed every T_o period and the ordered quantity, Q, is the difference between S and the the inventory position, which is the sum of all exisitng inventory or backordered inventory and the total ordered quantity in all outstanding orders, below this re-order level. Therefore, the stochastic model inherently increases inventory cost. Intuitively, the stochastic model is not dynamic enough because demand

is unpredictable due to its random nature. In this connection, quantity flexibility is introduced in the coordinated model in order to provide the flexibility to the retailer, as well as suppliers to react with system dynamics. In order to apply this coordination mechanism, the supply chain members must be coordination oriented, but no explicit information sharing is required. In the coordinated model, similar procedures are followed as in the stochastic model, with the following alteration:

In order to simply the following dicussion, the following discussions only foucs on a single product environemnet as in Chan and Chan (2006), but the same analysis applied to multi-product environemnt. When a job is announced, it consists of a range of quantities required instead of a fixed quantity. Equation (1) can be rewritten as the following equations:

$$S = \mu\,(T_o + L) + \upsilon\,\sigma\,\sqrt{(T_o + L)} = A + B \tag{2}$$

$$A = \mu\,(T_o + L) \tag{3}$$

$$B = \upsilon\,\sigma\,\sqrt{(T_o + L)} \tag{4}$$

The range of quantity Q is defined such that:

$$Q \in [A - B,\, A + B] \tag{5}$$

Equation (5) defines the "domain" of the variable "quantity" that the retailer requires the supplier to be shipped. In addition, the retailer will calculate a range of delivery dates so that supplier should ship the quantity as defined in equation (5) within the range of deliver due dates. The range can be defined as in equation (6):

$$\begin{aligned}&[\textit{Expected Delivery Due Date} - (\mathrm{B} / \mu),\\ &\text{Expected DeliveryDue Date} + (\mathrm{B} / \mu)]\end{aligned} \tag{6}$$

where *Expected Delivery Due Date* is given by equation (7):

$$\textit{Expected Delivery Due Date} = D_{it} + \frac{Q}{\textit{Mean Capacity}} \tag{7}$$

where D_{it} is the longest due date of supplier i at period t in its outstanding order
Q is the difference between S and inventory position

Refer to the above discussion, the range of quantity is set in relation to the safety stock, B, in the stochastic model. Therefore, an apple-to-apple comparison can be made between the proposed coordination mechanism and the stochastic model, which was employed as benchmark in later discussions. In fact, sensitivity analysis of this value (i.e. a small variation from B) has been conducted. In addition, different settings of the value of B (e.g. such as expressed as a percentage of the base quantity A) have been conducted as well. It was found that the results and trends of improvement are consistent with whatever value of B, with small difference in magnitude of the performance metrics, of course. Therefore, only the results with one-side wide equal to the safety stock is presented in this chapter.

The remaining procedure is the same as the stochastic model until lower bound of the due date in equation (6) of an outstanding order reaches. The retailer starts to coordinate with the supplier when and how many to be shipped. This turns out to define the final values of two variables – one is quantity Q, and the other is the date for shipment D. The two variables are distributed among the retailer and supplier under contract. Domain of Q is given by equation (5) and let Q_{low} (i.e. A – B in equation (5)) and Q_{high} (i.e. A + B in equation (5)) be the lower bound and upper bound respectively. Domain of the date for shipment is given by equation (6) and let D_{low} and D_{high} be the lower bound and the upper bound respectively. The objective is to solve this problem through a coordination mechanism. An outline of pseudo code is illustrated in Fig. 2.

```
t = t + 1;
coordination( )
    if (t∈[Dlow, Dhigh]) then … (i)
        if (t = Dhigh) then  … (ii)
            Iit = get supplier inventory( )
            if (Iit ∉ [Qlow, Qhigh]) then … (iii)
                    penalise supplier ( )
            exit( )
```

```
        else if (It = get my inventory ( ) > μ) then … (iv)
                exit ( )
            else
                Iit = get supplier inventory( )
                if (Iit ∉ [Qlow, Qhigh]) … (v)
                    penalise supplier ( )
                exit ( )
    else
        exit( )
end coordination ( )

get supplier's inventory( )
    if (t∈[Dlow, Dhigh]) then
        It = get my inventory ( )
        if (Iit > Qhigh)
        Iit = Qhigh
    else
        exit ( )
    return Iit
end get supplier's inventory( )
```

Figure 2. An outline of pseudo code for coordination (Source: Chan and Chan, 2006)

Condition (i) in Fig. 2 constrains the coordination to be taken place only if the due date is within the domain in equation (6). Condition (ii) ensures the coordination phase is ended when the upper bound of the due date in equation (6) reaches. In such case, outstanding order must be completed. Condition (iv) makes sure the retailer does have enough inventory if no shipment is made when D_{high} is not reached. Please note that conditions (iii) and (v) of the pseudo code allow the supplier to supply with quantity less than the defined domain, subject to penalty being incurred, if the inventory of the supplier less than the lower bound as stated in equation (5). This is a constraint relaxation and hence the new domain of Q is effectively become $(0, Q_{low}]$, i.e. any positive integer below Q_{low}. The reason to accept this argument is to ensure that the mechanism is complete and sound, i.e. the algorithm can always returns a solution. Of course, both the retailer and the supplier would not like to relax the constraint, if

if possible, because both will suffer – the retailer gets less product and the supplier makes a loss due to the penalty.

4.2 Simulation Results

Fig. 3 depicts the percentage improvement of the proposed coordination mechanism with quantity and due date flexibility as compared with the stochastic model in terms of total cost. Positive values mean the proposed coordination mechanism could reduce the total cost as compared with the stochastic counterpart under different settings. Please note that three groups of results could be found in Fig. 3. They are actually the results from different capacity level as sensitivity analysis. In fact, the results concur the results as in Chan and Chan (2006), which only study a supply chain with single product type.

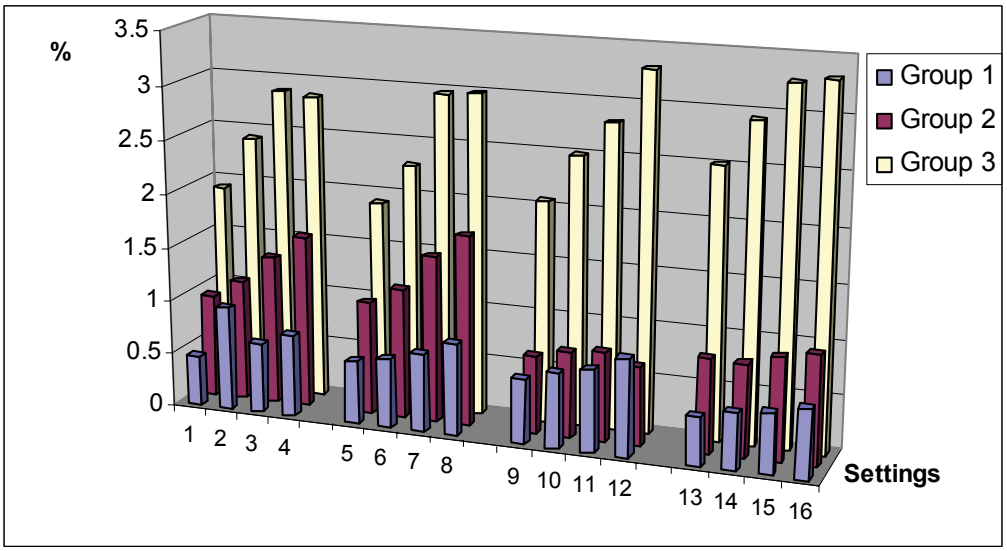


Figure 3. Percentage improvement of the coordination mechanism with flexibility against the stochastic model

5. Simulation Results with Information Sharing

As discussed in Section 2, information sharing is regarded as a solution facing system dynamics. The main objective of this section is to investigate whether the proposed coordination mechanism with flexibility could only perform better than the one with flexibility and information sharing together. Not surpris-

ingly, the answer is "no". However, the difference may not be so significant if the technical constraints of implementing information sharing (e.g. investment and trust) are taken into considerations. In fact, if we consider the stochastic model is the lower bound of the model under study we could assume the model with information sharing is the upper bound, in terms of improvement subject to system dynamics.

5.1 The Coordination Mechanism

The coordiation mechanism with flexilibity in Section 4 assumes no information sharing among agents. The main focus of this section is to relax this assumption and compare the effects of two information sharing schemes. The rationale of allowing information sharing together with the coordination mechanism with flexibility is due to the fact that supplier may not need to produce the upper bound of the quantity range of a certain product type for a particular contract. This is because the customer turns out may request the supplier to ship less and hence excessive inventory may produce. If a supplier can complete a contract at a proper level, though the supplier may not necessarily ship the product according to the contract terms as defined in the coordination mechanism, slack capacity for next order can then be "created".

Two negotiation-based information sharing schemes are studies. They are:

(i) NEG1

only the inventory information of the customer and the supplier who are involved in the negotiation can share information. When the middle of the quantity range reaches, the supplier sends a message to the customer to ask for inventory level. The supplier makes the decision based on the total inventory level of the customer and the supplier to decide stop production or not. In fact, decision is made based on the expected total cost in a short time horizon. Equations (8) and (9) give the total cost of a customer j (Z_j)and supplier i (Z_i) over a period of time T respectively:

$$Z_j = \sum_{t=1}^{T} \sum_{p} (h_{jp} I_{jpt} + b_{jp} B_{jpt}) \qquad (8)$$

$$Zi = \sum_{t=1}^{T} \sum_{p} h_{ip} I_{ipt} \qquad (9)$$

where h_{jp} is the unit inventory holding cost per period of product type p of customer j

h_{ip} is the unit inventory holding cost per period of product type p of supplier i

b_{jp} is the unit backorder cost per period of each product type p of customer j

I_{jpt} is the inventory level of product type p of customer j at period t

B_{jpt} is the backorder level of product type p of customer j at period t

I_{ipt} is the inventory level of product type p of supplier i at period t

Assume current period is at $t = 1$ and T is the deadline of the order or contract under consideration. In each negotiation cycle, the supplier develops a matrix of T x $T = \{C_{xy}\}$ such that x and $y = 1$ to T. Each element is the expected total cost (i.e. $Z_j + Z_i$) such that production is stopped at time x, and the contract is finished and delivered at time y. I_{jpt} and I_{ipt} are reduced or increased, if needed, according to the mean demand of the customer and mean capacity of the supplier respectively. Invalid elements are marked so that they are not eligible for later decision. From this matrix, the supplier can recognise the short term total cost and then is able to select the one with the lowest cost as the decision at this period. In other words, if it is not suggested to stop production at this period, the supplier will continue to produce a product and then reiterate the same negotiation at each period, i.e. update the matrix and reduce the size every period, until T is reduced to 1. Of course, the final delivery date depends on the retailer as well, which is governed by the original coordination mechanism.

(ii) NEG2

inventory information of all agents in the systems are sharable. Same as NEG1, when the middle of the quantity range reaches, the supplier sends a message to the customer for collecting all information on the inventory level of other agents. After the customer gathers all information, it is passed to the supplier. The supplier makes the decision based on the total inventory level of the all agents to decide stop production or not. Therefore, the cost equation is exactly the same to the

one in NEG1, but all agents are taking into consideration. Strictly speaking, this information sharing scheme is not really "full" information sharing because only inventory information is available. However, "full" is in respect of the inventory level. Decision making is the same as the one as in NEG1.

5.2 Simulation Results

Fig. 4 illustrates the percentage improvement of NEG1 and NEG2 as compared with the coordination mechanism with flexibility only. It was found that both information sharing scheme with flexibility outperforms the coordination mechanism with flexibility only. Although both NEG1 and NEG2 could reduce the total cost further, it could not be concluded that neither NEG1 nor NEG2 is the best one. In other words, both information sharing scheme with flexibility perform comparably in terms of total cost, and no single information sharing scheme is the dominant policy. In addition, the further cost reduction is not that significant, especially at the left hand side of the graph, at which demand is less uncertain. Some further improvement is even lower than 10%. Considering the investment that has to make to achieve information sharing, information sharing may not be that attractive because of its insignificant improvement in certain settings. However, if the demand variability is high (i.e. at the right hand side), it is still a good policy to overcome the impact of system dynamics.

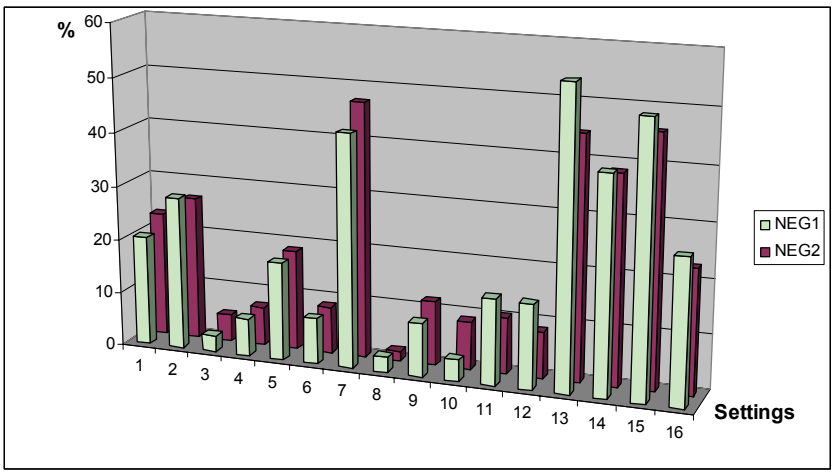


Figure 4. Percentage improvement of the two information sharing mechanism with flexibility against the coordination mechanism with flexibility only

The simulation results also support the argument at the very beginning of this section: If the stochastic model is the lower bound for benchmarking the performance of the coordination mechanism with flexibility, information sharing would be the upper bound.

6. Simulation Results with Flexibility and Adaptability

6.1 The Coordination Mechanism

This section summarises the principle of the adaptive coordination mechanism. As described in Section 5 above, the suppliers in fact have flexibility to allocate slack capacity for producing the next order to be processed on hand, as compared with a fixed quantity in the stochastic order-up-to policy. The rationale behind the proposed adaptive coordination mechanism is to "create" slack capacity artificially. In other words, production process of a product would stop before the maximum quantity is produced and switch to produce the product of the next planned order. One may argues if the supplier stop production of the current order at the minimum quantity of the range would results in more slack. However, this would only result in shorter and shorter ordering cycle because customers keep receiving less quantity in each ordering cycle. Therefore, a balanced scheme has to be designed in order to come up with a compromise between production quantity of the current order and the slack capacity for next order.

With information sharing as discussed in Section 5, this is relatively easy to achieve. However, without information sharing, an additional adaptive coordination mechanism is desired. In other words, the adaptive coordination mechanism helps the customers and suppliers to make the following decision: When should a supplier stop production of a product if the lower bound of the quantity range of the current order reaches, and then switch to production for next order? In the original coordination scheme, the customer takes the initiative to request completion of an order, unless deadline of a contract is due. In the adaptive coordination mechanism, this assumption is relaxed.

The supplier is able to send a similar request to the retailer once the supplier has produced middle of the quantity range in a contract, provided that the supplier has another outstanding. This is a signal to the customer that the supplier would like to stop production of the current order at a quantity lower than the upper limit of the contract. Since half of the range is equal to the

safety stock quantity, the customer then calculate the deviation of its current inventory level (i.e. I_{jpt}) from the safety stock and take one of the following actions:

(i) If the difference is positive which means customer's inventory level is higher than expected, then, the customer accepts the supplier's request. However, shipment is not made instantly. It still follows the original coordination mechanism because the customer still has the flexibility to request for shipment. In other words, the supplier who made the request is suffering from inventory cost for a short period of time.

(ii) In contrast, if the difference is negative, the customer would refuse the request and then shipment, as in the case (i) still governed by the original mechanism.

This scheme is adaptive because decision is based on the real-time situation, rather than on the planned schedule. Together with the quantity flexibility that is introduced, the overall scheme is flexible and adaptive.

6.2 Simulation Results

Fig. 5 depicts the simulation results in regard to the adaptive coordination mechanism.Basically, the proposed adaptive coordination mechanism with flexibility performs better than the one with flexibility only at different settings and different parameters.

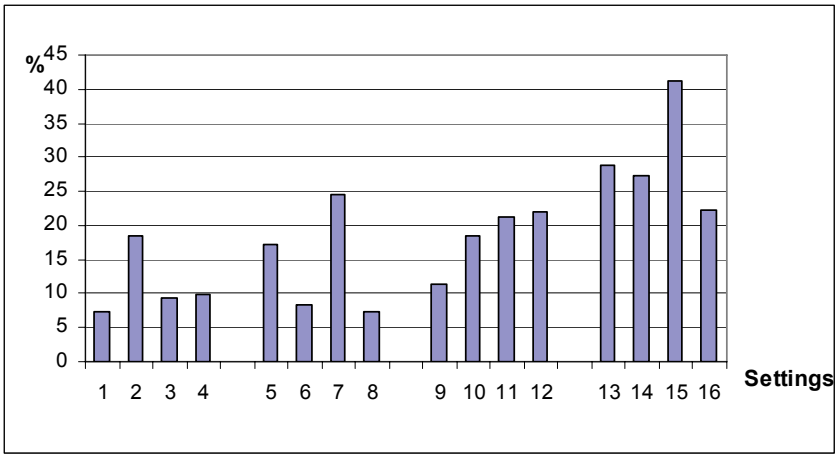


Figure 5. Percentage improvement of the adaptive coordination mechanism with flexibility against the coordination mechanism with flexibility only

However, only the percentage improvement in total cost of one instance is shown in Fig. 5 for simplicity. From Fig. 5, it is clear that the adaptive coordination mechanism outperform the one with flexibility only in all settings. In addition, results are even more promising at high demand uncertainty, i.e. the right-hand-side of Fig. 5.

7. Conclusions

A number of managerial implications could be drawn from this simulation study. They can be highlighted as below:

1. The core contribution of this study is introduction of flexibility and adaptability nature through a coordination mechanism for distributed supply chains in inventory management so that delivery decision (how many and when) of outstanding order is negotiable. This dynamic nature is proven, through simulation study, to be effective in reducing total system costs. Although traditional stochastic modelling is a means to reduce the total system cost by establishing safety stock in the system, it is not dynamic enough when the system is facing uncertainties. With the help of advanced information technology, the proposed mechanism is not difficult to implement.
2. By investigating the effects of information sharing as discussed in this paper, we found that information sharing with flexibility could perform even better in term of cost reduction as compared with the coordination mechanism with flexibility alone (and hence also better than the stochastic model). However, partial information sharing may perform considerably well as compared with full information sharing, subject to the same flexibility. By considering the investment and technical limitation of full information sharing (e.g. trust), it is not necessarily to pursue full information sharing all the time.
3. Regarding information sharing, another critical issue is to define the correct information to be shared for decision making. Of course, it is easier to say than to implement this in practice. However, the philosophy behind is intuitive.
4. Information sharing is in fact not the only solution. The performance of the adaptive coordination mechanism with quantity flexibility (i.e. the one in Section 6) is not worse than the one with information sharing (i.e. NEG1 and NEG2 in Section 5) subject to the same flexibility. Again, considering

the investment to achieve information sharing, the adaptive coordination mechanism or even the flexible coordination mechanism (i.e. the one in Section 4) would be a more feasible and economic solution.

The research findings can be strengthened in the future by employing more complex supply chain structures for testing. More sources of uncertainties could be added in the system for analysis. For example, unexpected events (e.g. supply interruption) can be modelled as another source of uncertainty in order to verify the research hypothesis regarding flexibility, information sharing, and adaptability in this paper under different scenarios. As a matter of fact, this simulation study is just a piece of proof-of-concept. It is worthwhile to use real data which can be obtained in real cases to verify the achieved simulation results as a future work.

8. References

Cachon, G. P. & Fisher, M. (2000). Supply chain inventory management and the value of shared information. *Management Sciences*, Vol. 46, No. 8 (August 2000), pp. 1032-1048, ISSN: 0025-1909.

Chan, F. T. S. & Chan, H. K. (2004). A new model for manufacturing supply chain networks: a multiagent approach. *Proceedings of the Institution of Mechanical Engineers Part B: Journal of Engineering Manufacture*, Vol. 218, No. 4 (April 2004), pp. 443-454, ISSN: 0954-4054.

Chan, F. T. S. & Chan, H. K. (2005). The future trend on system-wide modelling in supply chain studies. *International Journal of Advanced Manufacturing Technology*, Vol. 25, No. 7-8 (April 2005), pp. 820-832, ISSN: 0268-3768.

Chan, F. T. S. & Chan, H. K. (2006). A simulation study with quantity flexibility in a supply chain subjected to uncertainties. *International Journal of Computer Integrated Manufacturing*, Vol. 19, No. 2 (March 2006), pp. 148-160, ISSN: 0951-192X.

Chen, J. & Xu, L. (2000). Coordination of the supply chain of seasonal products. *IEEE Transactions on Systems Man, and Cybernetics - Part A*, Vol. 31, No. 6 (November 2000), pp. 524-531, ISSN: 1083-4427.

Gjerdrum, J.; Shah, N. & Papageorgiou, L. G. (2001). A combined optimisation and agent-based approach to supply chain modelling and performance assessment. *Production Planning and Control*, Vol. 12, No. 1 (January 2001), pp. 81-88, ISSN: 0953-7287.

Lee, H. L.; Padmanabhan, V. & Whang, S. (1997). Information distortion in a supply chain: The bullwhip effect. *Management Science*, Vol. 43, No. 4 (April 1997), pp. 546-558, ISSN: 0025-1909.

Lee, H. L.; So, K. C. & Tang, C. S. (2000). The value of information sharing in a two-level supply chain. *Management Science*, Vol. 46, No. 5 (May 2000), pp. 626-643, ISSN: 0025-1909.

Lin, F. R.; Huang, S. H. & Lin, S. C. (2002). Effects of Information Sharing on Supply Chain Performance in Electronic Commerce. *IEEE Transac-tions on Engineering Management*, Vol. 49, No. 3 (August 2002), pp. 258-268, ISSN: 0018-9391.

Lesser, V.R. (1998). Reflections on the Nature of Multi-Agent Coordination and Its Implications for an Agent Architecture. *Autonomous Agents and Multi-Agent Systems*, Vol. 1, No. 1 (March 1998), pp. 89-111, ISSN: 1387-2532.

Raghunathan, S. (2001). Information sharing in a supply chain: a note on its value when demand is nonstationary. *Management Sciences*, Vol. 47, No. 4 (April 2001), pp. 605-610, ISSN: 0025-1909.

Sadeh, N. M.; Hildum, D. W.; Kjenstad, D. & Tseng, A. (2001). MASCOT: an agent-based architecture for dynamic supply chain creation and coordination in the internet economy. *Production Planning and Control*, Vol. 12, No. 3 (April 2001), pp. 212-223, ISSN: 0953-7287.

Stevens, G.C. (1989). Integrating the Supply Chain. *International Journal of Physical Distribution & Materials Management*, Vol. 19, No. 8 (August 1989), pp.3-8, ISSN: 0269-8218.

Swaminathan, J. M.; Smith, S. F. & Sadeh, N. M. (1998). Modeling supply chain dynamics: a multiagent approach. *Decision Sciences*, Vol. 29, No. 3 (Summer 1998), pp. 607-632, ISSN: 0011-7315.

Tambe, M.; Adibi, J.; Al-Onaizan, Y.; Erdem, A.; Kaminka, G. A.; Marsella, S. C. & Muslea, I. (1999). Building agent teams using an explicit teamwork model and learning. *Artificial Intelligence*, Vol. 110, No. 2 (May 1999), pp. 215-239, ISSN: 0004-3702.

Tsay, A. A. (1999). The quantity flexibility contract and supplier-customer incentives. *Management Sciences*, Vol. 45, No. 10 (October 1999), pp. 1339-1358, ISSN: 0025-1909.

Viswanathan, S. & Piplani, R. (2001). Coordinating supply chain inventories through common replenishment epochs. *European Journal of Operational Research*, Vol. 129, No. 2 (March 2001), pp. 277-286, ISSN: 0377-2217.

28

An Autonomous Decentralized Supply Chain Planning and Scheduling System

Tatsushi Nishi

1. Introduction

For manufacturing industries, the integration of business processes from customer-order management to delivery (Supply Chain Management) has widely been received much attention from the viewpoints of agile and lean manufacturing. Supply chain planning concerns broad activities ranging network-wide inventory management, forecasting, transportation, distribution planning, production planning and scheduling, and so on (Jeremy, 2001; Simon et al., 2000). Various supply chain models and solution approaches have been extensively studied in previous literature (Vidal & Goetschalckx, 1997). These models are often divided into the following three categories:

1. Integration of production planning among several companies (Mckay et al., 2001)
2. Integration of production planning of multi-sites in a company (Bok et al., 2000)
3. Integration of production planning and distribution at a site from the procurement of raw materials, transportation to the distribution of intermediate or final products to the customers (Rupp et al., 2000).

The purpose of this work is to address an autonomous decentralized systems approach for integrated optimization of planning and scheduling for multi-stage production processes at a site with respect to material requirement planning, production scheduling and distribution planning. One conventional approach that has been used for planning and scheduling is a hierarchical decomposition scheme (Bitran & Hax, 1977). Planning concerns decisions about the amount of products to be produced over a given time so as to maximize the total profit. Scheduling involves decisions relating to the timing and sequencing of operations in the production processes so as to satisfy the production goal that is determined by the planning system. Tan (2001) developed a

hierarchical supply chain planning approach and a method of performance management.
Since in the hierarchical approach, there is practically no feedback loop from the scheduling system to the planning system, the decision made by the scheduling system does not affect the decision at the planning stage; however, the decision made by the planning system must be treated as a constraint by the scheduling system. Therefore, it becomes difficult to derive a production plan taking the precise schedules into account for the hierarchical systems. It is necessary to integrate the scheduling system and the planning system for global optimization of the supply chain (Wei, 2000).

A simultaneous multi-period planning and scheduling model has been proposed by Birewar & Grossmann (1990) where the scheduling decisions are incorporated at the planning level. It has been demonstrated that the planning profit is significantly increased when planning and scheduling decisions are optimized simultaneously. The disadvantage of their approach is that the planning and sequencing model is restricted to a certain class of simple problems, because an extremely large number of binary variables are needed to solve integrated planning and scheduling problems. Moreover, it is requested that the models of subsystems comprising an SCM system need to be flexible to deal with the dynamically changing environment in a practical SCM. The integrated large-scale models, however, often become increasingly complex. As a result, it becomes very difficult to execute the new addition of constraints and/or the modifications of the performance criterion so as to cope with unforeseen circumstances.

SCM systems must satisfy new requirements for scalability, adaptability, and extendibility to adapt to various changes. If the decisions taken at each subsystem are made individually while aiming to optimize the entire SCM system, it is easy for each subsystem to modify its own model in response to various requirement changes. Distributed planning and scheduling systems have been proposed as an architecture for next-generation manufacturing systems (NGMS). These architectures are often referred to as multi-agent systems, wherein each agent creates each plan locally within the shop floor and each agent autonomously resolves conflicts among plans of other agents in a distributed environment.
Hasebe et al. (1994) proposed an autonomous decentralized scheduling system that has no supervisory system controlling the entire plant with regard to creating schedules for multi-stage production processes. The system comprises a

database for the entire plant and some scheduling subsystems belonging to the respective production stages. Each subsystem independently generates a schedule for its own production stage without considering the schedules of the other production stages. However, a schedule obtained by simply combining the schedules of all production stages is impracticable in most cases. Therefore, the scheduling subsystem contacts the subsystems of the other production stages and obtains the schedule information of those stages to generate a new schedule. Schedules are generated at each stage and data are exchanged among the subsystems until a feasible schedule for the entire plant is derived. The effectiveness of the autonomous decentralized scheduling system for flowshop and jobshop problems is discussed by Hasebe et al. (1994).

An autonomous decentralized supply chain optimization system comprising three subsystems: material requirement planning subsystem, scheduling subsystem and distribution planning subsystem has been developed. A near-optimal plan for the entire supply chain is derived through the repeated optimizing at each subsystem and exchanging data among the subsystems. In Section 2 we briefly review distributed planning and scheduling approaches. Supply chain planning problem is stated in Section 3. The model structure and the optimization algorithm of the autonomous decentralized system are developed in Section 4. In Section 5 we compare the proposed method with a conventional planning method for a multi-stage production process. Section 6 summarizes conclusion and future works.

2. Distributed planning and scheduling

There have been several distributed planning and scheduling approaches under the international research program called Intelligent Manufacturing Systems (IMS). For example, the biological-oriented manufacturing system (BMS) is an evolutionary approach that contains DNA-type information and BN-type information acquired at each subsystem (Ohkuma & Ueda, 1996). For the holonic manufacturing system (HMS), intelligent agents called “holons” have a physical component as well as software for production planning and scheduling. A hierarchical structure is adopted to reduce complexity and to increase modularity (Gou et al., 1998; Fisher, 1999).

These distributed planning and scheduling approaches can be classified into hierarchical, non-heterogeneous, and heterogeneous algorithms according to the structure of the distributed systems (Tharumarajah & Bemelman, 1997).

Distributed Asynchronous Scheduling (DAS) is organized by three hierarchical agents: operational, tactical, and strategic agents. The constraints are propagated by the message passing through DAS schedulers (Burke & Prosser, 1990). The non-heterogeneous structure is used as a combination of distributed agents and the conflict coordinator when the coordination between the subsystems cannot be resolved. Maturana & Norrie (1997) addressed a mediator architecture where coordination of subsystems is dynamically achieved by employing the virtual systems created as needed for coordination. On the other hand, a heterogeneous structure resolves all conflicts among the subsystems without any other subsystems. Smith (1980) proposed a contract net protocol where each heterogeneous agent negotiates with another by receiving and awarding bids.

The algorithms of distributed planning and scheduling approaches can be classified into non-exhaustive or exhaustive approaches according to the conflict resolution and coordination method. In the non-exhaustive algorithm, the number of attempts at coordination is limited to the number of trials required for obtaining a feasible solution without consuming computational expenses (Shaw, 1987). The exhaustive algorithm is founded on the iterative-search based coordination method for obtaining a near-optimal solution, though the solution may only produce a locally optimal solution. The approach employed in this paper is an exhaustive approach with a heterogeneous structure having no supervisory system. The supply chain planning problem for a single-stage production system can be decomposed into a material requirement planning subproblem, a scheduling subproblem and a distribution planning subproblem following the principle of Lagrangian decomposition and coordination approach based on the mathematical programming method (Nishi et al., 2003). This method has been applied to planning and scheduling methods in many previous studies (Gupta et al., 1999; Gou et al., 1998; Hoitomt et al., 1993).

The autonomous decentralized approach features the characteristic that each subsystem has an optimization function for each subsystem based on the idea of decomposition and coordination. Most of the conventional distributed approaches have a hierarchical structure, where a supervisory system or a coordinator makes a decision by using the information obtained by the subsystem. Even though the decisions are created by each subsystem, it is still necessary to use some protocols for coordination. For conventional systems, it is necessary to reconstruct these protocols when the new constraints or the performance criterion is modified. By adopting the structure of the proposed system, the

proposed approach has a plenty of flexibility to accommodate various changes such as modification of constraints or performance criteria in each subsystem. In the following section, the supply chain optimization problem is stated. Then, the mathematical formulation of the problems is described.

3. Supply chain planning problem

The multi-stage flowshop production process is divided into multiple production stages by taking into account the technical and/or managerial relationships in the plant shown in Figure 1. In this study, we assume that the plant satisfies the following conditions.

1. Total planning period is divided into a set of time periods. For each time period, the lower and the upper bound of the production demand of products are given. If the amount of delivery is lower than the lower bound, some penalty must be paid to the customer.

2. Transportation time and transportation cost from supplier of raw material to the plant, and from the plant to customers are negligible.

3. The lead-time at the supplier of raw material is negligible. However, the ordered raw material arrives at the production process only on a prespecified date.

4. Production site has flowshop production line. Each production stage consists of a single batch machine. The amount of product per batch and the production time depend on the product type of the job, but they are fixed for each product.

5. Changeover costs at each stage depend on the product type of the operation executed successively.

6. The capacity of the storage space for raw materials and final products is restricted. Therefore, the amount of storage of each raw material or final product must be lower than its upper bound. The storage cost is proportional to the amount of stored material and the stored period.

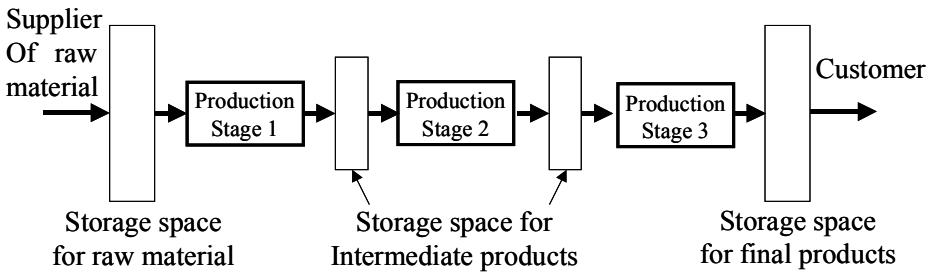


Figure 1. Supply chain for a multi-stage production processes

The supply chain optimization problem for a multi-stage production process is stated as:

The time horizon of planning and scheduling, the lower and upper bound of demand for products, the price of raw materials, inventory holding cost for raw materials, inventory holding cost for final products, the revenue of final product to customer, penalty cost for violating the lower of demand, processing time of operations for each products, changeover cost are given, the problem is to determine the arrival time and the amount of each raw material to storage space for each raw material, the production sequence of operations and their starting times at each production stage, the delivery time and the amount of each product to customers from the storage space for final products to optimize the objective function consisting of material cost, inventory holding cost for raw materials, sequence dependent changeover cost at the production stage, inventory holding cost for final products, production cost, penalty of production shortage.

To solve the above supply chain optimization problem, an autonomous decentralized supply chain optimization system is developed. The details of the proposed system are explained in the following section.

4. Autonomous decentralized supply chain planning and scheduling system

Supply chain optimization problems naturally involve the coordination of production, distribution, suppliers of raw material, and customers. Clearly, each of these sections has its own characteristic decision variables and an objective function relating to other sections. To achieve an efficient supply chain

management, a plan must be developed under the environment which each section is allowed to make independent decisions to its operation so as to optimize its own objective function while satisfying constraints of other sections. (Androulakis & Reklaitis, 1999). Taking this consideration into account, an autonomous decentralized supply chain optimization system for multi-stage production processes is developed. The supply chain planning problem is decomposed into a material requirement planning subproblem, a scheduling subproblem and a distribution planning subproblem when the material balancing constraints are relaxed following the principle of Lagrangian relaxation method (Nishi et al., 2003). Each subproblem is solved by the subsystem.

4.1 System structure

The structure of the system is shown in Figure 2. The total system consists of a database for the entire plant, a material requirement planning subsystem (MRP) and some scheduling subsystems (SS) for respective production stage, and a distribution planning subsystem (DP). The purpose of the MRP subsystem is to decide the material order plan so as to minimize the sum of the material costs and inventory holding costs of raw materials. The SS subsystem determines the production sequence of operations and the starting times of operations so as to minimize the changeover costs and due date penalties. The purpose of the DP subsystem is to decide the delivery plan of each product so as to maximize the profit including inventory costs for final products. The model structure of the decentralized supply chain optimization system is shown in Figure 3.

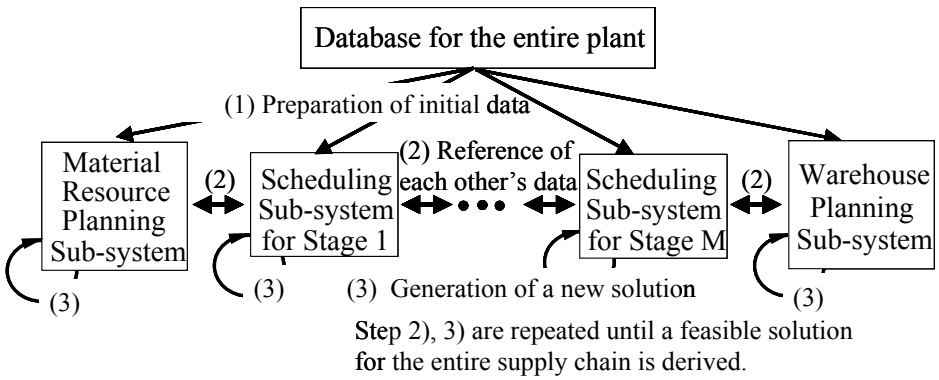


Figure 2. System structure of autonomous decentralized supply chain planning and scheduling system

Each sub-system has own local decision variables and an objective function. The decision variable and the objective function at each sub-system are also denoted in Figure 3.

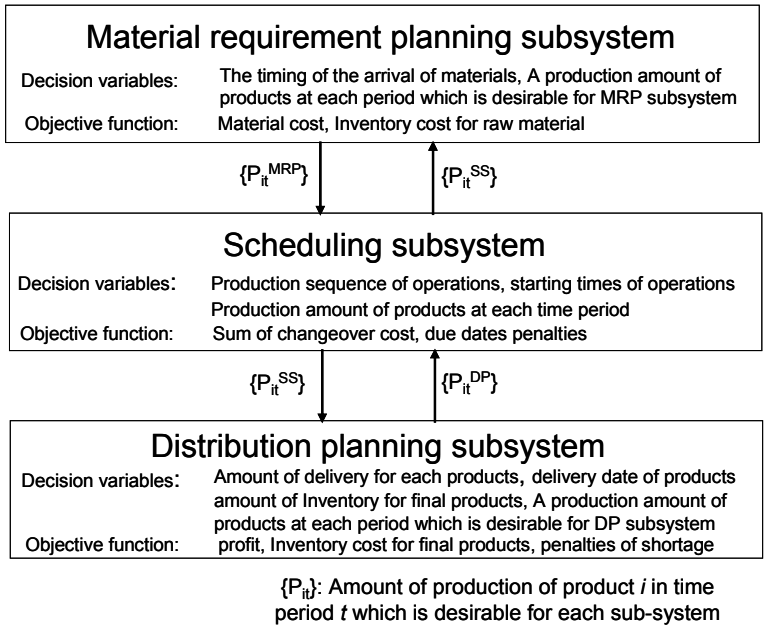


Figure 3. Model structure of the autonomous decentralized supply chain planning and scheduling system

Each subsystem generates a solution of own local optimization problem. However, if the solutions generated at sub-systems are combined, the obtained solutions are infeasible in most cases. To make the solution feasible, each subsystem contacts the other subsystems and exchanges the data among the sub-systems. The data exchanged between the subsystems is the amount of products produced in each time period $P_{i,t}$ derived at each sub-system. The superscripts MRP, SS and DP for $P_{i,t}$ indicate the data generated at the MRP subsystem, the scheduling sub-system, and the DP subsystem respectively. For the proposed system, both of data exchange and re-optimization at each subsystem are repeated several times until a feasible solution for the entire plant is derived. While repeating the data exchange and the re-optimization at each subsystem, penalties for violating the constraints among the subsystems are increased. If the solutions derived at subsystems satisfy feasible conditions for the entire plant, the proposed system generates a total plan and a schedule for the entire plant by combining the solution of all sub-systems. The detail of each subsystem is explained in the following section.

4.2 Material requirement planning subsystem

Material requirement planning subsystem determines the timing and amount of raw material arrived at the production process in each time period. $M_{r,t}$ represents the amount of raw material r arrived at the start of time period t, $C_{r,t}$ represents the amount of inventory for raw material r at the end of time period t, and $P_{i,t}^{MRP}$ represents the production amount of product i in time period t which is calculated by MRP subsystem. $Y_{r,t}$ denotes the 0-1 variables indicating whether material r is arriving at the start of time period t or not. Therefore, the optimization problem at the MRP subsystem is formulated as the following mixed integer linear programming problem (MILP).

$$(\mathrm{MRP})\ \min\left(\sum_{r,t} p_{r,t} M_{r,t} + \sum_{r,t} q_{r,t} C_{r,t} + \rho \sum_{i,t} PN_{i,t}\right) \tag{1}$$

$$C_{r,t} = C_{r,t-1} + M_{r,t} - \sum_{i \in U_r} P_{i,t} \quad (\forall r, \forall t) \tag{2}$$

$$M_{r,t} \le M_{r,t}^{\max} \cdot Y_{r,t} \quad (\forall r, \forall t) \tag{3}$$

$$PN_{i,t} \ge \left| P_{i,t}^{MRP} - P_{i,t}^{SS} \right| \quad (\forall i, \forall t) \tag{4}$$

$$C_{r,t} \le C_{r,t}^{\max} \quad (\forall r, \forall t) \tag{5}$$

$$\sum_{t} Y_{r,t} \le m_r \quad (\forall r)\ Y_{r,t} = \{0,1\}\ (\forall r, \forall t) \tag{6}$$

$$M_{r,t},\ C_{r,t},\ P_{i,t}^{MRP},\ PN_{i,t} \ge 0 \quad (\forall i, \forall r, \forall t) \tag{7}$$

where,

- I_r: set of products produced from material r,
- m_r: maximum number of the arrivals of raw material r in the total time horizon,

- $p_{r,t}$: price of the unit amount of raw material r from supplier to the pro duction process at the start of time period t,
- $P_{i,t}^{SS}$: amount of product i produced in time period t, which is obtained from the SS subsystem,
- $PN_{i,t}$: penalty for infeasibility of the schedule between $P_{i,t}^{MRP}$ and $P_{i,t}^{SS}$,
- $q_{r,t}$: inventory holding cost of unit amount of raw material r for the dura tion of time period t,
- U_r: set of products produced from material r,
- ρ: weighting factor of the penalty for violating the schedule derived at MRP subsystem and SS subsystem.

4.3 Scheduling subsystems

In this section, the scheduling algorithm of the SS subsystem is explained. The flowshop scheduling problem for the SS subsystem is formulated as:

$$\text{(SS)} \quad \min\left(\sum_k Ch^k + \rho\sum_{i,t}(\left|P_{i,t}^{MRP} - P_{i,t}^{SS}\right| + \left|P_{i,t}^{SS} - P_{i,t}^{DP}\right|)\right) \tag{8}$$

$$t_i^k - t_j^k \geq s_j^k \vee t_j^k - t_i^k \geq s_i^k \quad (\forall i, \forall j, \forall k, i \neq j) \tag{9}$$

Where

- Ch^k is the sequence dependent changeover cost at stage k,
- s_i^k is the processing time of job i at stage k.

The second and third terms in Eq. (8) indicate the penalty for the infeasibility of the schedule of SS subsystem with MRP subsystem, and with DP subsystem respectively. Eq. (9) indicates the sequence constraints of operations. The number of jobs for each product is not fixed in advance. Thus, at first, jobs are created by using the production data: $P_{i,t}^{DP}$ obtained from DP subsystem. The number of jobs for each product i is calculated by $\sum P_{i,t}^{DP} / V_l$, where V_l is the volume of the unit at the production stage l. The due date of each product is calculated so that the production amount of each product satisfies its due date. The earliest starting time of each job is calculated by $P_{i,t}^{MRP}$ in the same manner.

The above procedure makes it possible to adopt the conventional algorithms for solving the scheduling problem. In this paper, the simulated annealing method is used to solve the scheduling problem at each stage.
A scheduling subsystem belonging to each production stage generates a near-optimal schedule for respective production stage in the following steps:

1. Preparation of an initial data
   The scheduling subsystem contacts the database for the entire plant and obtains the demand data, such as product name. By using these data, each scheduling subsystem generates the list of jobs to be scheduled. Each job has its earliest starting time and due date. Each job is divided into several operations for each production stage. For each operation, the absolute latest ending time of job j for stage k, represented by ALET: F_j^k is calculated. Here, ALET is the ending time for the stage calculated under the condition that the job arrived at the plant is processed without any waiting time at each stage. ALET means the desired due date for each operation at each production stage.

2. Generation of an initial schedule
   Each scheduling subsystem independently generates a schedule of its own production stage without considering the schedules of other stages.

3. Data exchange among the subsystems
   The scheduling subsystem contacts the DP subsystem and MRP subsystem, and obtains $P_{i,t}^{DP}$: the production amount of products which is desirable for DP subsystem and $P_{i,t}^{MRP}$: the production amount of products which is desirable for MRP subsystem. By using these data, each scheduling subsystem modifies the list of jobs to be scheduled. Each job is divided into several operations for each production stage.
   The scheduling subsystem belonging to production stage contacts the other scheduling subsystems and exchanges the following data.
   a) The tentative earliest starting time (TEST) for each job j: e_j^k
      The ending time of job j at the immediately preceding production stage
   b) The tentative latest ending time (TLET) for each job j: f_j^k
      The starting time of job j at the immediately following production stage

Figure 4 illustrates the situation of scheduling for SS subsystem of production stage 2 on the condition that the schedules of the production stage 1 and stage 3 are fixed. TEST and TLET of job A at the production stage 2 are shown respectively. It is assumed that every job has the path from stage 1 through stage 3. TEST of job A at stage 2 indicates the ending time of job A at stage 1, and TLET of job A at stage 2 indicates the starting time of job A at stage 3. If the starting time of job A at stage 2 is earlier than TEST or the ending time of job A at stage 2 is later than TLET, the schedule is infeasible. Therefore, penalty of violating the feasibility of schedule is embedded in the objective function in the optimization at each scheduling subsystem for respective production stage.

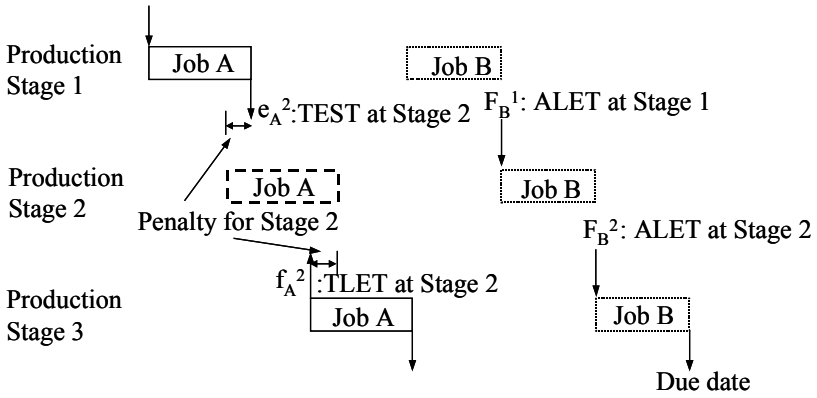


Figure 4. Tentative earliest starting time (TEST), tentative latest ending time (TLET) and absolute latest ending time (ALET)

4. Optimization of the schedule for each SS subsystem

Using the data obtained at step 3), the scheduling subsystem optimizes the production sequence of operations for that production stage. In order to include tardiness penalties in the objective function at every production stage, tardiness from ALET is embedded in the objective function. The scheduling problem for each scheduling subsystem is shown in Eq. (10). The term having weighting factor corresponds to the penalty for violating the precedence constraints with the preceding production stage and the following production stage respectively.

$$\text{(SS-}k\text{)}\ \min[\,Ch^k + \sum_j \max(0, t_j^k - F_j^k) + \rho\{\sum_j \max(0,\ e_j^k - t_j^k) + \sum_j \max(0,\ t_j^k - f_j^k)\} \tag{10}$$

s.t. Eq. (9)

where,

- t_j^k is the starting time of operation for job j at stage k.

The optimization problem (SS-k) for each SS subsystem is solved by using Simulated Annealing (SA) method combined with a neighbourhood search algorithm (Nishi et al., 2000b). The outline of the scheduling algorithm is composed of the following steps.

a) Generate an initial production sequence of operations and calculate the starting times of operations, and calculate the objective function.
b) Select an operation randomly and insert the selected operation into a randomly selected position, thereby change the processing order of operations.
c) For a newly generated production sequence, calculate the starting times of operations by the forward simulation and calculate the objective function. And then decide whether the newly generated schedule is adopted or not by using the criterion of simulated annealing method.
d) Repeat the procedure (b) to (c) for a predetermined number of times (N_S) at the same temperature parameter (T_{SA}), then the temperature parameter is reduced $T_{SA} \leftarrow \eta T_{SA}$, where η is annealing ratio. Then repeat (b) to (d) for a predetermined number of times (N_A).

A production schedule with a minimum objective function is regarded as the current optimal sequence. From the results of production sequence obtained by the simulated annealing method, the starting times of operations are calculated and the production amount of each products in each time period $P_{i,t}^{SS}$ is calculated by using the schedule generated by the simulated annealing method. In the proposed system, any scheduling model and any optimization algorithm can be adopted in the scheduling subsystem. Therefore, the proposed system can easily applicable to many types of scheduling problems such as jobshop problem (Hasebe et al., 1994), flowshop problem with intermediate storage constraints (Nishi et al., 2000c) by changing the algorithm of starting time calculation.

4.4 Distribution planning subsystem

When the lower and upper bound of the amount of production demand for the duration of group-time periods: $S_{i,m}^{\min}$, $S_{i,m}^{\max}$ are given at each group-time periods (week, or month), the DP subsystem determines the delivery plan to customers so as to maximize the profit taking the inventory cost and the penalties of product shortage. Thus, the optimization problem at the DPsubsystem is formulated as follows:

$$\text{(DP) } \min[\sum_{i,t} h_{i,t} I_{i,t} + \sum_{i,t} \nu_{i,t} P_{i,t}^{DP} + \sum_{i,m} \varsigma_{i,m} I_{i,m}^{-} - \sum_{i,t} \mu_{i,t} S_{i,t} + \rho \sum_{i,t} PN_{i,t}\ ] \tag{11}$$

$$I_{i,t} = I_{i,t-1} + P_{i,t}^{DP} - S_{i,t} \quad (\forall i, \forall t) \tag{12}$$

$$I_{i,m}^{-} \geq S_{i,m}^{\min} - \sum_{t' \in T_k} S_{i,t'} \quad (\forall i, \forall m) \tag{13}$$

$$PN_{i,t} \geq \left| P_{i,t}^{DP} - P_{i,t}^{SS} \right| \quad (\forall i, \forall t) \tag{14}$$

$$I_{i,t} \leq I_{i,t}^{\max} \quad (\forall i, \forall t) \tag{15}$$

$$\sum_{t \in T_m} S_{i,t} \leq S_{i,m}^{\max} \quad (\forall i, \forall m) \tag{16}$$

$$I_{i,t}, I_{i,m}^{-}, P_{i,t}^{DP}, S_{i,t}, PN_{i,t} \geq 0 \quad (\forall i, \forall m, \forall t) \tag{17}$$

where,

- $\mu_{i,t}$: revenue of product i sold in time period t,
- $h_{i,t}$: inventory cost for holding unit amount of final product i for the duration of time period t,
- $I_{i,t}$: inventory level of final product i at the end of time period t,
- $I_{i,m}^{-}$: amount of shortage of final product i in group-time periods m,
- $S_{i,t}$: amount of final product i delivered in time period t,
- T_m: set of group-time periods m,
- $\nu_{i,t}$: production cost of product i in time period t.

Eq. (11) is the objective function of DP subsystem which is the sum of the inventory holding cost for final products, production costs, penalty for product shortage, revenue of products and penalty for violating the constraints with SS subsystem. Eq. (12) indicates a material balance equation around the storage space for final product. Eq. (13) indicates the constraints on the minimum demand. Eq. (14) indicates the penalty value for violating the constraints imposed by the scheduling subsystem. Eq. (15) shows the capacity constraints of holding the final products in the storage space. Eq. (16) denotes the constraint of maximum amount of delivery to customer. Eq. (17) indicates the non-negative value constraints of all the decision variables.

4.5 Overall optimization algorithm

The total subsystem derives a feasible schedule by the following steps.

Step 1. *Preparation of the initial data.*
Each subsystem contacts the database and obtains the data and initializes the weighting factor of the penalty term, e.g. $\rho \leftarrow 0$.

Step 2. *Generation of an initial solution.*
Each subsystem independently generates a solution without considering the other subsystems.

Step 3. *Exchanging the data.*
Each subsystem contacts the other sub-systems and exchanges the amount of product data: $P_{i,t}^{MRP}$, $P_{i,t}^{SS}$, $P_{i,t}^{DP}$.

Step 4. *Judging whether the optimization at each subsystem is skipped or not.*
To avoid cyclic generation of same solutions, each subsystem skips Step 5 with a predetermined probability (see Hasebe et al., 1994).

Step 5. *Optimization at each subsystem.*
By using the data obtained at step 3, each subsystem executes the optimization of each subproblem.

Step 6. *Judging the convergence.*
When the solutions of all subsystems satisfy both of the following conditions, all of the subsystems stop the calculation, and the derived solution is regarded as the final solution.

- The solution generated at Step 5 is the same as that generated at Step 5 in the previous iteration.

- The value of the penalty function embedded in the objective function is equal to zero.

Step 7. *Updating the weighting factor.*
If the value of penalty function is positive, the derived solution is infeasible. Therefore, in order to reduce the degree of infeasibility, the weighting factor of the penalty term is increased. The value of weighting factor for penalty is updated by $\rho \leftarrow \rho + \Delta\rho$ at each iteration. Then return to Step 3. The incremental value $\Delta\rho$ is a constant. If the value of $\Delta\rho$ is larger, the performance index of solution derived by the proposed system becomes worse, on the other hand, the total computation time becomes shorter. On the contrary, if the value of $\Delta\rho$ is smaller, the total computation time is increased while the value of performance index has been improved. Our numerical studies show that when $\Delta\rho$ is less than 0.5, the computation time becomes exponentially larger even though the performance is not so improved. From these results, we have determined $\Delta\rho = 0.5$ as shown in Table 8.

By taking the above algorithm, it is easy for the proposed system to introduce the parallel processing system using multiple computers in which each subsystem execute its optimization concurrently. Figure 5 is a diagram showing the data exchange algorithm of the proposed system. Each square in Figure 5 illustrates steps of the data exchange algorithm in the iteration of the optimization at the subsystem and each arrow represents the flow of data. The total number of the processors required for solving the supply chain optimization problem is 5 processors for 3-stage production processes. The dotted arrow indicates the data of the production amount of each product in each time period determined by the MRP subsystem and DP subsystem. The thick arrow indicates the tentative earliest starting time (TEST) and tentative latest starting time (TLET) which are exchanged among the SS subsystems. In each iteration step, the data of the amount of products in each time period calculated in each subsystem are transmitted. Then, the data of TEST and TLST are exchanged. Therefore, jobs are generated at each iteration in the proposed system. While repeating the data exchange among each subsystem, the number of jobs and the starting time of operations are gradually satisfied with the constraints among each subsystem.

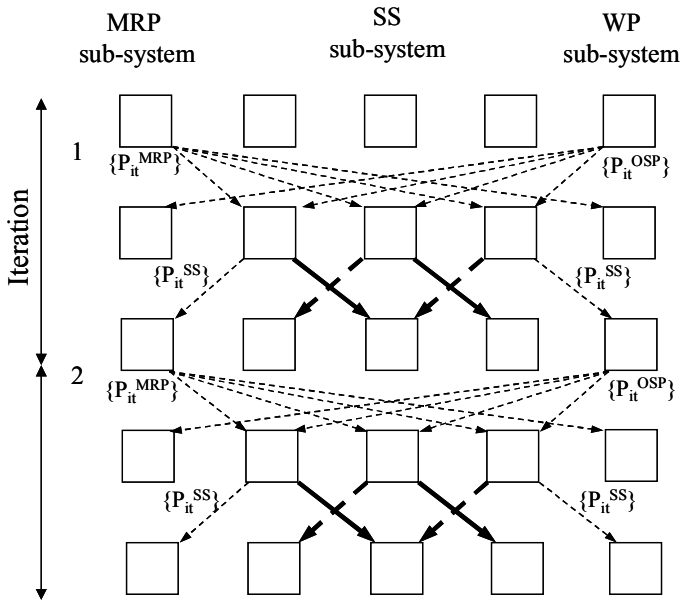


Figure 5. Data exchange algorithm

5. Computational results

The proposed scheduling system is applied to a supply chain optimization problem. The MRP subsystem and the DP subsystem is solved by a commercial MILP solver (CPLEX8.0 iLOG©). The algorithm used in the scheduling subsystem is coded by C++ language. Pentium IV (2.0AGHz) processor is used for computation.

5.1 Example problem

A batch plant treated in the example problem consists of three production stages shown in Figure 6. In this example, it is assumed that the production paths of all jobs are the same, meaning the each job is processed at stages 1 through 3. In this plant, four kinds of products are produced by each of two kinds of raw materials. Product A or B is produced from material 1, and product C and D is produced from material 2. The total planning horizon is 12 days, and it is divided into 12 time periods in the MRP subsystem and in the DP subsystem. The shipping of raw material for each material is available only two times in 4 days (1, 4, 7, 10). For each product, the lower bound and the upper bound of the production demand for each 4 days are given as the ag-

gregated value for each product. Thus, the delivery date can be decided by the DP subsystem. The plant is operated 24 hours/day. The available space for inventory for raw material and final products are restricted. The tables 1 to 8 show the data and parameters used in the example problem.

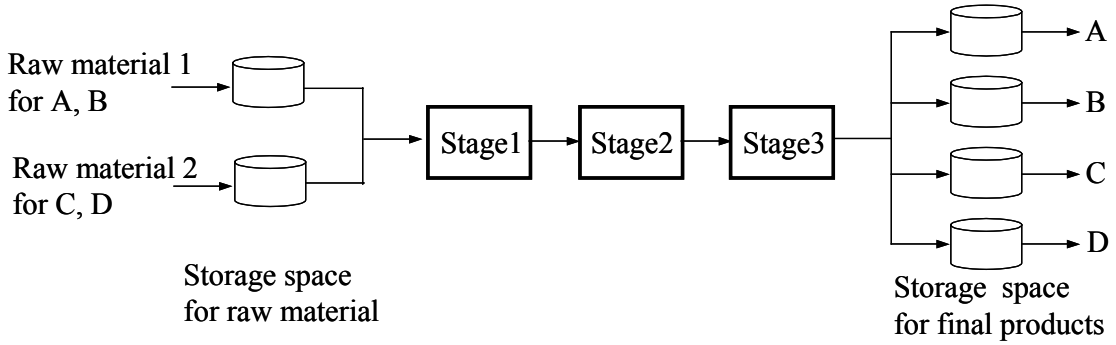


Figure 6. 3-stage production process for the example problem

J/T	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
A	1.90	2.00	1.80	1.95	2.00	1.95	2.10	2.00	1.80	1.90	2.20	2.30
B	2.20	2.30	2.15	2.15	2.25	2.30	2.15	2.20	2.35	2.40	2.25	2.50
C	2.00	2.20	2.20	3.20	3.25	2.25	1.10	1.25	2.05	2.10	1.35	2.30
D	2.20	2.50	2.20	2.50	1.30	2.20	3.50	1.20	2.20	3.70	2.60	2.40

Table 1. Revenue of products at each time period

J	*T: 1 - 4*	*T: 5 - 8*	*T: 9 - 12*
A	0.45	0.45	0.45
B	0.45	0.45	0.45
C	0.60	0.60	0.60
D	0.60	0.60	0.60

Table 2. Production costs

J	*T: 1 - 4*	*T: 5 - 8*	*T: 9 - 12*
A	100	200	100
B	100	200	400
C	200	300	300
D	100	100	400

Table 3. Minimum demand data

J	*T: 1 - 4*	*T: 5 - 8*	*T: 9 - 12*
A	1100	900	1200
B	1200	1400	1100
C	2200	2500	2000
D	1000	800	1000

Table 4. Maximum demand data

Stage	From/to	A	B	C	D
1	A	0	10	10	10
	B	40	0	10	10
	C	40	40	0	10
	D	40	40	40	40
2	A	0	0	30	30
	B	0	0	30	30
	C	30	30	0	0
	D	30	30	0	0
3	A	0	30	20	10
	B	10	0	30	20
	C	20	10	0	30
	D	30	20	10	0

Table 5. Sequence dependent changeover cost data

J/T	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
A	1.90	*2.00*	*1.80*	*1.95*	*2.00*	*1.95*	*2.10*	*2.00*	*1.80*	*1.90*	*2.20*	*2.30*
B	2.20	*2.30*	*2.15*	*2.15*	*2.25*	*2.30*	*2.15*	*2.20*	*2.35*	*2.40*	*2.25*	*2.50*
C	2.00	*2.20*	*2.20*	*3.20*	*3.25*	*2.25*	*1.10*	*1.25*	*2.05*	*2.10*	*1.35*	*2.30*
D	2.20	*2.50*	*2.20*	*2.50*	*1.30*	*2.20*	*3.50*	*1.20*	*2.20*	*3.70*	*2.60*	*2.40*

Table 6. Price and data of raw material

	A	*B*	*C*	*D*
Stage 1	4	5	2	5
Stage 2	5	5	2	4
Stage 3	3	5	4	4

Table 7. Processing time data

Other data
Unit Volume (V_1): 100, C^{max}: 3000 , I^{max}: 300, M^{max}:4000
Inventory holding cost for raw material: 0.001
Inventory holding cost for final product: 0.2
Penalty of product shortage: 4.0
Parameters for simulated annealing
Annealing times (N_A): 400
Search times (N_S): 200
Annealing Ratio: 0.97
Initial temperature: initial performance *0.1
Parameters for the proposed system
Skipping probability: 0.2%
Increment of weighting factor ($\Delta\rho$): 0.5

Table 8.Parameters used for computation

5.2 Results of coordination

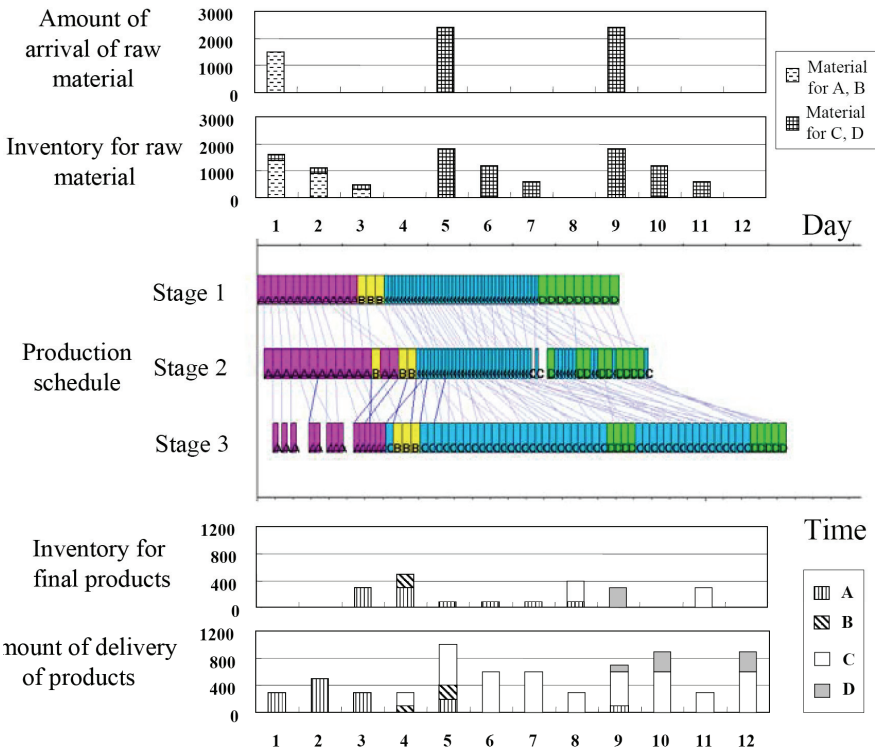


Figure 7. Intermediate result after 10 times of data exchange

Figure 7 shows an intermediate result obtained after ten times data exchange. This result is infeasible because several jobs at stage 3 are finished later than the scheduling horizon (Day 12).This is because the DP subsystem individually tries to generate a distribution planning so that the amount of delivery amount to the customer is maximized. Therefore, the planning result is also infeasible because the amount of production at stage 3 in each time period is not equivalent to the amount of the product delivery. Moreover, several operations at stage 2 are finished later than the starting times of operations at stage 3.
Figure 8 shows the final schedule obtained after 23 times of data exchanges. A feasible schedule is obtained by the proposed system. The transitions of the performance index in 23 times of iterations are shown in Figure 9.

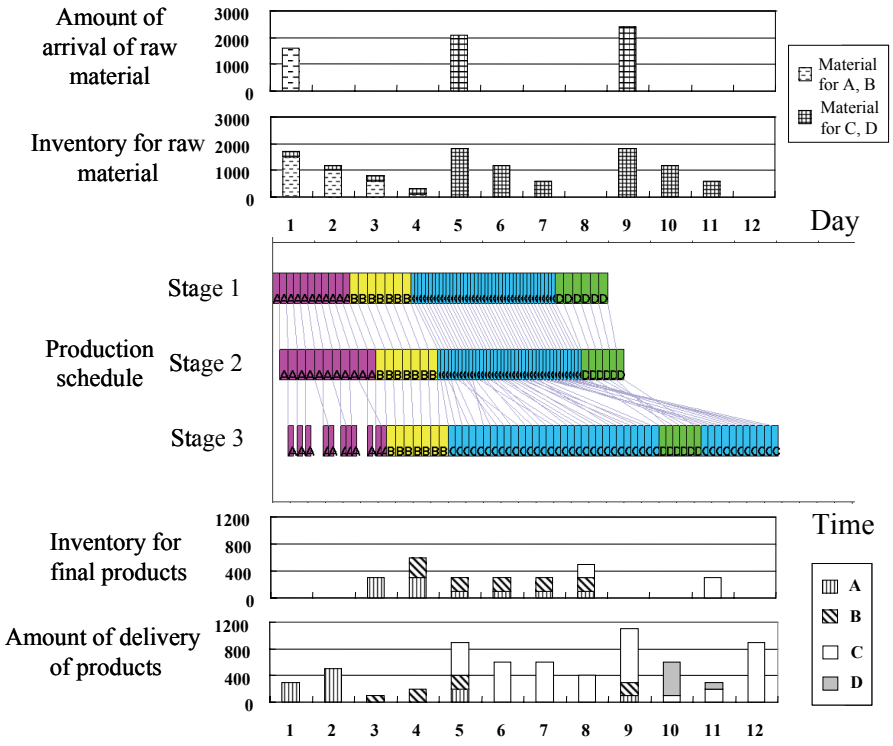


Figure 8. Final result after 23 times of data exchange

As the number of iteration increases, the value of penalty function decreases and it becomes close to zero. This indicates that the proposed system gradually generates a feasible solution by increasing the value of weighting factor of the penalty for violating the material balancing constraints although only the local information is used to optimize the objective function for each subsystem.

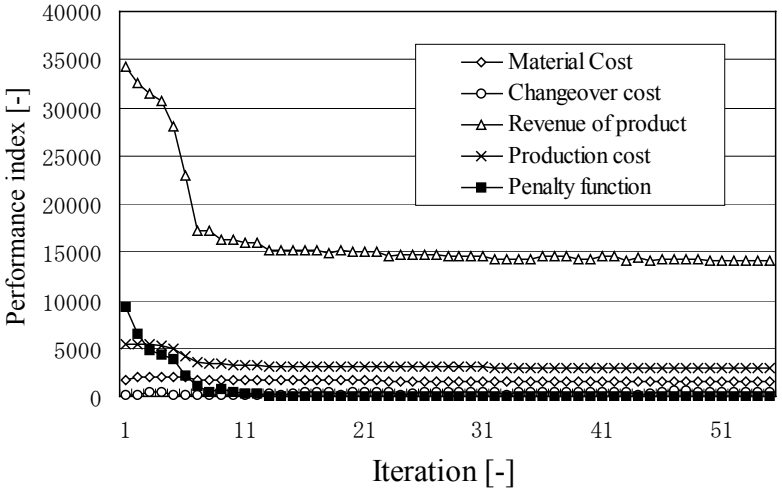


Figure 9. The transitions of the performance index

5.3 Comparison of the proposed system and the conventional system

In order to evaluate the performance of the proposed system, a hierarchical planning and scheduling system considering the entire plant is also developed. The planning problem (HP) is formulated as MILP problem by using the mid-term-planning model proposed by McDonald & Karimi (1997).

$$\text{(HP)}\ \min Z\ Z = \sum_{r,t} p_{r,t} M_{r,t} + \sum_{r,t} q_{r,t} C_{r,t} + \sum_{i,t} \nu_{i,t} P_{i,t} + \sum_{i,t} h_{i,t} I_{i,t} + \sum_{i,m} \varsigma_{i,m} I^{-}_{i,m} - \sum_{i,t} \mu_{i,t} S_{i,t} \tag{18}$$

$$C_{r,t} = C_{r,t-1} + M_{r,t} - \sum_{i \in U_r} P_{i,t} \quad (\forall r, \forall t) \tag{19}$$

$$C_{i,t} \le C^{\max}_{i,t} \quad (\forall i, \forall t) \tag{20}$$

$$\sum_{t} Y_{r,t} \le m_r \quad (\forall r) \tag{21}$$

$$\sum_{i} s^{l}_{i} \cdot P_{i,t} / V_l \le H_t \quad (\forall t, \forall l) \tag{22}$$

$$I_{i,t} = I_{i,t-1} + P_{i,t} - S_{i,t} \quad (\forall i, \forall t) \tag{23}$$

$$I_{i,t} \le I_{i,t}^{\max} \quad (\forall i, \forall t) \tag{24}$$

$$I_{i,m}^{-} \ge S_{i,m}^{\min} - \sum_{t' \in T_m} S_{i,t'} \quad (\forall i, \forall m) \tag{25}$$

$$P_{i,t}, M_{i,t}, C_{i,t}, I_{i,t}, I_{i,t}^{-}, S_{i,t} \ge 0 \quad (\forall i, \forall t) \tag{26}$$

$$Y_{i,t} \in \{0,1\} \quad (\forall i, \forall t) \tag{27}$$

where,

- H_t : amount of time available in the time period t ,
- T_m : set of group-time period m ,
- V_l : batch size of the machine at the production stage l .

For the hierarchical approach, the solution of the HP problem is transferred to the scheduling subsystem as the production request. The jobs are created by using the production amount: $P_{i,t}$ calculated in the planning system. The scheduling system obtains the amount of inventory for raw material, due dates for each job. And then the scheduling system is executed. For the scheduling system used in the hierarchical approach, a schedule considering the entire production stages is successively generated improving an initial schedule. The simulated annealing method is adopted so that the solution is not trapped in a local optimal.

Ten times of calculations are made with different seed numbers for generating random numbers in the simulated annealing method to compare the performance of the proposed system. The results of the performance index for the proposed system (DSCM1) and the hierarchical planning and scheduling system (CONV) are shown in Table 10. The average computation time for deriving a feasible schedule of the proposed algorithm is 198 seconds. The performance of the DSCM1 is lower than that of the hierarchical system (CONV). This is because some of the solutions of the schedule generated by each subsystem have been entirely different from that derived by other subsystem, which makes convergence of the proposed algorithm difficult. Thereby, the final solution of the proposed system has been trapped into a bad local optimum. In order to improve the efficiency of the proposed method, the capacity constraints for

each month are embedded into the DP subproblem. By adding the capacity constraints to DP subproblem, the number of generating meaningless solutions violating the constraints with the SS subsystem has been reduced. The results of computation for the improved method (DSCM2) are shown in Table 10. The profit of DSCM2 is successfully improved compared with that of DSCM1 without sacrificing the computational expenses. For DSCM2, the changeover cost is lower than that of CONV, though the profit for final products is higher than CONV. This is because both the changeover costs and the profit of the product greatly depend on the production schedule, and it is very difficult for CONV to determine the precise production schedule at the production planning level. It is demonstrated that the total profit of proposed system (DSCM2) is higher than that of the conventional system (CONV). The proposed system can generate a better solution than the conventional system even though only local information is used to generate the solution of each subsystem.

6. Conclusion and future work

An autonomous decentralized supply chain optimization system for multi-stage production processes has been proposed. The novel aspect of this paper is that we provide a novel distributed optimization system for a supply chain planning problem for multi-stage production processes comprising a material requirement planning (MRP) subsystem, scheduling subsystems for each production stage, and a distribution planning (DP) subsystem.

Methods	DSCM1	DSCM2	CONV
Number of data exchange	18	21	1
Profit [-]	7,533	9,538	9,418
Number of jobs [-]	66	66	62
Total revenue [-]	14,353	15,759	15,460
Costs of raw material [-]	1,540	1,688	1,677
Inventory cost for raw materials [-]	18	21	26
Inventory holding cost for final products [-]	504	720	360
Production cost [-]	2,970	3,006	2,790
Penalty of product shortage [-]	1320	360	400
Sequence dependent changeover cost [-]	468	426	1,030

Table 10. Comparison of the autonomous decentralized supply chain planning system (DSCM) and the conventional system (CONV)

Each subsystem includes an optimization function and repeats the generation of solutions for each subproblem and data exchange among the subsystems. The total system derives a feasible solution by gradually increasing the weighting factor for violating the infeasibility of the solution through repeated optimization at each subsystem and data exchanges among the subsystems. The data exchanged among the subsystems are tentative production amount of each product at each time period that is desirable for each subsystem. By adopting such a structure, it is easy to modify the subsystem when a new constraint is added or when performance evaluation criteria changes. Thus, the system can flexibly accommodate various unforeseen changes. he proposed system is successfully applied to a multi-stage supply chain optimization problem. The results demonstrate that feasible solutions could be obtained by the numerical examples. The performances of the proposed system are compared with those of the schedule derived by the conventional system. It has been shown that the proposed system can generate a better solution than the conventional system without sacrificing flexibility and computational resources. Future work should be investigated on how to optimize the entire supply chain under several uncertainties.

7. Nomenclature

- $C_{r,t}$: amount of inventory of raw material r at the end of time period t,
- $C_{r,t}^{\max}$: maximum amount of inventory for raw material r at the end of time period t,
- Ch^k: sequence dependent changeover cost at stage k,
- e_j^k: tentative earliest starting time of job i at stage k,
- f_j^k: tentative latest ending time of job i at stage k,
- F_j^k: absolute latest ending time of job j at stage k,
- $h_{i,t}$: inventory cost for holding unit amount of final product i for the duration of time period t,
- $I_{i,t}$: inventory level of final product i at the end of time period t,
- $I_{i,t}^{\max}$: maximum amount of inventory for final product i at the end of time period t,
- $I_{i,m}^{-}$: amount of shortage of inventory for final product i in group-time periods m,
- K: sufficiently large positive number,

- m_r: maximum number of the arrival of raw material r,
- $M_{r,t}$: amount of raw material r arrived from supplier at the start of time period t,
- N_A: annealing time for simulated annealing,
- N_S: search times at the same temperature for simulated annealing,
- $p_{r,t}$: price of the unit amount of raw material r from supplier to the plant at the start of time period t,
- $P_{i,t}^{MRP}$ tentative amount of production of product i in time period t, which is derived at MRP subsystem,
- $P_{i,t}^{DP}$: tentative amount of production of product i in time period t, which is derived at DP subsystem,
- $P_{i,t}^{SS}$: tentative amount of production of product i in time period t, which is derived at DP subsystem,
- $PN_{i,t}$: difference of production amount of product i in time period t,
- $q_{r,t}$: inventory holding cost of unit amount of raw material r for the duration of time period t,
- s_i^k: processing time of operation for job j at stage k,
- $S_{i,t}$: amount of final product i delivered in time period t,
- t_j^k: starting time of operation for job j at stage k,
- T_m: set of time in group-time periods m,
- $T_{available}$ periods of time when the material arrival is available,
- T_{SA}: annealing temperature for simulated annealing method,
- TEST tentative earliest starting time,
- TLET tentative latest ending time, U_r: set of products produced
- from material r,
- V_l: batch size of the machine at the production stage l,
- $Y_{r,t}$: binary variable indicating whether material r is arrived at the start of time period t or not.

Greek Letters

- η: annealing ratio (temperature reduction factor),
- $\mu_{i,t}$: revenue of of product i sold in time period t,
- $\nu_{i,t}$: production cost of product i in time period t,
- ρ: penalty parameter,
- $\varsigma_{i,m}$: penalty for unit amount of shortage of product in group-time
- periods m,

8. References

Androulakis, I. and Reklaitis, G. (1999), Approaches to Asynchronous Decentralized Decision Making, *Computers and Chemical Engineering*, Vol. 23, pp. 341-355.

Birewar, D. and Grossmann, I. (1990) Production Planning and Scheduling in Multiproduct Batch Plants, *Ind. Eng. Chem. Res.*, Vol. 29, 570-580.

Bitran, R., Hax, A. (1977) On the Design of Hierarchical Production Planning Systems, *Decision Sciences*, Vol. 8, pp. 29-55.

Bok, J.K., Grossmann, I.E., Park, S. (2000) Supply Chain Optimization in Continuous Flexible Process, *Ind. Eng. Chem. Res.*, Vol. 39, pp. 1279-1290.

Burke, P., Prosser, P., (1990) Distributed Asynchronous Scheduling, *Applications of Artificial Intelligence in Engineering V*, Vol. 2, pp. 503-522.

Gou, L., Luh, P.B. Luh, Kyoya, Y. (1998) Holonic manufacturing scheduling: architecture, cooperation mechanism, and implementation, *Computers in Industry*, Vol. 37, pp. 213-231.

Gupta, A., Maranas, C.D. (1999) Hierarchical Lagrangean Relaxation Procedure for Solving Midterm Planning Problems, Ind. Eng. Chem. Res., Vol. 38, pp. 1937-1947

Hasebe, S., Kitajima, T., Shiren, T., Murakami, Y. (1994) Autonomous Decentralized Scheduling System for Single Production Line Processes, *Presented at AIChE Annual Meeting*, Paper 235c, USA.

Hoitomt, D.J., Luh, P.B., Pattipati, R. (1993) A Practical Approach to Job-Shop Scheduling Problems, *IEEE Trans. Robot. Automat.*, Vol. 9, pp. 1-13.

Jeremy, F. S. (2001) Modeling the Supply Chain, *Thomson Learning*.

Fischer, K. (1999) Agent-based design of holonic manufacturing systems, *Robotics and Autonomous Systems*, Vol. **27**, pp. 3-13.

Maturana, F.P., Norrie, D.H., (1997) Distributed decision-making using the contract net within a mediator architecture, *Decision Support Systems*, Vol. 20, pp. 53-64.

McDonald, C. and Karimi, I. (1997), Planning and Scheduling of Parallel Semicontinuous Processes. 1. Production Planning, *Ind. Eng. Chem. Res.*, Vol. 36, pp. 2691-2700.

Mckay, A., Pennington, D., Barnes, C. (2001) A Web-based tool and a heuristic method for cooperation of manufacturing supply chain decisions, Vol. 12, pp. 433-453.

Nishi, T., Inoue, T., Yutaka, H., Taniguchi, S. (2000a) Development of a Decentralized Supply Chain Optimization System, *Proceedings of the International Symposium on PSE Asia*, 141-146.

Nishi, T., Konishi, M., Hattori, Y., Hasebe, S. (2003) A Decentralized Supply Chain Optimization Method for Single Stage Production Systems, *Transactions of the Institute of Systems, Control and Information Engineers*, 16-12, 628-636 (in Japanese)

Nishi, T., Sakata, A., Hasebe, S., Hashimoto, I. (2000b), Autonomous Decentralized Scheduling System for Just-in-Time Production, *Computers and Chemical Engineering*, Vol. 24, pp. 345-351.

Nishi, T., Sakata, A., Hasebe, S., Hashimoto, I. (2000c), Autonomous Decentralized Scheduling System for flowshop problems with storage cost and due-date penalties, *Kagaku Kogaku Ronbunshu*, Vol. 26, pp. 661-668 (in Japanese).

Ohkuma, K., Ueda, K., (1996) Solving Production Scheduling Problems with a Simple Model of Biological-Oriented Manufacturing Systems, *Nihon Kikaigakkai Ronbunshu C*, Vol. 62, pp. 429-435, 1996 (in Japanese).

Rupp, T.M., Ristic, M. (2000) Fine Planning for Supply Chains in Semiconductor Manufacture, *Journal of Material Processing Technology*, Vol. 107, pp. 390-397.

Shaw, M.J., (1987) A distributed scheduling method for computer integrated manufacturing: the use of local area network in cellular systems, *Int. J. Prod. Res.*, Vol. 25, No. 9, pp. 1285-1303.

Simon, C., Pietro, R., Mihalis G. (2000) Supply chain management: an analytical framework for critical literature review, *European Journal of Purchasing& Supply Management*, Vol. 6, pp. 67-83.

Smith, R.G., (1980) The Contract Net Protocol: High-Level Communication Control in a Distributed Problem Solver, *IEEE Transactions on Computers*, Vol. 29, No.12, pp. 1104-1113.

Tan, M., (2001) Hierarchical Operations and Supply Chain Planning, *Springer*.

Tharumarajah, A., Bemelman, R., (1997) Approaches and issues in scheduling a distributed shop-floor environment, *Computers in Industry*, Vol. 34, pp. 95-109.

Vidal, C.J. and Goetschalckx, M. (1997) Strategic production-distribution models: A critical review with emphasis on global supply chain models, *European Journal of Operational Research*, Vol. 98, pp. 1-18.

Wei, T., (2000) Integration of Process Planning and Scheduling - a Review, *Journal of Intelligent Manufacturing*, Vol. 11, pp. 51-63.

29

Simulation Modeling and Analysis of the Impacts of Component Commonality and Process Flexibility on Integrated Supply Chain Network Performance

Ming Dong and F. Frank Chen

1. Introduction

A supply chain can be defined as an integrated business process wherein a number of various business entities (i.e., suppliers, manufacturers, distributors, and retailers) work together. Supply chain configuration is concerned with determining supply, production and stock levels in raw materials, subassemblies at different levels of the given bills of material (BOM). End products and information exchange through (possibly) a set of factories, distribution centers of a given production and service network to meet fluctuating demand requirements. Through the evaluation of the supply chain network configurations, performance indicators of the supply chain such as fill rate, customer service level, associated cost and response capability can be obtained under different network configurations. Different network configurations include: (1) different stocking levels in raw materials, subassemblies and end products; (2) safety stock location; (3) production policy (make-to-stock or make-to-order); (4) production capacity (amount and flexibility); (5) allocation rules for limited supplies; and (6) transportation modes.

Reconfiguration of the supply chain network from time to time is essential for businesses to retain their competitive edge. Supply chain performance optimization consists of deciding on the safety stock level, reorder point, stocking location, production policy (make-to-stock or make-to-order), production capacity (quantity and flexibility), assignment of distribution resources and transportation modes while imposing standards on the operational units for performance excellence. Therefore, the aim of supply chain performance optimization is to find the best or the near best alternative configuration with which the supply chain can achieve a high-level performance.

In integrated supply chains, performance evaluation becomes more challenging since not only the distribution function but also the manufacturing function will be considered. In addition, there are many variables involved in the performance evaluation. More important, there exist interactions between some variables.

Problems with the integrated characteristics given above are difficult to be transformed into mathematical optimization models. When possible, often there are tens of thousands of constraints and variables for a deterministic situation. However, traditional deterministic optimization is not suitable for capturing the truly dynamic behavior of most real-world applications. The main reason is that such applications involve data uncertainties that arise because information that will be needed in subsequent decision stages is not available to the decision maker when the decision must be made (Beamon, 1998). Poorly integrated enterprise logistic system components and processes make it more difficult for firms to compete and differentiate themselves. Only with an integrated approach to supply network performance analysis and management can firms locate and remove sources of inefficiency and waste (Ross, Venkataramanan and Ernstberger, 1998). Through the performance evaluation, the impacts of different factors such as reorder point, safety stock, degree of component commonality and manufacturing flexibility can be investigated. Thus, simulation study can help us gain insight in network configuration problem. In turn, this can assist companies' decision-making in their supply chain management.

Due to the shortened product life cycle and the dynamics of the product market, a company has to improve current products and/or add new products to its existing product line. There are a few strategies available for a supply chain to simultaneously deal with product variety and keep high levels of productivity. Some of these are supply chain integration, component part commonality, and process flexibility. Different products may share common components (therefore, common inventories) and resources (facilities and capacities). Correspondingly, this requires that the company to reconfigure its supply chain network structure. The configuration of a supply chain network, including the links between entities and operational policies, is changeable and aimed at delivering products to customers in an efficient and effective way. The issue is how to evaluate and then change the structure of the network. The evolution aspect of the supply chain network structure provides the basis for the change.

The development of analytical measures describing product structure charac-

teristics is a prerequisite to understanding the relationships between product structure and supply chain performance. One characteristic of product structures is the degree of common components in a sub-assembly, a single product or any product family. The traditional MRP methodologies are completely blind to commonality and consequently are unable to exploit it in any way (Miguel, et al., 1999).

There exist a rich literature studying component commonality. However, the majority of work published so far has concentrated on the related effects of inventory and safety stock levels only. It has been clearly demonstrated in the literature that introducing a common component that replaces a number of unique components reduces the level of safety stock required to meet service level requirements.

Collier (1981) initiates an interest in taking advantage of the commonality situation. He finds that increased commonality reduces production costs through larger production lot sizes and reduces operation costs through increased standardization.

Eynan and Rosenblatt (1996) study the effects of increasing component commonality for a single-period model. They develop optimal solutions for the commonality and non-commonality models and provide bounds on the total savings resulting from using commonality. They demonstrate, under general and specific component cost structures, that some forms of commonality may not always be a preferred strategy. Furthermore, they present conditions under which commonality should not be used.

Hillier (1999) develop a simple multiple-period model with service level constraints to compare the effects of commonality in the single-period and multiple-period case. The results are drastically different for these two cases. When the common component is more expensive than the components it replaces, commonality is often still beneficial in the single-period model, but almost never in the multiple-period model.

Hong and Hayya's paper (1998) consider the effects of component commonality in a single-stage manufacturing system of two products manufactured in a single facility. They consider two economic lot schedules: the common cycle (CC) and basic period (BP) schedules. For each lot schedule, an expression for the total relevant cost for the system was given in their paper.

In an environment where demands are stochastic, it seems a good strategy to store inventory in the form of semi-finished products (vanilla boxes) that can serve more than one final product. However, finding the optimal configura-

tions and inventory levels of the vanilla boxes could be a challenging task. Swaminathan and Tayur (1998) model the above problem as a two-stage integer program with recourse. By utilizing structural decomposition of the problem and sub-gradient derivative methods, they provide an effective solution procedure.

Product structure (or bill of material) is a key input to an integrated supply chain design. The product structure may have a significant impact on component demand patterns, work-in-process inventory, and fill-rate performance. However, the effect of alternate product structures on integrated supply chains is not well understood. The simulation study in this chapter is designed to investigate the impacts of component commonality on the integrated supply chain network.

Process flexibility, whereby a production facility can produce multiple products, is a critical design consideration in multi-product supply chains facing uncertain demand. The challenge is to determine a cost-effective flexibility configuration that is able to meet the demand with high likelihood (Graves and Tomlin 2003). In a make-to-order environment, this flexibility can also be used to hedge against variability in customer orders in the short term (Bish, Muriel and Biller 2005). Graves and Tomlin (2003) present a framework for analyzing the benefits from flexibility in multistage supply chains. However, these analytical results are only suitable for simplified supply chains.

The remainder of this chapter is organized as follows. Section 2 provides an integrated modeling framework for multi-stage supply chains. In section 3, a state and resource based simulation modeling approach is proposed. Section 4 defines the new analytical measure for component commonality index. This commonality index is used to evaluate the impacts of component commonality on supply chain network performance in section 5. Section 6 investigates the effects of process flexibility on supply chain performance. Section 7 summarizes this research.

2. An Integrated Modeling Framework for Supply Chain Networks

Supply chains may differ in the network structure (serial, parallel, assembly and arborescent distribution), product structure (levels of Bill-Of-Materials), transportation modes, and degree of uncertainty that they face. However, they have some basic elements in common.

2.1 Sites and Stores

A supply chain network can be viewed as a network of functional sites connected by different material flow paths. Generally, there are four types of sites: (1) *Supplier sites*: they procure raw materials from outside suppliers; (2) *Fabrication sites*: they transform raw materials into components; (3) *Assembly sites*: they assemble the components into semi-finished products or finished goods; and (4) *Distribution sites*: they delivery the finished products to warehouses or customers. All sites in the network are capable of building parts, subassemblies or finished goods in either make-to-stock or make-to-order mode. The part that a site produces is a single-level BOM.

2.2 Links

All stores in the supply chain are connected together by links that represent supply and demand processes. Two types of links are defined: *internal link* and *external link*. Internal links are used to connect the stores within a site, i.e., they represent the material flow paths from input stores to output stores within a site. Associated with an internal link connecting an input store i to an output store j is a usage count, u_{ij}, which indicates the number of SKUs in the input store i required to produce a SKU in the output store j. Along with the usage counts, the internal links connecting input stores and output stores constitute the single-level BOM for that output store. A link connecting an output store of one site to an input store of another site is called an external link. This kind of link represents that the output store provides replenishments to the specified downstream input store. In the network topology, we define that a downstream input store has only one link between it and its upstream output store (Figure 1).

2.3 The Relationships Between Stores

Let ***ST*** be the collection of stores in a supply network and i be a store in ***ST***. The set of directly upstream supplying stores of store i is denoted as $UPST(i)$. The set of directly downstream receiving stores from store i is denoted as $DOWNST(i)$. If i is an input store, then $UPST(i)$ is a singleton set, i.e., it contains only one upstream supplying store. That is, each input store can obtain replenishment from only one supplier. On the other hand, $DOWNST(i)$ consists of one or more output stores at the same site. If i is an output store, then $UPST(i)$ is either empty, in which case i is a *source* store (e.g., a supplier), or

contains one or more stores, which are input stores at the same site. For *DOWNST*(*i*), it is either empty, in which case *i* is an *end* store, or contains one or more input stores at its downstream site.

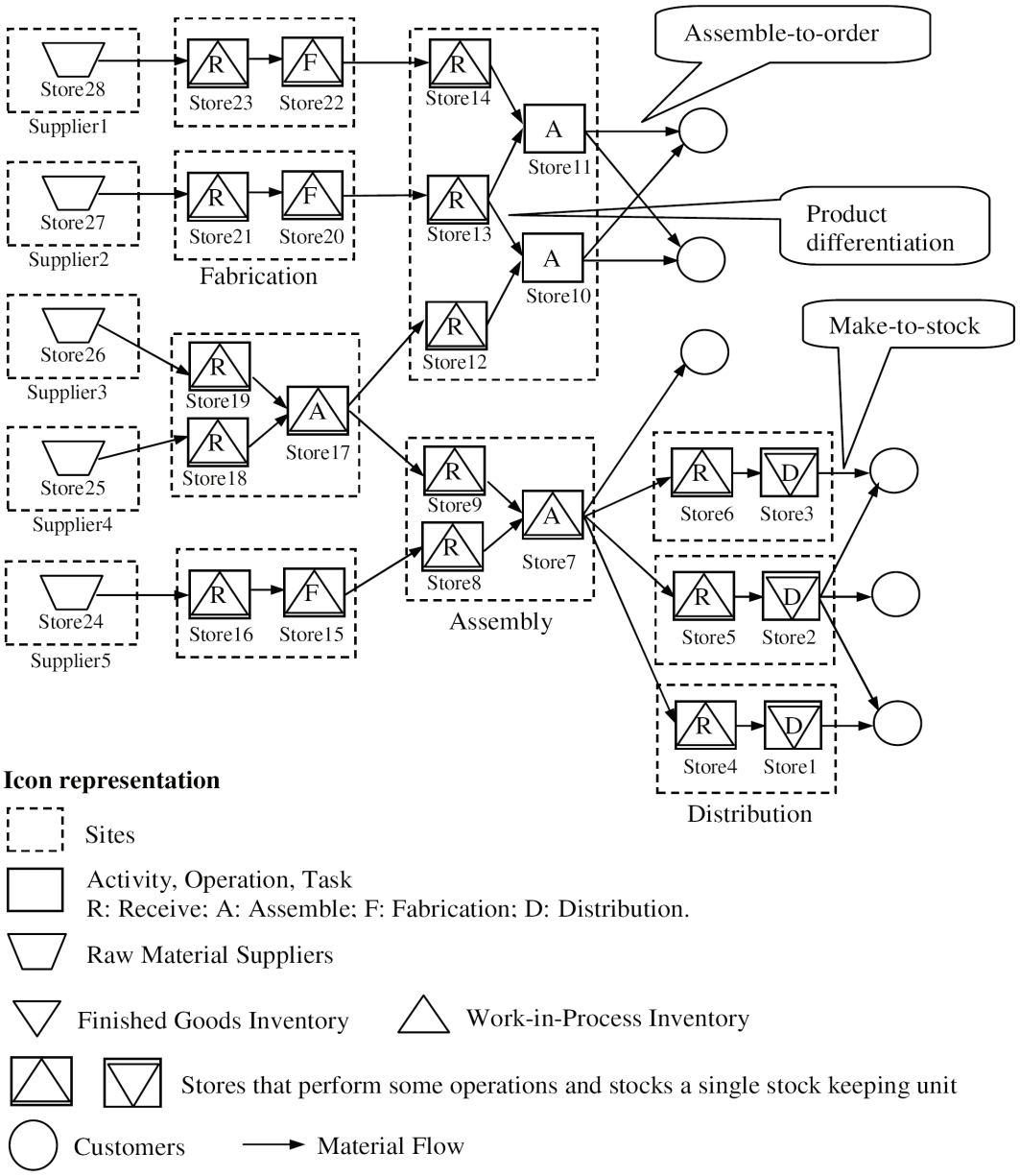


Figure 1. An Integrated Modeling Framework for Supply Chains

3. A Component based Simulation Modeling Approach

From a system perspective, a supply chain process consists of the flow of materials, information and services, and the monitoring and control of these flows. Typical activities include: raw material procurement, inventory management, order processing, warehousing, transportation, distribution and production. Supply chain management is concerned with the development of functions to support these activities.

Several methods to develop a model of a system have been proposed. Top down development starts with a model at a high abstraction level, this model is refined by a number of disaggregation (or decomposition) steps until the desired level of detail has been reached. Bottom up development starts with some subsystems that are detailed descriptions of some aspect or part of the systems. Then, these sub-models are composed into a model of the entire system. In this research, a mixture of top down and bottom up development is employed to build simulation models.

Practical experiences show that some supply chain networks have subsystems that have a lot in common. For example, a distribution center and a production unit have transportation subsystems for internal transport. To support the modeling process it is useful to reuse some typical subsystems, often called *components* or *building blocks*. Reusing these components reduces the modeling effort. And, from these reusable components, the rapid reconfiguration of a supply chain network can be achieved.

Some requirements on the components include:

1. they can be parameterized, which make them tailored for a specific situation;
2. they have to be robust in the sense that it can handle various inputs, i.e. the number of assumptions about the environment of the component is as few as possible.

Some typical components in a supply chain network are given as follows:

- Raw material supplier: the beginning of the chain
- Production unit: the manufacturing of goods (transforming, assembling, splitting up)
- Distribution center: the rearrangement and the distribution of goods
- Transportation center: the transportation of goods
- Consumer: the end of the chain

3.1 Stroboscope - A State and Resource based Simulation Language

STROBOSCOPE is a general-purpose discrete-event simulation language based on activity scanning and activity cycle diagrams (ACDs). A subset of the STROBOSCOPE modeling concepts are directly analogous to those used in timed stochastic colored Petri-nets, but use a different terminology (token=resource; place=queue; transition=activity; arc=link).
STROBOSCOPE tokens can be colored with any number of properties and methods. The entire state of the model (e.g., number of tokens in a place, number of times a transition has fired) and the colors of tokens are accessible via variables. Arcs can enable transition firing based on the truth of any expression; allowing arcs to be inhibitors, activators, or to take on any other role. Transition timing can be defined with any valid expression (functions that sample from various probability distributions are available). STROBOSCOPE also includes many powerful extensions not found in Petri-nets (Martinez 1996).

STROBOSCOPE's ability to dynamically access the state of the simulation and the properties of the resources involved in an operation differentiates it from other simulation tools. The state of the simulation refers to such things as the number of products in the inventory, the current simulation time, the number of times an activity has occurred, and the last time a particular activity started. Access to properties of resources means that operations can be sensitive to resource properties, such as quantity and holding cost, on an individual or an aggregate basis. The employment of state and resource in simulation will facilitate the implementation procedure since they are strong in modeling dynamic systems with highly interdependent components subject to activity startup conditions.

3.2 Network Elements

3.2.1 Resources

Resources are things required to perform tasks. These can be machinery, space, materials, labor, permits, or anything else needed to perform a particular task.
The most important characteristic of a resource is its type. The type of a resource places the resource within a category of resources that share common traits or characteristics.

3.2.2 Queues

Queues are nodes in which resources spend time passively (they are either stored there, or waiting to be used). Each queue is associated with a particular resource type. Queues that hold discrete resources have attributes that control the ordering of the individual resources within the Queue.

3.2.3 Activities

Activities are nodes that represent work or tasks to be performed using the necessary resources. Resources spend time in activities actively (performing a task). Resources involved in activities are productive, sometimes in collaboration with other resources.

Combi activities: represent tasks that start when certain conditions are met.

Normal activities: represent tasks that start immediately after other tasks end. Among all nodes in a network, only activity instances represent tasks that end and release resources. For this reason, only other activities can be predecessors to a *Normal Activity*.

3.2.4 Links

Links connect network nodes and indicate the direction and type of resources that flow through them. Links have many attributes that can be used to control the flow of resources from the predecessor node to the successor node.

4. Commonality Index (CI)

The commonality index is a measure of how well the product design utilizes standardized components. A component item is any inventory item (including a raw material) other than an end item that goes into higher-level items. An end item is a finished product or major subassembly subject to a customer order. The commonality index given by Collier (1981) cannot differentiate the product lines with same components but different quantities for each component.

Different from Collier, two types of commonality indexes are defined in this paper. One is called component-level (denoted as CI_i), which is to provide an indicator on the percentage of a component being used in different products. The other is called product-level (denoted as CI_P). There are three variables that will affect the commonality index, which are, number of unique compo-

nents (denoted as u), number of total components along the product line (denoted as c), and final number of product varieties offered (denoted as n). To get the appropriate product-level CI, all these three variables along with component-level CI should be considered. The basic idea is that, by ranking the different component-level CI values, the average for the differences of CI values is computed. Then, this average difference will be multiplied by a weight, which is the ratio of (c-n) and u. A special case appears when all component-level CI values are same, u<c and n<c. In this case, instead of the average difference, product-level CI is obtained by multiplying anyone component-level CI and the weight. Therefore, to calculate CI_P, we first find out the difference between the maximal component-level CI and the minimal component-level CI, which is same as the summation of differences among component-level CI values. Then, we divide the difference by number of unique components to get the average CI difference. Finally, the average CI difference is multiplied by (c-n) so that the information on how broad the components spread in product line is captured.

The following formula is used to calculate the component-level CI:

$$CI_i = \frac{\sum_j f_{ij} \cdot d_j}{\sum_{i,j} f_{ij} \cdot d_j} \qquad (1)$$

f_{ij} = number of component i in product j
d_j = demand of product j
$0 \le CI_i \le 1$

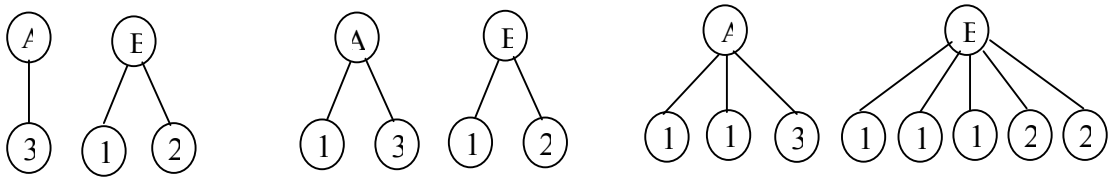
The lower bound of the component-level CI is 0 (no commonality). The upper bound on the degree of commonality is 1. Complete commonality results when the total number of distinct components (u) equals one.

In reality, it is reasonable to assume that number of total components along the product line is greater than final number of product varieties offered, i.e., $c > n$. The product-level CI is computed as follows:

u = number of unique components
n = final number of product varieties offered
c = total number of components along the product line

$$CI_p = \begin{cases} \dfrac{CI_i}{u} \times (c-n), & when \quad \max\limits_i \{CI_i\} = \min\limits_i \{CI_i\} \quad and \quad c > u \\ \left[\dfrac{\left(\max\limits_i \{CI_i\} - \min\limits_i \{CI_i\} \right)}{u} \right] \times (c-n), & otherwise \end{cases} \tag{2}$$

In general, a higher CI is better since it indicates that the different varieties within the product family are being achieved with more common components.



$$CI_1 = CI_2 = CI_3 = \frac{1}{3} \quad CI_1 = \frac{1}{2}, CI_2 = CI_3 = \frac{1}{4} \qquad CI_1 = \frac{5}{8}, CI_2 = \frac{2}{8}, CI_3 = \frac{1}{8}$$

$$CI_p = \left(\frac{1}{3} - \frac{1}{3} \right) \times 1 \div 3 = 0 \quad CI_p = \left(\frac{1}{2} - \frac{1}{4} \right) \times 2 \div 3 = \frac{1}{6} \qquad CI_p = \left(\frac{5}{8} - \frac{1}{8} \right) \times 6 \div 3 = 1$$

$$CI_1 = \frac{4}{7}, CI_2 = \frac{2}{7}, CI_3 = \frac{1}{7} \qquad CI_1 = \frac{5}{8}, CI_2 = \frac{1}{8}, CI_3 = \frac{1}{8}, CI_4 = \frac{1}{8}$$

$$CI_p = \left(\frac{4}{7} - \frac{1}{7} \right) \times 5 \div 3 = \frac{5}{7} \qquad CI_p = \left(\frac{5}{8} - \frac{1}{8} \right) \times 6 \div 4 = \frac{3}{4}$$

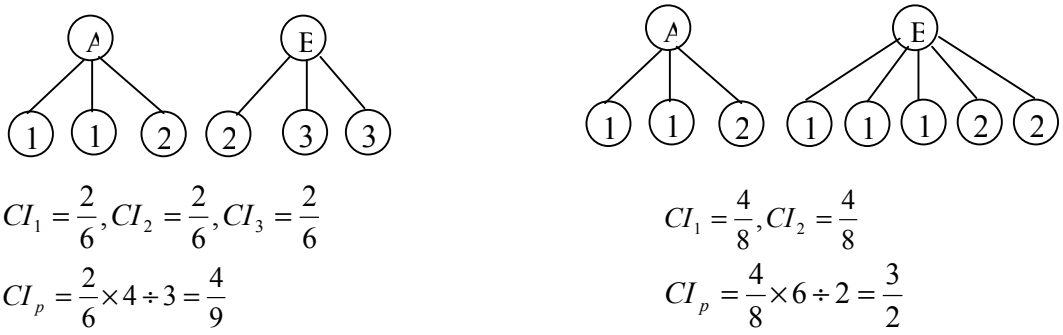


Figure 2. Computational examples for the degree of commonality index

Figure 2 illustrates the use of the CI measures for seven sets of two end products (labeled as A and B). Calculation of the CI is shown below each case. Here, we assume that all demands of products are same, i.e., $d_1 = d_2$.

5. Impact of Component Commonality on Integrated Supply Chain Performance

A multi-level inventory system is often controlled either by an installation stock reorder point policy or by an echelon stock reorder point policy. An *installation stock policy* means that ordering decisions at each installation are based exclusively on the inventory position at this installation. Here, *inventory position* means the stock on hand and on order minus the backlog. When using an echelon stock policy, ordering decisions at each installation are instead based on the echelon inventory position. The *echelon inventory position* is obtained by adding the installation inventory positions at the installation and all its down-stream installations. It is previously known that echelon stock policies dominate installation stock reorder point policies for serial and assembly multi-level inventory systems.

The purpose of the simulation study is to evaluate the performance of "integrated supply chain with component commonality" versus "integrated supply chain without component commonality." The simulation model for an integrated supply chain network with echelon stock policy and commonality index of 1 is shown in Figure 3. This simulation model is a comprehensive model since it contains raw material procurement, manufacturing processes, assembly operations, warehousing, and distribution functions.

Three different performance measures are employed in the experiment: order fill rate, delivery time and total cost. The experimental results for fill rate, delivery time, total cost and resource utilization rate are summarized in Table 1.

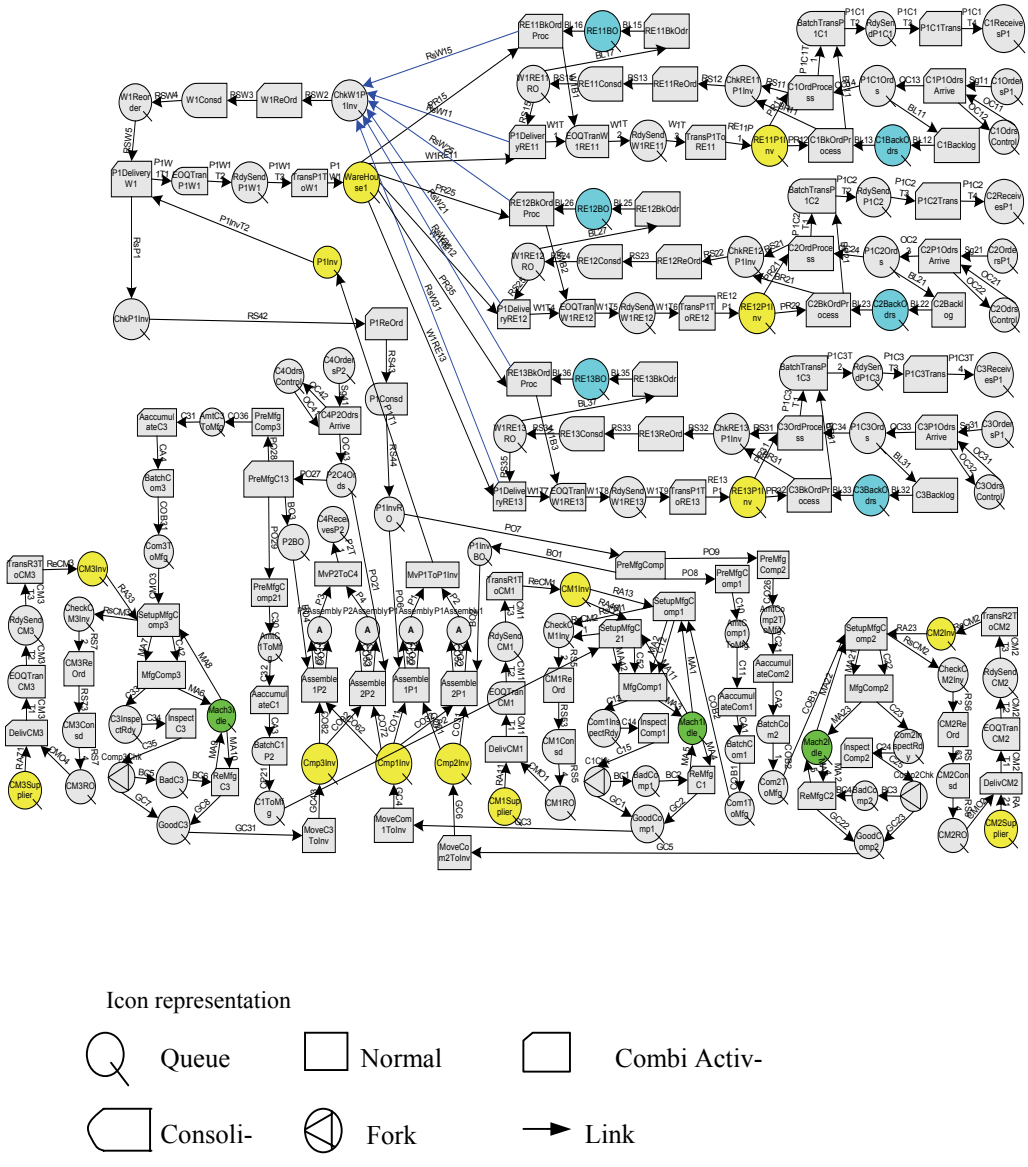


Figure 3. Simulation model for an integrated supply chain network with echelon stock policy and commonality index of 1

CI	Rep	Delivery	R1 Fill	R2 Fill	R3 Fill	M1UtilRate	M2UtilRate	M3UtilRate
	1	4223.63	0.918	0.952	0.918	0.981	0.952	0.901
	2	4250.13	0.882	0.963	0.940	0.976	0.947	0.903
	3	4222.55	0.915	0.964	0.897	0.981	0.952	0.867
	4	4230.74	0.911	0.956	0.934	0.979	0.950	0.917
	5	4240.48	0.881	0.952	0.942	0.977	0.949	0.911
	…	…	…	…	…	…	…	…
1	**496**	4175.87	0.918	0.939	0.971	0.991	0.962	0.888
	497	4207.11	0.960	0.953	0.888	0.984	0.956	0.889
	498	4228.98	0.929	0.966	0.888	0.980	0.952	0.879
	499	4260.51	0.934	0.960	0.888	0.973	0.944	0.899
	500	4249.00	0.956	0.945	0.916	0.976	0.947	0.916
	Me	4239.24	0.920	0.955	0.919	0.978	0.948	0.901
	SD	146.91	0.107	0.044	0.105	0.031	0.030	0.068
	1	8288.54	0.840	0.909	0.838	0.991	0.958	0.839
	2	8284.00	0.844	0.883	0.873	0.991	0.959	0.840
	3	8290.37	0.850	0.908	0.834	0.991	0.959	0.838
	4	8286.03	0.844	0.927	0.811	0.991	0.959	0.840
	5	8286.22	0.815	0.914	0.865	0.991	0.958	0.840
	…	…	…	…	…	…	…	…
1/6	**496**	8294.76	0.846	0.927	0.819	0.991	0.958	0.838
	497	8287.80	0.838	0.913	0.834	0.991	0.958	0.839
	498	8290.99	0.828	0.941	0.819	0.991	0.958	0.838
	499	8285.17	0.859	0.871	0.872	0.991	0.958	0.839
	500	8298.49	0.803	0.923	0.848	0.991	0.957	0.837
	Me	8294.51	0.816	0.939	0.829	0.991	0.958	0.838
	SD	18.14	0.072	0.069	0.077	0.001	0.003	0.003
	1	10211.95	0.758	0.908	0.775	0.686	0.778	1.000
	2	10208.29	0.728	0.904	0.785	0.686	0.778	1.000
	3	10198.74	0.761	0.895	0.761	0.687	0.779	1.000
	4	10206.93	0.762	0.907	0.767	0.686	0.779	1.000
	5	10202.91	0.760	0.904	0.778	0.687	0.778	1.000
	…	…	…	…	…	…	…	…
0	**496**	10212.09	0.745	0.906	0.767	0.685	0.777	1.000
	497	10208.49	0.722	0.896	0.802	0.686	0.778	1.000
	498	10202.85	0.747	0.902	0.802	0.686	0.779	1.000
	499	10203.70	0.763	0.894	0.781	0.686	0.778	1.000
	500	10201.42	0.746	0.918	0.772	0.686	0.779	1.000
	Me	10210.21	0.740	0.907	0.786	0.686	0.778	1.000
	SD	29.80	0.057	0.031	0.071	0.002	0.003	0.000

Table 1. Simulation results for fill rate, delivery time, and resource utilization rate

For each performance measurement, an analysis of variance (ANOVA) is conducted to compare the performance of "integrated supply chain with different component commonality indexes" and "integrated supply chain without component commonality." Here, the performance measures include delivery time and fill rates for different retailers. In the ANOVA, the level of confidence is set as $\alpha = 0.05$.

H_0: $\mu_1 = \mu_2 = \mu_3$.
H_1: At least two of the means are not equal.
The ANOVA are conducted as follows:

(1) Analysis-of-variance for delivery time

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
CI=1	500	2114450	4228.9	594.0537555
CI=1/6	500	4144618.5	8289.237	20.70920108
CI=0	500	5102868.5	10205.737	20.28362331

ANOVA

Source of Variation	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*	*Computed f*	*P-value*	*f critical*
Between Groups	186272964.43	2	93136482.21	439982.602	1.18E-61	3.00
Within Groups	316888.24	1497	211.6821933			
Total	186589852.67	1499				

Table 2. Analysis-of-variance for delivery time

Decision: Since $P<0.05$, or computed $f > f_{critical}$, reject H_0 and conclude that the average delivery time are not all the same.
However, we still don't know which of the delivery-time means are equal and which are different. We need to perform the further multiple comparison tests. Here, we adopt Tukey's test (Walpole et al., 1997). This test allows formation of simultaneous 100(1-α)% confidence intervals for all paired comparisons. The method is based on the studentized range distribution.

From the analysis-of-variance table, we know that the error mean square is s^2= 211.68 (1497 degrees of freedom). The sample means are given by (ascending order):

4239.24, 8294.51, 10210.21

With $\alpha = 0.05$, the value of $q(0.05, 3, 1497) = 3.32$. Thus all absolute differences are to be compared to

$$3.32\sqrt{\frac{211.68}{500}} = 2.16$$

As a result, the following represent means found to be significantly different using Tuksy's procedure:

1 and 2, 2 and 3, 1 and 3.

Therefore, we conclude that the delivery time of integrated supply chain with higher commonality index is significantly (with 95% C.I.) less than that of integrated supply chain with lower commonality index.

(2) Analysis-of-variance for retailers' fill rates

Anova: Single

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
CI=1	500	460.2	0.9204	0.00069449
CI=1/6	500	418.35	0.8367	0.00028468
CI=0	500	374.6	0.7492	0.00021218

ANOVA

Source of Varia-	*Sum of*	*Degrees of Freedom*	*Mean*	*Computed f*	*P-value*	*f criti-*
Between Groups	0.146571	2	0.073285633	184.545201	1.8E-16	3.00
Within Groups	0.59	1497	0.000397115			
Total	0.74	1499				

Table 3. Analysis-of-variance for retailer 1's fill rate

Decision: Since $P<0.05$, or computed $f > f_{critical}$, reject H_0 and conclude that the average fill rate for retailer 1 is not all the same.

The Tukey's test is conducted as follows.

From the analysis-of-variance table, we know that the error mean square is s^2= 0.000397 (1497 degrees of freedom). The sample means are given by (ascending order):

0.74, 0.816, 0.92

With $\alpha = 0.05$, the value of $q(0.05, 3, 1497) = 3.32$. Thus all absolute differences are to be compared to

$$3.32\sqrt{\frac{0.000397}{500}} = 0.00296$$

As a result, the following represent means found to be significantly different using Tuksy's procedure:

1 and 2, 2 and 3, 1 and 3.

Similarly, for retailer 2, we have:

Anova: Single

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
CI=1	500	477.5	0.955	7.4444E-05
CI=1/6	500	455.8	0.9116	0.00044027
CI=0	500	451.7	0.9034	5.2267E-05

ANOVA

Source of Varia-	*Sum of*	*Degrees of*	*Mean Square*	*Computed f*	*P-value*	*f criti-*
Between Groups	0.015377867	2	0.007688933	40.6837815	7.1E-09	3.00
Within Groups	0.283	1497	0.000188993			
Total	0.298	1499				

Table 4. Analysis-of-variance for retailer 2's fill rate

Decision: Since $P<0.05$, or computed $f > f_{critical}$, reject H_0 and conclude that the average fill rate for retailer 2 is not all the same.

The Tukey's test is conducted as follows.

From the analysis-of-variance table, we know that the error mean square is s^2= 0.000189 (1497 degrees of freedom). The sample means are given by (ascending order):

0.907, 0.939, 0.955

With $\alpha = 0.05$, the value of $q(0.05, 3, 1497) = 3.32$. Thus all absolute differences are to be compared to

$$3.32\sqrt{\frac{0.000189}{500}} = 0.00204$$

As a result, the following represent means found to be significantly different using Tuksy's procedure:

1 and 2, 2 and 3, 1 and 3.

For retailer 3, we have:

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
CI=1	500	459.1	0.9182	0.00080773
CI=1/6	500	420.65	0.8413	0.00050934
CI=0	500	389.5	0.779	0.00019733

ANOVA

Source of Variation	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*	*Computed f*	*P-value*	*f critical*
Between Groups	0.097238467	2	0.048619233	96.313147	5.1E-13	3.00
Within Groups	0.756	1497	0.000504804			
Total	0.853	1499				

Table 5. Analysis-of-variance for retailer 3's fill rate

Decision: Since $P<0.05$, or computed $f > f_{critical}$, reject H_0 and conclude that the average fill rate for retailer 3 is not all the same.

The Tukey's test is conducted as follows.

From the analysis-of-variance table, we know that the error mean square is s^2= 0.0005048 (1497 degrees of freedom). The sample means are given by (ascending order):

$$0.786, \qquad 0.829, \qquad 0.919$$

With $\alpha = 0.05$, the value of $q(0.05, 3, 1497) = 3.32$. Thus all absolute differences are to be compared to

$$3.32\sqrt{\frac{0.0005048}{500}} = 0.003336$$

As a result, the following represent means found to be significantly different using Tuksy's procedure:

$$1 \text{ and } 2, \qquad 2 \text{ and } 3, \; 1 \text{ and } 3.$$

From the above analysis, it can be shown that the fill rates of retailers 1, 2 and 3 of the integrated supply chain with higher commonality index are significantly (with 95% C.I.) higher than those of retailers 1, 2 and 3 of the integrated supply chain with lower commonality index, respectively.
Therefore, the fill rates of integrated supply chain with higher commonality index are significantly (with 95% C.I.) higher than those of integrated supply chain with lower commonality index. Furthermore, the relative benefits from component commonality increase with the difference of commonality index values for two supply chain commonality configurations.

(3) Resource utilization rates

By comparing the machines' utilization rates for the network configurations with different degree of commonality (see Table 1), it can be shown that the integrated supply network with higher commonality index will generate more balanced machines' utilization rates than the one with lower commonality index.

6. Production Capacity Flexibility in Integrated Supply Chain Networks

6.1 Manufacturing Flexibility in Supply Chains

In terms of graph theory, a chain is a connected graph. Within a chain, a path can be traced from any product or machine to any other product or machine via the product assignment links. No product in a chain is manufactured by a machine from outside that chain; no machine in a chain produces a product from outside that chain (Jordan and Graves 1995, Graves and Tomlin 2003, Bish, Muriel and Biller 2005). Figure 4 shows different flexibility configurations for a four-product four-machine stage.

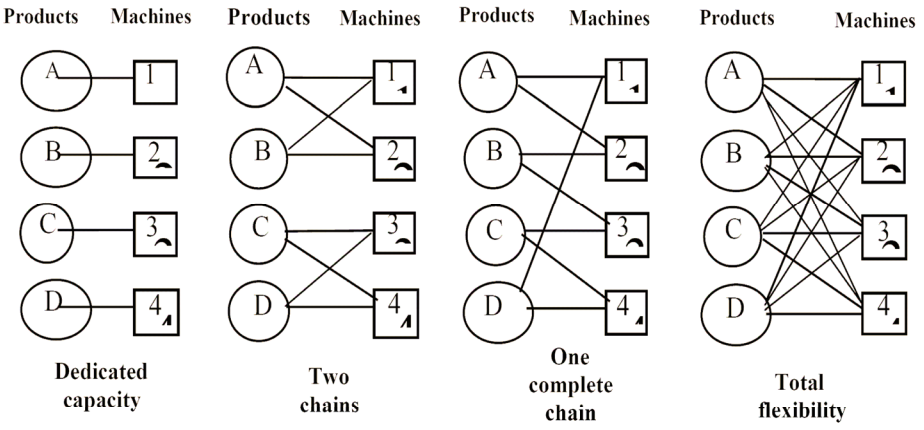


Figure 4. Configurations for manufacturing flexibility in supply chains

Jordan and Graves (1995) demonstrated that the complete chain configuration, in which all products and machines are contained in one chain and the chain is "closed," significantly outperforms the configuration with two distinct chains. If demands are uncertain, multi-stage supply chains face an issue that does not arise in single-stage systems; the bottleneck stage can vary with demand, where the bottleneck stage is that stage that limits throughput. Therefore, one important issue in this research is to examine to what extent the findings of Jordan and Graves apply to multi-stage supply chains. In addition, this chapter

will also investigate the impact of manufacturing flexibility in integrated supply chain networks with different degree of component commonality.

6.2 Design of Experiments

The simulation model for an integrated supply chain network with echelon stock policy and "one complete chain" is shown in Figure 5. Two factors are considered in the simulation study, i.e., manufacturing flexibility and degree of commonality. The design points are described as follows: (1) levels for factor 1 (commonality index): 0 (-), 5/8 (+); and (2) levels for factor 2 (manufacturing flexibility): dedicated capacity (-), one complete chain (+).

First, the manufacturing capacity is assumed to be less than or equal to 75% expected demand. After 500 replications of runs, the simulation results are given as follows:

CI	Flexibility	R1 Fill Rate	R2 Fill Rate	R3 Fill Rate	M1UtilRate	M2UtilRate	M3UtilRate
5/8	One Complete Chain	0.906	0.952	0.944	0.679	0.653	0.663
	Dedicated Capacity	0.92	0.955	0.919	0.978	0.948	0.901
0	One Complete Chain	0.736	0.905	0.785	0.613	0.688	0.699
	Dedicated Capacity	0.74	0.907	0.786	0.686	0.778	1

Table 6. Simulation results for integrated supply chains with "one complete chain" and "dedicated capacity

The 2^k factorial design matrix is shown in the following Table:

Points	Factor 1 (C) Commonality	Factor 2 (F) Flexibility	Responses R1 Fill Rate	R2 Fill Rate	R3 Fill Rate
1	-	-	0.74	0.907	0.786
2	-	+	0.736	0.905	0.785
3	+	-	0.92	0.955	0.919
4	+	+	0.906	0.952	0.944
		e_C =	0.175	0.0475	0.146
		e_F =	-0.009	-0.0025	0.012
		e_{CF} =	-0.005	-0.0005	0.013

Table 7. 2^k factorial design matrix with "one complete chain" and "dedicated capacity" (low demand)

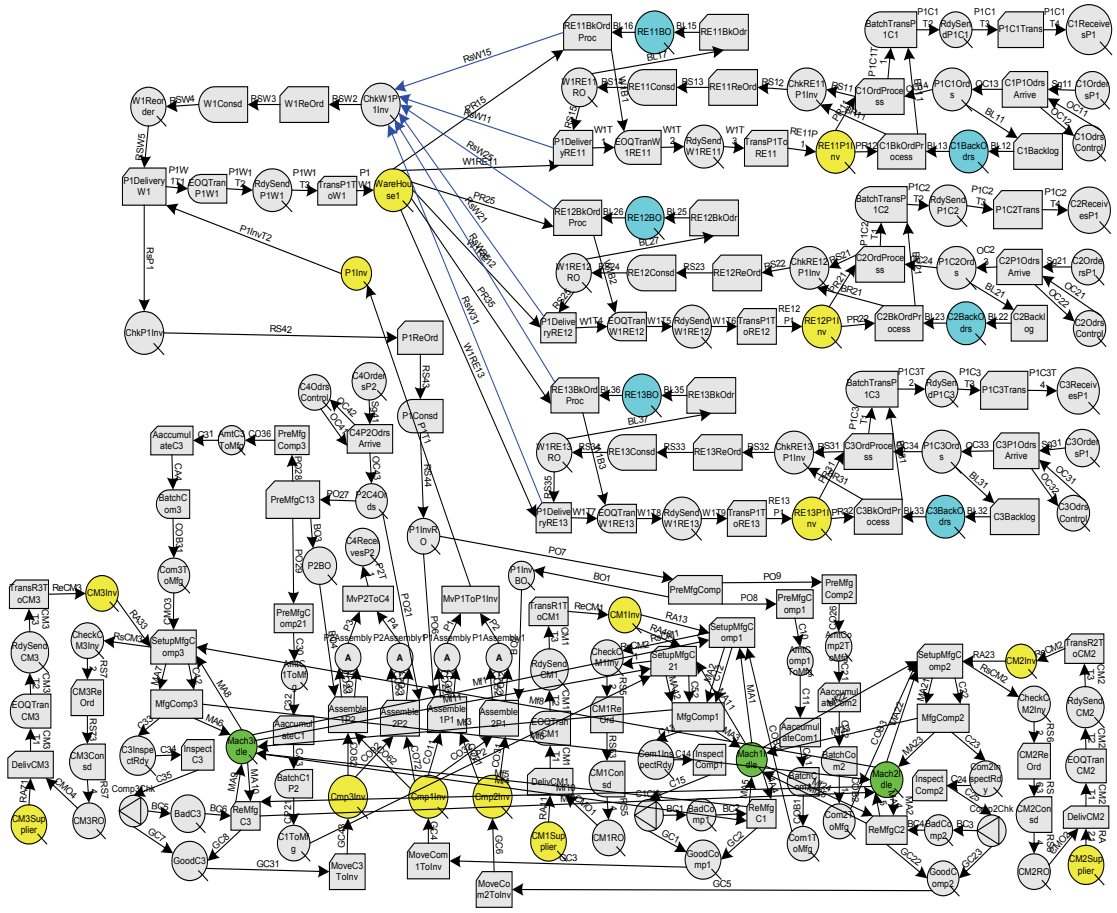


Figure 5. Simulation model for an integrated supply chain network with echelon stock policy and "one complete chain"

The average effect of increasing degree of commonality from 0 to 5/8 is to increase the retailer 1's fill rate by 0.175 (23.7%), increase retailer 2's fill rate by 0.0475 (5.24%) and increase retailer 3's fill rate by 0.146 (18.6%).

On the other hand, the average effect of changing manufacturing flexibility from "dedicated capacity" to "one complete chain" is to decrease the retailer 1's fill rate by 0.009 (1.1%), decrease retailer 2's fill rate by 0.0005 (0.27%) and increase retailer 3's fill rate by 0.013 (1.4%). Therefore, it can seen that, when manufacturing capacity is less than or equal to 75% expected demand, the effect of changing the manufacturing flexibility is not significant as changing the degree of commonality. The t-test shows that there is no significant (with 95% C.I.) difference on fill-rate performance between an integrated supply chain with "one complete chain" and an integrated supply chain with "dedicated capacity."

The interaction effect can be used to judge whether the effect of one factor depends on the levels of the others. The values of the interaction effect e_{CF} are very small and the corresponding t-test shows that 95% confidence interval for $C \times F$ contains zero. So degree of commonality and manufacturing flexibility are not interacting.

Similarly, the performance of integrated supply chains with "total flexibility" and "one complete chain" can be evaluated and compared as follows.

CI	Flexibility	R1 Fill Rate	R2 Fill Rate	R3 Fill Rate	M1UtilRate	M2UtilRate	M3UtilRate
5/8	One Complete Chain	0.906	0.952	0.944	0.679	0.653	0.663
	Total Flexibility	0.92	0.942	0.948	0.997	0.997	0.997
0	One Complete Chain	0.736	0.905	0.785	0.613	0.688	0.699
	Total Flexibility	0.747	0.906	0.776	1	1	1

Table 8. Simulation results for integrated supply chains with "one complete chain" and "total flexibility"

The design points are described as follows: (1) levels for factor 1 (commonality index): 0 (-), 5/8 (+); and (2) levels for factor 2 (manufacturing flexibility): one complete chain (-), total flexibility (+).

Design Points	Factor 1 (C) Commonality	Factor 2 (F) Flexibility	Responses R1 Fill Rate	R2 Fill Rate	R3 Fill Rate
1	-	-	0.736	0.905	0.785
2	-	+	0.747	0.906	0.776
3	+	-	0.906	0.952	0.944
4	+	+	0.92	0.942	0.948
		e_C =	0.1715	0.0415	0.1655
		e_F =	0.0125	-0.0045	-0.0025
		e_{CF} =	0.0015	-0.0055	0.0065

Table 9. 2^k factorial design matrix with "one complete chain" and "total flexibility" (low demand)

The average effect on fill-rate performance by changing manufacturing flexibility from "one complete chain" to "total flexibility" is less than 2%. Therefore, when manufacturing capacity is less than or equal to 75% expected demand, the effect of changing the manufacturing flexibility is not significant. The corresponding t-test shows that there is no significant (with 95% C.I.) difference on fill-rate performance between an integrated supply chain with "one complete chain" and an integrated supply chain with "total flexibility".
Furthermore, it can be observed that the utilization rates of machines become more balanced with the increase of manufacturing flexibility.
In the following, the manufacturing capacity is assumed to be approximately equal to expected demand. After 500 replications of runs, the simulation results are given as follows:

CI	Flexibility	R1 Fill Rate	R2 Fill Rate	R3 Fill Rate	M1UtilRate	M2UtilRate	M3UtilRate
5/8	Total Flexibility	0.913	0.942	0.972	1	1	1
	One Complete Chain	0.892	0.938	0.968	0.899	0.64	0.461
	Dedicated Capacity	0.835	0.884	0.888	0.668	1	0.184
0	Total Flexibility	0.811	0.903	0.834	1	1	1
	One Complete Chain	0.803	0.894	0.826	0.63	0.627	0.743
	Dedicated Capacity	0.744	0.863	0.769	0.999	0.687	0.992

Table 10. Simulation results for integrated supply chains with different flexibility configurations

The 2^k factorial design matrix is shown in the following Table:

Design Points	Factor 1 (C) Commonality	Factor 2 (F) Flexibility	Responses R1 Fill Rate	R2 Fill Rate	R3 Fill Rate
1	-	-	0.744	0.863	0.769
2	-	+	0.803	0.894	0.826
3	+	-	0.835	0.884	0.888
4	+	+	0.892	0.938	0.968
		e_C =	0.09	0.0325	0.1305
		e_F =	0.058	0.0425	0.0685
		e_{CF} =	-0.001	0.0115	0.0115

Table 11. 2^k factorial design matrix with "one complete chain" and "dedicated capacity" (equal demand)

The average effect of increasing degree of commonality from 0 to 5/8 is to increase the retailer 1's fill rate by 0.09 (11.6%), increase retailer 2's fill rate by 0.0325 (3.7%) and increase retailer 3's fill rate by 0.1305 (16.4%).

The average effect of changing manufacturing flexibility from "dedicated capacity" to "one complete chain" is to increase the retailer 1's fill rate by 0.058 (7.35%), increase retailer 2's fill rate by 0.0425 (4.87%) and increase retailer 3's fill rate by 0.0685 (8.27%). Therefore, it can seen that, when manufacturing capacity is approximately equal to expected demand, the effect of changing the manufacturing flexibility is significant. The t-test shows that there is a significant (with 95% C.I.) increase on fill-rate performance by changing from an integrated supply chain with "dedicated capacity" to an integrated supply chain with "one complete chain."

The values of the interaction effect e_{CF} are very small and the corresponding t-test shows that 95% confidence interval for $C \times F$ contains zero. So the degree of commonality and the manufacturing flexibility are not interacting.

Similarly, for equal demand situation, the performance of integrated supply chains with "total flexibility" and "one complete chain" can be evaluated and compared as follows.

Design Points	Factor 1 (C) Commonality	Factor 2 (F) Flexibility	Responses R1 Fill Rate	R2 Fill Rate	R3 Fill Rate
1	-	-	0.803	0.894	0.826
2	-	+	0.811	0.903	0.834
3	+	-	0.892	0.938	0.968
4	+	+	0.913	0.942	0.972
		e_C =	0.0955	0.0415	0.14
		e_F =	0.0145	0.0065	0.006
		e_{CF} =	0.0065	-0.0025	-0.002

Table 12. 2^k factorial design matrix with "one complete chain" and "total flexibility" (equal demand)

The average effects (for three retailers) on fill-rate performance by changing manufacturing flexibility from "one complete chain" to "total flexibility" are less than 2%. Therefore, when manufacturing capacity is approximately equal to expected demand, the effect of changing the manufacturing flexibility is not significant. The corresponding t-test shows that there is no significant (with 95% C.I.) difference on fill-rate performance between an integrated supply chain with "one complete chain" and an integrated supply chain with "total flexibility."
Same as the low demand situation, it can be observed that the utilization rates of machines become more balanced with the increase of manufacturing flexibility.

7. Conclusions

Effective configuration of the supply chain networks is nowadays recognized as a key determinant of competitiveness and success for most manufacturing organizations. This paper focuses on the simulation study of integrated supply chain network configurations and performance analysis.
First, this paper presents an integrated modeling framework for supply chains that can be used to model the different network topologies such as serial, parallel, assembly and arborescent structures. Second, a component-based simulation modeling approach is suggested. The advantage of the component-based simulation framework is that the reconfiguration of supply chain networks for different design alternatives can be easily achieved. To keep the op-

erations at a high level of efficiency, in the presence of a large product variety, companies resort to certain strategies, important of which are product component standardization and machine flexibility. Component commonality can greatly reduce the inventory of a supply chain and improve its performance. Similarly machine flexibility would enable the machine process different operations and components, to keep a low machine idle time. In this research, design of experiments and Tukey's test are employed to investigate the effects of component commonality and manufacturing flexibility on supply chain performance criteria such as delivery time, fill rate and cost in an integrated environment.

8. References

Beamon, B.M., 1998, Supply Chain Design and Analysis: Models and Methods, *International Journal of Production Economics*, 55, pp.281-294

Bish, E.K., Muriel, A. and Biller, S. Managing flexible capacity in a make-to-order environment, *Management Science*, v 51, n 2, February, 2005, p 167-180.

Collier, D., 1981, The Measurement and Operating Benefits of Component Part Commonality, *Decision Sciences*, Vol.12, No.1, January 1981, pp.85-97

Dong, M., 2001, Process Modeling, Performance Analysis and Configuration Simulation in Integrated Supply Chain Network Design, Ph.D. dissertation, Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

Eynan, A. and Rosenblatt, M J., 1996, Component Commonality Effects on Inventory Costs, *IIE Transactions*, Vol.28 n.2, pp.93-104

Graves, S.C. and Tomlin, B.T., Process flexibility in supply chains, *Management Science*, v 49, n 7, July, 2003, p 907-919.

Hillier, M.S., 1999, Component Commonality in a Multiple-period Inventory Model with Service Level Constraints, *International Journal of Production Research*, v.37, n.12, pp.2665-2683

Hong, Jae-Dong and Hayya, Jack C., 1998, Commonality in an Integrated Production-Inventory Model for a Two-product Single-facility System, *Proceedings - Annual Meeting of the Decision Sciences Institute*, v 3, Atlanta, GA, USA, pp.1287-1289

Jordan, W.C. and Graves, S.C. 1995, Principles on the Benefits of Manufacturing Process Flexibility, Management Science, Vol. 41, No. 4, pp. 577-594.

Martinez, J.C., 1996, *STROBOSCOPE: State and Resource Based Simulation of Construction Processes*, Ph.D. Dissertation, University of Michigan, Ann Arbor, MI.

Miguel, F.R., F. Xavier, G. and Enrique, L.T., 1999, *3C: A Proven Alternative to MRPII for Optimizing Supply Chain Performance*, CRC Press - St. Lucie Press.

Ross, A., Venkataramanan, M.A. and Ernstberger, K.W., 1998, Reconfiguring the Supply Network Using Current Performance Data, *Decision Sciences,* 29:(3), pp.707-728

Swaminathan, J.M. and Tayur, S.R., 1998, Managing Broader Product Lines through Delayed Differentiation using Vanilla Boxes, *Management Science*, v.44, n.12, pp.S161-S172

Walpole, R.E., Myers, R.H., and Myers, S.L., 1997 (6 edition), *Probability and Statistics for Engineers and Scientists*, Prentice Hall.

30

On Direct Adaptive Control for Uncertain Dynamical Systems - Synthesis and Applications

Simon Hsu-Sheng Fu and Chi-Cheng Cheng

1. Introduction

In the rapidly growing research on nonlinear control theory, much work has been focused on the problems of uncertainties exist in the system model or systems with unknown disturbances and nonlinearities. A direct adaptive control framework for adaptive stabilization, disturbance rejection, and command following of multivariable nonlinear uncertain systems with exogenous disturbances, where the bounded disturbances were assumed to be a known vector, has developed in (Haddad & Hayakawa, 2002) and guarantees partial stability of the closed-loop system. However, it is worth to note that the disturbances may be the result of unmodeled dynamics, noisy measurements, parameter uncertainty, or non dissipative forces affecting the plant, and most of time not available for the control design.

There are considerable amount of literatures published the area of adaptive control synthesis for uncertain systems. However, the application of Lyapunov stability theory along this track still shown relative limited results, especially for discrete-time systems. The major difficulty encountered concerns the proof of the global stability of the overall adaptive control loop. The main reason is that the Lyapunov candidate cannot easily be constructed, such that the negative definiteness of the Lyapunov difference could not easily shown (Zhao & Kanellakopoulos, 1997).

For direct adaptive control gains are adjusted without explicit parameter identification. In this Chapter, we are investigating the problem of direct adaptive control of uncertain systems, where both discrete-time and continuous-time systems are considered. For continuous time case, motivated by the result of robust stabilization of nonlinear systems affected by time-varying uniformly bounded affine disturbances (Loria et al., 1998), where a passive-based control framework has formulated and achieved global uniform convergence. Facilitating the direct adaptive scheme, our framework guarantees that the closed-

loop system is Lyapunov stable under the assumption of matched disturbances. In addition, the asymptotic stable of solution x with respect to origin can be proved.

There were considerable amount of discrete-time adaptive results have been published. For example, discrete-time neural net adaptive controller was depicted in (Levin & Narendra, 1996), the MIT rule for adaptive control refers to the combination of model reference control together with a gradient type parameter update law (Mareels & Polderman, 1996), and a stable and convergent direct adaptive control has been developed in (Johansson, 1989). An ARMARKOV model for MIMO uncertain systems achieved adaptive disturbance rejection and traction (Venugopal & Bernstein, 1999). In addition, Shibata et al. proposed a simplified adaptive control scheme based on Lyapunov analysis while the system satisfies the so called almost strictly positive real (ASPR) condition (Shibata et al., 1996). Bar-Kana (Bar-Kana, 1989) also used ASPR assumption and presented a robust discrete-time adaptive algorithm subjected to the condition of BIBO and the boundedness of the residual term. Guo (Guo, 1997) examined the global stability for a class of discrete-time adaptive nonlinear control systems and proved critical stability for least square-based adaptive control systems.

Furthermore, several most recent works were published and the results were close to our results presented in this Chapter. A direct adaptive control for reachable linear discrete-time systems with exogenous disturbances (Fu & Cheng, 2003, a) and ℓ_2 disturbances (Fu & Cheng, 2003, b), direct adaptive control application to a class of linear discrete-time systems, where the nominal system A is known and the deviation of $|A - A_c| = |BK_g|$ is bounded, were investigated by (Fu & Cheng, 2004, a); (Fu & Cheng, 2004, b), and direct adaptive control for a class of nonlinear normal discrete-time systems were presented in (Fu & Cheng, 2004, c), all results above satisfied Lyapunov stability theory. In addition, robust direct adaptive control of nonlinear uncertain systems with unknown disturbances were proposed in (Fu & Cheng, 2005, a); (Fu & Cheng, 2005, b). However, these solutions were limited by the hypothesis of trajectory dependence. In this paper we successfully release this limitation and obtain stability results, such that the discrete-time system stability theory (Hitz & Anderson, 1969) can be applied.

The contents of this paper are as follows. In Section 2, we present the adaptive control framework for uncertain continuous-time nonlinear systems with matched disturbances and discrete-time systems with exogenous and ℓ_2 dis-

turbances. Next, several numerical examples are presented in Section 3, which include van der Pol oscillator, one linked rigid robot, and active suspension systems, to demonstrate the efficacy of the proposed frameworks. Finally, we illustrate the results of this paper and future research in Section 4.

2. Adaptive Control for Uncertain Continuous-Time Nonlinear Systems with Matched Disturbances

Our main concern in this paper is to deal with uncertain nonlinear systems perturbed by affine disturbances. We begin by considering the problem of characterizing adaptive feedback control laws for nonlinear uncertain MIMO systems ***G*** given by

$$\dot{x} = f(x(t)) + G(x(t))u(x(t)) + J(x(t))w(t, x(t)) \qquad (1)$$

where $x(t) \in R^n$ is the state vector, $x(0) = x_0$, $u(t) : R^n \to R^m$ is the control vector, $f : R^n \to R^n$ characterize system dynamics with uncertain entries, and $f(0) = 0$. $G : R^n \to R^{n\times m}$ and $J : R^n \to R^{n\times d}$ are the input and disturbance weighting matrix functions, respectively, with unknown entries. In addition, the disturbance vector $w : R \times R^n \to R^{d\times d}$ satisfies **Assumption 2.1** illustrated next.

Assumption 2.1 (Loria et al., 1998)
The vector function $w(t, x(t))$ is bounded, and can be characterized by

$$w(t, x(t)) \le \overline{w}(t, x(t))\theta_1 + \theta_2 \qquad (2)$$

where $\theta_1 \in R^d$ and $\theta_2 \in R^d$ are unknown constants, and $\overline{w} : R \times R^n \to R^{d\times d}$ is a known continuous matrix function.

It is important to note that the disturbance $w(t, x(t))$ may be the result of unmodeled dynamics, noisy measurement, parameter uncertainty or exogenous disturbances. For the nonlinear system ***G***, we assume that the existence and uniqueness of solutions are satisfied and zero-state observability of (1) while

$w(t,x(t)) \equiv 0$. Furthermore, assume there exits $F : R^n \to R^s$ with $F(0)=0$, $K_g : R^{n\times s}$, and $\overline{G} : R^n \to R^{m\times m}$ such that

$$f_c(x(t)) \underline{\underline{\Delta}} f(x(t)) + G(x(t))\overline{G}(x(t))K_g F(x(t)) \qquad (3)$$

is globally asymptotically stable, where a scalar function $V_s : R^n \to R$ is Lyapunov function, and $\ell : R^n \to R^t$. Then

$$V_s'(x)\, f_c(x) = -\ell^T(x(t))\ell(x(t)), \forall x : R^n. \qquad (4)$$

Theorem 2.1 (Fu & Cheng, 2005)
Consider the nonlinear uncertain system ***G*** given by (1) is zero state observable with $w(t,x(t)) \equiv 0$, where the disturbances $w(t,x(t))$ satisfy **Assumption 2.1**. In addition, let that the zero solution of (1) defined in (3) is globally asymptotically stable. Furthermore, there exists matrix functions $\Psi : R^{m\times d}$ and $\overline{J} : R^n \to R^{m\times m}$, such that the matching condition $G(x)\overline{J}(x)\Psi = J(x)$ is satisfied. Then the adaptive feedback control law

$$u(x) = \overline{G}(x)K(t)F(x) + \overline{J}(x)\Phi(t)(\overline{w}(x,t)\hat{\theta}_1 + \hat{\theta}_2), \qquad (5)$$

where $K(t) : R^{m\times n}$, $\Phi(t) : R^{m\times d}$, $\tilde{\theta}_1 \underline{\underline{\Delta}} \hat{\theta}_1 - \theta_1$, and $\tilde{\theta}_2 \underline{\underline{\Delta}} \hat{\theta}_2 - \theta_2$. Now, let the design matrices $P_1 > 0$, $P_2 > 0$, $Q_1 > 0$, $Y > 0$, $Q_2 > 0$, and $Z > 0$ with the update laws

$$\dot{K} = -\frac{1}{2}Q_1\overline{G}^T(x)G^T(x)V_s'^T(x)F^T(x)Y, \qquad (6)$$

$$\dot{\Phi} = -\frac{1}{2}Q_2\overline{J}^T(x)G^T(x)V_s'^T(x)(\overline{w}(x,t)\hat{\theta}_1 + \hat{\theta}_2)^T Z, \qquad (7)$$

And

$$\dot{\hat{\theta}}_1 = \frac{1}{2}P_1^{-1}\overline{w}^T(x,t)J^T(x)V_s'^T(x), \qquad (8)$$

$$\dot{\hat{\theta}}_2 = \frac{1}{2}P_2^{-1}J^T(x)V_s'^T(x), \qquad (9)$$

where $V_s'(x) \underline{\underline{\Delta}} \frac{\partial V_s(x)}{\partial x}$, guarantees that the closed-loop system, given by (1) and (5) to (9), is Lyapunov stable. In addition, if (4) is applied and let the output $y(t) \underline{\underline{\Delta}} \ell(x)$, then $\ell(x) \to 0$ as $t \to \infty$. Furthermore, the asymptotic stable solution x with respect to origin will arrive when $\ell^T(x)\ell(x) > 0$.

Proof

To show Lyapunov stability of the closed-loop system (1) and (5) to (9), we first consider the Lyapunov function candidate

$$\begin{aligned} &V(x, K, \Phi, \theta_1, \theta_2) = \\ &V_s(x) + trQ_1^{-1}(K - K_g)Y^{-1}(K - K_g)^T \\ &+ trQ_2^{-1}(\Phi - \Psi)Z^{-1}(\Phi - \Psi)^T + \tilde{\theta}_1^T P_1 \tilde{\theta}_1 + \tilde{\theta}_2^T P_2 \tilde{\theta}_2 \end{aligned} \tag{10}$$

where $V_s(x)$ satisfies the condition of (4) and tr represents trace operator. Note that the Lyapunov candidate $V(0, K_g, -\Psi, 0, 0) = 0$ and $V(x, K, \Phi, \theta_1, \theta_2) > 0$ for all $(x, K, K_g, \theta_1, \theta_2) \neq (0, K_g, -\Psi, 0, 0)$. In addition, $V(x, K, \Phi, \theta_1, \theta_2)$ is radially unbounded. Furthermore, $V(\bullet, K, \Phi, \theta_1, \theta_2)$ and K are continuous in x for $t \geq 0$. The corresponding Lyapunov derivative is given by

$$\begin{aligned} \dot{V} &= V_s'(x)[f(x) + G(x)u(t) + J(x)(\overline{w}(x,t)\theta_1 + \theta_2)] \\ &+ 2\dot{\tilde{\theta}}_1^T P_1 \tilde{\theta}_1 + 2\dot{\tilde{\theta}}_2^T P_2 \tilde{\theta}_2 + 2trQ_1^{-1}(K - K_g)Y^{-1}\dot{K}^T \\ &+ 2trQ_2^{-1}(\Phi + \Psi)Z^{-1}\dot{\Phi}^T \\ &= V_s'(x) f(x) + V_s'(x)G(x)[u(t) - \overline{G}(x)KF(x) \\ &- \overline{J}(x)\Phi(\overline{w}(x,t)\hat{\theta}_1 + \hat{\theta}_2)] + 2trQ_2^{-1}(\Phi + \Psi)Z^{-1}\dot{\Phi}^T \\ &+ V_s'(x)G(x)\overline{G}(x)(K - K_g)F(x) + 2\dot{\hat{\theta}}_1^T P_1 \tilde{\theta}_1 \\ &+ 2trQ_1^{-1}(K - K_g)Y^{-1}\dot{K}^T + V_s'(x)G(x)\overline{J}(x)\Phi(\overline{w}(x,t)\hat{\theta}_1 + \hat{\theta}_2) \\ &- V_s'(x)J(x)\overline{w}(x,t)\tilde{\theta}_1 - V_s'(x)J(x)\overline{w}(x,t)\tilde{\theta}_2 + 2\dot{\hat{\theta}}_2^T P_2 \tilde{\theta}_2 \\ &= V_s'(x) f_c(x) \end{aligned} \tag{11}$$

Next, since the condition (4) is satisfied, the resulting Lyapunov derivative along the system trajectory is

$$\dot{V}(x, K, \Phi, \theta_1, \theta_2) = -\ell^T(x)\ell(x) \leq 0. \tag{12}$$

This completes the proof. Furthermore, if $\ell(x) \to 0$ as $t \to \infty$, and the asymptotic stable solution x with respect to origin will arrive when $\ell^T(x)\ell(x) > 0$.
We further extend the above result to the case where the entries of the system matrix and the input matrix are uncertain. Note that the adaptive control law (5) does not require explicit knowledge on the desire gain matrix K_g, disturbances $w(t, x(t))$, system dynamics $f(x)$, and matching matrix Φ.
Theorem 2.1 also requires that the zero solution to (3) is globally asymptotically stable. Next, we consider the case where $f(x)$, input weighting matrix $G(x) = B$, and disturbance weighting matrix $J(x) = D$ are uncertain. Specifically, given as the following

$$\dot{x} = f(x) + Bu(t) + Dw(x,t), \tag{13}$$

where $w(t, x(t))$ satisfies **Assumption 2.1**, and

$$f_c(x) \triangleq f(x) + BK_g F(x), \tag{14}$$

is global asymptotically stable. Next, let $B_s : R^{m\times m}$ is the sign definite matrix with unknown entries; such that (Fu & Cheng, 2004)

$$B = [0_{m\times(n-m)}, B_s]^T \; B_0 \triangleq \begin{cases} [0_{m\times(n-m)}, I_m]^T, B_s > 0 \\ [0_{m\times(n-m)}, -I_m]^T, B_s < 0 \end{cases} \tag{15}$$

and

$$B_s \triangleq UD_B U, \; |B_s| = \sqrt{B_s^2}, \tag{16}$$

Where U is orthogonal and D_B is real diagonal. Similarly, assume that $D_s : R^{d\times d}$ is the sign definite matrix with unknown matrix; that is

$$D = [0_{d\times(n-d)}, D_s]^T \; D_0 \triangleq \begin{cases} [0_{d\times(n-d)}, I_d]^T, D_s > 0 \\ [0_{d\times(n-d)}, -I_d]^T, D_s < 0 \end{cases} \tag{17}$$

and

$$D_s \underline{\underline{\Delta}} \hat{U} D_D \hat{U}, \ |D_s| = \sqrt{D_s^2}, \tag{18}$$

Corollary 2.1
Consider the nonlinear uncertain system given by (13) is zero state observable. Let B and D satisfy (15) and (17), respectively. Then, the adaptive feedback control law

$$u(x) = K(t)F(x) + \Phi(t)(\overline{w}(x,t)\hat{\theta}_1 + \hat{\theta}_2), \tag{19}$$

with the update laws

$$\dot{K} = -\frac{1}{2} B_0^T V_s'^T(x) F^T(x) Y, \tag{20}$$

$$\dot{\Phi} = -\frac{1}{2} B_0^T(x) V_s'^T(x)(\overline{w}(x,t)\hat{\theta}_1 + \hat{\theta}_2)^T Z, \tag{21}$$

and

$$\dot{\hat{\theta}}_1 = \frac{1}{2} \overline{w}^T(x,t) D_0^T V_s'^T(x), \tag{22}$$

$$\dot{\hat{\theta}}_2 = \frac{1}{2} D_0^T V_s'^T(x), \tag{23}$$

guarantees that the closed-loop system, given by (13), (19), and (20) to (23) is Lyapunov stable. Furthermore, if (14) is applied and let the output $y(t) \underline{\underline{\Delta}} \ell(x)$, then $\ell(x) \to 0$ as $t \to \infty$. Furthermore, the asymptotic stable solution x with respect to origin will arrive when $\ell^T(x)\ell(x) > 0$.

Proof
The result is a direct extension of **Theorem 2.1**. Let $G(x) = I_m$ and $\hat{J}(x) = I_m$, and the matching condition be $B\hat{J}(x)\Psi = D$. In addition, let $P_1 > 0$, $P_2 > 0$, $Q_1 > 0$, $Q_2 > 0$, $Y > 0$, $Z > 0$, and assume that $P_1^{-1}\overline{w}^T(x,t) = \overline{w}^T(x,t)P_1^{-1}$, and let

Q_1 be replaced by $q_1|B_s|^{-1}$, Q_2 be replaced by $q_2|B_s|^{-1}$, P_1 be replaced by $q_3|D_s|$, and P_2 be replaced by $q_4|D_s|$, where $q_i > 0$, $i = 1,2,3,4$ are arbitrary real. Next, let $q_1 Y$ and $q_2 Z$ be replaced by Y and Z respectively. Finally, let $q_3 = q_4 = 1$, then the resulting update laws (20) and (21) are obtained.

Note that the frameworks of **Theorem 2.1** and **Corollary 2.1** can extend to linear systems such that $f(x) = Ax$ and $f_c(x) = A_c x$, where $A_c = A + BK_g$ is an asymptotically stable matrix. Also, applied to nonlinear time-varying uncertain systems given by

$$\dot{x} = f(x(t),t) + G(x(t),t)u(x(t)) + J(x(t),t)w(t,x(t)) \qquad (24)$$

and tracking problems given by

$$\dot{e} = f(e(t)) + G(e(t))u(e(t)) + J(e(t))w(t,e(t)) \qquad (25)$$

where $e(t) = x(t) - r_d(t)$ is tracking error, and $r_d(t)$ is reference. Next, we present the discrete-time counterpart of direct adaptive control for Uncertain Nonlinear Systems given as Section 3.

3. Adaptive Control Designs for Nonlinear Uncertain Discrete-Time Systems

3.1 Discrete-Time Systems with Disturbance Measurement

In this section, we extend the results of **Theorem 2.1** to nonlinear uncertain discrete time MIMO systems with disturbances measurement given by

$$x(k+1) = f(x(k)) + G(x(k))u(x(k)) + J(x(k))w(k), \qquad (26)$$

which is zero state observable when $w(k) \equiv 0$, where $x(k) \in R^n$ is the state vector, $u(k): R^n \to R^m$ is the control vector, $f: R^n \to R^n$ characterize system dynamics with uncertain entries, and $f(0) = 0$. $G: R^n \to R^{n\times m}$ and $J: R^n \to R^{n\times d}$ are the input and disturbance weighting matrix functions, respectively. In ad-

dition, let the disturbance vector $w: R \times R^n \to R^{d \times d}$ is measurable, then the feedback law given by

$$u(x(k)) = \overline{G}(x(k))K(k)F(x(k)) + \overline{J}(x(k))\Phi(k)w(k), \tag{27}$$

where $K(k): R^{m \times n}$, $\Phi(k): R^{m \times d}$, and $F: R^n \to R^s$.

Theorem 3.1 (Fu & Cheng, 2005)
Consider the nonlinear discrete time MIMO systems with exogenous disturbances given by (26). Next, assume there exist $\hat{J}: R^n \to R^{n \times d}$ and $\Psi: R^{m \times d}$, such that the matching condition $G(x)\hat{J}(x)\Psi = J(x)$ is satisfied. Furthermore, let $P_{1u}: R^n \to R^{1 \times m}$, $P_{2u}: R^n \to R^{m \times m}$, $P_{uw}: R^n \to R^{m \times d}$, $P_{1w}: R^n \to R^{1 \times d}$, and $P_{2w}: R^n \to R^{d \times d}$, the Lyapunov function V_s is defined as

$$\begin{aligned} &V_s(x(k+1)) \underline{\underline{\Delta}} V_s(f(x)) + P_{1u}(x)u(x) + u^T(x)P_{2u}(x)u(x) \\ &+ u^T(x)P_{uw}(x)w(k) + P_{1w}(x)w(k) + w^T(k)P_{2w}(x)w(k), \end{aligned} \tag{28}$$

In addition, let $\Gamma: R^n \times R^d \to R$ is a positive scalar function, $\ell: R^n \to R^p$ is output vector, then

$$0 = V_s(f_c(x)) - V_s(x) + \ell^T(x)\ell(k) + \Gamma(x, w). \tag{29}$$

The adaptive feedback control law (27), with the measurable disturbances, and the update laws

$$\begin{aligned} &K(k+1) = K(k) - Q_1\hat{G}^T(x(k))G^T(x(k))PG(x(k)) \\ &[2K(k)F(x(k))F^T(x(k)) + \hat{J}(x(k))\Phi(k)w(k)F^T(x(k))]Y, \end{aligned} \tag{30}$$

$$\begin{aligned} &\Phi(k+1) = \Phi(k) - Q_2\hat{G}^T(x(k))G^T(x(k))PG(x(k))G^T(x(k))P \\ &[J(x(k)) - 2G(x(k))\hat{J}(x(k))\Phi(x(k))]w(k)w^T(k)Z, \end{aligned} \tag{31}$$

where $Q_1 > 0$, $Y > 0$, $Q_2 > 0$, and $Z > 0$, guarantees that the closed-loop system given by (26), (27), (30) and (31) is Lyapunov stable.

Proof

To show Lyapunov stability of the closed-loop system, given by (26), (27), (30) and (31). We first consider the Lyapunov function candidate

$$\begin{aligned} &V(x(k), K(k), \Phi(k)) = \\ &V_s(x(k)) + trQ_1(K(k) - K_g)Y^{-1}(K(k) - K_g)^T \\ &+ trQ_2(\Phi(k) - \Psi)Z^{-1}(\Phi(k) - \Psi)^T, \end{aligned} \tag{32}$$

Note that the Lyapunov candidate $V(0, K_g, \Psi) = 0$, and $V(x, K, \Phi) > 0$ for all $(x, K, \Phi) \neq (0, K_g, \Psi)$. In addition, $V(\bullet, K, \Phi)$ and K are continuous with respect to x, $V(x, K, \bullet)$ and Φ are continuous with respect to w for $k \geq 1$. Let $x(k)$, $k \geq 0$, denotes the solution of the closed-loop system (26) and (27), and is global asymptotic stability when $w(k) \equiv 0$. The corresponding Lyapunov difference is given by

$$\begin{aligned} &\Delta V(k) = \Delta V(x(k), K(k), \Phi(k)) = \\ &V(x(k+1), K(k+1), \Phi(k+1) - V(x(k), K(k), \Phi(k)), \end{aligned} \tag{33}$$

and follow the similar proof of **Theorem 2.1** with the following adaptive laws

$$K(k+1) = K(k) - Q_1 \hat{F}(x(k), w(k)) F^T(x(k)) Y, \tag{34}$$

$$\Phi(k+1) = \Phi(k) - Q_2 R_w(w(k)) w^T(k) Z, \tag{35}$$

where

$$\begin{aligned} &\hat{F}(x, w) = \frac{1}{2} \hat{G}^T(x)[P_{1u}^T(x) + P_{2u}(x)\hat{G}(x)K(k)F(x) \\ &+ P_{2u}(x)\hat{J}(x)\Phi(k)w(k)], \end{aligned} \tag{36}$$

$$\begin{aligned} &R_w(w(k)) = -\frac{1}{2} \hat{J}^T(x(k))[P_{uw}(x(k)) \\ &+ G^T(x(k))PJ(x(k))]w(k), \end{aligned} \tag{37}$$

Next, let $P = N^T N$, $N : R^{n \times n}$, and chose the following

$$P_{uw}(x(k)) = -4 P_{2u}(x(k)) \hat{J}(x(k)) \Phi(k), \tag{38}$$

$$P_{1u}(x(k)) = 2 F^T(x(k)) K^T(k) \hat{G}^T(x(k)) P_{2u}(x(k)), \tag{39}$$

$$P_{2u}(x(k)) = G^T(x(k)) P G(x(k)), \tag{49}$$

$$\begin{aligned} &P_{1w}(x(k)) = \\ &2 w^T(k) \Phi^T(k) \hat{J}^T(x(k)) P_{2u}(x(k)) \hat{J}(x(k)) \Phi(k), \end{aligned} \tag{41}$$

$$\begin{aligned} &P_{2w}(x(k)) = [G^T(k) \hat{J}^T(x(k)) \Phi(k) - J(x(k))]^T \\ &P[G^T(k) \hat{J}^T(x(k)) \Phi(k) - J(x(k))], \end{aligned} \tag{42}$$

by substituting (34) and (35) into (33), after some manipulations yields

$$\begin{aligned} &\Delta V(k) = [w^T(k) Z w(k)] R_w^T(x) Q_2 R_w(x) + \\ &[F^T(x) Y F(x)] \hat{F}^T(x) Q_1 \hat{F}(x) + V_s(f_c(x)) - V_s(x) \\ &- w^T(k) \Phi^T(k) \hat{J}^T(x) P_{2u}(x) \hat{J}(x) \Phi(k) w(k) \\ &2 w^T(k) \Phi^T(k) \hat{J}^T(x) P_{2u}(x) \hat{J}(x) \Phi(k) F(x) \\ &- \Gamma_F^T(x) G^T(x) P G(x) \Gamma_F(x), \end{aligned} \tag{43}$$

$$\Gamma_F(x(k)) = \hat{G}^T(x(k))[K(k) - K_g] F(x(k)), \tag{44}$$

Since (29) is satisfied, and let

$$\begin{aligned} &\Delta V(k) = -\left| NG(x) \hat{G}(x)[K(k) - K_g] F(x) + NG(x) \hat{J}(x) \Phi(k) w(k) \right|_2^2 \\ &- \ell^T(x) \ell(x) \le -\ell^T(x) \ell(x). \end{aligned} \tag{43}$$

where $x(k)$ denotes the solution to the closed-loop dynamical system (26) and (27). Then the resulting Lyapunov difference becomes

This completes the proof. If $\ell(x) \neq 0$, $k \geq 0$, then $x \to 0$ as $k \to \infty$. Furthermore, if $\ell(x) \to 0$ as $k \to \infty$, and the asymptotic stable solution x with respect to origin will arrive when $\ell^T(x)\ell(x) > 0$.

Note that the adaptive control laws (30) and (31) do not require explicit knowledge of the matrix K_g, the disturbance matching matrix Ψ and system dynamics $f(x(k))$. Next, we extend the solution of **Theorem 3.1** to the following dynamic system

$$x(k+1) = f(x(k)) + Bu(x(k)) + Dw(k), \tag{44}$$

where the entries of B and D are unknown and satisfy the conditions given in (15) and (17), respectively.

Corollary 3.1
Consider the nonlinear discrete time system given by (44). Next, let $F: R^n \to R^s$ and there exists $K_g: R^{n\times s}$ such that $f_c(x) \triangleq f(x) + BK_g F(x)$ is exponentially stable. In addition, let $\Psi: R^{m\times d}$ and the matching condition $B\Psi = D$ is satisfied. Then the feedback law

$$u(x(k)) = K(k)F(x(k)) + \Phi(k)w(k), \tag{45}$$

with the adaptive gain matrices

$$\begin{aligned} K(k+1) &= K(k) - q^2 B_0^T P B_0 [2K(k)F(x(k)) \\ &+ \Phi(k)w(k)]F^T(x(k))Y, \end{aligned} \tag{46}$$

$$\Phi(k+1) = \Phi(k) - q^2 B_0^T P[2B_0\Phi(k) - D_0]w(k)w^T(k), \tag{47}$$

where $K(k) = |B_s|K(k)$, $\Phi(k) = |B_s|\Phi(k)|D_s|^{-1}$, $w(k) = |D_s|w(k)$, and $q > 0$, guarantees that the closed-loop system given by (44), (45), (46), and (47) is Lyapunov stable, and equivalent to the following

$$x(k+1) = f(x(k)) + B_0K(k)F(x(k)) + (B_0\Phi(k) + D_0)w(k), \tag{48}$$

Proof

The proof is a direct extension of **Theorem 3.1**. First, we consider the Lyapunov candidate given by (32), the feedback law (45), with the assumptions that (15), (17), (28) and (29) are satisfied. Next, consider the following adaptive laws

$$\begin{aligned}|B_s|K(k+1) = |B_s|K(k) + |B_s|Q_1|B_s|B_0^T PB_0\Phi(k)|D_s|^{-1}|D_s|w(k)F^T(x(k))Y \\ -2|B_s|Q_1|B_s|B_0^T PB_0K(k)F(x(k))F^T(x(k))Y,\end{aligned} \tag{49}$$

$$\begin{aligned}|B_s|\Phi(k+1)|D_s|^{-1} = |B_s|Q_2|B_s|B_0^T PD_0|D_s||D_s|^{-1}\Phi_w(k)|D_s|^{-1}Z|D_s|^{-1} \\ +|B_s|\Phi(k)|D_s|^{-1} - 2|B_s|Q_2|B_s|B_0^T PB_0\Phi(k)|D_s|^{-1}\Phi_w(k)|D_s|^{-1}Z|D_s|,\end{aligned} \tag{50}$$

Where

$$\Phi_w(k) = |D_s|w(k)w^T(k)|D_s|,$$

$$Q_1 = Q_2 = q^2|B_s|^{-1}|B_s|^{-1},$$

$$Z = |D_s||D_s|,$$

The resulting Lyapunov difference becomes

$$\Delta V(k) \le -\ell^T(x)\ell(x). \tag{51}$$

Then (49) and (50) reduce to (46) and (47), respectively. This complete the proof. Finally, since the adaptive gains we obtained are actually $|B_s|K(k)$ and $|B_s|\Phi(k)|D_s|^{-1}$, and the measured disturbance is $|D_s|w(k)$. The closed-loop system given by (44), (45), (46) and (47) can be rewritten as (48).

Lastly, we propose a robust adaptive solution to the linear uncertain systems given as following

$$x(k+1) = Ax(k) + Bu(x(k)) + Dw(k), \tag{52}$$

where B and D matrices satisfy the conditions given by (15) and (17), pair (A,B) is controllable, and there exists a gain matrix $K_g : R^{m\times n}$, such that $A_c = A + BK_g$ is exponentially stable. In addition, let $\Delta A = A_c - A$ is bounded, and the norm $|\Delta A|$ indicates the system dynamics A deviates from the stable solution A_c (Fu & Cheng, 2004).

Corollary 3.2
Consider the nonlinear discrete time system given by (52). Assume that B and D satisfy (15) and (17), respectively. Next, let $\Psi : R^{m\times d}$ and the matching condition $B\Psi = -D$ is satisfied. The feedback law given by

$$u(x(k)) = K(k)x(k) + \Phi(k)w(k), \tag{53}$$

where $K(k) = |B_s|K(k)$, $\Phi(k) = |B_s|\Phi(k)|D_s|^{-1}$, $w(k) = |D_s|w(k)$, and $q > 0$. Furthermore, the adaptive gain matrices

$$\begin{aligned} K(k+1) &= K(k) - q^2 B_0^T P[(B_0 K(k) + A_c)x(k) \\ &+ (B_0\Phi(k) + D_0)w(k)]x^T(k)Y, \end{aligned} \tag{54}$$

$$\begin{aligned} \Phi(k+1) &= \Phi(k) + q^2 B_0^T P[(B_0\Phi(k) \\ &- D_0)w(k) - A_c x(k)]w^T(k), \end{aligned} \tag{55}$$

guarantees that the closed-loop system given by (52), (53), (54), and (55) is Lyapunov stable, and equivalent to the following form

$$x(k+1) = (A + B_0 K(k))x(k) + (B_0\Phi(k) + D_0)w(k). \tag{56}$$

Proof
The proof is a direct extension of **Corollary 3.1**. First, we consider the Lyapunov function candidate

$$\begin{aligned} V(x(k),K(k),\Phi(k)) &= x^T(k)Px(k) + trQ_1(K(k)-K_g)Y^{-1}(K(k)-K_g)^T \\ &+ trQ_2(\Phi(k)-\Psi)Z^{-1}(\Phi(k)-\Psi)^T, \end{aligned} \tag{57}$$

Next, consider the Lyapunov difference (33), and assume that (15), (17), (28) and (29) are satisfied. Then the feedback control (53) with the adaptive laws given by

$$\begin{aligned}|B_s|K(k+1) &= |B_s|K(k) - |B_s|Q_1|B_s|B_0^T P[B_0|B_s|K(k) \\ &+ A_c]x(k)x^T(k)Y \quad - |B_s|Q_1|B_s|B_0^T P[B_0|B_s|\Phi(k)|D_s|^{-1} \\ &+ D_0]|D_s|w(k)x^T(k)Y,\end{aligned} \tag{58}$$

$$\begin{aligned}|B_s|\Phi(k+1)|D_s|^{-1} &= -|B_s|Q_2|B_s|B_0^T PA_c x(k)w^T(k)|D_s|D_s|^{-1}Z|D_s|^{-1} \\ &- |B_s|Q_2|B_s|B_0^T P[D_0 + B_0|B_s|\Phi(k)|D_s|^{-1}]|D_s|w(k)w^T(k)|D_s|D_s|^{-1}Z|D_s|^{-1} \\ &+ |B_s|\Phi(k)|D_s|^{-1},\end{aligned} \tag{59}$$

After some manipulations, the Lyapunov difference ΔV reduced to

$$\begin{aligned}\Delta V(k) &= x^T(k)[A_c^T PA_c - P + K_g^T B^T PBK_g \\ &+ R_w^T(x(k),w(k))Q_2 R_w(x(k),w(k))Z + \\ &\hat{F}^T(x(k),w(k))Q_1\hat{F}(x(k),w(k))]x(k) \\ &- x^T(k)K^T(k)B^T PBK(k)x(k) - |NB\Phi(k)w(k) + NDw(k)|_2^2,\end{aligned} \tag{60}$$

where

$$\begin{aligned}\hat{F}(x(k),w(k)) &= B^T P[(BK(k) + \\ &A_c)x(k) + (B\Phi(k) + D)w(k)],\end{aligned} \tag{61}$$

$$\begin{aligned}R_w(x(k),w(k)) &= B^T P[A_c x(k) \\ &+ (B\Phi(k) + D)w(k)],\end{aligned} \tag{62}$$

Since $x(k)$ be the solution of the closed-loop system, and the following conditions are satisfied

$$|\Delta A|^2 P \geq \Delta A^T P \Delta A = K_g^T B^T PBK_g \tag{63}$$

$$R \geq \hat{F}(x(0),\hat{w})Y + R_w^T(x(0),\hat{w})Q_2 R_w(x(0),\hat{w})Z + |\Delta A|^2 P \tag{64}$$

The resulting Lyapunov difference becomes

$$\Delta V(k) \le -\ell^T(x)\ell(x). \tag{65}$$

Next, let $Q_1 = Q_2 = q^2|B_s|^{-1}|B_s|^{-1}$ and $Z = |D_s\|D_s|$, then (58) and (59) reduce to (54) and (55), respectively. In addition, since a normalized adaptive gains $|B_s|K(k)$ and $|B_s|\Phi(k)|D_s|^{-1}$ are obtained through this design, with measured disturbance $w(k)$. The closed-loop system given by (52), (53), (54) and (55) can be rewritten as (48). This completes the proof.
Note that, the framework of of **Corollary 3.1** and **Corollary 3.2** do not require the knowledge of $|B_s|$ and $|D_s|$.

3.2 Discrete-Time Systems with ℓ_2 Disturbances

In this section we propose an adaptive feedback control solution for nonlinear uncertain discrete time MIMO systems with bounded ℓ_2 disturbances given by

$$\begin{aligned} x(k+1) = & f(x(k)) \\ & + G(x(k))u(x(k)) + J(x(k))w(k), \end{aligned} \tag{66}$$

where $w: R^d$, $k \ge 1$, is the unknown bounded energy ℓ_2 disturbance, $x(k) \in R^n$ is the state vector, $u(k): R^n \to R^m$ is the control vector, $f: R^n \to R^n$ characterize system dynamics with uncertain entries, and $f(0) = 0$. $G: R^n \to R^{n\times m}$ and $J: R^n \to R^{n\times d}$ are the input and disturbance weighting matrix functions, respectively. and the feedback law

$$u(x(k)) = \hat{G}(x(k))K(k)F(x(k)), \tag{67}$$

guarantees nonexpansivity condition given as **Theorem 4.1**.

Theorem 4.1
A nonlinear discrete-time system (66) is nonexpansive when $x(0) = x_0$, if the solution $x(k)$, $k \ge 0$, satisfies the following

$$\sum_{i=0}^{k} z^T(i)z(i) \le \hat{\gamma}^2 \sum_{i=0}^{k} w^T(i)w(i) + V(x(0), K(0)), \tag{68}$$

where $z(k)$ is output signal, and the Lyapunov candidate

$$\begin{aligned} &V(x(k), K(k)) = V_s(x(k)) + \\ &trQ_1(K(k) - K_g)Y^{-1}(K(k) - K_g)^T, \end{aligned} \tag{69}$$

for all $k : \mathrm{N}$, $w(\bullet) \in \ell_2$, $D : R^{n\times d}$, $\hat{\gamma}$ and γ be positive reals such that $\hat{\gamma}^2 I_d \ge \gamma^2 I_d + 2D^T PD$.
Next, we state and prove the discrete-time adaptive result for nonlinear system with bounded energy ℓ_2 disturbances.

Theorem 4.2
Consider the nonlinear discrete time system ***G*** given by (66), where the system dynamics f is uncertain. Next, We assume that there exists a gain matrix $K_g \in R^{m\times s}$, $\hat{G} : R^n \to R^{m\times m}$, and vector $F : R^n \to R^s$, such that

$$f_c(x(k)) = f(x(k)) + G(x(k))\hat{G}(x(k))K_g F(x(k)), \tag{70}$$

Furthermore, there exist $P_{1u} : R^n \to R^{1\times m}$, $P_{2u} : R^n \to R^{m\times m}$, $P_{uw} : R^n \to R^{m\times d}$, $P_{1w} : R^n \to R^{1\times d}$, and $P_{2w} : R^n \to R^{d\times d}$, the Lyapunov function V_s is defined as

$$\begin{aligned} &V_s(x(k+1)) \triangleq V_s(f(x(k))) + P_{1u}(x(k))u(x(k)) + u^T(x(k))P_{2u}(x(k))u(x(k)) \\ &+ u^T(x(k))P_{uw}(x(k))w(k) + P_{1w}(x(k))w(k) + w^T(k)P_{2w}(x(k))w(k), \end{aligned} \tag{71}$$

Let $\Gamma : R^n \to R$ be a positive scalar function and $\ell : R^n \to R^p$ is output vector, the following is assumed to be true

$$0 = V_s(f_c(x(k))) - V_s(x(k)) + \ell^T(x(k))\ell(x(k)) + \Gamma(x(k)), \tag{72}$$

Then the adaptive feedback control law

$$u(x(k)) = \hat{G}(x(k))K(k)F(x(k)), \tag{72}$$

with the update law

$$K(k+1) = K(k) - \frac{1}{2} Q \hat{G}(x(k)) P_{1u}(x(k)) F(x(k)) Y - Q \hat{G}(x(k)) P_{2u}(x(k)) \hat{G}(x(k)) K(k) F(x(k)) F^T(x(k)) Y, \quad (74)$$

where $Q > 0$ and $Y > 0$, guarantees that the closed-loop system, given by (66), (73), and (74), satisfies the nonexpansivity constraint given in **Theorem 4.1**.

Proof

The proof is a direct extension of **Theorem 2.1** and **Theorem 4.1**. We first consider the Lyapunov function candidate (69), such that $V(0, K_g) = 0$, and $V(x(k), K(k)) > 0$ for all $(x(k), K(k)) \neq (0, K_g)$, then $V(x(k), K(k))$ is radially unbounded. Furthermore, assume that $V(\bullet, K(k))$ and $K(k)$ are continuous in $x(k)$ for $k \geq 1$. The corresponding Lyapunov difference is given by

$$\Delta V(k) = V(x(k+1), K(k+1)) - V(x(k), K(k)). \quad (75)$$

Next, consider the update law

$$K(k+1) = K(k) - Q \hat{F}(x(k)) F^T(x(k)) Y, \quad (76)$$

$$\hat{F}(x(k)) = \frac{1}{2} \hat{G}^T(x(k)) P_{1u}(x(k)) + \hat{G}^T(x(k)) P_{2u}(x(k)) \hat{G}(x(k)) K(k) F(x(k)), \quad (77)$$

we then add and subtract $\gamma^2 w^T(k) w(k)$ to and from (75), and apply the fact $trxy^T = y^T x$, $\forall x, y \in R^n$, then (75) becomes

$$\begin{aligned} \Delta V(k) = {} & V_s(f_c(x(k))) - V_s(x(k)) - F^T(x) K_g^T \hat{G}^T(x) P_{2u}(x) \hat{G}(x) K_g F(x) \\ & - F^T(x) K^T(x) \hat{G}^T(x) P_{2u}(x) \hat{G}(x) K(x) F(x) + w^T(k) P_{2w}(x) \\ & + 2 F^T(x) K_g^T \hat{G}^T(x) P_{2u}(x) \hat{G}(x) K(x) F(x) + P_{1w}(x) w(k) \\ & + F^T(x) K^T(x) \hat{G}^T(x) P_{uw}(x) w(k). + [F^T(x) Y F(x)] \hat{F}^T(x) Q \hat{F}(x) \end{aligned} \quad (78)$$

Furthermore, let

$$\begin{aligned}
&\Gamma(x(k)) = [F^T(x)YF(x)]\hat{F}^T(x)Q\hat{F}(x) \\
&\qquad + F^T(x)K^T(k)\hat{G}^T(x)P_{2u}(x)\hat{G}(x)K(x)F(x), \\
&P_{2u}(x) = G^T(x)PG(x), \\
&P_{2w}(x) = J^T(x)PJ(x), \\
&P_{uw}(x) = 2G^T(x)PJ(x), \\
&P_{1w}(x) = \gamma^2 w^T(k), \quad \gamma^2 I_d - P_{2w}(x) \geq 0,
\end{aligned}$$

and $P = N^T N$. After some manipulations, the resulting Lyapunov difference becomes

$$\begin{aligned}
&\Delta V(k) \leq -\ell^T(x(k))\ell(x(k)) + \gamma^2 w^T(k)w(k) \\
&\rightarrow V(x(k),K(k)) - V(x(0),K(0)) \leq -\sum_{i=0}^{k}\ell^T(x(i))\ell(x(i)) + \gamma^2\sum_{i=0}^{k}w^T(i)w(i) \\
&\rightarrow -\sum_{i=0}^{k}\ell^T(x(i))\ell(x(i)) \leq \gamma^2\sum_{i=0}^{k}w^T(i)w(i) + V(x(0),K(0)) - V(x(k),K(k)) \\
&\qquad\qquad \leq \gamma^2\sum_{i=0}^{k}w^T(i)w(i) + V(x(0),K(0))
\end{aligned} \tag{79}$$

This proves that the closed-loop trajectory satisfies the nonexpansivity constraint given in Theorem 4.1. In addition, if $\ell(x(k)) \neq 0$, $k \geq 0$, then $x(k) \rightarrow 0$ as $k \rightarrow \infty$, $\forall x(0) \in R^n$. Finally, combining (78) and (76), (74) can therefor be obtained.

Next, let $G(x(k)) = B$ is sign definiteness matrix and satisfies (15). Specifically, the nonlinear system given by

$$x(k+1) = f(x(k)) + Bu(x(k)) + J(x(k))w(k). \tag{80}$$

We state without proof the following Corollary, since this is a direct extension of **Theorem 4.2**.

Corollary 4.1

Consider the nonlinear discrete time system given by (80). Assume that $F: R^n \rightarrow R^s$ and $\Gamma: R^n \times R^d \rightarrow R$, such that (72) is applied, and V_s is defined as (71). The feedback law

$$u(x(k)) = K(k)F(x(k)), \tag{81}$$

with the normalized adaptive gain matrices

$$K(k+1) = K(k) - 2q^2 B_0^T PB_0 K(k)F(x(k))F^T(x(k))Y, \tag{82}$$

where $K(k) = |B_s|K(k)$, and $q > 0$, guarantees that the closed-loop system given by (80), (81), and (82), equivalent to

$$x(k+1) = f(x(k)) + B_0 K(k)F(x(k)) + J(x(k))w(k), \tag{83}$$

satisfies the nonexpansivity constraint given in **Theorem 4.1**.
Note that the solution of adaptive gain matrix (82) is given by the selection of

$$P_{1u}(x(k)) = 2B^T PBK(k)F(x(k)). \tag{84}$$

Specifically, if $P_{1u}(x(k)) = 2B^T PBx(k)$, then the adaptive gain matrix can be given by

$$\begin{aligned} K(k+1) &= K(k) - q^2 B_0^T P(B_0 K(k) \\ &F(x(k)) + x(k))F^T(x(k))Y, \end{aligned} \tag{85}$$

Finally, we consider the linear discrete-time system G, where $J(x(k)) = D$ is a sign definiteness matrix and $f(x(k)) = Ax(k)$. Specifically, given by

$$x(k+1) = Ax(k) + Bu(x(k)) + Dw(k), \tag{86}$$

where $A \in R^{n\times n}$ is the time-invariant uncertain system matrix, $B \in R^{n\times m}$ is the input matrix, and $D \in R^{n\times d}$ is the disturbance weighting matrix. Let (A, B) be controllable pair, and B and D satisfy (15) and (17), respectively. We then state and prove the robust adaptive control design for linear uncertain systems as following.

Corollary 4.2
Consider the reachable linear discrete time system G given by (86). Assume there exists a gain matrix $K_g : R^{m\times n}$, such that $A_c = A + BK_g$ is exponentially stable, and let $\Delta A = A_c - A$ is bounded, and the norm $|\Delta A|$ indicates the system dynamics A deviates from the stable solution A_c. Next, let γ be a posi-

tive real, $L \in R^{n\times d}$, $W \in R^{d\times d}$, $R > 0$, $\hat{R} > 0$, $\hat{\Gamma} > 0$, and $P \in R^{n\times n}$ be the positive definite solution to the discrete-time Lyapunov equation given as

$$A_c^T P A_c - P = -\hat{\Gamma} - R\hat{R}, \tag{87}$$

$$A_c^T P D = LW, \tag{88}$$

$$\gamma^2 I_d - 2D^T P D = W^T W, \tag{89}$$

Then the adaptive feedback control as (81), with the update law

$$K(k+1) = K(k) - q^2 B_0^T P(B_0 K(k) + A_c) x(k) x^T(k) Y, \tag{90}$$

guarantees that the closed-loop system, given by (86), (81), and (90), satisfies the nonexpansivity constraint given in **Theorem 4.1.**

Proof

We first consider the Lyapunov function candidate given by

$$\begin{aligned} V(x(k), K(k)) &= x^T(k) P x(k) \\ &+ trQ^{-1}(K(k) - K_g) Y^{-1} (K(k) - K_g)^T, \end{aligned} \tag{91}$$

The corresponding Lyapunov difference is given by

$$\Delta V(k) = V(x(k+1), K(k+1)) - V(x(k), K(k)). \tag{92}$$

During the manipulations, we let

$$\begin{aligned} K(k+1) &= K(k) - Q\hat{F}(x(k)) x^T(k) Y, \\ \hat{F}(x(k)) &= B^T P(BK(k) + A_c) x(k), \end{aligned} \tag{93}$$

Next, add and subtract $\gamma^2 w^T(k) w(k)$ and $x^T(k) K_g^T B^T P B K_g x(k)$ to and from (92), apply the conditions (87) to (89), and the fact $trxy^T = y^T x$, $\forall x, y \in R^n$. In addition, assume that

$$\hat{R} \geq \hat{F}^T(x(k))Q\hat{F}(x(k))Y - 2\left|\Delta A\right|_2^2 P, \qquad (94)$$

where is a symmetric positive definite matrix. The resulting Lyapunov difference then becomes

$$\Delta V(k) \leq -x^T(k)Rx(k) + \gamma^2 w^T(k)w(k). \qquad (95)$$

Now, by summing (92) over $k \geq 0$ meets the nonexpansivity constraint given in Theorem 4.1. This completes the proof. Next, (93) could be rewritten as

$$K(k+1) = K(k) - QB^T P(BK(k) + A_c)x(k)x^T(k)Y, \qquad (96)$$

Furthermore, let $Q = q^2\left|B_s\right|^{-1}\left|B_s\right|^{-1}$, $K(k) = \left|B_s\right|K(k)$, and apply (15), (17). By similar procedure as in Corollary 3.2, (96) becomes (90). The closed-loop system, given by (86), (81), and (90), equivalent to

$$x(k+1) = (A + B_0K(k))x(k) + Dw(k), \qquad (97)$$

3.3 Adaptive Stabilization for Nonlinear Discrete-time Uncertain Systems

The Lyapunov direct method gives sufficient conditions for Lyapunov stability of discrete-time dynamical systems. In this section, we begin by characterizing the problem of adaptive feedback control laws for nonlinear uncertain discrete time MIMO systems given by (Fu & Cheng, 2004)

$$x(k+1) = f(x(k)) + G(x(k))u(x(k)), \qquad (98)$$

where $w \in R^d$, $k \geq 1$, is the unknown exogenous disturbance, $x(k) \in R^n$ is the state vector, $u(k): R^n \to R^m$ is the control vector, $f: R^n \to R^n$ characterize system dynamics with uncertain entries, and $f(0) = 0$. $G: R^n \to R^{n\times m}$ is the input weighting matrix function. We assume that there exists a gain matrix $K_g \in R^{m\times s}$, $\hat{G}: R^n \to R^{m\times m}$, and vector $F: R^n \to R^s$, such that

$$f_c(x(k)) = f(x(k)) + G(x(k))\hat{G}(x(k))K_g F(x(k)), \qquad (99)$$

is exponentially stable. We hereby state the main results of adaptive stabilization for nonlinear discrete-time uncertain systems.

Theorem 5.1
Consider the nonlinear discrete time system *G* given by (98), where the system dynamics f is uncertain, such that there exists a gain matrix K_g and (99) is applied. Next, let $P_{1u} : R^n \to R^{1\times m}$, $P_{2u} : R^n \to R^{m\times m}$, and the Lyapunov function V_s is given by

$$\begin{aligned} V_s(x(k+1)) &= V_s(f(x(k))) + P_{1u}(x(k))u(x(k)) \\ &+ u^T(x(k))P_{2u}(x(k))u(x(k)), \end{aligned} \tag{100}$$

In addition, let $\Gamma : R^n \to R$ is a positive scalar function, $\ell : R^n \to R^p$ is output vector, then

$$0 = V_s(f_c(x(k))) - V_s(x(k)) + \ell^T(x(k))\ell(x(k)) + \Gamma(x(k)), \tag{101}$$

The adaptive feedback control law

$$u(x(k)) = \hat{G}(x(k))K(k)F(x(k)), \tag{102}$$

with the update law

$$\begin{aligned} K(k+1) &= K(k) - \frac{1}{2}Q\hat{G}^T(x(k))P_{1u}(x(k))F^T(x(k))Y \\ &- Q\hat{G}(x(k))P_{2u}^T(x(k))\hat{G}(x(k))K(k)F(x(k))F^T(x(k))Y, \end{aligned} \tag{103}$$

where $Q > 0$ and $Y > 0$, guarantees that the closed-loop system, given by (98), (102), and (103), is globally asymptotically stable.

Proof
We first consider the Lyapunov function candidate

$$\begin{aligned} V(x(k), K(k)) &= V_s(x(k)) \\ &+ trQ^{-1}(K(k) - K_g)Y^{-1}(K(k) - K_g)^T, \end{aligned} \tag{104}$$

such that $V(0,K_g)=0$, and $V(x(k),K(k))>0$ for all $(x(k),K(k))\neq(0,K_g)$. In addition, $V(x(k),K(k))$ is radially unbounded. Furthermore, assume that $V(\bullet,K(k))$ and $K(k)$ are continuous in $x(k)$ for $k\geq 1$. The corresponding Lyapunov difference is given by

$$\Delta V(k)=V(x(k+1),K(k+1))-V(x(k),K(k)). \tag{105}$$

Next, consider the update law

$$\begin{aligned}
&K(k+1)=K(k)-Q\hat{F}(x(k))F^{T}(x(k))Y,\\
&\hat{F}(x(k))=\frac{1}{2}\hat{G}^{T}(x(k))P_{1u}^{T}(x(k))\\
&+\hat{G}^{T}(x(k))P_{2u}(x(k))\hat{G}(x(k))K(k)F(x(k)),
\end{aligned} \tag{106}$$

and apply the fact $trxy^{T}=y^{T}x,\ \forall x,y\in R^{n}$, then the Lyapunov difference be comes

$$\begin{aligned}
&\Delta V(k)=V_s(f(x))-V_s(x)-F^{T}(x)K_g^{T}\hat{G}(x)P_{2u}(x)\hat{G}(x)K_gF(x)\\
&-F^{T}(x)K^{T}(x)\hat{G}(x)P_{2u}(x)\hat{G}(x)K(x)F(x)\\
&+2F^{T}(x)K_g^{T}\hat{G}(x)P_{2u}(x)\hat{G}(x)K(x)F(x)\\
&+[F^{T}(x)YF(x)]\hat{F}(x)Q\hat{F}(x).
\end{aligned} \tag{107}$$

Furthermore, we select

$$\begin{aligned}
&\Gamma(x(k))=[F^{T}(x(k))YF(x(k))]\hat{F}^{T}(x(k))Q\hat{F}(x(k)),\\
&P_{2u}(x(k))=G^{T}(x(k))P_n^{T}P_nG(x(k)),\ \ P_n\in R^{n\times n}.
\end{aligned}$$

After some manipulations, the resulting Lyapunov difference becomes

$$\begin{aligned}
&\Delta V(k)=-\ell^{T}(x(k))\ell(x(k))-\\
&\left|P_nG(x(k))\hat{G}(x)(K(k)-K_g)F(x)\right|_2^2\\
&\qquad\leq-\ell^{T}(x(k))\ell(x(k)).
\end{aligned} \tag{108}$$

where $|\bullet|_2^2$ is Euclidean norm. This proves that the closed loop system is asymptotically stable, if $\ell(x(k)) \neq 0$, $k \geq 0$, then $x(k) \to 0$ as $k \to \infty$, $\forall x(0) \in R^n$. Finally, combining (106), (103) can therefore be obtained.
Specifically, if $P_{1u}(x(k)) = 2P_{2u}(x(k))\hat{G}(x(k))K(k)F(x(k))$ then (103) can be otained

$$K(k+1) = K(k) - 2Q\hat{G}^T(x(k))P_{2u}(x(k))\hat{G}(x(k))K(k)F(x(k))F^T(x(k))Y, \tag{109}$$

Note that the adaptive control law (103) or (109) do not require explicit knowledge of the matrix K_g and the system dynamics. Next, we extend the above result to the uncertain system given by

$$x(k+1) = f(x(k)) + Bu(x(k)), \tag{110}$$

where B satisfies (15) is the sign definite matrix with unknown entries. We state without proof the following results.

Corollary 5.1
Consider the nonlinear discrete-time uncertain system G given by (110). Assume that there exists a gain matrix K_g, such that $f_c(x(k)) = f(x(k)) + BK_gF(x(k))$ is exponentially stable. Next, $P_{1u}: R^n \to R^{1\times m}$ and $P_{2u}: R^n \to R^{m\times m}$, such that Lyapunov function V_s is given by

$$V_s(x(k+1)) = V_s(x(k)) + P_{1u}(x(k))u(x(k)) + u^T(x(k))P_{2u}(x(k))u(x(k)), \tag{111}$$

Furthermore, let $\Gamma: R^n \to R$ is a positive scalar function, $\ell: R^n \to R^p$ is output vector, and (101) is satisfied. The adaptive feedback control law

$$u(x(k)) = K(k)F(x(k)), \tag{112}$$

with the normalized update law

$$K(k+1) = K(k) - q^2 B_0^T P[x(k) + B_0 K(k)F(x(k))]F^T(x(k))Y, \tag{113}$$

where $K(k) = |B_s|K(k)$, $q > 0$ and $Y > 0$, guarantees that the closed-loop system, given by (110), (112), and (113), can be rewritten as

$$x(k+1) = f(x(k)) + B_0 K(k) x(k), \tag{114}$$

is Lyapunov stable.

Note that **Corollary 5.1** implies we may have different update law by different choice of P_{1u}. By the end of this section, we can further extend the results from above to linear uncertain systems given as following

$$x(k+1) = Ax(k) + Bu(x(k)), \tag{115}$$

where (A, B) be controllable pair. Next, assume there exists a gain matrix $K_g : R^{m \times n}$, such that $A_c = A + BK_g$ is exponentially stable, and let $\Delta A = A_c - A$ is bounded, and the norm $|\Delta A|$ indicates the system dynamics A deviates from the stable solution A_c.

Corollary 5.2

Consider the linear discrete-time uncertain system given by (115). Furthermore, let $R \in R^{n \times n}$ and $P \in R^{n \times n}$ are positive definite matrices, $\Gamma : R^n \to R$ is a positive scalar function, such that the Lyapunov function

$$P = A_c P A_c + R - \Gamma(x(k)), \tag{116}$$

with the assumption that $R \geq |\Delta A|^2 + \Gamma(x(k))$, where $x(k)$ is the solution. Then the adaptive feedback control law $u(x(k)) = K(k)x(k)$ with the normalized update law

$$K(k+1) = K(k) - q^2 B_0^T P[A_c + B_0 K(k)] x(k) x^T(k) Y, \tag{117}$$

where $K(k) = |B_s|K(k)$, $Q > 0$ and $Y > 0$ guarantees that the closed-loop system, given by (115), (117), can be rewritten as

$$x(k+1) = Ax(k) + B_0 K(k) x(k), \tag{118}$$

is Lyapunov stable.

Proof

The result is a direct extension of **Theorem 5.1** and **Cororllary 5.1**. Specifically, we consider the Lyapunov candidate

$$V(x(k), K(k)) = x^T(k)Px(k) + trQ^{-1}(K(k) - K_g)Y^{-1}(K(k) - K_g)^T, \quad (119)$$

Next, let $R \geq |\Delta A|^2 + \Gamma(x(k))$, normalized adaptive law $K(k) = |B_s|K(k)$, $Q = q^2|B_s|^{-1}|B_s|^{-1}$, $\Gamma(x(k)) = \hat{F}^T(x(k))Q\hat{F}(x(k))$, and

$$\hat{F}(x(k)) = B^T PA_c x(k) + B^T PBK(k)x(k). \quad (120)$$

Furthermore, we can substitute (117) into (115), the closed-loop form can be rewritten as (118).

4. Numerical Examples

In this section we illustrate the utility of the proposed direct adaptive control frameworks, both discrete-time and continuous-time, in the control problems of chaotic oscillator (Loria et al. 1998), one-link rigid robotic manipulator given by (Zhihong et al., 1998), and flexible joint robot manipulator (de Leòn-Morales et al., 2001), (Haddad & Hayakawa, 2002).

4.1 The van der Pol oscillator

The first example is a well known perturbed van der Pol equation used to model electrical circuit with triode valve (Loria et al. 1998), and given as following

$$\ddot{v} + \mu(1 - v^2)\dot{v} + v = u + q\cos(\omega t), \quad (121)$$

where the parameters specifically chosen as $\mu = 5$, $q = 5$, and $\omega = 2.463$, which exhibits chaotic behaviour, and u is control input. Next, let state space form with $x = [v, \dot{v}]^T = [x_1, x_2]^T$, (121) be rewritten as

$$f(x) = \begin{bmatrix} x_2 \\ -\mu(1-x_1^2)x_2 - x_1 \end{bmatrix},\ G(x) = \begin{bmatrix} 0 \\ 1 \end{bmatrix},\ J(x) = 1,\ \theta_1 = \begin{bmatrix} 0 \\ q \end{bmatrix},\ \overline{w}(x,t) = \cos(\omega t).$$

Next, let

$$F(x) = \begin{bmatrix} x_2 \\ x_1^2 x_2 \end{bmatrix},\ K_g = \begin{bmatrix} \mu - \beta \\ -\mu \end{bmatrix},\ A_c = \begin{bmatrix} 0 & 1 \\ -1 & -\beta \end{bmatrix}$$

Specifically, we chose

$$R = \begin{bmatrix} 0.005 \\ 12.5 \end{bmatrix},\ Z = 0.5,\ \beta = 0.8,\ Y = 1,\ P_1 = \begin{bmatrix} 100 & 0 \\ 0 & 25 \end{bmatrix},\ x(0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix},\ K(0) = \begin{bmatrix} 0 & 0 \end{bmatrix},$$

and P is the solution of Lyapunov equation

$$A_c^T P + PA_c + R = 0, \qquad (121)$$

By Corollary 2.1, the closed-loop system guarantees $x \to 0$ as $t \to \infty$, if $\overline{w}(x,t) = 0$. Figure 1 shows the phase portrait of the controlled system. The adaptive controller regulate the perturbed system to the origin under no knowledge of system dynamics, matrix K_g, and disturbance, while the disturbance exist. Figures 2 illustrates the time response of the feedback gain K and the control inputs.

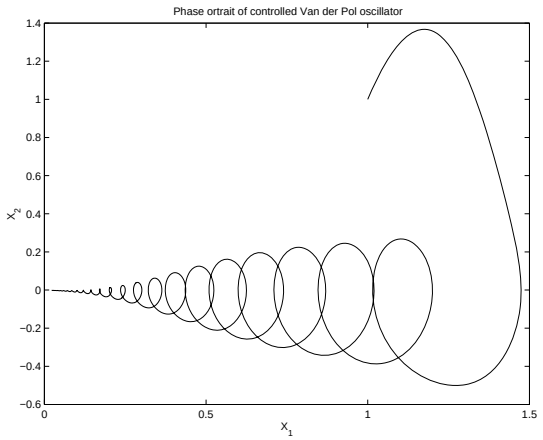


Figure 1 Phase Plot of perturbed van der pol equation

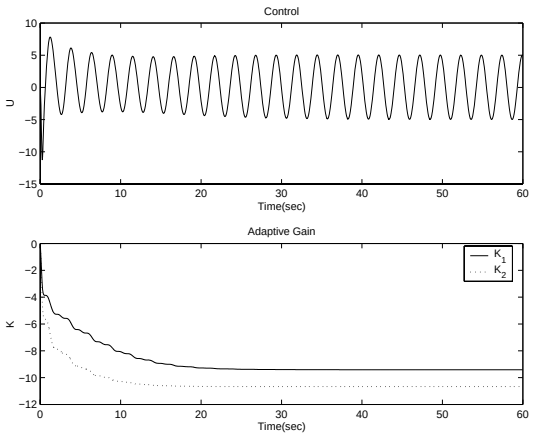


Figure 2 Control Signal and Adaptive gains

4.2 One-Link Rigid Robot under Gravitation Field

The dynamic equation of the one-link rigid robot placed on a tilted surface with an fixed angle θ is given by (Zhihong et al, 1998)

$$\begin{bmatrix} \dot{q} \\ \ddot{q} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & -\frac{d}{ml^2} \end{bmatrix} \begin{bmatrix} q \\ \dot{q} \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{ml^2} \end{bmatrix} u - \begin{bmatrix} 0 \\ \frac{g}{l} \end{bmatrix} \cos(q + \theta), \tag{122}$$

and the reference model is defined as

$$\begin{bmatrix} \dot{q}_r \\ \ddot{q}_r \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -16 & -8 \end{bmatrix} \begin{bmatrix} q_r \\ \dot{q}_r \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} r, \tag{123}$$

Next, since the tracking error is defined as

$$\begin{bmatrix} e \\ \dot{e} \end{bmatrix} = \begin{bmatrix} q \\ \dot{q} \end{bmatrix} - \begin{bmatrix} q_r \\ \dot{q}_r \end{bmatrix}. \tag{124}$$

the tracking model can be formulated as

$$\begin{aligned} \begin{bmatrix} \dot{e} \\ \ddot{e} \end{bmatrix} &= \begin{bmatrix} 0 & 1 \\ 0 & -\frac{d}{ml^2} \end{bmatrix} \begin{bmatrix} q \\ \dot{q} \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{ml^2} \end{bmatrix} u \\ &+ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 16 & 1 & 1 & 1 & 1 \end{bmatrix} \overline{w} \begin{bmatrix} 1 \\ 8 - \frac{d}{ml^2} \\ -\frac{g}{l}\cos(\phi) \\ \frac{g}{l}\sin(\phi) \\ 1 \end{bmatrix}, \end{aligned} \tag{125}$$

where

$$\overline{w} = \begin{bmatrix} q_r & 0 & 0 & 0 & 0 \\ 0 & \dot{q}_r & 0 & 0 & 0 \\ 0 & 0 & \cos(q) & 0 & 0 \\ 0 & 0 & 0 & \sin(q) & 0 \\ 0 & 0 & 0 & 0 & r \end{bmatrix}. \tag{126}$$

Specifically, we chose

$$R = \begin{bmatrix} 100 & 0 \\ 0 & 500 \end{bmatrix},\ Z = 8000\ ,\ Y = 6000\ ,\ P_1 = \begin{bmatrix} 10 & 0 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0 & 0 \\ 0 & 0 & 0.02 & 0 & 0 \\ 0 & 0 & 0 & 0.05 & 0 \\ 0 & 0 & 0 & 0 & 0.02 \end{bmatrix},$$

$$r = \sin(\ 20\,t\,)$$

and P is the solution of Lyapunov equation (121). In addition, let $m = l = d = 1$ and $g = 9.8$. Since (125) fits (13), and Corollary 2.1 can be directly applied. The initial conditions given $e(0) = \begin{bmatrix} 0 & 0 \end{bmatrix}^T$ and $K(0) = \begin{bmatrix} 0 & 0 \end{bmatrix}$. To demonstrate the robustness of the controller handle the uncertainty of the system dynamics, we introduce a changed to $m = 0.8$ at time $t = 0.5$ second. The simulation results, Figure 3 shows the states for each time step. The adaptive controller regulate the perturbed system to the origin under no knowledge of system dynamics, matrix K_g, and disturbance, while the disturbance exist. Figures 4 illustrates the time response of the control input, a constant force is applied to compensate the gravitation field. It shows that the controller can readapt the sudden change and stabilize the system.

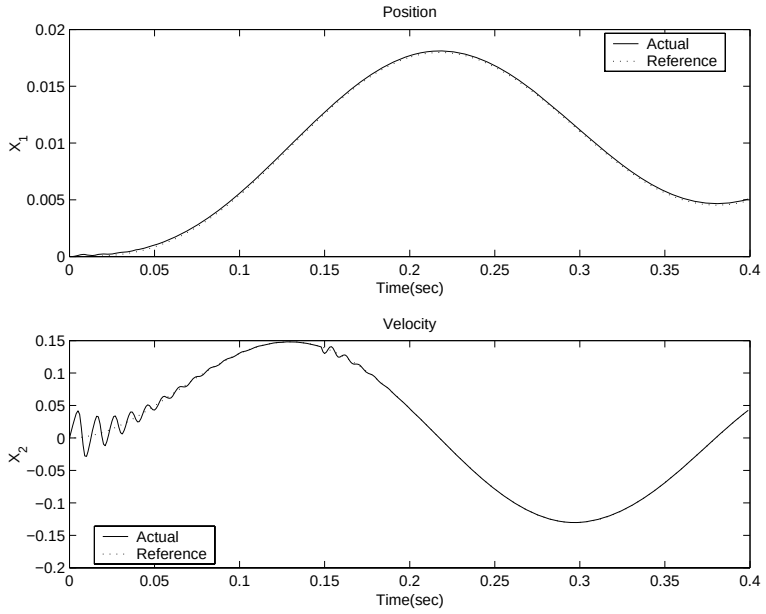
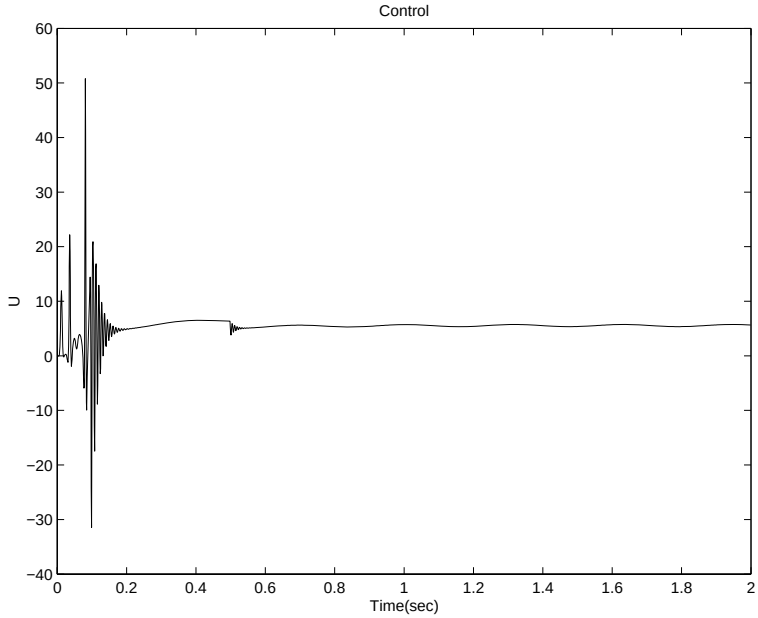


Figure 3. The states of one-link Rigid Robot



Figures 4. Control input

4.3 Continuous-time Active Suspension System

The dynamic equation for this quarter-car suspension is (Chantranuwathana & Peng, 1999)

$$\dot{x} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & -1 & 1 \\ 0 & \frac{K_s}{m_s} & -\frac{C_s}{m_s} & \frac{C_s}{m_s} \\ -\frac{K_{us}}{m_{us}} & -\frac{K_s}{m_{us}} & \frac{C_s}{m_{us}} & -\frac{C_{us}+C_s}{m_{us}} \end{bmatrix} x - \begin{bmatrix} 0 \\ 0 \\ \frac{1}{m_s} \\ -\frac{1}{m_s} \end{bmatrix} F_a + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 1 \end{bmatrix} \overline{w} \begin{bmatrix} -\frac{K_{us}}{m_{us}} \\ \frac{C_{us}}{m_{us}} \end{bmatrix}, \tag{127}$$

where

$$x = \begin{bmatrix} x_w \\ x_w - x_c \\ \dot{x}_c \\ x_w \end{bmatrix}, \quad \overline{w} = \begin{bmatrix} x_r & 0 \\ 0 & \dot{x}_r \end{bmatrix}, \quad x_r = \begin{cases} 0.01\sin(10t), & t \le 0.8 \\ 0.07\sin(5t), & 2 < t \le 2.2 \cdot \\ 0, & 0.8 < t \le 2. \end{cases}$$

x_w, x_c, and x_r are displacements of wheel, vehicle, and road, $x_w - x_c$ is hydraulic piston displacement, $m_s = 253kg$ is sprung mass, $m_u = 26kg$ is unsprung mass, $C_s = 348.5\frac{N}{m \cdot \sec}$ is suspension damping, $C_{us} = 10\frac{N}{m \cdot \sec}$ is tire damping, $K_s = 12000\frac{N}{m}$ is suspension stiffness, $K_{us} = 90000\frac{N}{m}$ is tire stiffness, and F_a is force of suspension actuator. Next, let A_c is asymptotically stable.

$$A_c = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & -1 & 1 \\ 0 & 0.05 & -1 & -0.05 \\ -0.1 & -5 & 1 & -5 \end{bmatrix}, \quad B_0 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \end{bmatrix},$$

First, we apply the framework of Corollary 2.1 and choosing the design matrices

$$Y = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.25 & 0 \\ 0 & 0 & 0 & 0.5 \end{bmatrix}, \quad R = 0.01 \cdot \begin{bmatrix} 1 \\ 5 \\ 7 \\ 5 \end{bmatrix}, \quad Z = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.001 \end{bmatrix},$$

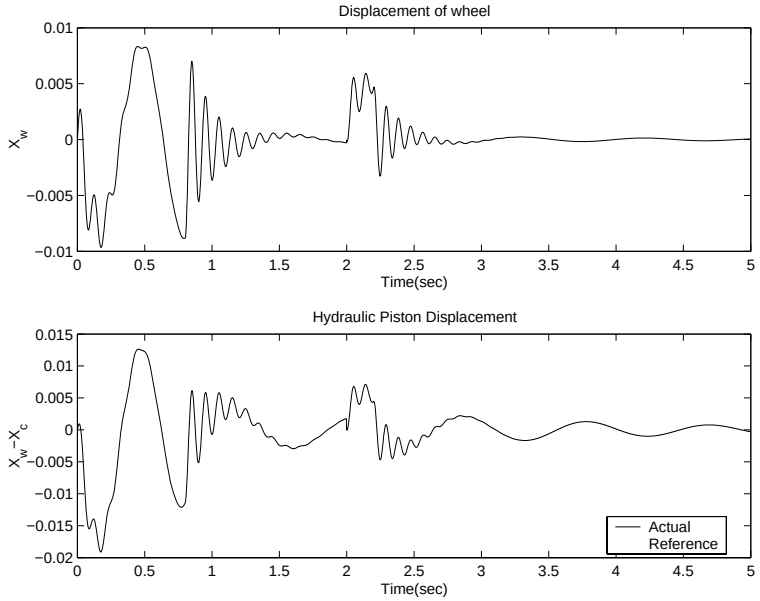


Figure 5 Displacement of wheel and Hydraulic pistion displacement

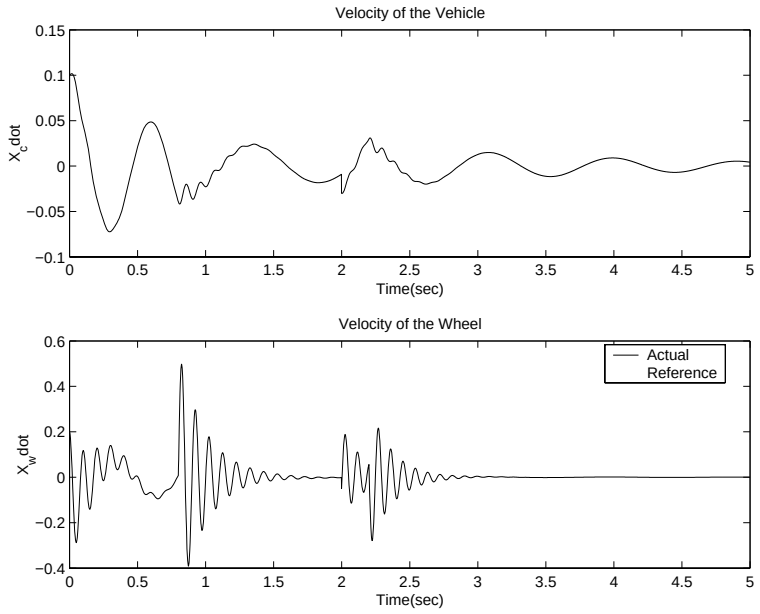


Figure 6 Velocity of vehicle and wheelhydraulic piston

where P satisfies the lyapunov condition (121). The simulation start with $x(0) = [0 \quad 0 \quad 0.1 \quad 0.2]^T$. At time $t = 2\sec$, the states are perturbed $x(2) = [0 \quad 0 \quad -0.03 \quad -0.05]^T$, and the system parameters are changed to $m_s = 213kg$, $m_u = 20kg$, $C_s = 320\frac{N}{m \cdot \sec}$, $C_{us} = 9\frac{N}{m \cdot \sec}$, $K_s = 11500\frac{N}{m}$, and

$K_{us} = 85000\frac{N}{m}$. The controller can re-adapt and stabilize the system in $5\sec$ under no information of the system parameters, either the perturbation of the states. Figure 5 depicts displacement of wheel and hydraulic piston displacement versus the time, Figure 6 shows the velocity of vehicle and wheel versus time, Figure 7 and Figure 8 illustrate the control inputs and adaptive gains at each time step.

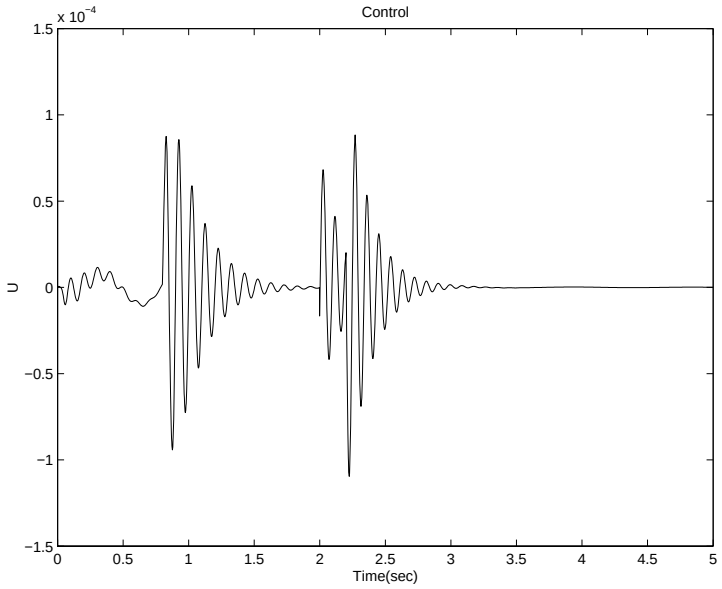


Figure 7 Control Input

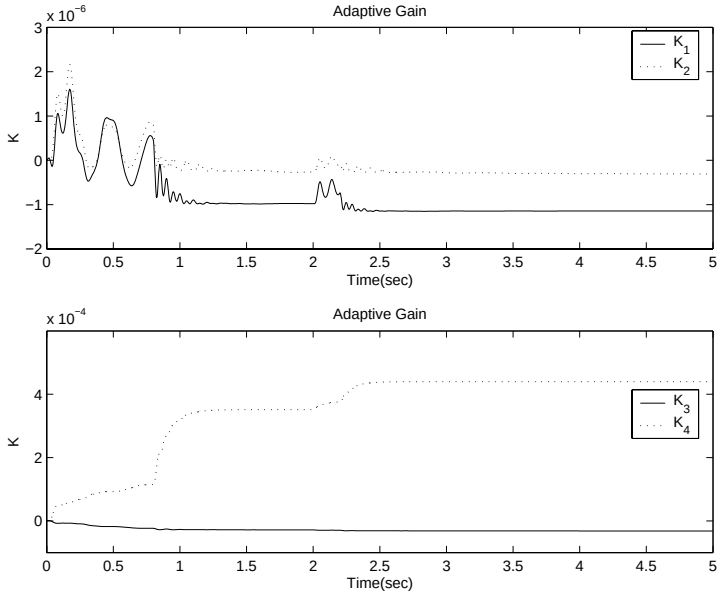


Figure 8 Adaptive Gains

4.4 Discrete-time Active Suspension System

We use the quarter car model as the mathematical description of the suspension system, given by (Laila, 2003)

$$x(k+1) = \left(\begin{bmatrix} 1 & T & 0 & 0 \\ -\dfrac{T\omega^2}{\rho+1} & 1 & \dfrac{T\rho\omega^2}{(\rho+1)^2} & 0 \\ 0 & 0 & 1 & T \\ T\omega^2 & 0 & -\dfrac{T\rho\omega^2}{\rho+1} & 1 \end{bmatrix} + \Delta(k) \right) x(k) - \begin{bmatrix} T \\ 0 \\ 0 \\ 0 \end{bmatrix} d(k) - \begin{bmatrix} 0 \\ 0 \\ 0 \\ T(1+\rho) \end{bmatrix} u(x(k)), \tag{128}$$

where

$$d(k) = \begin{cases} 0, & k \le 0 \\ 10\pi \sin(20\pi k), & 0 < k \le 100 \\ 0, & k > 100 \end{cases}, \quad \Delta(k) = \begin{cases} \begin{bmatrix} -10 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 5 & 1 & 1 & 1 \end{bmatrix} \Theta(k), & \mathrm{k} = 800 \\ 0, & \mathrm{k} \ne 800 \end{cases},$$

$$\Theta(k) = \left| 0.01 \sin(0.3k) \right|.$$

$x(k) = [x_1(k) \quad x_2(k) \quad x_3(k) \quad x_4(k)]^T$, and x_1 is tire defection, x_2 is unsprung mass velocity, x_3 is suspension deflection, x_4 is sprung mass velocity, $\omega = 20\pi \dfrac{rad}{\sec}$ and $\rho = 10$ are unknown parameters, $T = 0.001$ is sampling time, $d(k)$ is disturbance modeling the isolate bump with the bump height $A = 0.01m$, and $\Delta(k)$ is the perturbation on system dynamics. Next, let A_c is asymptotically stable

$$A_c = \begin{bmatrix} 1 & -1 & 0.75 & 1 \\ 0.75 & 0.3 & -1 & -0.1 \\ 0 & 0 & -0.1 & -0.5 \\ -0.1 & 0 & 0 & 0.1 \end{bmatrix}, \quad B_0 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix},$$

We apply the framework from Corollary 4.2 and choosing the design matrices

$$Y = 0.03 \cdot \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad R = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 9 & 0 \\ 0 & 0 & 0 & 0.82 \end{bmatrix}, \quad q = 0.05,$$

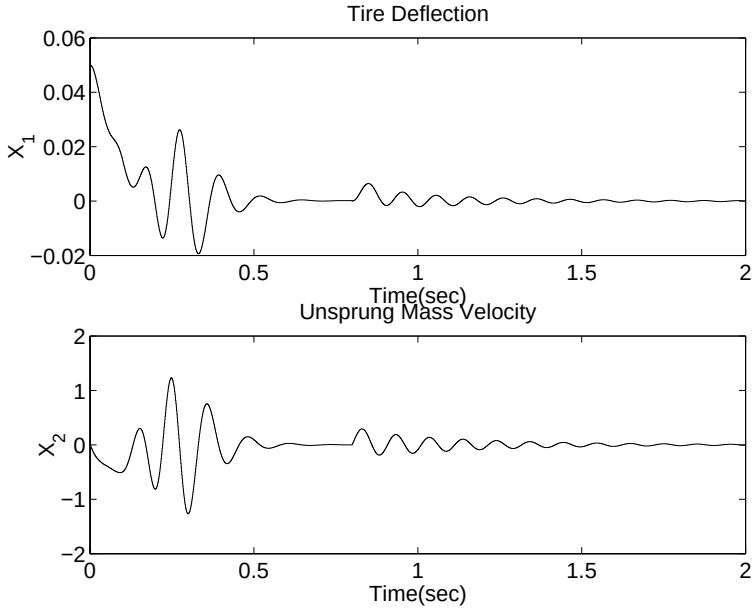


Figure 9 Tire defection and unsprung mass Velocity

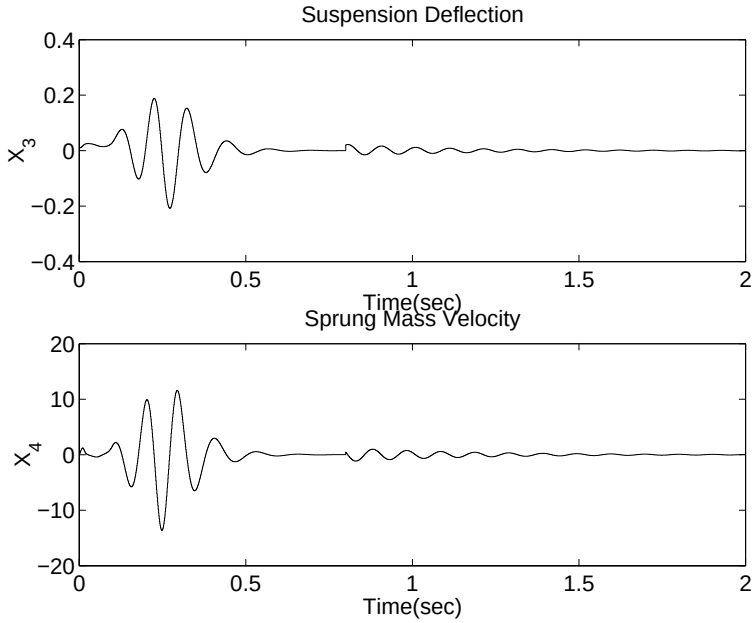


Figure 10 Suspension deflection and mass velocity

P satisfies the Lyapunov equation (121). The simulation start with $x(0) = [0.05 \quad 0 \quad 0.01 \quad 0]^T$. To demonstrate the efficacy of the controller, the states are perturbed to $x(800) = [0 \quad 0 \quad 0.02 \quad 0.5]^T$ at $k = 800$, and the system parameters are changed to $\rho = 4$. The controller stabilizes the system in 2 sec under no information of the system changes, either the perturbation of the states. Figure 9 depicts tire defection and unsprung mass velocity versus the time steps, Figure 10 shows the suspension deflection and sprung mass velocity versus the time step, Figure 11 and Figure 12 illustrate the control inputs and adaptive gains at each time step.

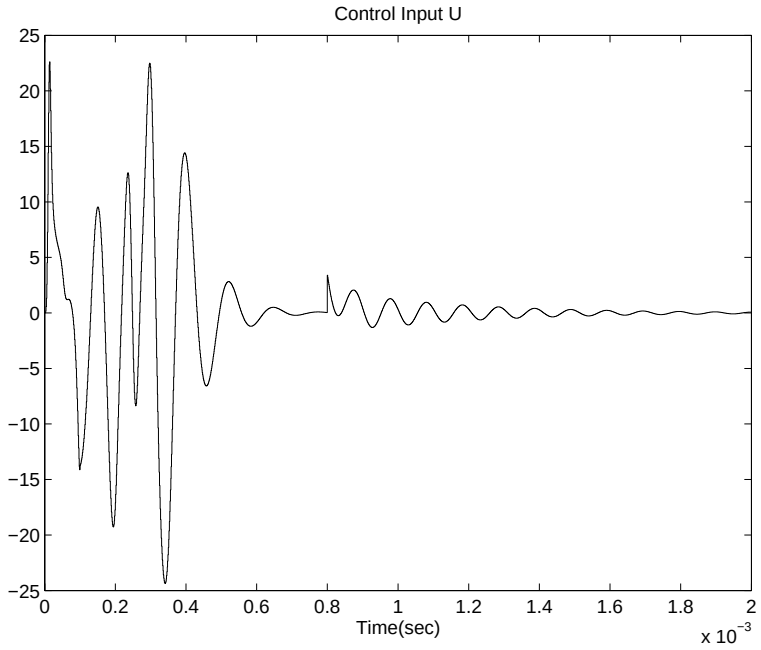


Figure 11 Control Input

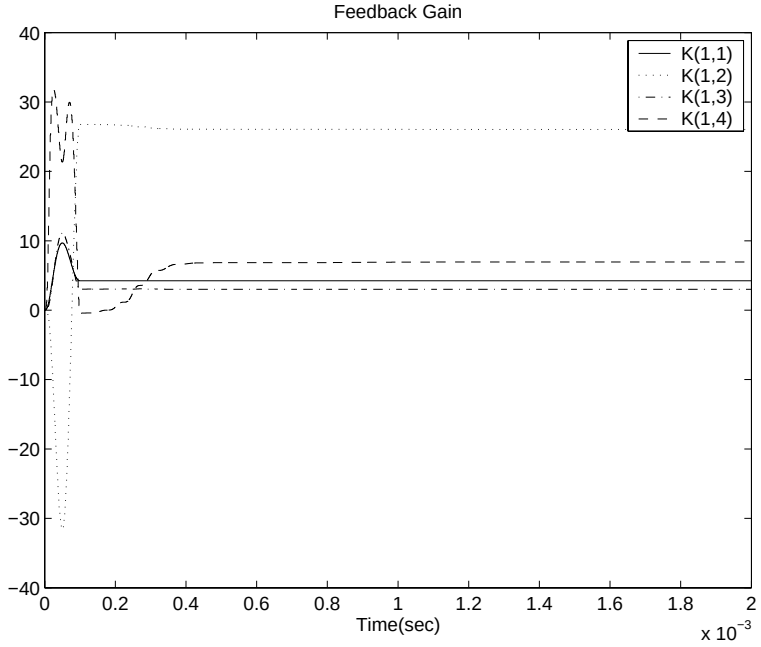


Figure 12 Adaptive Gains

4.4 Nonlinear Discrete-time Uncertain System

We consider the uncertain nonlinear discrete-time system in normal form given by (Fu & Cheng, 2004); (Fu & Cheng, 2005)

$$x(k+1) = \left(\begin{bmatrix} x_2(k) \\ ax_1^2(k) + bx_2(k)\cos(x_2(k)) \\ cx_3(k) + dx_1^3(k) \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} u(x(k)), \qquad (129)$$

where a, b, c, and d are unknown parameters. Next, let $f_c(x(k))$ to be

$$f_c(x(k)) = A_0 x(k) + \begin{bmatrix} 0 \\ ax_1^2(k) + bx_2(k)\cos(x_2(k)) \\ cx_3(k) + dx_1^3(k) \end{bmatrix} + \begin{bmatrix} 0 \\ B_s \end{bmatrix} \left| B_s \right|^{-1} (\Theta_n f_u(x(k)) - \Theta f_u(x(k)) + \Phi_n \hat{f}_u(x(k))), \qquad (130)$$

$$f_u(x(k)) = \begin{bmatrix} x_1^2 \\ x_2(k)\cos(x_2(k)) \\ x_3(k) \\ x_1^3(k) \end{bmatrix}, \quad \hat{f}_u(x(K)) = \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix},$$

$$F(x(k)) = \begin{bmatrix} x_1^2(k) \\ x_2(k)\cos(x_2(k)) \\ x_3(k) \\ x_1^3(k) \\ x_1(k) \\ x_2(k) \end{bmatrix}$$

and Θ_n and Φ_n are chosen such that

$$\Theta_n f_u(x(k)) + \Phi_n \hat{f}_u(x(k)) = \hat{A}x(k). \tag{131}$$

where $\hat{A} \in R^{2\times 3}$ is arbitrary, such that

$$f_c(x(k)) = \begin{bmatrix} \tilde{A}_0 \\ \hat{A} \end{bmatrix} x(k) = A_c x(k), \tag{132}$$

and A_c is asymptotically stable, specifically, chose

$$A_c = \begin{bmatrix} 0 & 1 & 0 \\ -0.5 & 0.4 & 0.1 \\ 0.3 & -0.5 & 0.9 \end{bmatrix}, \quad B_0 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix},$$

First, we apply the update law (113) and choosing the design matrices $Y = 0.1I_6$, $R = 0.2I_3$, and $q = 0.005$, where P satisfies the Lyapunov condition $P = A_c^T P A_c + R$. The simulation start with $x(0) = [1 \quad 0.5 \quad -1]^T$, and let $a = 0.5$, $b = 0.1$, $c = 0.3$, and $d = 0.5$. At time $k = 19$, the states are perturbed $x(19) = [1 \quad -0.5 \quad 0.5]^T$, and the system parameters are changed to $a = 0.65$, $b = 0.25$, $c = 0.45$, and $d = 0.55$. The controller does not have the information of the system parameters, either the perturbation of the states. Figure 13 – Figure 15 show the states versus the time step, Figures 16 shows the control inputs at each time step, and Figure 17 shows the update gains. The results indicate that the proposed controller can stabilize the system with uncertainty in

the system parameters and input matrix. In addition, re-adapt system while perturbation occurs. The only assumption required is sign definiteness of the input matrix and disturbance weighting matrix.

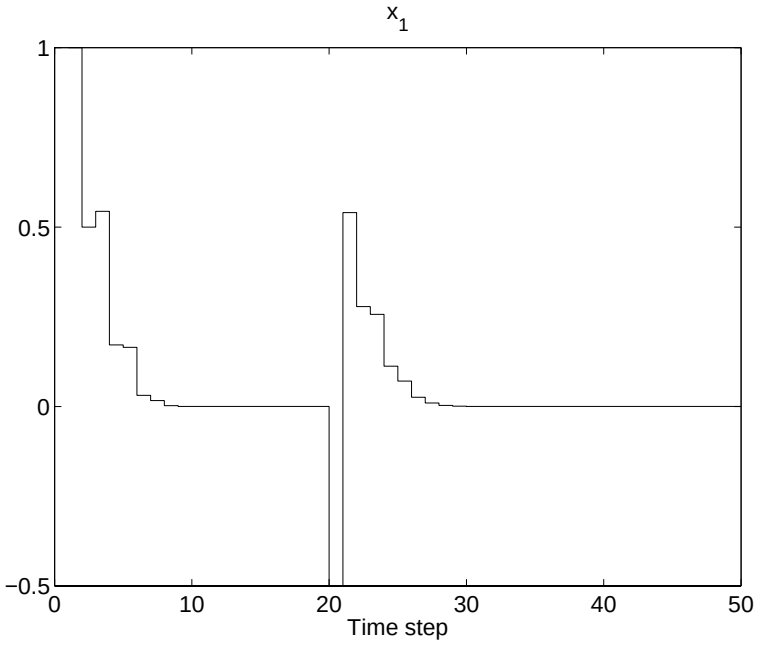


Figure 13 x_1

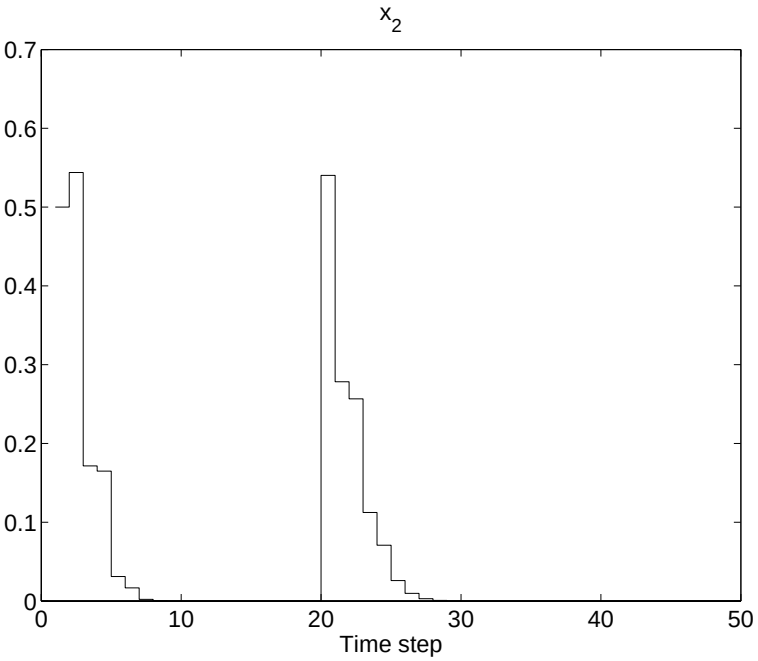


Figure 14. x_2

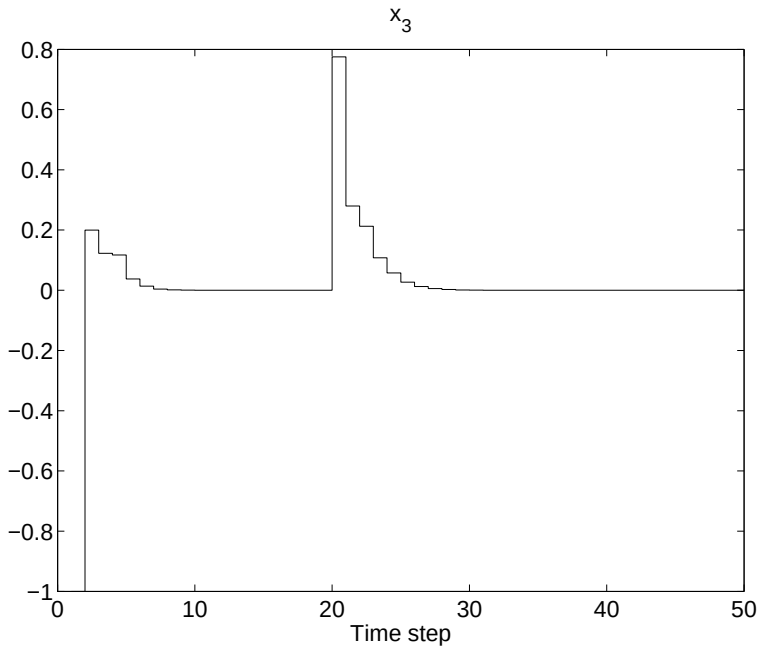


Figure 15. x_3

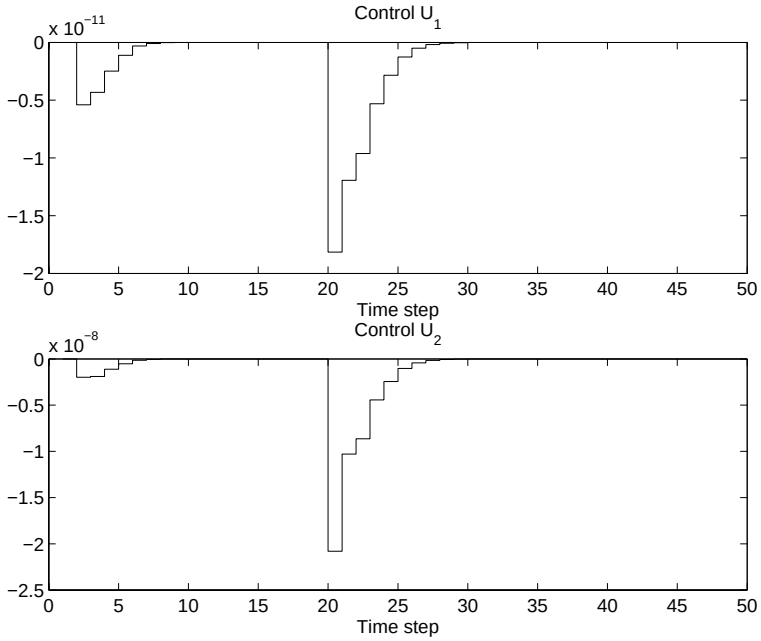


Figure 16. Control Signal

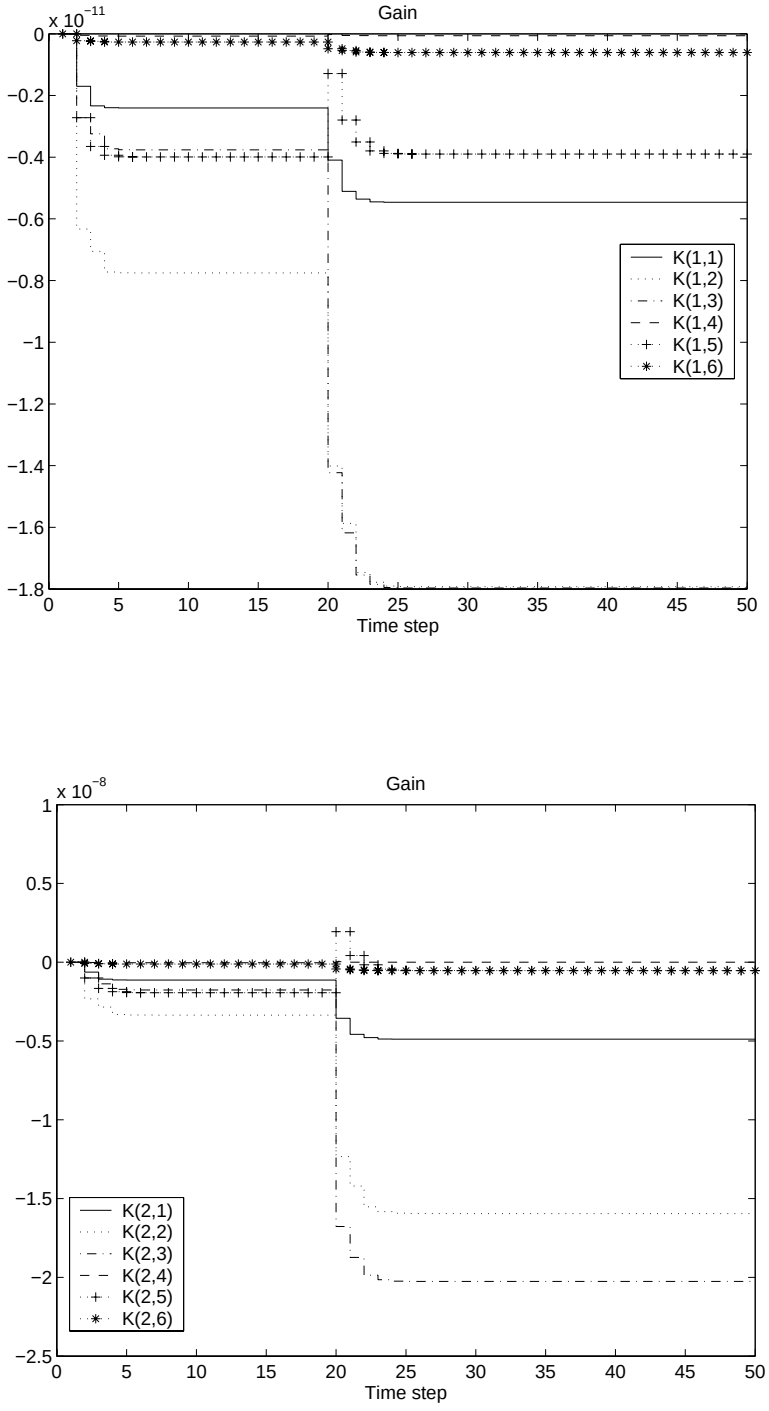


Figure 17. Update Gains

5. Conclusion

In this Chapter, both discrete-time and continuous-time uncertain systems are investigated for the problem of direct adaptive control. Noted that our work were all Lyapunov-based schemes, which not only on-line adaptive the feedback gains without the knowledge of system dynamics, but also achieve stability of the closed-loop systems. We found that these approaches have following advantages and contributions:

1. We have successfully introduced proper Lyapunov candidates for both discrete-time and continuous-time systems, and to prove the stability of the resulting adaptive controllers.
2. A series of simple direct adaptive controllers were introduced to handle uncertain systems, and readapt to achieve stable when system states and parameters were perturbed.
3. Based on our research, we claim that a discrete-time counterpart of continuous-time direct adaptive control is made possible.

However, there are draw backs and require further investigation:

1. The nonlinear system is confined to normal form, which restrict the results of the proposed frameworks.
2. The assumptions of (63), (64), and (72) still limit our results.

Our future research directions along this field are as following:

1. Further investigate the optimal control application, i.e. to seek the adaptive control input $u \in L_2$ or $u \in l_2$, minimize certain cost function $f(u)$, such that not only a constraint is satisfied, but also satisfies Lyapunov hypothesis.
2. Stochastic control application, which require observer design under the extension of direct adaptive scheme.
3. Investigate alternative Lyapunov candidates such that the assumptions of (63), (64), and (72) could be released.
4. Application to ship dynamic control problems.
5. Direct adaptive control for output feedback problems, such as

$$x(k+1) = f(x(k)) + G(x(k))u(x(k)) + J(x(k))w(k),$$
$$y(k) = H(x(k))x(k) + I(x(k))u(x(k)),$$
$$u(k) = K(k)y(k)$$

or

$$\dot{x} = f(x(t)) + G(x(t))u(x(t)) + J(x(t))w(t),$$
$$y = H(x(t))x(t) + I(x(t))u(x(t)),$$
$$u(t) = K(t)y(t)$$

6. References

Bar-Kana, I. (1989), Absolute Stability and Robust Discrete Adaptive Control of Multivariable Systems, *Control and Dynamic Systems*, pp. 157-183, Vol. 31.

Chantranuwathana, S. & Peng, H. (1999), Adaptive Robust Control for Active Suspension, *Proceedings of the American Control Conference*, pp. 1702-1706, San Diego, California, June, 1999.

de Leòn-Morales, J.; Alvarez-Leal, J. G.; Castro-Linares, R. & Alvarez-Gallego, J. A. (2001), Control of a Flexible Joint Robot Manipulator via a Nonlinear Control-Observer Scheme, *Int. J. Control*, vol. 74, pp. 290—302 .

Fu, S. & Cheng, C. (2003, a), Direct Adaptive Control Design for Reachable Linear Discrete-time Uncertain Systems, *in Proceedings of IEEE International Symposium on Computational Intelligence in Robotics and Automation*, pp. 1306-1310, Kobe, Japan, July, 2003.

Fu, S. & Cheng, C. (2003, b), Direct Adaptive Control Design for Linear Discrete-time Uncertain Systems with Exogenous Disturbances and ℓ_2 Disturbances, *in Proceedings of IEEE International Symposium on Computational Intelligence in Robotics and Automation*, pp. 1306-1310, Kobe, Japan, July, 2003.

Fu, S. & Cheng, C. (2004, a), Adaptive Stabilization for Normal Nonlinear Discrete-time Uncertain Systems, *in Proceedings of Fifth ASCC*, pp. 2042—2048, Melbourne, Australia, July, 2004.

Fu, S. & Cheng, C. (2004, b), Direct Adaptive Control for a Class of Linear Discrete-time Systems, *in Proceedings of Fifth ASCC*, pp. 172—176, Melbourne, Australia, July, 2004.

Fu, S. & Cheng, C. (2004, c), Direct Adaptive Feedback Design for Linear Discrete-time Uncertain Systems, *Asia Journal of Control*, Vol. 6, No. 3, pp. 421-427.

Fu, S. & Cheng, C. (2005, a), Robust Direct Adaptive Control of Nonlinear Uncertain Systems with Unknown Disturbances, *Proceedings of American Automatic Control Conference*, pp. 3731-3736, Portland, Oregon, June , 2005.

Fu, S. & Cheng, C. (2005, b), Direct Adaptive Control Designs for Nonlinear Discrete-time Systems with Matched Disturbances, Proceedings of IEEE International Conference on Mechatronics, pp. 881-886, Taipei, Taiwan, July 2005.

Fukaom, T.; Yamawaki, A. & Adachi, N. (1999), Nonlinear and H_∞ Control of Active Syspension Systems with Hydraulic Actuators, *Proceedings of the 38th IEEE CDC*, pp. 5125—5128, Phoenix, Arizona, December, 1999.

Guo, L. (1997), On Critical Stability of Discrete-time Adaptive Nonlinear Control, *IEEE Transactions on Automatic Control*, pp. 1488-1499, Vol. 42.

Haddad, W. & Hayakawa, T. (2002). Direct Adaptive Control for Nonlinear Uncertain Systems with Exogenous Disturbances, *Journal of Signal Processing and Adaptive Control*, Vol. 16, pp 151-172.

Haddad, W.; Hayakawa, T. & Leonessa, A. (2002), Direct Adaptive Control for Discrete-time Nonlinear Uncertain Dynamical Systems, *in Proceedings of American Control Conference*, pp. 1773-1778, Anchorage, May 2002.

Hitz, L. & Anderson, B. D. O. (1969), Discrete Positive-real Functions and Their Application to System Stability, *Proc. IEE*, pp. 153-155, Vol. 116.

Johansson, R. (1989), Global Lyapunov Stability and Exponential Convergence of Direct Adaptive Control, Int. J. Control, pp. 859-869, Vol. 50.

Laila, D. S. (2003), Integrated Design of Discrete-time Controller for an Active Suspension System, Proceedings of the 42th IEEE CDC, pp. 6406-6411, Maui, Hawaii, December 2003.

Levin, A. & Narendra, K. (1996), Control of Nonlinear Dynamical Systems using Neural Networks-Part II: Observability, Identification, and Control, IEEE Trans. Neural networks, Vol. 7, pp. 30-42.

Loria, A.; Panteley, E.; Nijmeijer, H. & Fossen, T. (1998), Robust Adaptive Control of Passive Systems with Unknown Disturbances, *IFAC NOLCOS*, pp. 866-872, Enschede, The Netherlands, 1998.

Mareels, I. & Polderman, J. (1996), *Adaptive Systems An Introduction*, Birkhauser.

Venugopal, R. & Bernstein, D. (1999), Adaptive Disturbance Rejection Using ARMARKOV System Representations, *Proc. of the 36th IEEE CDC*, pp. 1654-1658, San Diego, CA, December 1999.

Shibata, H.; Li, D.; Fujinaka, T. & Maruoka, G. (1996), Discrete-time Simplified Adaptive Control Algorithm and its Application to a Motor Control, *IEEE Industrial Electronics*, pp. 248-253, Vol. 1.

Zhao, J. & Kanellakopoulos I. (1997), Discrete-Time Adative Control of Output-Feedback Nonlinear Systems, *in IEEE Conference on Decision and Control*, pp. 4326-4331, San Diego, CA, December 1997.

Zhihong, M.; Wu, H. R. & Palaniswami, M. (1998), An Adaptive Tracking Controller Using Neural Network for a Class of Nonlinear Systems, *IEEE Transactions on Neural Network*, Vol. 9, pp 947-954.

Corresponding Author List

Kazem Abhary
School of Advanced Manufacturing and Mechanical Engineering
University of South Australia
Australia

Jose Barata
Universidade Nova de Lisboa – DEE
Portugal

Thierry Berger
LAMIH, University Valenciennes
France

Felix T. S. Chan
Department of Industrial and Manufacturing Systems Engineering
The University of Hong Kong
P.R. China

Che-Wei Chang
Graduate Institute of Business and Management
Yuan-Pei University of Science and Technology
Taiwan, ROC

Fan-Tien Cheng
Institute of Manufacturing Engineering, National Cheng Kung University
Taiwan, ROC

Cheng Siong Chin
Nanyang Technological University
Singapore

Jorge Corona-Castuera
CIATEQ A.C. Advanced Technology Centre, Queretaro
Mexico

Alexandre Dolgui
Division for Industrial Engineering and Computer Sciences
Ecole des Mines de Saint Etienne
France

Ming Dong
Shanghai JiaoTong University
P.R. China

Jerry Fuh Ying Hsi
Department of Mechanical Engineering, National University of Singapore
Singapore

Shih-Wen Hsiao
Department of Industrial Design
National Cheng Kung University
Taiwan, ROC

Pau-Lo Hsu
Information & Communications Research Labs, Industrial Technology Research Institute, Taiwan R.O.C.

Meifa Huang
Department of Electronic Machinery and Transportation Engineering, Guilin University of Electronic Technology
PR China

Che Ruhana Isa
Faculty of Business and Accountancy
University of Malaya
Malaysia

Tritos Laosirihongthong
Department of Industrial Engineering, Faculty of Engineering
Thammasat University-Rangsit Campus
Thailand

Ismael Lopez-Juarez
CIATEQ A.C. Advanced Technology Centre, Queretaro
Mexico

Carlos G. Mireles P.
MEI WCR,
Pasadena, USA

Jean Luc Marcelin
Joseph Fourier University Grenoble
France

Koshichiro Mitsukuni
Business Solution Systems division, Hitachi, Ltd
Japan

Tatsushi Nishi
Department of Systems Innovation
Graduate School of Engineering Science
Osaka University
Japan

Mario Pena-Cabrera
Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas
IIMAS-UNAM,
Mexico

Maki K. Rashid
Mechanical and Industrial Engineering, Sultan Qaboos University
Sultanate of Oman

Mehmet Savsar
Department of Industrial & Management Systems Engineering
College of Engineering & Petroleum
Kuwait University

Cem Sinanoglu
Erciyes University Engineering Faculty
Department of Mechanical Engineering
Turkey

Bill Tseng
Department of Mechanical and Industrial Engineering
The University of Texas at El Paso
USA

Joze Tavcar
Faculty of Mechanical Engineering
University of Ljubljana
Slovenia

A.M.M. Sharif Ullah
Department of Mechanical Engineering
College of Engineering
United Arab Emirates University

Yong Yin
Department of Public Policy & Social Studies
Yamagata University
Japan

Chao Zhang
Department of Mechanical & Materials Engineering
University of Western Ontario
Canada