# Hierarchical production management : the flow-control layer

Camille Libosvar

N° d'Ordre :

Metz BU
S/MZ

# THESE



présentée à

**LA FACULTE DES SCIENCES**
**DE L'UNIVERSITE DE METZ**

pour obtenir le grade de

**DOCTEUR DE L'UNIVERSITE DE METZ**

par

**Camille LIBOSVAR**

Sujet de la thèse :

# HIERARCHICAL PRODUCTION MANAGEMENT : THE FLOW-CONTROL LAYER

Soutenue le 15 Avril 1988, devant la commission d'examen composée de

| | | |
|---|---|---|
| **Mme** | **F. CHATELIN** | **Rapporteur** |
| **MM.** | **S.B. GERSHWIN** | |
| | **J.B. LASSERRE** | |
| | **B. MUTEL** | **Rapporteur** |
| **Melle** | **M.C. PORTMANN** | |
| **M.** | **J.M. PROTH** | |

## ABSTRACT

Production Management is concerned with a class of decisions to be made in a manufacturing firm in order to gear it towards its objective. Since this decision making problem is very large, it must be approached hierarchically. Hierarchical production management systems are characterized by several decision levels operating in a coordinated fashion. Designing such systems means defining the models to be used at each level (entities, objective, horizon), and a coordination procedure. The models studied in this work are devised for the higher levels of a hierarchy; the production system is represented as a network of subsystems with limited capacity and the objective sought is to minimize the flow time of product families. It is proved that under certain assumptions concerning the inventory holding costs, a very simple algorithm exists to solve this deterministic optimization problem. It is then shown that it is possible to relax this assumption by using dynamic programming but the amount of computations required increases dramatically.



key words: hierarchical control, flow control, finite capacity, inventory, deterministic optimizing.



## RESUME

La gestion de production s'intéresse à une classe de décisions à prendre dans une entreprise de production de façon à lui faire atteindre son objectif. Comme le problème à résoudre est très vaste, il faut l'aborder au moyen d'une approche hiérarchisée. Les sytèmes de gestion hiérarchiques se caractérisent par plusieurs niveaux de décision coordonnés. Concevoir de tels systèmes suppose de définir les modèles à utiliser à chaque niveau (entités, objectif, horizon), et une procédure de coordination. Les modèles étudiés dans ce mémoire sont destinés au niveau haut d'un système hiérarchique; l'outil de production est représenté comme un réseau de sous-systèmes à capacités finies et l'objectif à atteindre est la production à flux tendus de familles de produits. On démontre que pour certaines structures de coûts de stockage, il existe un algorithme très simple pour résoudre ce problème d'optimisation déterministe. On montre également qu'il est possible de relaxer cette contrainte et d'utiliser la programmation dynamique, mais le volume de calcul requis s'en trouve considérablement augmenté.



mots-clés: gestion hiérarchisée, controle de flux, capacité finie, stocks, optimisation déterministe.

Cette thèse est le résultat de la recherche menée par l'auteur pendant son séjour à l'INRIA Lorraine, puis au Centre de Recherche du groupe Péchiney à Voreppe et enfin au Massachusetts Institute of Technology. Elle a été entièrement financée par la Société Aluminium Péchiney, et rédigée durant l'année 87 au Laboratory for Information and Decision Systems du MIT.

*Dans les différentes équipes où j'ai eu la chance de travailler, j'ai rencontré de nombreuses personnes qui, d'une façon ou d'une autre, ont contribué à faire de mon séjour une expérience à la fois agréable et enrichissante. Je leur en suis extrêmement reconnaissant. Je tiens à remercier tout particulièrement:*

*M. GAUDILLERE, Directeur des Flux à Aluminium Péchiney, qui a fourni le thème et le financement de cette étude,*

*Philippe VARIN, Directeur d'Aluval Péchiney, pour m'y avoir accueilli et avoir remis mon travail en perspective quand j'en avais besoin,*

*Stanley GERSHWIN, Senior Research Scientist au MIT, pour m'avoir accueilli dans son équipe et pour tout ce que j'ai appris à son contact,*

*et enfin, surtout, je remercie M. PROTH, directeur de l'INRIA Lorraine, mon patron 'et néanmoins ami' (sic), pour sa confiance et son constant soutien, et pour m'avoir enseigné le métier de chercheur.*

# TABLE OF CONTENTS

# INTRODUCTION

Production Management is one of the fields in which there is a significant gap between theory and practice. This gap has even seemed to widen as researchers strived to make their models more 'credible' -at the expense of an increasing complexity-, whereas industry favored solutions based on simple concepts, sometimes more philosophical than concrete.

In fact, what seems to be the objective of production management, i.e. to improve the 'performance' -in the broadest sense of the term- of a production system, can be achieved through a variety of actions:

- → improving the product design, the production process or the tooling performance;
- → improving the work organization or the workforce's attitude;
- → improving the information system;
- → improving the planning of activities related to production.

These are, at least, the 'levers' that a manufacturing engineer is typically aware of, when he thinks of ways to 'do his job'.

Production management is mostly concerned with the last type of actions and, in particular, with the development of models to that effect; modelling has proved to be difficult in two respects:

On the one hand, the on-going controversy about the adequateness of the optimizing models developed to date suggests that designing good models for production planning or scheduling is not an easy task. The interactions between different ways to improve the performance of a production system, as well as the existence of many criteria to evaluate these actions partially account for this difficulty.

On the other hand, most of the models designed to date (and in particular those based on combinatorial optimization) require a very large, usually excessive, amount of computations.

The gap between theory and practice in production management can thus be partially accounted for by the fact that it would seem pointless to struggle to achieve a given gain in performance by a better production planning if a better design or a better work organization yielded the same gain at a lower cost. This explains the widespread use of 'manual' scheduling aimed at feasability, as opposed to computer-based optimizing scheduling.

Another factor that accounts for this situation is the lack of data concerning the events that take place in a manufacturing system: a substantial gain is still to be realized in many companies through the implementation of an information system and also, which is more difficult, through the definition of a set of data analysis procedures. In fact, it is essential to have both a reliable information concerning the state of the production system at any point in time, and a historical information from which different aspects of its efficiency can be assessed.

Also, the existence of interactions between the different possible 'levers' available for performance improvement has motivated the search for a certain synergy and for consistency in the decision process (see PARNABY [PN] and [IE]). Hence the success of such 'global' concepts as the Just In Time / Zero Default implemented in the Japanese industry (see SHOENBERGER [SO]), as well as the growing interest in Computer Integrated Manufacturing: all the decisions related to manufacturing made consistently through the use of a computer…

Unfortunately, JIT is more a philosophy than a technique, more the statement of a goal than a method to achieve it. Therefore, although the previous improvements of management systems claimed to be implementations of it exemplify some methods to achieve JIT, it is often necessary to find new methods for each new application.

Similarly, CIM is mostly a research objective; in fact, if none of the decision making problems related to manufacturing (i.e. design, process planning or scheduling) has yet been solved on a computer individually (at least in a satisficing manner and for a wide variety of cases), solving all of them jointly is not currently possible, except maybe for very particular manufacturing systems.

At the same time, the competition in many industrial fields has increased -due especially to the internationalization of the markets- to the point where it is not possible any more, for a firm, to rely exclusively on one type of action (e.g. improving the process) to keep its competitive edge:

Taking all possible actions has become a must, and there is a strong demand, in these industries, for methods to improve production planning, especially when this function was not considered essential in the past, and the only scheduling tools currently used are Gantt charts... The questions are then: *which models will satisfy this demand?*, and its counterpart: *why should such models be found now if they were not in the past?*

The answer to this latter question is that the field has matured and the tools to use are better than they were: on the one hand, the recent developments in computer technology and information processing allow to consider solving problems of increasing complexity in a 'reasonable' amount of time; on the other hand, more and more emphasis has been placed, in recent research, on the practicality of solutions, and good suboptimal solutions have been sought where the optimization methods were not applicable.

As for the initial question, the main thrust of the teams with which the author has collaborated is that hierarchical models for production management could satisfy the needs of industry.

Such models are characterized by a coordinated resolution of several subproblems and they require the specification of:

→ a method to decompose the original decision problem,

→ formulations for each of the subproblems identified, and algorithms to solve them consistently.

A new approach to hierarchical production management is currently investigated by PROTH and co-workers [HE], [MP]; the decomposition method is a product- and machine aggregation algorithm devised to generate a flowshop-like aggregate model of the system, that is, a model in which product families flow between subsystems, so that no two flows are in opposite directions. Coordination is achieved through the top-down transmission of constraints, the decisions made at the upper levels defining constraints for lower levels decisions.

The object of this thesis is to study the type of aggregate models resulting from the previous decomposition method, which are called flow control models. The organization of the work presented is described hereunder.

In Chapter 1, several arguments concurring to justify the use of a hierarchical approach for management are presented, and the contribution of this thesis is outlined in the perspective of the framework proposed by PROTH and co-workers. Chapter 2 is a survey of both the research conducted in Control Theory on hierarchical systems and of the different hierarchical models studied in production management. The flow control model that is the object this study is introduced in Chapter 3, and the assumptions imbedded in its formulation are justified; analytical results are then derived in Chapter 4, and two applications of these results are presented in Chapter 5.

# chapter 1:

# PRODUCTION MANAGEMENT IN PERSPECTIVE,

# OR WHY IT IS HIERARCHICAL.

## I OBJECTIVE STATEMENT

In our economic system, surviving, for a firm, is synonymous with making a profit. -Although, for a short period of time, or just when starting its activity, a company is in fact allowed to lose a limited amount of money-. Furthermore, the higher the profit, the better the competitive position of the firm, as long as present profits are not realized at the expense of future ones. Therefore, the objective of a manufacturing firm -stated in mathematical terms-, is to maximize the expected value of its discounted profit over a certain period of time called strategic horizon:

- expected value, because profit usually depends on some factors such as, for example, the evolution of the market, which are never known with certainty when decisions are to be made,

- discounted, because, depending on the type of firm (and in particular on what could be called its response time), the relative importance of present and future results will be different,

- horizon, because the performance can be dangerously lowered by the absence of anticipation, but also because, beyond a certain point, the uncertainty of the forecasts is such that taking them into account no longer improves the performance of the firm.

In order to achieve this objective, the various actors of the firm are constantly taking actions. The question is then: how should they decide what action to take next, given that there is an objective that the firm as a whole must meet? or, in other words, how to design a efficient decision-making organization, efficient meaning enabling the firm to achieve its objective. (Some work has been devoted to numerical methods to assess the ability of a system to meet its goals: see for example WASHINGTON and LEVIS [WL] for the decision of whether or not to implement an F.M.S.).

## II DECISION-MAKING ORGANIZATION

There is usually a large number of decisions to be made concurrently at each point in time to run a firm. Also, if the strategic horizon is to be long enough to allow for reaction to the evolution of the environment, then it is generally of several orders of magnitude longer than the duration of elementary actions. This means that the problem of finding the optimal plan of action in its most detailed form and updating it each time the realizations differ from the plan cannot be dealt with by currently available computers. Moreover, even if there were computers to solve problems of this size, using them to that effect would result in a waste, because by the time actions would actually be taken, they would have been re-planned thousands of times.

The approach generally adopted in human organizations consists, on the one hand, of lumping elementary actions (or 'steps') into more global ones for the purpose of decision-making and, on the other hand, of identifying classes of weakly coupled actions that can be decided independently. For example, building a new plant can be considered as a single action, and there will be only one decision concerning it as such, although its implementation clearly involves a wealth of more detailed steps. On the other hand, many decisions made by the Sales department, for example, are independent, to a large extent, from the decisions made in Design. The processes resulting in these simplifications are referred to as *aggregation* and *specialization*.

## III AGGREGATION

The aggregation of actions is performed along two dimensions: time and scope, which means that an aggregate decision can be a sequence of steps, or a group of simultaneous steps, or any combination. There is thus a complete spectrum of actions with different scopes and/or different durations.

Following from ANTHONY's classification, this spectrum is commonly broken down into three classes: strategic, tactical, and operational actions. Launching a new product, stopping a line for preventative maintenance or casting an ingot are actions that belong respectively to each of these classes.

These examples illustrate how the specification of an action is all the more abstract as the action is 'aggregate'. This fact has two implications: on the one hand, decisions concerning aggregate actions will materialize only if they actually trigger a set of elementary actions, and, on the other hand, the exact set and sequence of these elementary actions is not known a priori, which means that there is a need for more decisions after an aggregate action is decided.

Also, an action actually materializes only at the last level of decision-making, when all its physical contents has been carried out. Therefore, an aggregate action takes on different 'meanings', depending the level of abstraction at which it is considered: for example, for a shop-floor manager, stopping a line for preventative maintenance does not mean the action of pushing a red button, but rather: scheduling this loss of production, making sure that the maintenance personnel will be there, that the line personnel will be kept busy, etc.

## IV DECISION PROCESS

This example casts a new light on the role of a decision-maker (DM). As stated previously, it is to 'choose among a set of actions or sequences of actions the one that will best realize an objective, and then implement it'; this statement is now made more specific.

First, a DM is assigned an objective and commensurate resources to achieve it. It must then be possible to assess whether or not the objective has been achieved, and if it has not, by how much it was missed.

Secondly, a DM must have some knowledge about the actions that he can take to achieve his objective, more precisely, he must have a model that will allow him to predict the impact of different actions on his objective. This model must in particular specify the horizon of the problem, that is, the period for which the impact of the actions on the objective must be considered.

Thirdly, a DM needs some feedback concerning both the resources that he has been assigned, and the actions taken by all other DMs. This information is essential to the DM because the set of actions he can possibly take and their impact on the objective both depend on it, as will become clear when the previous example of the preventative maintenance is considered.

What this example illustrates is that implementing an aggregate action generally means, for the DM:

→ identifying those elementary actions that he will perform himself (e.g. call the subcontractor in charge of the maintenance),

→ defining more concrete objectives or subgoals to pass to other decision makers for them to implement the corresponding actions and alloting them some resources to that effect; (usually, the more 'aggregate' the action, the more important this component of it),

→ coordinating the action of these other DMs, that is, constraining their choices of actions in order that each of them is in a position to achieve his objective, even if it depends on the other DMs' actions.

Also, if the action being implemented is part of more aggregate plan of action, then the DM must generate feedback concerning the realization of his objective, and often evaluate -based on his model- his resource needs before he decides on his action. At that point, the model he has will allow him to do so. These interactions are illustrated in figure 1.

Note that the model that a D.M. has of the impact of his decisions is very dependent on the type of actions that he decides. In particular, the more aggregate and abstract the actions, the more aggregate and abstract the model also.

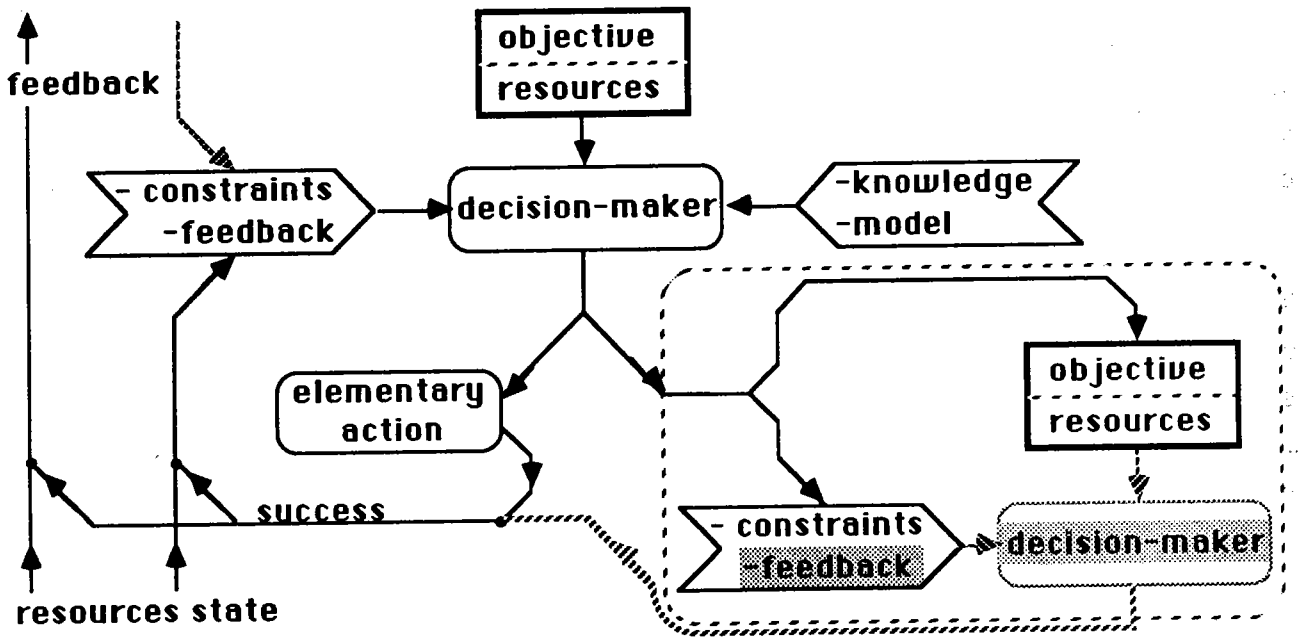



fig. 1

## V SPECIALIZATION

The other process used to reduce the size of the decision problem, *specialization*, is based on the assumption that the set of actions to take in a firm can be partitioned so that actions belonging to two different classes can be decided relatively independently.

This partitioning can be based on different factors as the type of product, the type of process, the location of the facilities, the market covered, etc. For example, a large corporation has usually one division per product line: jet engines, power generators, plastics… but it could also have a division 'heavy industry', a division 'high tech.', and a division 'consumer goods'.

The decision domain, that is, the type of knowledge required to make the decisions, is also used as a partitioning criterion. Roughly speaking, four types of knowledge and information can be identified in a firm: financial, technical, managerial and market-related (knowledge and information differ only in their life time).

→ The financial information can be found summarized in any annual financial report. It is mainly used in strategic decisions because directly related to the global objective. The corresponding knowledge is for example the knowledge of the stock market.

→ The market information comprizes qualitative and quantitative answers to such questions as "what is the position of the firm with respect to its competitors?", or "at which stage of their life-cycle are the products of the portfolio?", but it also includes less elaborate informations, such as the contents of a customers data-base.

→ The technical knowledge is the domain of researchers, engineers and technicians; some of it is proprietary and protected by pattents but most of it is in the know-how of experienced personnel. Technical information can consist of drawings, finite elements analyses, etc.

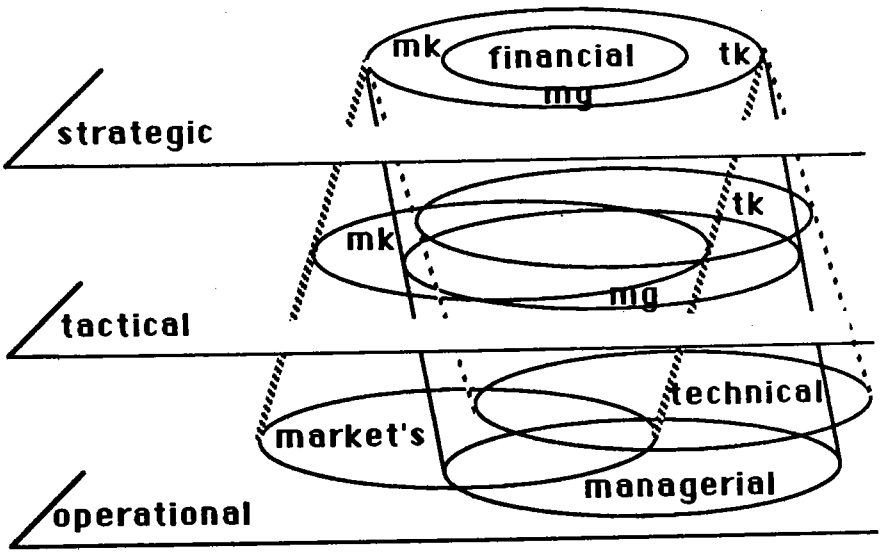
→ The managerial knowledge is what is required for (but also results from) the understanding of the behavior of the production system. It includes a representation of the production processes, of the machines' characteristics and of workforce skills.

A single decision-maker is unable to process all this information or to master this knowledge, unless the firm under consideration is particularly small or if several of the components of the decision knowledge are atrophied. For instance, if the firm is a consulting company working exclusively on defense contracts, most of the knowledge it needs is technical. In general, however, the decision process is carried out by separate entities, specilalized in only one of the classes of knowledge, like the Marketing and Sales (MS), Development and Design (DD) or Manufacturing (M) divisions.

## VI KNOWLEDGE PARTITION

The idea behind this partition is that decisions requiring specific knowledge should be made by the competent decision makers. In fact, as figure 2 illustrates it, the partition based on the nature of the knowledge or information needed for decision is all the better as the decisions are elementary. The more aggregate a decision, and the more likely it is to require elements of knowledge or information drawn from different domains. Strategic decisions, in particular, cannot be associated with any one type of knowledge exclusively. On the other hand, the financial knowledge is used predominantly at the strategic level.

To simplify the exposition, it is assumed that all the operational or tactical decisions are made by three 'operational' divisions: Marketing and Sales (MS), Development and Design (DD) and Manufacturing (M), and they are based mostly on knowledge of one kind. The strategic decisions, on the other hand, are made by the 'Strategy and Finance' (Sφ) division, based on knowledge of the four kinds.




**types of knowledge required for decisions of different levels**

**fig.2**

## VII DECISION-MAKING STRUCTURE

Given the partition of the decision-making organization adopted, its structure is entirely determined. In fact, the Strategy/Finance division translates the overall objective into local objectives for the three 'operational' divisions, and share the resources of the firm among them. The operational divisions in turn make local decisions to achieve their objective and constantly feed back information concerning their level of performance.

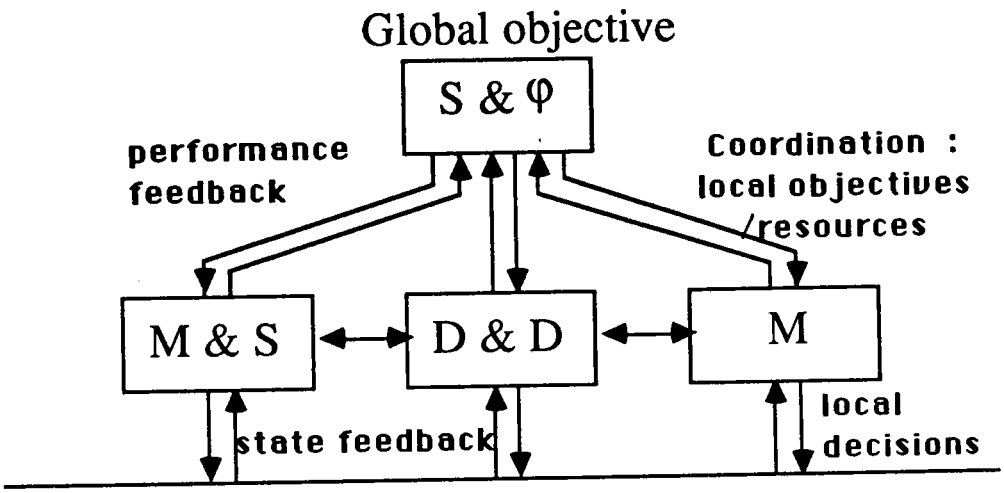


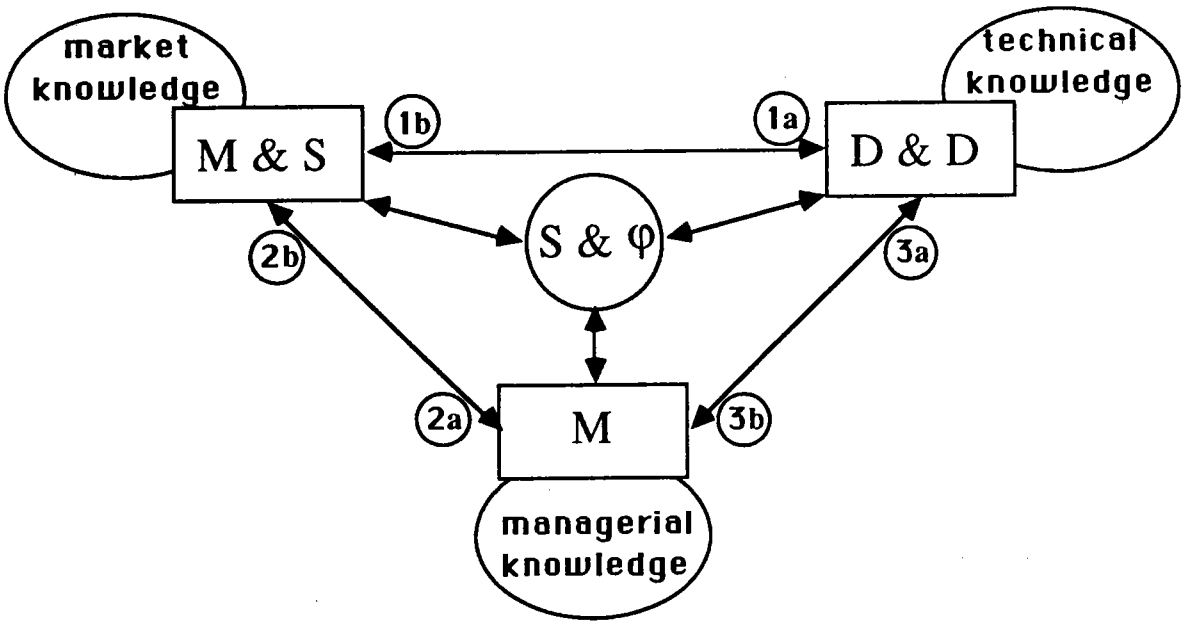

**fig.3**

As represented on figure 4, the actions that each of the three operational divisions can take depend on the decisions made by the two others. Therefore, since each of them is striving to achieve its own objective and their objectives are different, their decisions would be antagonistic if it were not for the coordination function of the Sφ division.

In that context of competition for resources, the coordination problem consists of finding the resource allocation that will yield an equilibrium between the operational divisions, and the concept of shadow prices can be used to find that equilibrium.

But this is not the only aspect of coordination: when the ability of one division to achieve its objective depends on the output of a different one, coordination for feasibility is required. This means that the outputs of each division must be sufficiently constrained for the other divisions' problem to be solvable. For example, the set of possible designs for a part will be restricted to those that allow for production in small batches if the objective of the manufacturing division is to reduce the volume of work-in-process…

The major interactions between the divisions of a firm are represented on the following flow-chart:




1a market's desiderata
1b firm's technical expertise
2a sales and forecasts
2b production capability
3a performance of the tools
3b process plan

fig.4

## VIII MANUFACTURING

Whatever its objective, the manufacturing division's ability to achieve it will depend, to some extent, on two ouputs of the other operational divisions: the process plans of the products, which specify the needs of machine-time per part to produce, and the sales plan, from which production requirements are drawn. The problem of the Manufacturing division can thus then be stated:

*Given the necessary information about the production system, find the actions to take so as to best achieve the objective assigned with the resources allocated for that purpose and in such a way that the constraints imposed by the Sales and Technology divisions (e.g. the production volumes and technical specifications) be satisfied.*

Although this is still a very general statement, the previous illustrations suggest that, for any system of realistic size, the dimensionality of this planning problem is still prohibitively large. The concepts of specialization and aggregation introduced previously should thus be used again iteratively until the planning problem is translated into a set of tractable sub-problems. At the outset, the structure of the planning system will be a recursion of the pattern represented on figure 5, i.e. a multi-level, hierarchical structure.

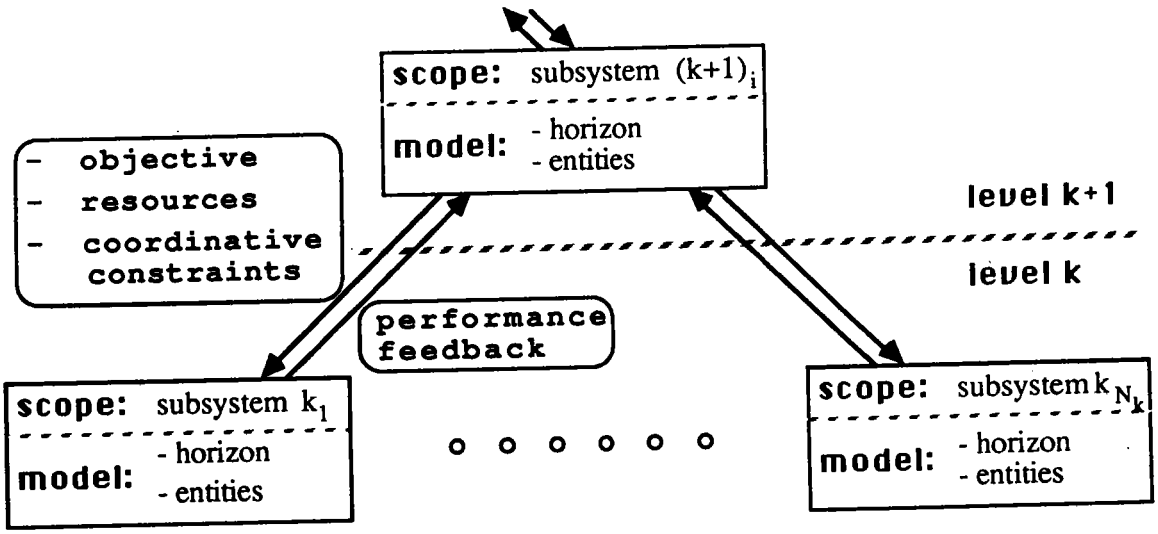



fig. 5

In practice, the design of this structure can be conducted through a bottom-up procedure, the input data being the process plans of the different products, the performance of the machines and the strategic demand forecasts. Such a procedure is described in HILLION, MEIER and PROTH [HE], and its major steps are:

→ Identify groups of machines constituting subsystems and such that for any two subsystems, all products visiting both of them will visit them in the same order,

→ Find the routings that best balances the workload of the machines within each subsystem,

→ Aggregate the products into product families so that products belonging to the same family have comparable processing times on all the subsystems,

Iterate these steps until either there is only one product family left, or the last subsystem obtained is the entire system,

→ Determine the horizons associated with each level by assuming it first infinite and then reducing it to bring the system within the bounds of tractability,

→ In a top-down approach (that is, starting from the objective of the entire division), determine the objectives of all the subproblems; these will be chosen such that the objective assigned to a given level is actually likely to be achieved within the time framework defined by the horizon of that level.

Notes: 1- This design procedure is only one of the procedures devised for that purpose (other frameworks are described in Chapter 2) and the model studied in this work is consistent with this approach. However, models of this type, i.e. flow-control models, would also be relevant in different frameworks for production management; their characteristic properties are described in the following section.

2- The constraint on the flows between sub-systems is meant to force the aggregate models to be of the 'flow-shop' type, for which there are more algorithms to solve optimal control problems; this assumption is not particularly restrictive in most heavy industries.

3- The procedure to determine the product families is basically a clustering technique: the products being represented by the tuple of their processing times on the different subsystems, the idea is to find classes of these tuples so as to minimize the total intra-classes inertia.

4- The planning horizon of an optimal control problem (P) is the smallest $h$ such that $\forall z > h$, the restriction to $[0,h]$ of the solution to (P) on $[0,z]$ is equal to the solution to (P) on $[0,h]$. For certain types of problems, like the concave costs problems (see PROTH [PR]), there exist analytical results that charaterize the planning horizon. More generally, the concern expressed in the constraint previously stated for the search of horizons is to have them shortened by means of analytical results and not merely just enough for the resulting problems to be tractable for a given computational power.

## IX PROBLEM STATEMENT

The previous sections of this chapter have been devoted to an analysis of the decisions to be made in a manufacturing firm. It has been argued that controlling a system of such nature and size, that is, gearing it towards the achievement of its objective (maximizing the expected profit) requires that the control process, also called decision making process be structured. A structure based on the concepts of aggregation and specialization was constructed, and production management was defined as a part of this decision-making structure.

As such, production management was itself described as a hierarchically structured decision-making process. One of the problems to solve in this process consists of *deciding, over a certain horizon, the production rates of the machines in the system*. This problem is called flow-control problem and it can appear at different levels of a hierarchical production management system. Depending on the level, the 'machines' considered are physical machines, or work centers, or even entire plants, whereas the 'products' whose flow rates must be determined are either physical parts, or part families. More generally, the characteristics of the physical system to be retained in the model will depend on the decision level concerned.

The flow-control model studied in this work is consistent with the the hierarchical design procedure outlined in the previous section. It is also particularly well suited for the corporate level of a large manufacturing firm: the entities considered are production subsytems and part families, and the objective is to minimize the accumulation of these 'products' between production subsystems, given that there is an external demand to satisfy and that the production subsytems have a finite capacity.

Given the problem addressed -i.e. flow control- and given that it is addressed as one level of a hierarchical framework, the work presented is organized as follows.

## X OUTLINE OF THE WORK

Chapter 2 is a survey of the work on hierarchies in production management and control. The literature on hierarchical control is surveyed in the first part; model aggregation techniques, both for static and dynamic systems, are presented first, and the remaining work is classified consistently with the notions introduced in MESAROVIC et al. [MC] of multilevel and multilayer hierarchies.

Multilayer hierarchies are introduced in correlation with the notion of time-scale decomposition, whereas multilevel hierarchies are shown to be based on the concept of coordination.

The literature on hierarchical production management is surveyed in the second part of the chapter; the first two sections are devoted to work that has directly influenced the research on hierarchical production management: monolithic models for decisions belonging to different levels, economics models of the 'resource allocation' type, and decomposition techniques for mathematical programming. Then, the seminal work of HAX and co-workers is described, together with different improvements and applications. Finally, the novel approaches to hierarchical control motivated by the interest in Flexible Manufacturing -and in particular the approach of GERSHWIN and co-workers- are surveyed.

Chapter 3 justifies all the assumptions imbedded in the fromulation of the flow-control model studied: in particular, the choice of a continuous representation of time and the absence of randomness are addressed, as well as the definition of the objective and the formulation of the capacity constraint, shown to be an approximation. (In fact, these characteristics of the model are cause that it is better suited for a high level of a hierarchical control system). The mathematical problem corresponding to this model is then formulated for a generalized flow-shop. It consists of minimizing the integral of the inventory holding cost incurred at all production subsystems, subject to a constraint on the minimal output of the system, and a constraint on the maximal flow through each subsystem.

Some analytical results concerning the solution to this problem are derived in Chapter 4; first, the solution for a single stage production system and a single product is characterized. It is shown that the optimal control is of the 'switching' type and that the production at any time is equal either to the demand or to the capacity at this time.

More precisely, it is proved that the production rate is equal to the capacity if and only if the inventory is strictly increasing or decreasing; in other words, the optimal control consists of producing as late as possible, and to overproduce only to hedge against a future peak in demand.

It is then shown that the first result also holds for a multi-stage mono-product system; in fact, if the inventory holding costs increase with the stage, the optimal control can even be obtained by solving first the single-stage problem for the most downstream subsystem and then, iteratively, for all the previous subsystems. In other words, under this cost structure, the global optimum is achieved if each stage optimizes its local criterion while satisfying its constraints, and in particular the demand it faces. However, it is shown on a counter-example that this result does not hold if the cost assumption is not satisfied, and that no 'simple' algorithm could possibly be found in that case.

The single-stage multi-product system is studied next, and it is proved that the optimal control is obtained by solving first a single-stage, mono-product problem for the most 'expensive' product -i.e. the one with highest inventory holding cost- and depleting the capacity left for the other productions accordingly, and then iterating this procedure for all the products. This result means in particular that the production rate at any time is either equal to the demand, or to the capacity, or to zero, and that if there is a product kept in inventory, then no cheaper product is being produced, and the system is working at capacity.

In order that these results be extended to the most general case (multi-stage, multi-product), a restrictive assumption on the cost is required: it is necessary that the ranking of the products by cost be the same at all stages and the cost increments for any two products between stages be in the same order as their costs.

However, whether or not this assumption holds, the optimal control must satisfy some necessary conditions which limit the 'candidates' to a finite number. The question of whether this allows to use dynamic programming to solve the problem is thus investigated in the first part of Chapter 5: the two-stage conter-example of Chapter 4 is solved by this technique, but it appears that the amount of computations required explodes as the size of the problem increases and that it becomes excessive, even for probems of realistic size.

On the other hand, the algorithm of Chapter 4 -proved to yield the optimal control under certain cost conditions- is extremely simple and can even be programmed on a spreadsheet. An example is also developed in the second part of Chapter 5, and possible extensions of this work are proposed in the conclusion.

# chapter 2:

# HIERARCHIES IN PRODUCTION MANAGEMENT AND CONTROL: A SURVEY

# INTRODUCTION

The present chapter is an attempt to survey the work that introduces the concept of hierarchy in production management and to review a representative set of techniques developed in control theory that are related to this same concept.

Obviously, the statement of this objective is somewhat fuzzy, since the word "hierarchy" can assume a wide variety of meanings: the management science literature alone provides several types of work related to hierarchies, ranging from the hierarchical decision process described in SAATY [SA] to the hierarchical production planning of HAX and MEAL [HM] or the decentralization by pricing of BAUMOL and FABIAN [BF]. Similarly, in MESAROVIC et al. [MC] -which sets the theoretical foundations of hierarchies in the context of Large Scale Systems- three different definitions are proposed for hierarchical systems, whereas the DANTZIG and WOLFE decomposition method, which is essential to the work on decentralization by pricing, is excluded from the hierarchical techniques.

That is to say this chapter certainly does not survey all the work it should but also reviews papers that might not be considered as contributions to the target field.

The work surveyed falls in three classes that can be roughly characterized by an emphasis on one of the following concepts:

1- decomposition of a physical system and coordination of the subsystems control units: the issue is to provide these units with enough information for them to achieve a global optimum.

2- layering of the decisions or control applying to a given physical system and consistency issues.

3- aggregation and disaggregation of a mathematical model: reduction of the dimensionality with least loss of information.

The first concept is mainly developed in the control theory literature, the second in the management science literature and it seems that the third one has been devoted an equally limited effort in both fields.

## OUTLINE OF THE CHAPTER

The first part of this chapter reviews the portion of the control theory literature that can be related to the notion of hierarchy; to retain the classification of hierarchies suggested in MESAROVIC et al. [MC], the concepts of multilayer and multilevel control systems are surveyed in Sections 1.II and 1.III. Section 1.I surveys the work concerning aggregation of control models.

In the second part, although the same classification could have been adopted, the work surveyed is divided in five sections: the first presents some management systems in which a hierarchical decomposition of the managerial decision process is acknowledged but does not result in a decomposition of the associated control models. Section 2.II introduces the work related to the multilevel concept, namely the decentralization of resource allocation through pricing.

The last three sections describe different multilayer management systems. In Section 2.III, the most substantial work aimed at designing hierarchical systems is surveyed: the models presented feature a multi-horizon structure and a top-down constrained decision process. In this section are also reviewed several papers addressing different issues that arise in hierarchical control, namely temporal aggregation and disaggregation, consistency of decisions at different levels, evaluation of the systems and multi-stage production systems.

The work in Section 2.IV focuses on the difficult problem of integrating detailed scheduling in a hierarchical system: the coordination scheme described is iterative. Finally, section 2.V presents the results obtained by applying a control approach to Flexible Manufacturing Systems management.

# 1. HIERARCHIES IN CONTROL THEORY

All the work presented hereunder could perfectly be considered as a collection of mathematical decomposition techniques applied to control problems. MESAROVIC et al. [MC] give it a specific identity: a theory of hierarchical, multilevel systems (as a subset of large scale systems theory), by developing a mathematical formalism for the qualitative concepts of hierarchy and by showing that these decomposition techniques fit in the provided framework.

Three types of hierarchies are identified which should account for all existing hierarchies:

. descriptive hierarchies: the lower the "stratum", the more focussed and detailed;

. decisional or multilayer hierarchies: the higher the "layer", the more complex and global the decision function;

. organizational or multilevel hierarchies: infimal (i.e. lower-level) units control subsystems and are coordinated by a supremal unit.

Common characteristics of the last two types of hierarchies are that a higher level unit is concerned with the slower aspects and with a larger portion (or broader aspects) of the system behavior and that the decision period at a higher level is longer than that of lower level units.

However, the literatures corresponding to multilayer and multilevel systems do not intersect and represent very different amounts of work. Therefore, they are reviewed separately in Sections 1.II and 1.III respectively, whereas Section 1.I introduces the concept of aggregation which is underlying in various hierarchical approaches, particularly in management.

Prior to entering the detailed description of these concepts, I bring to the attention of the interested reader that most of the work reviewed in this part is summarized in Sections II and V of the excellent survey of decentralized control methods by SANDELL et al. [SV]. Moreover, [WS] (especially chapters 1,4,5 and 8) gathers a representative selection of the work concerning multilevel systems and aggregation, whereas SINGH [SI] presents an extensive synthesis of the work in dynamic multilevel systems and FINDEISEN et al. [FB2] reviews both multilevel and multilayer systems.

The volume of the relevant work deserves these three books and explains why present part can at most claim to review a representative selection of papers in a production management perspective.

## 1.I MODEL AGGREGATION

Implicit in the notion of multilayer hierarchy (and explicit in the definition of descriptive hierarchy) is the idea that different levels require different models of the system considered and, in particular, that lower level models need to be more detailed and closer to the physical system whereas higher level models are more aggregate.

Note that the multilevel concept avoids this idea that appears to be more intuitively appealing than easy to translate quantitatively. In fact, in the multilevel systems parlance, the supremal unit's task is to coordinate the infimal units and therefore does not necessarily require an aggregate model of the physical system; moreover, Section 1.II will provide enough evidence that there is no systematic procedure to design models that would satisfy the requirement stated hereabove.

AOKI [AO1][AO2][WS] proposes a concrete but restricted formulation for the concept of aggregation in control and explores the problems arising when one tries to reduce the dimensionality of a model (i.e. if one tries to determine a control based on a reduced-size model).

## Static systems

In the static case [AO2] a model can be viewed as a mapping f between the sets X and Y of exogenous and endogenous variables. Aggregation consists of mapping these two sets on reduced dimension sets $X^*$ and $Y^*$ by means of aggregation procedures $g: X \to X^*$ and $h: Y \to Y^*$, and to define the aggregate model $f^*$ as a mapping between $X^*$ and $Y^*$. Aggregation is perfect when $h \circ f = f^* \circ g$.

$$\begin{array}{ccc} X & \xrightarrow{f} & Y \\ {\scriptstyle g}\downarrow & & \downarrow{\scriptstyle h} \\ X^* & \xrightarrow{f^*} & Y^* \end{array}$$

When perfect aggregation cannot be achieved, two types of approximate aggregations are sought; these approximations consist either of restricting the "perfection" constraint to a subset of X or approximating the equality $h \circ f = f^* \circ g$. For instance, if X is a vector space and its elements can be modelled as random vectors of known first and second moments, then $f^*$ will be determined so as to minimize the expectation of $| h \circ f - f^* \circ g |$. This same type of technique is applied by AXSÄTER in the context of manufacturing (see [AX]).

## Dynamic systems

For linear dynamic systems, the objective of aggregation is to reduce the dimension of the state vector. If the real system is described by equation $\dot{x} = Ax + Bu$ (where $\dot{x}$ stands for $dx/dt$) and the aggregate model is also sought as a linear differential equation $\dot{x}^* = Fx^* + Gu$, then a linear aggregation procedure $x^* = Cx$ yields this type of model provided that F and G satisfy $FC = CA$ and $G = CB$. In that case, it is shown that F inherits some of the eigenvalues of A. (However, in the general case, the stability of the system cannot be deduced from that of the aggregate model).

This result means that $x^*(t)$ is a combination of the modes of $x(t)$ retained by the aggregation procedure. Thus, these modes must be chosen among the dominant ones if the dynamics of the aggregate model should closely approximate those of the original one.

The particular case of a quadratic objective and a feedback control law $u = K x^*$ based on the aggregate state vector is then investigated by AOKI. The aggregate objective function is derived from that of the real system and the matrix $K$ that would yield an optimal feedback control for the aggregate model is determined. Bounds on the (suboptimal) value of the real system objective when the control $u = K x^*$ is applied are found. The aggregation matrix $C$ can then be determined so as to minimize the difference between upper and lower bounds.

In general, the condition for perfect aggregation is not satisfied but two particular cases are described in which perfect or almost perfect aggregation can be achieved. In the first case (restricted dynamics), it is assumed that there exist two matrixes $D$ and $E$ of appropriate dimension and rank such that $A = DEC$. Then $F = CDE$ will automatically satisfy $FC = CA$. Reciprocally, disaggregation is always feasible in that case, that is, the value of the original state vector can always be derived from that of the aggregate state vector and from the past values assumed by the control. The second case can be illustrated by means of a geometric interpretation of the aggregation procedure as a projection over a (possibly time-varying) subspace $S$ of the state space. Perfect aggregation means that the path generated by the real system lies in the subspace $S$.

If a feedback control of the type $u = Lx$ is applied, two conditions can be derived for $A$,$B$ and $L$ to yield a good aggregation, namely that if the initial state vector $x_0$ is in $S$, the trajectory must remain in $S$ and that if $x_0$ is not in $S$, the distance between $x(t)$ and $x^*(t)$ has to tend to zero. Unfortunately, in this case, disaggregation will never be achieved exactly but modulo a subspace (the subspace along which the projection is performed).

Finally, if the linear relation between aggregate and real state vector cannot be maintained because the condition for perfect aggregation is not satisfied, alternative aggregate models can be investigated, which represent the real system with enough accuracy for an aggregate model-based control to yield a good behavior of the real system. For instance, a model described by $x^* = Fx^* + Gu + Dy$, where y is the observed output of the real system is proposed in AOKI [AO1].

This alternative approach highlights the fact that in this work, the structure of the aggregate model is assumed given. As pointed out in SANDELL et al. [SV], the theory that would make it possible to determine the structure of the aggregate model from the description of the real system (detailed model, objective function) is lacking.

The definition of multilayer systems shows that the concept of aggregation is essential to this type of hierarchical control: if the control function is layered and "higher" layers have to make more complex and global decisions, it is very likely that the models to be used by these layers are not as detailed as those used to make local decisions. The aggregation techniques proposed by AOKI are mostly intended to retain the dominant modes of the detailed model in the aggregate one and, in that respect, they are perfectly well suited to the needs of multilayer systems.

In production systems models, the dynamics are usually represented by linear equations like those considered by AOKI; unfortunately, the variables are bound to lie in a constraint set and the aggregation and disaggregation of these sets is still an unresolved issue (Section 2.III gives evidence of this satement).

Another remark concerning AOKI's work is that it focuses on the aggregation procedure itself and not at all on the "direction" along which aggregation is performed: in [AO1] and [AO2], the aggregate variable sets are arbitrary. The fraction of the multilayer literature reviewed in next section present one of the possible direction for aggregation, namely the time behavior of the variables.

## 1.II TIME-SCALE DECOMPOSITION FOR MULTILAYER HIERARCHICAL CONTROL

Time-scale decomposition generally refers to a technique developed for the analysis of dynamic systems in which different components of the state vector have very different dynamics, that is, when the modes of the system can be partitioned in such a way that, for any two modes belonging to different classes, one is fast compared to the other (see CHOW and KOKOTOVIC [CK] or SANDELL et al. [SV]).

In the case of a singular perturbation (i.e. "a perturbation to the left-hand side of a differential equation" [SV]) the model can be simplified insofar that, when the system is considered in a given frequency band, the state variables corresponding to lower frequencies (i.e. slower modes) can be considered constant, whereas those corresponding to higher frequencies can be discarded. One of the major reproaches to these works is that the structure of the system (which modes are "fast", which ones are slow) has to be given.

CODERCH et al. [CW] consider the class of linear systems defined by $\dot{x}(t) = A(\varepsilon)\, x(t)$, where the matrix $A(\varepsilon)$ is analytic in the small parameter $\varepsilon$. Under necessary and sufficient conditions on $A(\varepsilon)$, the system exhibits a multiple time-scale behavior, which means that $\exp(A(\varepsilon)\, t/\varepsilon^k)$ can be approximated by different matrices depending on the exponent $k$.

This "descriptive" decomposition "automatically" yields a set of reduced order models, each representing the behavior of the system accurately at a given time-scale. If $A(\varepsilon)$ is the transition matrix of a Markov process, the aggregate models can be interpreted as being obtained by collapsing states between which the transitions are frequent compared to the transitions between states lumped in two different aggregates.

Although the qualitative notions related to this technique appear in the work reviewed in this section, the term time-scale decomposition will assume a much looser interpretation in the following description of multilayer hierarchies.

MESAROVIC et al. [MC] first introduce the concept of multilayer decisional hierarchies and exemplify it by the early work of ECKMANN and LEFKOWITZ [EL]. These authors suggest a decomposition of the control task in several sub-tasks of different "natures" in order to accomodate the concept of adaptive control. More precisely, they state that the task of updating the parameters of an optimizing model for automatic control can itself be automated and introduced in the controller as an additional layer. In this setting, the higher layers do not affect the system under control but only the lower control layers. This structure is called "functional multilayer hierarchy" in FINDEISEN et al.[FB2].

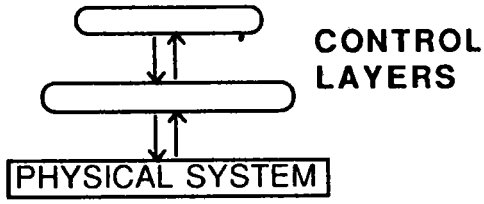
The other multilayer concept (termed "multi-horizon" in FINDEISEN et al.[FB2]) is also introduced in MESAROVIC et al.[MC] and related to the hierarchical management systems in which the controller is decomposed "into algorithms operating at different time intervals" [FB2],[FV].

All the layers of the controller directly affect the process but the higher ones control only its slower aspects : they intervene less frequently, with a longer optimization horizon and based on models that retain only the variables of interest (i.e. aggregate models).

The interference with the terminology of time scale decomposition is evident. However, the type of techniques used in singular perturbation analysis do not apply, for there is not necessarily a partition of the state vector. Actually, the variables manipulated by the higher layers may be aggregates of the lower layer variables, which in turn raises the issue of consistency between the decisions made at different levels. This aspect is examined in the management literature (see Section 2.III).

Although the multilayer concept characterizes one of the two fundamental classes of hierarchical systems (the other being characterized by the multilevel concept), SANDELL et al. [SV] point out that the literature in this field is rather scanty and qualitative.

## FUNCTIONAL MULTILAYER HIERARCHIES




In [EL], ECKMAN and LEFKOWITZ first insist on the advantages of model-based control, then define the concept of adaptive control and propose a conceptual decomposition of the controller in four layers: the lowest layer is a set of ordinary servo-loops devised to keep the system at an operating point set by the second layer. These servo-loops are designed to cope with the disturbances, thereby allowing for deterministic modelling of higher layers. The second layer (optimization) solves an optimal control problem and sets the operating point for the servo-loops.

The next higher layer (model adaptation) is in charge of the periodic adjustment of the optimizing model parameters; it basically "forces a best fit of this model to the past system behavior in the vicinity of the operating point" [EL]. The highest layer (system evaluation) includes human intervention to modify the criteria or structure of the models built in the optimizing and model adaptation layers.

The advantages of this approach according to the authors are the consistency of the objectives of the different layers with the overall performance criterion, the flexibility in the choice of the optimization method for each layer, and the fact that each layer operates in a given time domain and thus can operate independently of the next higher one.

All these ideas are considerably refined in LEFKOWITZ [LE]. The suggested controller design procedure is built on the assumption that the control task is divided in four functions (regulation, optimizing control, adaptive control and self organizing control) operating at decreasing frequencies and with different information sets.

The notion of multilevel decomposition is then introduced in that framework. The author points out that if the process under control can be split in subprocesses, each of these can be controlled separately by a multilayer controller assuming that the interactions between subprocesses remain constant. A higher level supervisor would then coordinate the controllers to cope with deviations in the interactions, assumed much slower than the variations of other variables. This idea of combining multilayer and multilevel concepts was already present in MESAROVIC et al.[MC] and is also reminiscent of the singular perturbation theory.

These notions are then mathematically formulated in the case of a continuous process and an insightful result is pointed out, namely that the first and third layers simplify the second layer model by dealing with certain classes of disturbances, which introduces the idea of a tradeoff in the amount of computation required from the different layers. For instance, if the optimizing layer is sufficiently accurate in its representation of the system, it will obviate the adaptive layer but require more computations.

Finally, some "ordering" features are provided as design guidelines and, besides the obvious ordering in the sampling periods of the different layers, it is suggested that these periods should be determined in order to balance the loss in performance that they originate and the computation cost.

A different application of the multilayer concept can be found in the control of systems modelled as large Markov chains. In [FV], FORESTIER and VARAYIA characterize a two-layer structure by three essential features, namely that the upper layer must have a longer sampling period, must use less information and solve a "higher level" problem. They point out that, under these assumptions, the system performance should increase if the supervisor's interventions are more frequent or based on a larger information set.

In the problem investigated, a process described by a series $(s_t)$ taking values in a finite state set S is considered. For each state reached, there is a nonempty set of feasible controls that the regulator can apply and a cost is associated to each pair (state, control). A strategy is defined as a function assigning a control to each state and, for any given strategy, the process $(s_t)$ is Markovian. Thus to each strategy corresponds a cost defined as the expected value of the state-control costs cumulated along a random path.

The multilayer concept is introduced in that a boundary set B included in the state space S is defined so that whenever a boundary state is reached, a new strategy can be defined. Thereby the controller is divided into a regulator and a supervisor, where the regulator applies the strategy imposed by the supervisor and the supervisor controls a process $(b_t)$ slower than $(s_t)$ since the jumps occur only when $s_t$ is a boundary state. In these conditions, $(b_t)$ is a Markovian process, while $(s_t)$ is not any more. Namely, the transition probabilities depend on the control applied and, whereas in the case of a single strategy the control was entirely determined by the state $s_t$, in the two-layer case it also depends on the current strategy.

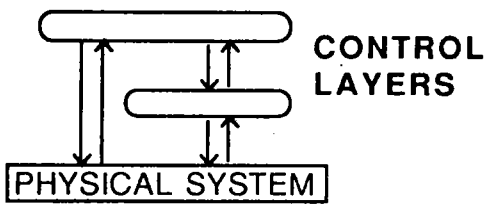
The supervisor's problem consists of choosing a regulator strategy to assign to each possible state of the process it observes: thus a meta-strategy can be defined as a function assigning a strategy to each boundary state and the cost criterion has to be "lifted" to the supervisor's level. It is proved that the expectation of the cost relative to a meta-strategy exists and necessary and sufficient conditions for the optimality of a meta-strategy are given.

Besides its specificity due to the particular nature of the model considered, this application of the multilayer concept is interesting in that it features some characteristics of the two types of multilayer structures. The choice of a regulator by the supervisor when the system reaches a certain type of state can be interpreted as an example of adaptive control, or the Markovian process observed by the supervisor can be viewed as an aggregate model of the system and

the interactions between regulator and supervisor compared to the ones that arise in management.

It can be noted that the ordering in intervention frequencies is retained (the process observed by the supervisor is slower than the regulator's), even though the interventions are imposed by feedback of the stochastic process state. FORESTIER and VARAYIA note that all the computational burden is on the supervisor. Since this is an undesirable feature, the work reviewed hereunder is directly aimed at balancing the computations required from the different layers.

## MULTI-HORIZON HIERARCHIES




The concluding idea proposed in LEFKOWITZ [LE] of finding a trade-off between loss in performance and computation cost to determine the sampling rate is further developed in DONOGHUE and LEFKOWITZ [DL], and substantiated by the results in sensitivity analysis: the objective is to design a multilayer and multivariable controller for a static system facing disturbances, under the assumption that each higher layer will update a larger number of variables at a lower frequency.

The problem is then to determine the number of layers, the sampling rates and variable partitioning so as to minimize the weighted sum of the expected loss in performance due to the effect of disturbances and of periodical action on the one hand and, on the other hand, the costs of computation and implementation of the results. Each of the layers will then affect directly the control variables but the higher layer models will be of higher dimension than the lower ones.

The approach adopted consists of assigning to the lower layers those variables to which performance is more sensitive. Therefore the partitioning reduces to determining the number of variables to assign to each layer, the variables being ranked by decreasing sensitivity. Similarly, the sampling periods are chosen so that higher layer periods are multiples of lower layer ones. A heuristic search procedure is described and a numerical example presented.

It then appears that this approach achieves a decomposition of the control and still retains an interesting characteristic, namely that the sampling rate depends on the disturbance frequencies. The sensitivity analysis is devised to partition the controlled variables for that purpose. Furthermore, the actions of the different layers are consistent since they are all aimed at keeping the system at the operating point (the system is assumed static).

However, this approach is implicitly based on the assumption that there is no "natural" frequency for any type of decision (natural in the sense of "resulting from its nature or characteristics"). In management, the systems modelled are seldom static and some of the decisions must be made at at given frequencies, be it for organizational reasons (annual contracts) or physical reasons: response-time of the system.

FINDEISEN et al. [FB2] describe a multi-layer control structure that is very similar to those presented in Section 2.III and have become typical for hierarchical management: the models (controlled variables, state variables and disturbances) manipulated by the different layers are more aggregated for higher layers and the optimization horizons are longer, whereas the objective functions are qualitatively the same for all the layers. The decisions are made according to statistical models of the disturbances but generally, the effects of these disturbances are controlled by repetitive open loop optimization, the values of observed variables being updated for each computation.

Moreover, all the variables being inter-related by constraints, the values assumed by those controlled by the higher layers influence the values that can be taken by lower layer variables. Finally, the minimal optimization horizon of each layer can be determined as the settling time of the system described by the model adopted.

This approach is illustrated by the control model of a water supply system with retention reservoirs. The top layer determines the optimal levels of the main reservoirs (or groups of small ones) over a long horizon. The lowest layer determines all the flows over a short horizon in order to optimize an objective function whilst reaching a final state consistent with that determined by next higher layer and still meet the "external" requirements.

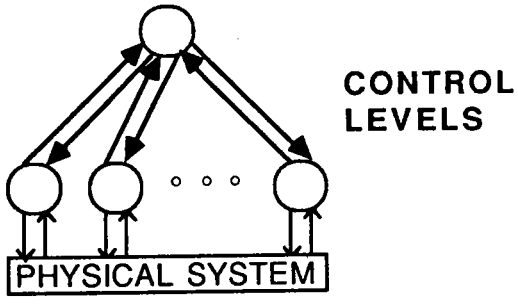
This example illustrates the difference between ordinary multi-variable systems and multi-variable systems with significantly different response times to changes in the control variables. For this latter class of systems (and the water supply system belongs to it since changing the level of one of the main reservoirs will take much more time than regulating the flows), the multi-horizon structure is particularly well suited for two main reasons: first, the "slow" variables must be modified less frequently if the effect of a decision is to be observed before the next one is made and second, the decision of altering these variables must be made over a longer horizon.

The work reviewed in Section 2.III in particular suggests that manufacturing sysytems are particularly amenable to a multihorizon control structure. Along this vein, the approach suggested by DONOGHUE and LEFKOWITZ [LE] would be particularly appealing if the sensitivity analysis that they refer to could yield "automatically" a ranking of the variables that would reflect these characteristics. Paradoxically, in manufacturing systems, the decisions made the least frequently (e.g. building a new plant) are supposedly the ones with the highest influence on the performance of the whole system, whereas in the approach of DONOGHUE and LEFKOWITZ the variables to which performance is most sensitive are those to update most frequently.

This contradiction suggests that the performance sensitivity cannot be the only factor considered to determine the frequency of a decision, and that such factors as the response time of the system should be taken into account in the design of the control unit. Unfortunately, according to FINDEISEN et al.[FB2], the quantitative techniques to design multi-horizon systems have not been developed yet. Therefore, all the multilayer management systems described in Part 2, although designed to reap full advantage of these features, have been developed empirically. This, however, is not their major defect. Problems of consistency between decisions made at different layers arise and the optimization problems to solve at each layer are different in nature. This results in an unbalanced computational effort, even when the horizons are chosen to counterbalance the difference in computational difficulty.

The alternative hierarchical decomposition of a control problem, namely multilevel decomposition, avoids the first of these problems and could be a solution to the second one.

## 1.III DECOMPOSITION-COORDINATION: MULTILEVEL HIERARCHICAL CONTROL




Unlike the multilayer systems literature, the literatur introducing multilevel systems is vast; most of MESAROVIC et al. [MC already addresses the central issue in multilevel systems, namel coordination.

For a two-level organizational hierarchy, mathematical meanings are given to the notions of coordinability (i.e. existence of a supremal coordination control for which the infimal units can solve their local control problem) and consistency. The Consistency postulate states that whenever the supremal and infimal units can solve their respective problems, then an overall solution exists.

Furthermore, two coordination principles are derived. The Interaction Prediction Principle states that if the supremal unit predicts the values of the interactions between the subprocesses controlled by the infimal units in order to coordinate their action, then the overall solution is reached when the value of these interactions resulting from the controls suggested by the infimal units is equal to the predicted value. The Interaction Balance Principle states that if the interaction variables are let free, then the overall solution is reached when the values they are given independently by the infimal units are consistent.

As will appear along this section, these principles provide the qualitative interpretation for two wide classes of hierarchical decomposition techniques, generally referred to as model decomposition (or feasible) techniques and goal coordination (or dual-feasible) techniques. (In this terminology, "coordination" and "decomposition" are used interchangeably since they represent the two dual phases of a same method and "unfeasible" refers to the fact that while an iterative procedure is used to solve the problem, only the final solution satisfies the constraints).

The first algorithms for multilevel systems were found in open loop dynamic control and further extended to closed loop, static and dynamic control. SMITH and SAGE [SS] give an excellent tutorial introduction to the multilevel concepts and techniques: they consider a system described by the equations:

$$\dot{x} = f[x(t),u(t),t] \qquad \text{and} \qquad x(t_0) = x_0$$

and the optimal control problem consists of minimizing the cost function:

$$J = \theta[x(t_f),t_f] + \int_{t_0}^{t_f} \phi\,[x(t),u(t),t]dt$$

It is assumed first that this system can be decomposed into N subsystems described by equations:

$$\dot{x}_i = f_i\ [x_i,u_i,\pi_i,t] \qquad \text{and} \qquad x_i(t_0) = x_{i_0}$$

where the variables $\pi_i$ represent the interactions between subsytems, and moreover, that the overall performance criterion is separable, that is, local criteria can be found for each subsystem so that their sum is equal to the overall criterion:

$$\theta = \sum_{i=1}^{N} \theta_i\ [x_i(t_f),t_f] \quad \text{and} \quad \phi = \sum_{i=1}^{N} \phi_i\ [x_i,u_i,t]$$

When the optimum is reached, the values assumed by the interaction variables $\pi_i$ must be consistent, which is expressed by the constraint $\pi_i = g_i(x,u)$. The Pontryagin maximum principle is applied to determine necessary conditions for optimality; the Hamiltonian for the overall system can be defined in terms of the subsystem variables as:

$$H\ [x,u,\mu,\beta,\pi,t] = \sum_{i=1}^{N}\{\ \phi_i[x_i,u_i,t] + \mu_i'(t) f_i[x_i,u_i,\pi_i,t] + \beta_i'(t)[\pi_i(t)-g_i(x,u)]$$

With the additional assumption that the functions $g_i$ are separable, this Hamiltonian becomes separable too:

$$\text{if} \quad g_i(x,u) = \sum_{j\neq i} g_{ij}\ (x_i,u_j) \qquad \text{then} \qquad H = \sum_{i=1}^{N} H_i$$

$$\text{where} \quad H_i = \phi_i + \mu_i' f_i + \beta_i'\pi_i - \sum_{j\neq i} \beta_j' g_{ij}(x_i,u_i)$$

Each term $H_i$ is itself the Hamiltonian of an "infimal" problem that can be formulated as:

$$\underset{u_i}{\text{Min}}\ J_i = \theta_i + \int_{t_0}^{t_f} [\ \phi_i + \beta_i'\pi_i - \sum_{j \neq i} \beta_j' g_{ij}(x_i,u_i)\ ]\ dt$$

$$\text{s.t.} \quad \dot{x}_i = f_i\ [x_i,u_i,\pi_i,t] \quad \text{and} \quad x_i(t_0) = x_{i_0}$$

At that point, a theorem due to MACKO justifies the decomposition. If there exist solutions both to the global and to the infimal problems, then those that satisfy the necessary conditions for optimality relative to the subproblems also satisfy the necessary conditions for the overall problem.

The original problem (of finding controls which satisfy necessary conditions for optimality) has thus been decomposed in a set of lower dimension problems that will yield the overall solution provided that their resolution can be coordinated. Several coordination techniques are described, including those corresponding to the two principles suggested in MESAROVIC et al. [MC].

**1. The prediction principle**: the supremal unit predicts the values of the interactions $\pi(t)$ and supplies the infimals with both $\beta(t)$ and $\pi(t) = \Pi(\beta)$. The infimals then satisfy their local problems and determine the optimal $x_i^*(t)$ and $u_i^*(t)$. The global solution is reached when $\Pi(\beta) = g(x^*,u^*)$ , which means that the interactions resulting from the optimal controls determined by the infimals are equal to their predicted value.

This coordination technique is termed feasible because in an iterative procedure to determine the optimal solution, the interconnection constraints would be satisfied at each step, since the values of the interaction variables are determined by the supremal unit. Unfortunately, as pointed out in PEARSON [WS], this positive feature has its disadvantage when constraints ($R_i\ [x_i,u_i,\pi_i,t] \geq 0$) are considered, namely that the infimal feasible sets may be empty for given values of the interaction variables.

This difficulty would not arise if the constraints were linear and the number of interaction variables were less than the number of control variables, but this is generally not the case. Therefore, the goal coordination technique has more applications.

**2. The balance principle**: the coordination variables are only the $\beta(t)$; the supremal unit supplies the infimals with the values of these variables and the infimals in turn solve their respective problems and determine $u^*$, $x^*$ and $\pi(\beta,u^*)$ i.e. the values of the interaction variables that would maximise their objective (the variables $\pi(t)$ are treated by the infimals as additional control variables). Since the interaction variables are actually determined by the control and state variables, consistency is achieved when $\pi(\beta,u^*) = g(x^*,u^*)$.

For each value $\beta$ of the coordination variables the problem to be solved by the infimals consists of determining the values of $u_i, x_i$ and $\pi_i$ which minimize $J_i(u_i,x_i,\pi_i,\beta)$; hence to each value of $\beta$ will correspond a value $J(\beta)$ of the overall objective.

It is proved that if $\beta^*$ is the value of the coordination vector that solves the overall problem and, for a given value $\beta$, each subproblem has a unique solution $u_i(\beta), x_i(\beta), \pi_i(\beta)$ then $J(\beta) < J(\beta^*)$ or $\beta = \beta^*$. Qualitatively, this result means that since the overall solution is more constrained than the solution to the set of unrelated subproblems, the overall minimal cost is not less than the sum of the infimal minimal costs.

This result is stated in a more general form in LASDON [LA] and PEARSON [WS], and related to the theory of duality. If $f$ and $g$ are linear, the necessary conditions for optimality are also sufficient. In that case, the following relation holds:

$$J_d(x,m,\pi) \leq J_d(x^*,m^*,\pi^*) = J(x^*,m^*,\pi^*) \leq J(x,m,\pi)$$

where $J_d$ stands for the cost related to the solution of the dual problem. Moreover, under adequate assumptions, $J_d(\beta)$ is concave over convex subsets of the feasible set. Therefore, the minimization

problem is transformed in a max-min problem, that is, the search for a saddle point. Three methods are described to iteratively determine the values of the coordination variables: gradient, conjugate gradient, and variable metric methods.

This scenario is common to a number of algorithms reviewed in SINGH, DREW and COALES [SD]. TAKAHARA's algorithm is an application of the method introduced above to the linear-quadratic case (the strong duality theorem does not hold in the general case).

The dual optimization method of LASDON consists of reformulating the original problem by means of Lagrangean relaxation. Since the Lagrangean is separable, the dual problem is split and coordination is achieved by means of the multipliers. TAMURA adapts this algorithm to the discrete time case and simplifies its resolution by adding a third level in which each subunit corresponds to a time period. This enhancement yields analytically solvable lower level problems and replaces a dynamic problem by a static one, as pointed out in SCHOEFFLER [WS]. The same approach is adopted to decompose a time-delay system and an application in traffic control is presented in [SD].

**3. The use of a penalty function** The objective function is modified to include a quadratic term penalizing the difference between actual interactions and interactions that are optimal for the subsystems. This method has the defect that it slows down the convergence of the gradient search algorithm used by the supremal unit.

Several applications of these principles are described in the literature and extend these results along different directions. Since the use of conditions derived from the maximum principle will actually yield the optimal control only in restricted cases (basically in the linear case), SINGH [SI] surveys the methods required in the nonlinear case: besides the techniques that avoid the emergence of a duality gap -like that of squaring the interconnection constraint- a method consisting of forcing the controlled system to "follow" a hierarchically controlled linear system is presented.

Similarly, the introduction of feedback to the coordinator is investigated by FINDEISEN et al. [FB1] in the static case i.e. for a fast system facing slow disturbances. Both direct instruments and price instruments are showed to be usable. In the "direct" case, the coordination variables are the subsystems outputs, which directly determine the interactions. The method is then a particular case of interaction prediction and the study is aimed at adapting this method to a case in which the set of possible coordinations is not known.

When price instruments and feedback to the coordinator are used, an enhancement of the interaction balance method is required. Not only must balance be achieved beween the model-based interaction values determined independently by the subsystems, but these values have to be consistent with the interactions actually observed. Finally, since use of feedback requires that the intermediate controls be implemented and the methods presented previously are infeasible, the concept of safe control is introduced. The solution proposed to prevent the control from violating the constraints is a projection on the set of safe controls.

SINGH [SI] formulates the control problem as that of optimally bringing back a system to its steady-state after a disturbance has removed it from that state, "optimally" meaning "so as to minimize a given function of the state and control trajectories". The open-loop controllers derived by a hierarchical decomposition require a time-consuming calculation that makes them impractical for the on-line control of any large-scale system, except those with very slow dynamics. Namely, that computation must be performed each time a disturbance is observed, and the initial conditions may have changed by the time the control is determined.

Therefore, a closed-loop solution is highly desirable, especially if the parameters of the feedback law do not depend on the initial conditions, that is on the disturbance, for in that case, a measurement of the state determines the optimal control. This type of

feedback law can only be computed for linear quadratic systems and SINGH [SI] modifies the interaction prediction approach used in TAKAHARA's algorithm by introducing the open loop compensation (O.L.C.) vector and showing first that the control vector can be written as a function of the state vector and the O.L.C. vector, and then that the O.L.C. vector and the state vector are related by a time invariant transformation.

A computational method is also proposed in SINGH and TITLI [ST] to determine a feedback law to be applied by the infimal units when the objective is nonlinear. The infimal open loop control problem resulting from PEARSON's goal coordination method is reformulated as a two point boundary value problem in terms of the control and coordination vectors. This problem can be solved by quasilinearization and yields a relation between the coordination and the state vectors. After substitution of the coordination vector in the expression of the control, this becomes a function of the state and time; a feedback law has thus been derived, but which is unfortunately initial-state dependent.

Some sub-optimal multi-level control methods are investigated, in particular in SINGH [SI] for the case of serially interconnected subsystems. These subsystems are characterized by the fact that the dynamics of a system depend on the state it has reached and the control it is subjected to, but also on the state reached by next "upstream" system a given time in advance.

A classical hierarchical method applied to a system consisting of a series of such interconnected subsystems would imply an enormous computational burden. An intuitively appealing sub-optimal method consists of solving first the control problem corresponding to the most upstream subsystem, and then solve the problems corresponding to next downstream subsystems sequentially, the near optimal state trajectories being fed forward. An implementation is described, in river pollution control.

Some stochastic control considerations also are introduced through the estimation aspect. Any estimator or filter can be cascaded with a deterministic controller to constitute a control structure capable of operating in a stochastic environment.

DELEBECQUE and QUADRAT [DQ] consider a controlled Markov process with generator $B(u) + \varepsilon A(u)$ and motivate this study with the example of a power plant operation. The control problem consists of finding a strategy (mapping the state space on the control domain) that minimizes the expected discounted cost associated with the evolution of the state. If the matrix $B$ has $N$ ergodic sets, an accurate approximation of the optimal control is found by solving an aggregate N-state problem. In this aggregate problem, the costs associated with each state are determined by solving the optimal stochastic control problems corresponding to the related ergodic sets. The main advantage of the approach is a reduction of the dimensionality of the problem and a decomposition of the solution.

## EVALUATION

To replace this work in the management perspective, three remarks can probably account for the interest of hierarchical control.

First, as regards the mathematical aspect, the sample of literature reviewed shows that many algorithms are variants of a reduced number of basic ones and it is not really surprising to see COHEN [CO] or LOOZE [LZ] claim a unified approach for decomposition-coordination techniques.

Furthermore, the evaluation of multilevel techniques in SANDELL et al. [SV] amounts to saying that the computational requirements are not reduced, except for the special cases in which the problems to be solved are not only of reduced dimension but of a simpler type. Moreover, the only obvious reduction in information requirements concerns the knowledge of the model, especially in human organizations.

Finally, whether or not this can be considered a criterion to evaluate the applicability of a work, there seem to be very few implementations of multilevel techniques reported in the literature, and the models actually implemented (e.g. SINGH et al. [SD]) cannot be adapted to management problems. The only exception known to the author is LASDON [WS], that is, the application of decomposition to a very specific problem, namely determine the number of machines to set up for each of N products during each of T time periods, in order to satisfy demand while minimizing inventory-holding and setup costs.

Hence the conclusion is that the techniques reviewed in this first part on hierarchical control cannot be directly applied to production systems. However, all the concepts characterizing a hierarchical management system appear in the control literature and such references as [DQ], [DL], or [MC] will provide a systems designer with valuable insights.

# 2. HIERARCHIES IN MANAGEMENT

"The purpose of manufacturing system control is not different in essence from many other control problems: it is to ensure that a complex system behave in a desirable way." (GERSHWIN et al. [GH]). It seems however that management and control have not reached the same degree of maturity.

For instance, already in 1960, LEFKOWITZ [LE] describes the considerable advantages of model-based optimizing control over the direct control method (i.e. the method that consists of manipulating the input of a physical system in a direction observed to improve its performance). In the context of manufacturing systems, however, the lack of adequate models still requires the "direct method" to be used in 1987.

In the wafer fabrication industry, where random yields, failures and reentrance complicate the process, draining all the buffers of a whole facility before resuming the loading at a controlled rate was the only policy that some production managers found to reach a state in which a reasonable throughput would be achieved. (It has been found empirically -see [CH]- that in this particular industry, the ratio of average throughput time to average processing time increases at an increasing rate and tends to infinity when the output approaches a critical value which determines the effective capacity of the plant). This example shows that the behavior of the system is simply not understood and "ensuring that it behaves in a desirable way" may not be an easy task.

Production systems have the additional particularity that the decisions to be made in order to influence their behavior (or, in other words, the control variables) are not all of the same kind. For instance, the decision to machine part 71 before part 53 on lathe 7, and the decision to increase by 10% the production of the plant over

the next two years are essentially different, at least in the sense that they have different scopes, different response times, require different types of informations and represent a different risk for the firm. And yet, both should be aimed at ensuring that the system behaves in a determined way, considered desirable. In that sense, they are somewhat redundant, which adds to the difficulty of the problem, because they have to be made consistently.

Because breaking down a problem into more easy to handle subproblems is one of the fundamental processes of human reasoning, the management of a production system is divided in a number of different functions.

### 2.0.1 How the Managerial Decision Process is Decomposed

HAX [HA2] reviews several frameworks to classify the logistic decisions, with a particular interest in the hierarchical taxonomy described by ANTHONY. In [AN], this author proposes a classification of decisions as strategic, tactical, and operational, based on their horizon and scope, as well as level of information detail, degree of uncertainty and level of management involvement, according to the most common practices in enterprises.

Strategic decisions are defined as the decisions related to long term marketing and financial policies as well as with facilities design. Tactical decisions consist of deciding the work-force and overtime levels, as well as production rates of aggregate products. Typically, the problem to solve in order to make these decisions is called Aggregate Capacity Planning. ZOLLER [ZO] defines this problem as that of "adapting production processes to fluctuating demand". Operational decisions (or detailed scheduling) concern typically the assignment of workers to machines where they will perform given jobs so that a number of requirements be met.

Another type of problem arising in production control as well as ordering (or inventory control) is the lot sizing and scheduling problem. It is that of determining the best compromise between setup costs and inventory holding costs. This problem can be seen as intermediate between aggregate capacity planning and detailed scheduling since it requires that products be considered at a low level of aggregation but does not model the sequence of operations these products have to undergo. In other terms, the production system is modelled as a global set of "resources".

### 2.0.2 Coordination / Integration of Decisions

These problems encountered in production management cannot be solved independently. The need for an integration of the tools developed to solve each of them was thus felt very early. HOLSTEIN [HO] argues that inefficiencies that appear at the short term control level can be due to bad longer term decisions and he describes the information flows required by an integrated system that would link long term capacity planning, master scheduling and short term scheduling, and in which the necessary flexibility would be kept by use of feedback.

Note: in the terminology introduced above, the equivalent would be strategic decisions, aggregate planning and detailed scheduling; master scheduling is however a standard term and it will be used again in the remainder of this work.

Twenty years later, the information systems to support an integrated approach to management exist, but the models to ensure a coherent multi-level control still require some research. Most of the work reviewed hereunder propose models to deal with the interactions between decisions concerning two (at least) of the levels corresponding to aggregate planning (AP), lot sizing and scheduling (LSS) and detailed scheduling (DS) but very few of these models have actually been implemented.

### 2.0.3 The Computational Aspect

One of the guidelines suggested in [LE] for the design of multilayer systems is that the lower level problems, that have to be solved more frequently, should require less computations than the higher level ones. Unfortunately, the computational difficulty of the models typically associated with the three problems AP, LSS and DS increases as the degree of information detail increases.

Aggregate planning can generally be formulated as a reasonably solvable linear or non-linear program; lot-sizing involves discontinuous variations and requires more sophisticated algorithms; and if the physical system considered has no structure that can be exploited, detailed scheduling models result in a combinatorial search, i.e. an NP-hard problem.

Since the difficulty inherent to the optimizing methods that could solve these models stems from dimensionality, one can interpret the increase in "sub-optimality" of the solutions found for these three problems as resulting from the failure to reduce their scope. Very few results have been found concerning decomposition of a production system for managerial purposes and the attempts to consider multi-stage systems have resulted in a considerable increase of the complexity of the models and in the loss of the interesting properties featured by single-stage systems (some thoughts about this issue can be found for example in CANDEA [CA]).

The "solution" found for this problem has often been a decomposition of the mathematical problem stated in the model (e.g. the efficient lot-sizing algorithm of LASDON and TERJUNG [LT]). Unfortunately, Section 2.1 gives extensive evidence of the gap there can be between mathematical decomposition and managerial decentralization.

## 2.I MONOLITHIC MODELS FOR MULTI-LEVEL DECISIONS

The first work entering this category is probably GELDERS and KLEINDÖRFER [GK1],[GK2]. The authors consider the problem of finding the optimal trade-off between aggregate planning and detailed scheduling costs. In their setting, the aggregate decision variable is the level of overtime, whereas the detailed level decisions consist of finding a schedule that minimizes the costs related to tardiness and flow-time.

Since the detailed problem is constrained by the aggregate decision, the model proposed includes the criteria of both levels and yields a globally optimal solution obtained by branch and bound. However, a significant result of the computational experience is that the level of overtime determined by a Fibonacci search on the lower bounding function is very close to the optimal solution. This result therefore means that it is near optimal to make the aggregate and detailed decisions sequentially.

Although the authors did not emphasize this point this is typically the kind of idea that triggered all the work reviewed in section 2.III. However, given the specificity of the model and the assumptions made, the result could by no means be considered an analytical proof of the near-optimality of the top-down constrained models of section 2.III.

As pointed out previously, lot sizing and scheduling can be considered as a problem related to an intermediate level between aggregate planning and detailed scheduling. Note that it can also not be considered a distinct level. In particular, HAX reviews the work described hereunder in his survey of aggregate capacity planning [HA3]: However, the basic assumption in HAX and MEAL [HM] and subsequent work is that lot sizing and aggregate capacity planning have to be performed at two different levels…

Several monolithic models have been devised to solve aggregate planning and lot sizing and scheduling problems jointly, and a particularly interesting series of technical improvements to an initially rich model can be found in the works performed by MANNE [MA], DZIELINSKI and GOMORY [DG], KLEINDÖRFER and NEWSON [KN], LASDON and TERJUNG [LT] and NEWSON [NE1],[NE2].

All these works are based on MANNE's result on "dominant" schedules (independently found by WAGNER and WHITIN [WW]), which makes it possible to reformulate, with a good degree of approximation, an intrinsically nonlinear problem (setup times and costs disrupt the linearity of the objective function and of the constraints) as a linear program. Since this reformulation involves increasing dramatically the number of variables (these now represent all the alternative sequences), all the works reviewed address one of the difficulties that hierarchical systems claim to tackle, namely dimensionality; the approaches, however, differ considerably: MANNE addresses the problem from a managerial point of view and proposes a product aggregation consistent with the type of physical system he considers (assembly) and with his model; DZIELINSKI and GOMORY address it from a mathematical point of view and use the type of column generation technique suggested by DANTZIG and WOLFE.

This technique is described in detail in DIRICKX and JENNERGREN [DJ] together with applications and other decomposition methods. The authors' interest is in what they term "multilevel systems analysis" and define as an approach to solve a problem by decomposing it into subproblems and coordinating the solution of these subproblems by an interactive exchange of information. The relation with MESAROVIC's work on multilevel systems is clear; however, there is a basic distinction between this type of work and that reviewed in the next section, namely that in [DJ], the decomposition is only viewed as a way to make the solution easier, and the problem is still solved by one single Decision-Maker. Elsewhere, the decomposition is a means to define the respective problems of several DMs in a hierarchy.

Plus, by using a column generation technique, DZIELINSKI and GOMORY confront the problem of infeasible methods. To tackle this difficulty, LASDON and TERJUNG [LT] address the problem directly and propose an efficient algorithm. In the implementation they describe, these authors consider a production system in which controlling the final stage accounts for most of the management task. However, they feel the need to modify the objective and the constraints in order to take into account the effect of the final stage decisions on the upstream stages.

This essentially pragmatic approach clearly illustrates a common feature of all the work reviewed in this section, namely that aggregate decision variables are plugged into the lot sizing model as a straightforward enhancement. This makes sense from the computational point of view, since linear terms in the objective do not increase dramatically the complexity of the problem.

It is interesting to note that LASDON [WS] proposes a multilevel decomposition of the problem formulated in DZIELINSKI and GOMORY [DG]. This decomposition is based on the same results as in the continuous case: Lagrangean relaxation and duality. However, the author's conclusion is that "there is no theoretical guarantee that discrete problems of the type considered can be solved using duality".

Except for GELDERS and KLEINDÖRFER [GK1] [GK2], none of these works claims any contribution to coordination between decision levels. Nevertheless, the models adopted are intrinsically similar in all these works and a product aggregation as proposed in MANNE [MA] prefigures HAX and MEAL's. Moreover, the issue of decomposition technique versus management hierarchies needed to be pointed out. It is further addressed in next section.

## 2.II RESOURCE ALLOCATION IN DECENTRALIZED ORGANIZATIONS

The work reviewed in this section was initiated in the field of economics as an attempt to find a coordination method by pricing in a multi-sector economy in which each sector strives to maximize its own profit by using communal scarce inputs and subject to a set of constraints relating the output levels to the input levels, whereas the final goal should be to maximize the profit of the economy as a whole.

The initial idea, based on an observation of the supply and demand law, was that a "central unit" could determine output prices in order that the sectors' drive to individual profit result in an overall optimum. In that context, DANTZIG and WOLFE decomposition seemed to provide a suitable numerical tool in the case of linear costs and constraints. Unfortunately, (as pointed out in MESAROVIC et al. [MC]) decentralization cannot be achieved by prices alone. Except under very restrictive assumptions, the central unit has to transmit some other kind of information to the sectors of the economy (or the divisions of a multi divisional firm) to make them determine their optimal resource utilization that would also optimize the overall objective.

In [BF] BAUMOL and FABIAN describe the setting of the problem both in the qualitative terms used here and in terms of the structure of the linear program used to model it. They explain then in detail the economic interpretation of the decomposition method. The divisional optimization problems are solved iteratively with an "executive" program which determines the best convex combination of all the plans submitted by the divisions, that is, the one that will maximize the corporation's benefit subject to the corporate constraints.

This program subsequently modifies the output prices for the divisions to re-determine their optimal plans. These prices subtract

from the actual corporation's profit the value to rest of the firm of the scarce input the divisions used, (i.e. the scalar product of the dual price vectors associated with the corporate constraints by these constraint coefficients). The authors acknowledge at the end of this description that when the iterative process has yielded the equilibrium prices, the central unit still has to transmit the convex coefficients of the last plans submitted by the divisions in their optimal solution. That is, it has to impose the divisional plans.

The same conclusion is reached by RUEFLI who nevertheless proposes an interesting model in [RU]: this model can be described as a three-level tree in which the root represents a central unit that splits a resource (or assigns goal levels, in an alternative interpretation of the model) to the first-level nodes symbolizing management units that in turn split their share of resource among a number of operating units (second-level nodes).

The contribution of this work to the field of decentralization through pricing lies in the fact that prices are generated by the management units (i.e. intermediate units) and not by the cental unit. The objective of these management units is to determine the activity levels of their subordinates in order to minimize the deviation between the amounts of resources required by the lower level and allocated by the upper level.

The technique of introducing in the objective function what could otherwise be considered a constraint (namely that the total resource used by the management units be equal to the amount allocated by the central units) is referred to as goal programming.

The dual prices associated with the resource constraints measure the potential improvement that a relaxation of these constraints would allow. Hence the objective of the central unit consists of maximizing the sum of the amounts of resource allocated to the different

management units weighted by their shadow prices and subject to a volume constraint on the total resource. Conversely, the operating units' objective is to "shift" their resource requirement vector to its cheaper components, subject to technological requirements.

The algorithm outlined is a price-adjusting mechanism based on an iterative solution of the three models, the management units fixing the shadow prices as a result of their computations and the central and operating units determining respectively the share of resources and their needs. This process will converge in a finite number of steps (possibly in a very inefficient way) but the set of prices reached will not be sufficient for the management units to determine their optimal share of resources.

Since this shortcoming restricts the utility of this type of model for decentralization, some work has been aimed at determining what information should be delegated along with the prices in order to achieve coherent decentralization and still leave enough autonomy to the lower-level units. CHARNES, CLOWER and KORTANEK [CC] propose to delegate what are called preemptive goals, namely either additional constraints in the divisional problems that relate the division activity level to an objective determined by the central unit, or an additional term in the minimand of the divisional problem, that serves the same purpose, in a goal programming approach.

The conditions that these goals have to satisfy in order that the solutions to the modified divisional problems, taken together, form an optimal solution to the overall problem are derived. Moreover, it is proved that the method is robust in the sense that small errors in the goals will yield a profit that is only slightly sub-optimal.

KYDLAND [KY] determines a class of situations in which the divisions will achieve the global optimum while striving to satisfy their individual problems if the central unit provides them with the

equilibrium prices and with the order in which they are to solve their problem (and thus deplete the amount of resource available to the following divisions). Moreover, for situations where this hierarchical ordering has to be supplemented by preemptive goals, the author provides a rule to determine the minimal number of goals required to achieve coherent decentralization.

This paper seems to indicate the climax of the research effort intended to achieve decentralization in resource allocation systems through use of the DANTZIG WOLFE decomposition.

KORNAI and LIPTAK [KL] adopt a different approach to solve the kind of resource-allocation problem that arises in the Hungarian national planning. It consists of determining the different sectors' activities while taking their interactions into account. The model initially adopted maximizes a linear function of the sectors' activity levels subject to linear constraints on these same variables. The constraints arise from the fact that the sectors share a number of common resources, including the products they supply.

Since solving the linear program that represents this "Overall Central Information" problem is computation-intensive, the authors reformulate it by considering the subproblems each sector would have to solve if its resource share were fixed. The conditions for the two formulations to be equivalent are investigated but since the problem of determining the optimal resource share (also called a central program) is difficult, it is showed to be equivalent to a strategic game for which a solution can be found.

In this game, the players are the central unit, which submits central programs and the sectors team, which return the set of shadow prices that minimize the "cost" of the plan. The objective of the central unit consists of maximizing this cost and thus an iterative procedure of "fictitious play" is devised to determine the optimal plan.

The central unit proposes a guessed initial resource share (in the case of the Hungarian economy, this initial program is generated by traditional methods) and, subsequently, each "player" proposes a solution (program or prices) that is a convex combination of his previous proposal and of the optimal response to the last adversary's proposal. (The weight on the first term increases with the iteration index.) The essential property in this scheme is that the components of the shadow-price vector can be determined independently by the sectors.

At each step, a lower and an upper bound of the optimal cost can be determined and thus the process can be stopped at an arbitrary degree of sub-optimality. When that point is reached, the sectors are able to determine their activity levels by solving the dual of the last problem they have solved to compute their components of the shadow-price vector. An application of this model to the Hungarian economy is presented subsequently.

As can be seen from the description of the iterative exchange of information yielding an equilibrium between central unit and infimal units, this planning system matches the definition of a two-level hierarchical system proposed in MESAROVIC et al [MC]. However, the methods described in current section have had a limited impact on management techniques, essentially because the efficiency of DANTZIG and WOLFE's decomposition is counterbalanced by the fact that it does not allow for a real decentralization. The mainstream in the hierarchical management literature is actually based on the concept of multilayer hierarchies and presented in the following section.

## 2.III HIERARCHICAL PRODUCTION PLANNING

All the work reviewed in this section is related to the class of multilayer hierarchical systems and more precisely to the type introduced as "multi-horizon" in part one. This type of system is characterized by the fact that the controller is split in several layers, so that the higher ones are concerned by the slower aspects of the system and have a longer optimization horizon. However, for stylistic reasons, the term "level" will often be preferred to the term "layer" in the remainder.

It seems that only two papers include a survey of the work done in hierarchical production planning, namely GELDERS and VAN WASSENHOVE [GW] and DEMPSTER, FISHER et al. [DF]. Both groups of authors consider the work initiated by HAX and MEAL [HM] and developed at M.I.T. in the seventies as the most substantial contribution in the area of hiearchical management. Therefore, the chronological evolution of this work is described in the first part of this section.

### 2.III.1 HAX and MEAL's and Derived Models

Although designed for a particular implementation, the model described in HAX and MEAL [HM] -along with the analysis that warrants it- was considered sufficiently general for its structure to be retained in the work derived subsequently. Several characteristics make this structure representative of a hierarchical management:

- first, four decision levels are considered, each of them related to a different horizon and articulated in such a way that the longer range decisions provide the constraints for shorter range decision-making.

- moreover, since the system is designed for a multiple plant firm, the highest level of the management model determines what products should be supplied by the different plants and thus decomposes the problem into decoupled sub-problems. The scope of the three lower decision levels is then narrower (a single plant) than that of the highest.

As was argued previously, this is a highly desirable characteristic for a management system. It could be objected, however, that the highest level appears to be different from the lower ones insofar that the decisions are to be made only once, and not repeatedly (which breaks the recurrence of the model) and are also more case-dependent. In that sense, they are closer to design-type decisions than to control. This observation also holds for HAX [HA1], which presents an implementation of a "hierarchical" system in an aluminum company.

In the system described, a mathematical program is used in a "static" way at the strategic level to help make such decisions as whether or not to build a new plant or how much to produce in the existing plants, and the results obtained then constrain the tactical level model designed to assign the orders to the different casting machines.

The remainder of this section will therefore focus on the three lower levels of HAX and MEAL's model:

- based on an analysis of the production process, three levels of aggregation are considered for the products:
  - . product families group items sharing the same major setup.
  - . product types group families that have similar seasonal patterns and inventory cost per hour of production;
- each production-planning level is assigned a model consistent with the horizon decomposition and the product aggregation scheme:
  - . At the higher level -aggregate production planning- a linear model is proposed, to determine the production level of the different product types over a 15-month horizon. The only costs considered are incurred for production and inventory holding. Setup costs cannot be taken into account within the model because they would be incurred each time there is a production of a family in the period considered; since types group several families, there is no information at the aggregate planning level concerning the number of setups incurred.

In the top-down constrained approach, the assumption is that higher-level decisions have a bigger impact on the objectives. In the implementation considered, the analysis showed that the major issue was to deal economically with seasonal demands. Thus the higher level model is intended to determine the optimal trade-off between inventory holding and overtime work (i.e. production cost), whereas setup costs, of secondary importance, are not considered. Experimental results show that the level of performance of the system decreases when setup costs increase.

The aggregate plan is updated every month over a rolling horizon (in a "repetitive open-loop optimization" process, according to the terminology introduced in FINDEISEN [FB2]) in order for the evolution in forecasts to be taken into account.

- Setup costs are first considered in the second level decisions through a heuristic family disaggregation over the first period of the aggregate production plan, based on economic order quantity, safety stock and overstock computation techniques. The capacity allocated to the product type is split between the families for which the inventory level falls under the safety threshold during the period considered. For each of these families, the production volume is chosen as close as possible to the EOQ, provided that it does not lead to an overstock at the end of the period.

- The third decision level consists of a heuristic item disaggregation based on equalizing of run-out times (EROT). As the setup costs are incurred whenever any of the items in a family has to be produced, it seems desirable that all the items in a family run out at the same time. KARMARKAR [KA] gives a proof of the optimality of this decomposition technique. As in the family disaggregation model, the capacity allocated to a product family has to be split among the different items.

It appears then that consistency between decisions made at the different levels is ensured by the constraints that a given level's decisions impose on the next lower level. Still these constraints sometimes yield an empty feasible set at a lower level. In other terms, disaggregation of an aggregate schedule is not always feasible. This was the first issue addressed in further research.

GABBAY [GA] considers an aggregate plan for which a feasible detailed plan (i.e. a plan meeting detailed demands without backordering) exists. He shows that under certain conditions, disaggregation performed over the first period only will lead to a state for which there will be no feasible disaggregation in subsequent periods. A qualitative interpretation of this phenomenon is that whereas in the single product problem, capacity and inventory are equivalent for meeting demands, this result does not hold if the "product" is an aggregate, since the inventory of one item cannot be used to meet the demands of a different one.

Therefore, the aggregate plan must be drawn from "net" demands, i.e. the aggregation of detailed demands net of the initial inventories. When disaggregation is performed over one period at a time only, it must be in such a way that this property can be retained for the remaining horizon. GABBAY derives some necessary and sufficient conditions for consistent disaggregation. Unfortunately, the detailed demands must be known for all the planning horizon if these conditions are to be satisfied, whereas the main advantage claimed for the hierarchical approach is that it reduces the detailed data requirements. He thus refines the result by proving that the time intervals for which the cumulative production capacity is sufficient to satisfy the cumulative demand can be treated separately and so the detailed demands are required "only" over the consistency horizon, that is until the first period in which the aggregate inventory is zero.

These results are then extended: first, the single-echelon, single-product, capacitated problem is showed to be solvable by a siple backward dynamic program even under a quite general cost structure. This model can then be used as the aggregate level in a multiproduct problem. If the detailed demand is known over the consistency horizons, the disaggregation scheme studied previously will still yield the optimal plan. In the case of multiechelon systems, the same results can be retained at the expense of a very restrictive cost-structure.

GOLOVIN [GO] also acknowledges the issue of disaggregation consistency and proposes to solve both aggregate and detailed production planning problems by means of a single mixed integer program featuring two levels of product aggregation, two time scales and setup costs considered for the "short-term" production. Hence disaggregation is "automatically" achieved and setup costs are still taken into account. The computational gain is reaped from the use of a shorter horizon for detailed production.

This model is complicated by the need to penalize the difference between expected "aggregate" production and the corresponding detailed production and it seems that this approach was neither implemented nor further developed. GOLOVIN then explores the problems arising when the periods considered at the higher level correspond to several lower-level periods. In that case, the detailed plan obtained by disaggregation of the aggregate plan is only feasible on average.

BITRAN and HAX [BX2] propose a computational improvement to HAX and MEAL's model in that they reformulate the family and item disaggregation problems as knapsack problems, for which they previously provided an efficient solution algorithm [BX1]. It is shown that whenever setup costs are low, the results obtained by using the resulting system are very close to optimal and quite unsensitive to forecast errors.

WINTERS [WI] investigated three methods for coupling "inventory control" (reorder point / reorder quantity decisions) and "production smoothing" (aggregate planning): constrain the detailed inventories, constrain the production levels, or adjust the reorder points while keeping the reorder quantities at their infinite-horizon, unconstrained values. This last method, although highly heuristic, was shown experimentally to give good results and require little computation.

HAAS, HAX and WELSCH [HH] compare the results of four heuristic disaggregation methods: HAX and MEAL's, WINTER's, BITRAN and HAX's knapsack method and the equalizing of run-out times (EROT) method. Results of the Wilcoxon test show that HAX and MEAL's heuristic performs very well under a wide range of assumptions and outperforms the other methods.

This result motivated the search for some improvements to the knapsack-based system. BITRAN, HAAS and HAX [BS1] prove that the EROT method is an optimal disaggregation scheme to minimize the cost of initial inventory. Insight gained from this result as well as from previous work enabled improvement of the knapsack-based system.

First, at the family disaggregation level, instead of minimizing the number of setups expected for the <u>whole</u> aggregate planning horizon (according to demand forecasts) one does it over a shorter horizon. This allows the system to be responsive to seasonal variations in demands. The second improvement is a "one step look ahead" procedure to ensure that disaggregation will be feasible over two periods instead of one. The last one consists of modulating the families' production volumes in order to keep them close to the Economic Order Quantities, especially in case of high setup costs.

The enhanced system was then shown to outperform all previous ones on a set of simulation runs and to yield close-to-optimal results (within 3% of optimum) even when setup costs were relatively high.

ERSHLER, FONTAN and MERCE [EF] first synthesize all the previous results concerning the issues of feasibility of an aggregate plan and consistency of the rolling-horizon disaggregation, and derive two sets of necessary and sufficient conditions for consistency (these results are based on the mass balance equations and do not depend on the cost structure).

Thence, they consider the system proposed in BITRAN, HAAS and HAX [BS1]. In the "one step look-ahead" procedure of [BS1], the families to be scheduled during the coming period are determined in order that a disaggregation be also possible at least for the subsequent period. ERSHLER et al. propose to extend this procedure to "look ahead" at all the periods for which detailed demands are known.

Moreover, in [BS1], after the families to be scheduled are determined, a knapsack problem is solved to determine the quantities to schedule. ERSHLER et al. point out that introducing the necessary and sufficient conditions for consistency as additional constraints would break the "knapsack" structure. They therefore propose to introduce only the (necessary) conditions that retain the structure of the problem -in order to keep it efficiently solvable- and they show that the re-enhanced system performs better.

This was the most elaborate system derived directly from HAX and MEAL's. It keeps the basic features of the original system, namely the open-loop, top-down constrained approach.

### 2.III.2 Extensions of the Model

#### Introduction of Feedback

In GRAVES [GR], a different approach to the problem is adopted, that introduces feedback between the decision layers. Based on the product-aggregation scheme proposed by HAX and MEAL the problem to solve is formulated as a monolithic mixed integer program (similar in its principle to GOLOVIN's [GO]), which is then decomposed by means of a technique that is widely used in hierarchical control, namely Lagrangean relaxation.

This decomposition yields a linear program on one hand, that happens to be an aggregate planning model, and a set of uncapacitated lot-sizing problems for each product-type on the other hand. Interaction between these models results from the presence of the Lagrange multipliers in the "inventory holding costs". The problem then consists of determining the values of the multipliers that yield consistency in the families' inventory levels computed in the lot-sizing models and in the aggregate planning model. This result is achieved through iteratively solving the two models and updating the multipliers.

Multistage Systems

Other enhancements of HAX and MEAL's model were devised to adapt hierarchical planning to multistage systems. CANDEA [CA] reviews the theoretical results existing in production planning of multistage systems and identifies a need for further research. Thereupon, several issues raised by the application of HAX and MEAL's approach to multi-stage fabrication and assembly systems are addressed (e.g. the need for a new concept of product aggregation taking into account the composition of assembled products).

The author proposes an algorithm to reduce the computational size of the aggregate planning problem as well as two heuristic methods to determine the economic lot-sizes at the different stages of the system. His conclusion is that extending HAX and MEAL's approach to multistage systems appears to be much more difficult than expected.

Nevertheless, BITRAN, HAAS and HAX [BS2] propose an extension of their previous work to a two-stage fabrication/assembly system and successfully compare the results of their hierarchical planning system to results obtained by use of an MRP-based system, on a test-bed built from data supplied by a pencil manufacturer.

Several difficulties pointed out by CANDEA are tackled in the hierarchical system through very pragmatic approximations. For example, the product families take a very restrictive definition (they group products that share both a common setup and bill of materials) and the aggregate mass balance equation is approximated: the composition of product types in part types is not constant but has to be derived from the volume ratios of the different products in the type, based on their demand forecasts…

Evaluation of HAX and MEAL's based models

A critical analysis of both the advantages claimed by the authors and the shortcomings of the approach can account for this outcome.

Advantages claimed in [HA2] are:

- reduction of the computational size, compared to a monolithic optimization with detailed data over the entire horizon.
- enabling of managerial interaction,
- reduction of data requirements since detailed data are needed only for the short term decisions (namely for the first period of the aggregate planning).

Positive features of the systems described are:

- capacity constraints are explicitly considered (whereas they are not in MRP systems, for example),
- the structure of the models used at different levels is consistent with the product aggregation scheme,
- since the emphasis has been set, in manufacturing, on the reduction of change-over times and costs, models based on the assumption that setup costs are relatively low now fit a reasonably wide range of production systems.

Major shortcomings of the approach are:

- the product aggregation proposed in HAX and MEAL [HM] fits a given type of production systems and the models suggested for each management level are based on a particular cost structure. That means that, although the resulting system can be retained for a fairly wide range of applications, other aggregation schemes and hierarchical models featuring a similar consistency could have been investigated where the initial ones fell short. For the implementation in the aluminum industry, the product aggregation was empirical.
- the detailed data requirements are reduced only if backordering is allowed.
- no randomness is taken into account and forecast errors have to be absorbed by means of safety stocks.
- even though in [ME], MEAL still emphasizes the need for delegating the decisions, no "spatial" decomposition of a system is proposed.

It is interesting to notice that the hierarchical models proposed tend to lack generality. As a consequence, the implementations have been designed to be consistent only with the qualitative ideas on which the theory is based. However, they make use of models that are totally different from the ones worked out in theoretical settings, because they need to represent the specific issues raised by the structure of the systems they are designed for.

### 2.III.3 Implementations

A good illustration of this statement is given in MACKULAK, MOODIE and WILLIAMS [MM] in which the implementation described is intended for the steel industry. After two management systems were simulated and provided disappointing results (one based on a production-to-order concept and real time control, the other producing to inventory with a fixed product-mix), a hierarchical model is proposed that combines their advantages.

The highest level is forecasting and requires a specific product aggregation. The authors point out the difference with assembly industries in which the number of final products is much lower than the number of components and thus forecasting can be performed on the final products. In the steel industry, the number of end-products is very large and forecasts are accurate only for groups of these products.

The next lower level is master scheduling which consists of determining the production level for the different product families (steel grades) defined for forecasting purposes. A goal programming technique is used.

The model proposed combines an inflexible capacity constraint and three goal constraints. One aims at setting the production as close as possible to the actual requirements (forecast and backorders net of inventory). The second tends to minimize the volume of unassigned product. The last one smoothes the variations of the weekly production levels. The lowest level is a heuristic scheduling-to-slabs model that assigns the heats planned in the master schedule to slab orders.

In this implementation, the product aggregation is determined by the forecasting constraints which are characteristic of the steel industry. The planning model is chosen to combine hard and soft constraints, which is another very desirable feature in this context.

M.R.P.(Material Requirements Planning or Manufacturing Resources Planning) systems have undoubtedly gained some recognition from practitionners in particular industries. However, as explained in MAXWELL et al. [MU], such systems just "look at the effect of a master schedule on the detailed plan rather than developing a plan that lies within the bounds of capacity". This means that an MRP system must be supplemented with an aggregate production planning sofware to generate the master schedule, and also that MRP systems do not integrate capacity constraints.

BAKER and COLLINS [BC] point out that a prerequisite for any management system is an information system, that is, a data-base. Therefore, more benefit is reaped from using a sound database and a poor algorithm (as in MRP) than from running sound algorithms on erroneous data. However, combining the advantages of both approaches would be better.

ANDERSSON, AXSÄTER and JONSSON [AA] consider an implementation in an assembly industry. They choose to adapt an MRP system in order that it satisfies some capacity constraints. A tailor-made aggregate planning model is described, which takes into account the multi-stage structure of the production considered. Two disaggregation procedures are proposed to fit the schedule provided by the MRP system to the aggregate plan. One consists of modifying the order release times and the other consists of modifying the order quantities. Both procedures are tested by simulation and appear to reduce the cost of overtime labor, which is the evaluation criterion. Namely, in a classical MRP system, whenever the production time required by the master schedule exceeds the regular work time, it is resorted to overtime.

MAXWELL et al. [MU] review the existing production control systems and their respective weaknesses and list five necessary improvements, namely the consideration of multiple-stage systems, load-dependent lead-times, capacity limitations, uncertainty of demand and supply, and setup time and cost. The framework they propose is a hierarchical set of three models: master production planning, planning for uncertainty, and resource allocation. They illustrate the use of this model for a stamping plant in the US automotive industry.

LASSERRE, MARTIN and ROUBELLAT [LM] address the production planning of a photomultiplier plant, in which the machines are not specialized. The solution they adopt is a hierarchical control: the medium term planning level determines the number of operations of each type to perform weekly over a given horizon and for each product type; the short term scheduling level allocates to the in-process inventory the operations planned for the first week of the medium term horizon.

This particular formulation of the planning-scheduling problem arises because the standard mass balance equations allow a given product to undergo several operations (possibly the whole process) during a single period, which is physically impossible for the products considered. Therefore, an additional set of constraints is introduced to limit the number of operations per period at each production phase. A heuristic resolution is proposed for the resulting scheduling problem.

The planning constraints being linear, the objective is sought as a convex (piecewise linear) function. The procedure used to solve the resulting program is based on DANTZIG-WOLFE decomposition and on the efficient algorithm proposed by LASSERRE for linear programs with a special structure. Several decomposition schemes are investigated.

Closer to the theory described in previous sections, PENDROCK [PE] applies HAX's design technique -described in [HA2] - to the case of a production and distribution firm, and OLSON [OL] proposes a hierarchical three level system for a specific two-stage/two-product enterprise, based on simple mathematical models and "hand tuning" of their results.

GELDERS and VAN WASSENHOVE [GW] describe qualitatively the hierarchical approach and acknowledge the theoretical work performed to address the issues of consistency, infeasibility and suboptimality. However, in the implementations reported, the mathematical models are very specific heuristics designed primarily to provide some numerical basis for the decision process. A "crossfunctional managerial committee" is actually in charge of coordinating the higher and lower decisions in order to ensure consistency and avoid infeasibilities by inserting slack time or safety stocks in a "thoroughly controlled" way. The issue of suboptimality is not considered to be of primary importance in these implementations.

## 2.III.4 Software Systems

HAX and GOLOVIN [HG] describe the implementation of the hierarchical planning concepts in a "Computer based Operations Management System" (COMS). This system accomodates the three levels of product aggregation introduced in HAX and MEAL [HM]. However, the user is not limited to the algorithms described in [HM], [BX2] and [BS1]; he is given the choice of the procedure (optimizing or heuristic) to use for each of the decision levels: aggregate planning, type to family disaggregation, and family to item disaggregation. The management system built by COMS is thus customized to meet the user's needs, provided that he can determine what algorithms are best suited to his case. It seems that COMS has mainly been used for research purposes.

Another hierarchical scheduling system, PATRIARCH, is currently developed at Carnegie Mellon University. MORTON and SMUNT [MS] and LAWRENCE and MORTON [LO] describe it as a decision support system requiring a variable degree of human intervention and expertise, depending on the type of decision considered: strategic decisions are mostly manual, scheduling is entirely automated. The system combines accurate suboptimization algorithms with simulation capabilities and rules of thumb. Two issues considered of significant importance (e.g. in [MU],[BC]) are addressed: the use of "shadow costs" for evaluation of alternative solutions and integration of "soft" constraints, and the variability of lead-times due to the load of the system.

## 2.III.5 Suboptimality of the Hierarchical Approach

This issue has been addressed in DEMPSTER, FISHER et al. [DF]. A framework is proposed to compare the performance of a hierarchical control algorithm with that of a stochastic model. Namely, one of the main reasons to implement a hierarchical control is that it allows one to make long-term decisions based on aggregate forecasts when the detailed data are not known beyond a short horizon. Therefore the performance of a hierarchical system cannot be equitably compared to the performance of a deterministic model in which all the future data are known with certainty.

The authors then illustrate their evaluation method on a simplified version of the design-and-scheduling system proposed in ARMSTRONG and HAX [AH]. The job shop is reduced to a set of parallel identical machines and the higher level decision consists of determining the optimal number of machines, **m**. The lower level is concerned with the problem of scheduling **n** jobs on these **m** machines in order to minimize the completion time. There is a cost associated with the purchase of a machine and a cost propportional to the completion time. It is also assumed that the job processing times become known only after the number of machines has been chosen. The data on which the higher level decision is based is a forecast of the sum of the **n** processing times.

The performance of this hierarchical decision scheme is compared to that of a stochastic model in which the two costs involved are included in a single model and the vector of processing times is supposed to be random with known mean value. It is proved that when the number of jobs tends to infinity, the performance of the two systems become equal.

This interesting evaluation approach is claimed by the authors to apply to any of the four types of hierarchical systems they review, that is "aggregate/detailed scheduling", "job shop design and scheduling", "distribution system design and scheduling", and "vehicle routing and scheduling".

However, the analytical results presented in the application depend strongly on the criteria and models chosen. As the authors point out, hierarchical systems are preferred to monolithic systems (especially stochastic) for computational reasons. This means that one cannot expect to evaluate a hierarchical system by comparing its solution to that of a multistage stochastic programming model of the problem. Instead, lower bounding techniques should be used. Therefore, the numerical results will necessarily depend on the nature of the models and no general statement can be made concerning the quality of hierarchical systems. Besides, it seems that this evaluation method has not been applied to other models.

## 2.III.6 Aggregation - Disaggregation

Another issue pointed out as essential by GELDERS and VAN WASSENHOVE is that of infeasibility or, in other words, of disaggregation. KRAJEWSKI and RITZMAN [KR] is supposed to be a survey of the research work on the issue: according to its abstract, this paper is aimed at drawing attention "to the lack of an interfacing mechanism [which] diminishes the utility of solution procedures for aggregate planning, inventory control and scheduling".

Actually, the definitions subsequently adopted for the concept of disaggregation suit all planning models. Hence a wide range of production planning models are surveyed and the unifying disaggregation model initially suggested is, in all respects, a planning model.

The aggregation or disaggregation schemes considered here are of three sorts: over time, products, and machines. While an important part of the work performed by HAX et al. was aimed at solving the problems raised by time and product disaggregation, apparently no work has addressed the issue of triple aggregation and disaggregation.

The first work in disaggregation concerns only a product disaggregation. ZOLLER [ZO] considers a two level economic model in which the aggregate production is determined by minimizing a cost function. The product-mix and sales price are determined at the lower level in order to maximize the profit, assuming that the demand volume depends on the sales price and that the function binding the two variables is known. The author provides an algorithm to solve this lower level problem and, like GELDERS and KLEINDÖRFER [GK1],[GK2]. He chooses an iterative process in order to reach the optimal solution, although he acknowledges the alternative solution of a sequential top-down decision process.

In the field of project-oriented production (shipyard), HACKMAN and LEACHMAN consider the case when several concurrent projects compete for scarce resources. These resources must be allocated to the project managers who in turn schedule the operations required to complete their project within these capacity constraints.

In order to reduce the dimension of the problem, the operations in a project that require the same resource-mix are aggregated. The issue investigated is that of reformulating the operations precedence constraints at the aggregate level; a continuous time representation is adopted for the production functions, that indicate the cumulative resource consumption of aggregate operations. Given the early and late start-times for each detailed operation, the production function must be inside a "window" defined by two extreme scenari: all operations starting at their early start-time versus all operations starting at their late start-time. For two consecutive aggregate operations, the precedence constraint is approximated by a condition on the production functions with respect to their time windows.

Besides these very specific disaggregation models and the work summarized in ERSHLER, FONTAN and MERCE [EF], one can find the issue of double aggregation/disaggregation over parts and machines addressed in AXSÄTER [AX]. Solving an aggregate optimizing problem, in terms of product-families and machines subsytems yields an "aggregate" control that it may not be possible to disaggregate.

The author derives a necessary and sufficient condition on the aggregation matrixes (whose $(i,j)^{th}$ element is 1 if product (resp. machine) j is in product- (resp machine-) group i, and 0 otherwise) for the aggregation to be perfect, i.e. in order that it be possible to disaggregate any aggregate plan. Moreover, since perfect aggregation is not always attainable, AXSÄTER provides a method to build the aggregate model, assuming that the products and machines families are given, and that the control can be modelled as a random vector of known first and second order statistics.

Whatever the difficulties encountered in hierarchical production planning models reviewed so far -and the gap between theory and practice shows there are difficulties-, they still are not comparable with those that arise when the lower decision level is that of detailed scheduling. The next section reviews the few models developed to coordinate detailed scheduling and aggregate planning.

## 2.IV INTERACTION BETWEEN MODELS FOR AGGREGATE- AND DETAILED- SCHEDULING

GREEN [GN] is probably one of the first papers that address directly the issue of coordination of two separate models, of which one is related to detailed scheduling. The author's approach is based on a double observation :

- on the one hand, planning of the workforce and production levels through use of HOLT, MODIGLIANI, MUTH and SIMON's linear decision rule (HMMS) is straightforward but requires that assumptions be made concerning the parameters of the rule. And this leads to a substantial sub-optimality;

- on the other hand, a detailed simulation of the system for a given control will yield accurate values for the HMMS cost function and could be used to find a good, if not optimal, solution. However, this would imply an excessively heavy computational burden in the absence of a guiding procedure to improve the solution.

The procedures explored are different iterative "couplings" of the two models, in which some initial guesses about parameters (such as the productivity factor) are made to apply HMMS rule and then a simulation is run with the workforce and production levels determined, which yields the actual value of the "coupling" parameter. The process is iterated until consistency is reached.

It is interesting to point out that GREEN does not consider this hierarchical-type decision process as a managerial necessity but only as a solution to the computational problem. In his opinion, one would ideally be able (in a not too distant future...) to run the simulation instantaneously and at a very low cost, which would make it possible to determine the optimal control by trial and error. In the light of current simulation users' opinion, it seems that this future is slightly more distant than GREEN thought it was.

On the contrary, SHWIMER [SW] clearly aknowledges the theoretical foundation for hierarchical decision-making, as well as the intractability of a monolithic job-shop scheduling model formulated as a mixed integer program. He therefore proposes to split the planning-scheduling problem in aggregate capacity planning and detailed scheduling. The first problem is formulated as a mixed integer program that can readily be approximated by a linear program but the second one becomes intractable whenever solved by means of the optimization model initially suggested.

Hence the method investigated consists of iteratively solving the aggregate planning problem and running a simulation of the system where scheduling is performed by means of standard priority rules. A number of procedures for passing the information between the two models are proposed. Hence the aggregate decisions can be made consistently with the lower level constraints. That is, one is assured that the feasible decision-set they yield for the lower level contains a control (namely the decision rule previously simulated) compatible with the constraints that will appear at the lower level only.

ARMSTRONG and HAX [AH] use the same type of approach in a model devised to plan the workforce levels -by skills- as well as the replacement of conventional machine tools by numerically controlled machines in a naval tender job shop. The mixed integer program modelling the higher level decisions and a simulation of the detailed scheduling are run iteratively until a satisfactory machine occupation is achieved. This iteration ensures that these design decisions will yield an efficient production system. The coupling between models is assumed to be principally based on "managerial interaction".

More recently, IMBERT [IM] proposed to replace the use of dispatching rules suggested by SHWIMER for the lower level model by an analytical approach to scheduling ("constraint analysis"). The machine loads are determined by running the simulation according to this scheduling method. They are then fed back to the linear program that

determines the aggregate production plan. The two models are run iteratively until enough manpower is allocated for the schedule both to be feasible and to require the same amount of manpower as allocated. The applicability of the resulting management system is then tested on data representative of a small job-shop.

Since this seems to be the last work on the subject, it appears that the conclusions drawn at the end of Section 2.III hold for the systems including detailed scheduling in its traditional combinatorial formulation. Also, a unifying framework that would be of practical interest for the development of management systems in cluding detailed scheduling in a wide range of settings is still lacking. However, this is not fatal: when the control policies are sought in a different, more restricted set, detailed scheduling can be handled in a hierarchical system. Such a special case is described in the following section.

## 2.v HIERARCHICAL SYSTEMS FOR FLEXIBLE MANUFACTURING

The advent of so-called Flexible Manufacturing Systems has generated a fresh approach to the planning and control theory based on the fact that these new systems require a higher degree of automation of the decision process. In other words, it may be possible, in traditional manufacturing systems, to rely on humans' ability to make decisions when unexpected events occur. However, this is not acceptable in FMSs, especially if they are supposed to run unmanned for one shift a day.

Consequently, two attitudes are adopted by manufacturers and control theorists. On the one hand, the flexible systems implemented tend to perform a restricted set of operations, which considerably simplifies the management but results in a poor use of flexibility (see [JA]). On the other hand, new management structures are being progressively developed in order to match these new requirements.

O'GRADY and MENON [OM] define the scheduling and control function as one of translating broad goals for a whole firm into specific instructions to workers or automated resources, and explain why this function is more critical in automated manufacturing systems. They also describe three very similar hierarchical scheduling frameworks: the AMRF's (Automated Manufacturing Research Facility of the National Bureau of Standards), the CAM-I's (Computer Aided Manufacturing International Inc.) and their own one. They point out that none of them has been entirely implemented yet.

VILLA et al. [VO] also propose a hierarchical framework to model and control FMSs. They first define an FMS as a structure composed of a physical system, an information system and a decision-and-control system. The tasks performed by the latter can be divided into periodic planning and event-driven control, which can involve re-planning in response to rare large-scale events or just some noise-control in response to frequent small-scale events.

Since the complexity of these tasks makes a global approach impractical, a decomposition procedure has to be found. The authors leave aside the option of using a mathematical technique to achieve this decomposition and choose to investigate a decomposition based on physical insight. In fact, they assume that a "natural" tree-like structure of the physical system exists, in which each subsystem -starting with the FMS itself- can be viewed as a set of lower-level subsystems. They analyze the management system one could build by assigning a decision-maker to each node of the tree.

Prior to proposing any quantitative model, they infer some necessary conditions for the control structure to operate, namely **(1)** that each decision-maker must be assigned an objective function and a horizon consistently with the global system's objectives, **(2)** that information about the state variables must be available at each level and **(3)** that the conjunction of constraints due to higher-level decisions and constraints arising locally must never yield an empty feasible decision-set for any decision-maker. Consequently, the decision and information systems will consist respectively of a top-down constraint flow and a bottom-up information flow.

Then, VILLA et al. identify the problem to be solved by each decision maker (DM) and assert that it is essentially the same for all of them. This is fortunately consistent with the fact that the decision structure fits the physical structure and thus can display an arbitrary number of levels. Each DM determines the size and sequence of the batches to load in the subsystem under control, in order to meet the requirements set by the demand rate, and assigns each element of this subsystem both a flow rate target and a service rate target.

This definition of the problem in turn provides insight about the control system. Since it happens that the higher the level, the larger the decision scope and the longer the settling time of the subsystem under control, it follows that the horizon must also be longer for higher levels if the system is supposed to operate in steady-state.

Moreover, the DMs' objectives will include optimizing some economic criterion like the number of "tooling" changes, but will principally consist of minimizing the settling time. It is finally pointed out that each lower level DM will gain the extra degree of freedom required for this optimization by adopting a higher control frequency. This framework thus fits the models described in FINDEISEN et al. [FB2].

According to the terminology used in management science, the two problems to be solved in the hierarchical framework proposed belong respectively to the classes of lot-sizing-and-scheduling problems and routing problems. The scheduling problem is further addressed in VILLA and ROSSETTO [VR], whereas the routing problem is addressed in VILLA, CANUTO and ROSSETTO [VC]. The FMS considered in both references can be modelled as a set of cells physically connected by a material-handling system (MHS) and decoupled -from a managerial point of view- by buffers. Each of the cells is a set of workstations and buffers connected by an internal MHS.

Both the scheduling and the routing are split into a deterministic open-loop "planning" function, and a "control" function triggered by unexpected events and aimed at minimizing their effect. The formulation of scheduling problem is the same at the FMS and FMC levels: given a production objective and the state of the system (capacity of the cells -resp. workstations- and intercell -resp. intracell- buffer levels), determine next level's production target over a shorter period, so as to maximize the throughput and minimize the WIP (work in process). At the workstation level, the objective is to sequence the jobs in order to minimize the queues clearing time. The in-the-cell routing problem is formulated in [VC], and its solution outlined.

The conclusion could be that it is theoretically possible to conceive a hierarchical control structure for an FMS based on the tools developed in Large Scale Systems theory. However, the tools currently used in management are substantially different and generally address only subproblems arising in production control.

The question that arises is then how to "integrate" existing tools in order to build a global control system. VILLA, MOSCA and MURARI [VM] suggest to use a framework called "integrated control structure" based on the same spatial decomposition of the physical system and frequency-band partition of the events as in [VC] but in which the decision-making units would use Artificial Intelligence tools to solve their problems.

This idea has been quite successful in the past years and there have been several attempts to use generic A.I. tools to solve scheduling problems. For example, SHAW [SH] proposes a two-level scheduling system for an FMS consisting of a network of cells connected by a local area network: the tasks to perform in order to complete the jobs released into the system are assigned through a bidding procedure to the cell that can complete them in the shortest time, just before their predecessor is completed. The operations required to complete the tasks assigned to a cell are then sequenced by a general-purpose non-linear planner: XCELL.

It seems however that there is little hope for general-purpose tools such as planners to be applicable to problems that raise the issue of dimensionality even when they are addressd through ad-hoc procedures. The inability of the renowned system ISIS to deal with the plant it was designed for corroborates this opinion (see PAPAS [PA]).

In VILLA et al. [VM], the control units are modelled as a knowledge base and an inference engine, and each "layer" of control units is related to a frequency band. Hence the horizon ratios of different control layers must be consistent with the event frequency ratios. The importance of an event is defined as the index of the higher control layer at which its effects are likely to influence the control. Thus the optimal control strategy for each decision module consists of solving an open-loop-feedback problem to update its policy each time an event occurs with an importance greater than its level index. This updating process includes ensuring consistency between the policies determined by the lower level decision modules.

Solving the planning problem each time an event of given importance occurs requires an excessive computational capacity. The authors thus make the assumption that each module's knowledge base contains a model of all events likely to affect the module, as well as the possible dynamics consequent to these events. Hence the computational problem is replaced by one of retrieving information by a form of pattern-matching. (This idea was already developed in BECHLER et al. [BO]).

The inference engine performs three tasks: (1) select the best policy according to the state of the system and the type of event; (2) coordinate the lower level actions based on the coordination rules retrieved from the knowledge-base with the control policy; and (3) feed this knowledge-base with a description of the consequences of the controls applied, in terms of the resulting dynamics of the system.

This framework combines many interesting ideas about hierarchical control but it is yet to be applied to design a control system. The hierarchical model of a work-center controller initially proposed by KIMEMIA [KI] and improved in subsequent work has definitely come much closer to the implementation phase, although built around a rather complex stochastic feedback control model.

In [KI], the manufacturing system is flexible to the extent that it can be set up to process different types of parts with a changeover time negligible in comparison with the processing times. Moreover, the machines are failure-prone, and the mean time between changes in machine state is assumed to be much longer than the processing times (which justifies a continuous model of the part flows). The parts requirements are stated as production rates to be met over a horizon that is an order of magnitude longer than the mean time between changes in machine state. Finally, the failure/repair process is supposed to be memoryless and the machine state is thus modelled as a Markov chain.

Under these assumptions, a three-level controller is devised, combining input parts flow control, routing (i.e. splitting of the parts flows along the different possible routes) and sequencing of the individual parts. The main assumption in this approach is that whenever the parts loading rate is within the capacity of the system, there is a solution to the routing and sequencing problems. Since the processing time for each operation performed on a given part type and a given machine is fixed, the capacity set in steady state is a convex (and machine-state dependent) polyhedron in the routes flow-rate space. Under additional assumptions, a polyhedral capacity-set can also be defined in the parts flow-rate space.

Consequently, KIMEMIA's work focuses on the flow-control problem. The controller is penalized for deviations of actual production from the target rates. More precisely, a time-variant vector called the buffer state (surplus state in more recent work) measures the cumulative difference between loading rate and demand for the different parts. The control policies are sought among feedback laws (i.e. as functions of the surplus state, machine state, and time) which, for each machine state, divide the surplus state space in a finite number of regions within which the control remains constant at an extremum point of the capacity set.

This means that the optimal control policy consists of loading parts at one of the maximal feasible rates in order to drive the surplus state towards a point called the "hedging point" at which the inventory accumulated allows one to hedge against future failures at minimal cost.

This hedging point as well as the optimal paths to reach it depend on the relative costs and "vulnerabilities" of the different parts, the vulnerability of a part measuring the ability of the system to recover from a deficit of this type of part subsequent to a failure. When the hedging point has been reached, the optimal control consists of keeping the surplus state invariant and thus to load parts at the demand rate until a failure makes it impossible to do so.

For a given state reached at time t, the cost-to-go is defined as the expectation of the cost over the rest of the horizon. If the optimal cost-to-go function has been determined, the optimal control can be derived by solving a linear program. Unfortunately, obtaining the exact cost-to-go requires solving a system of coupled partial differential equations, which is impossible for a problem of realistic size. One of the sub-optimal control schemes proposed consists of approximating the cost-to-go function, based on the result that if the capacity-set is a hypercube, the differential equations are decoupled (i.e. become ordinary differential equations) and can be solved separately for each part-type. Since these computations are still substantial, they are performed off-line, the "estimate-based" cost-to-go functions being stored in decision tables accessed by the on-line layer of the controller, which is in charge of solving the LP.

KIMEMIA and GERSHWIN [KG] , which summarizes the most innovative results of [KI], also presents the off-line generation of decision tables as a fourth control layer. This interpretation is consistent with the concept of adaptive control introduced in LEFKOWITZ [LE], since the decision tables have to be up-dated if the values of the machines failures parameters come to change.

Both in [KI] and [KG] the two lower control levels are not described in great detail. In the routing algorithm, the FMS is modelled as a network of queues and the objective is to minimize congestion and delays, whereas the sequencing algorithm only attempts to maintain the flow rates set by the routing. Note that the routing level is omitted in subsequent work because the additional assumption that make it possible to state the capacity constraint in terms of the parts flow-rates, namely that a given operation can be performed on a given part only by identically performing machines, basically obviates the routing. The modifications to the flow control algorithm that allow the consideration of alternative routing have been presented only recently in MAIMON and GERSHWIN [MG].

In GERSHWIN, AKELLA and CHOONG [GC], three major improvements of the hierarchical controller are presented, one for each of the levels considered: generation of the decision parameters, computation of the loading rates and sequencing of individual parts. First, the cost-to-go function being approximated by a quadratic function, the hedging point is estimated by use of an intuitive model of the optimal behavior of the system, which considerably simplifies its computation. Additionally, whereas in KIMEMIA [KI] and KIMEMIA and GERSHWIN [KG] the loading rates were determined at a constant frequency, the improved algorithm aims at keeping the surplus state trajectory on the optimal path to the hedging point. It thus suppresses the chattering observed in the previous setting when the buffer-state crossed an attractive boundary, making the control "jump" between two vertices of the capacity at each re-computation of the control. The third improvement consists of sequencing the parts so as to achieve the conditional future trajectory, i.e. the optimal trajectory that the surplus state will follow if no change occurs in the machine state; again, the computation of this trajectory is greatly simplified by the quadratic assumption.

The modified controller is tested in a simulation of a printed circuit card assembly facility, and the resulting performance of the line is successfully compared in AKELLA, GERSHWIN and CHOONG [AC] to the performance achieved by using other policies. GERSHWIN [GE1] places this whole work in the context of a new approach to management based on "a discipline that, at each level of a hierarchy, keeps material flow within capacity, even in the presence of uncertainty, by the use of feedback."

The idea underlying this work is that such concepts as capacity appear at all levels of a control hierarchy though with different meanings because different time-scales are considered. This was pinpointed in the implementation of the hierarchical controller and suggests a time-scale decomposition and a recursion in the models to use at different levels. An illustration is proposed in GERSHWIN [GE2] with the addition of a control level to determine the setup frequency.

GERSHWIN [GE3] synthesizes and extends these ideas in a novel hierarchical framework for production scheduling. Production is represented as the occurrence of different <u>events</u> (some controllable, others random) affecting the <u>resources</u> of the system and indicating the beginning or the end of <u>activities</u>. The state of resource i is represented by the time-varying vector $[\alpha_{ij}(t)]_j$ where $\alpha_{ij}(t) = 1$ if resource i is used by activity j at time t, and 0 otherwise. Every activity j has a characteristic duration $\tau_{ij}$ and frequency $u_{ij}$ on each resource i.

Two assumptions are made:

→ <u>flow conservation</u>: the frequency $u_j$ of an activity is the same for all the resources it affects.

Hence the formulation of a capacity constraint for each resource when the system is in steady-state: $\forall$ i resource, $\sum_j u_j \tau_{ij} < 1$

→ <u>frequency separation</u>: the set of all activities can be partitioned into subsets $J_1,..,J_k,..$ of activities with "very different" frequencies.

More precisely, each subset $J_k$ is assigned a characteristic frequency $f_k$ such that activity j is in $J_k$ iff $f_{k-1} << u_j << f_{k+1}$; k is then called the <u>level</u> of activity j, and a level is termed "high" if it corresponds to low frequency activities.

A quantity is said to be observed at level k, and noted with a superscript k, if the observer cannot distinguish the occurrence of events with frequency higher than $f_k$. Therefore an activity has three different aspects: it appears as a pair of discrete events (start, end) at its own level, is a constant for a lower level observer, and evolves at a continuous rate for a higher level observer. A simple relation ties the frequencies of an activity observed at two consecutive levels:

$$E_{k-1}(u_j^k) = u_j^{k-1} \qquad \textbf{(T)}$$

where $E_{k-1}$ is the conditional expectation, assuming that the state of the system remains constant for an observer at a level $m < k-1$.

For a controllable activity, **(T)** gives a guideline to translate objectives from the top level down the hierarchy to its own level. For a non-controllable activity, **(T)** indicates how to aggregate the information collected at the activity level, for higher levels. At each level **k**, the capacity constraint can be rewritten:

$$\forall i, \quad \sum_{L(j) > k} u_j^k \tau_{ij} < 1 - \sum_{L(j) \leq k} \alpha_{ij}^k$$

Two strategies are proposed to translate the objectives down the hierarchy:

- → the hedging point strategy is used to translate rates and is a generalization of the policy described in KIMEMIA [KI] and subsequent work. The idea is to define a surplus for each activity and to keep it at a given value (the hedging point), that balances the expected costs incurred for faling behind the target activity rate or for being ahead of it.
- → the staircase strategy (basically, the loading policy of [AC] and [GC]) is used to determine when to start an activity, given an objective expressed in terms of rate. The idea is to keep the cumulated number of starts close to the product of the target frequency by the elapsed time.

This framework is analyzed in GERSHWIN [GE4] in the simple case of a two-part, two-machine system, one machine being totally flexible (no setup time to switch between parts) but fallible, and the other one being totally reliable but requiring a setup to switch production.

Three unresolved issues arise from this application: (1) there is no hint about how to determine the objectives at the highest level, from which the lower-level objectives will be drawn; (2) the structure of the hierarchy (what must be decided at which level) can depend on the highest level computations if these include determining activity frequencies, and (3) there can be interrelations between strategies that are not captured by the framework: for example production and setup rates are not completely independent.

However, the numerical results reported demonstrate the good performance of the hierarchical controller and show that the design framework of [GE3] can be successfully applied.

# CONCLUSION

The objective of this chapter was to survey, in the perspective of an application to manufacturing systems, the work focusing on the concept of hierarchy, both in the field of control and the field of management science.

Two main structures of control/management systems have been investigated. Multilayer structures are characterized by a partitioning of the decision/control variables affecting a single system whereas in multilevel structures the system under control is divided into subsystems and the controller consists of several infimal units coordinated by a supremal unit.

In the control literature, multilayer models have a minute share, closely related either to time-scale decomposition or to adaptive control. They feature an approach that is very similar to that adopted in the most common hierarchical production planning models. However, some of the ideas developed in this work have not been adapted to manufacturing systems yet (see [DL]), which means they can still suggest new models for production management.

The aggregation techniques directly address one of the issues that arise both in control and management science, namely what Bellman terms "the curse of dimensionality". These techniques are essentially mathematical tools used to reduce the dimension of a control model with a minimal loss of performance and their applications to manufacturing problems are very scarce. This can be attributed to the lack of large scale models of production systems that would account for all the relevant phenomena and keep a structure amenable to aggregation techniques. The aggregate models based on the physical insight of their designer seem to be more satisfactory and obviate the use of aggregation techniques.

Multilevel models result from a mathematical decomposition of typical control models in the case when the physical system has a special structure. Unfortunately, very few real systems have this structure. Plus, problems in manufacturing are generally not perfectly structured, and when they are, it can be more efficient to use a heuristic decomposition rather than multilevel techniques (see [SD]).

It thus seems that the work on aggregation and multilevel systems is mostly likely to provide mathematical tools if some of the models appear to be relevant in the context of manufacturing. [KG] is an example of a successful transfer of model between control and management science.

Identifying the work relevant to the concept of hierarchy in the management literature was not as straightforward as in control. This is because the existence of a hierarchy in managerial decision making: is widely understood. So-called strategic decisions require more time than tactical decisions to become effective, they modify the system more deeply and, therefore, they will constrain the decisions to make at the lower level. The same type of relationship holds between tactical and operational decisions and this hierarchical structure directly follows from the definition of the different classes of decisions. Hence, any work addressing a managerial problem will fit in this hierarchical framework. However, a number of questions remain, and the answers determine the extent to which any paper should be considered "hierarchical".

The first question arises immediately when the "strategic, tactical, operational" classification is applied to a company where there are intermediate decision levels. The question is how to derive a partitioning of the decisions from the qualitative taxonomy described above? Two answers are examined in this chapter: one is a "static" answer ([HA1],[HM]), namely that, in general, the tactical and operational decision levels can be divided in four standard problems: aggregate planning, lot sizing and sequencing, detailed scheduling, and some sort of shop-floor real time control. The other

originated in the control literature ([VO],[VM],[GE3],[GE4]) and consists of a decomposition of the decisions based on their frequency.

The second question is to what extent decisions related to different levels can or must be made independently. More precisely, are there specific criteria to optimize at each level, and how is it possible to ensure that decisions made independently are consistent? Very different answers are given to this question in the work surveyed.

A first answer consists of obviating the question ([DG], [GK1], [GO]): when the objective chosen is directly affected by the decisions relative to two levels, a monolithic model is proposed and the decisions are made jointly. The next question, of course, is then whether the resulting system is hierarchical.

When, on the contrary, the objective can be split and different criteria are associated with several decision levels, two coordination schemes are proposed. If there is no guarantee that the decisions made at the higher level will result in a feasible decision set for the lower level, an iterative procedure is adopted and the issue of dimensionality appears. In the more fertile case where the constraints of the lower level can be taken into account (even though not perfectly) for the higher level decision making, a top-down constrained scheme is proposed ([BX2],[EF],[HM]). However, there is a need for further work to identify models in which lower level constraints can be transmitted to higher decision levels.

The third question is partially related to the second, since it concerns situations in which, due to unforecasted events, the feasible domain at a given level becomes empty. In that case, the higher level decisions have to be altered. Very little work has dealt with this feedback problem; the most general assumption is that the controls are recomputed according to the new conditions. In [KG], however, a feedback control law is proposed in a case where the state of the system can be described by two types of variables.

The last question is that of spatial decomposition: the decisions to be made in a manufacturing system are also hierarchical in that they have different scopes depending on their level. Very little work has addressed the issue of coordinating the decisions to make for different subsystems of a single global system.

The contribution of this thesis is an answer to this question within the framework defined by HILLION, MEIER and PROTH [HE].

# chapter 3:

# THE FLOW-CONTROL LAYER

Chapter I introduced the reader to one particular approach to the design of a hierarchical production planning system. One of the characteristics of this approach is that the machine aggregation process forces the aggregate models to satisfy the condition to be a generalized flow-shop. Other characteristics of the aggregate models analyzed in this work are described and justified hereunder.

## I CONTINUOUS* FLOWS

Typically, although not necessarily, the models obtained by this technique could represent the 'corporate' image of a large manufacturing firm and the subsystems would then be plants or shops, the product families being complete product lines. At that level, the 'amounts' of product-entities are usually represented by continuous variables, because it is not really relevant to measure them as numbers of items. For example, in the detergent industry, it is certainly more relevant to consider the volume, or the weight (volumes are usually convenient for liquids, weights are for solids) of detergent X produced during the week, rather than the number of boxes, especially if there are different box sizes.

But there is more to this example than meets the eye: first of all, it suggests that the characteristic by which one measures 'amounts' of a product should be adapted to the use that is made of this information. This means that the weekly production of detergent does not <u>have</u> to be measured in tons: it can be measured in KF (kilo-francs), or man-hours if one of these measures is more useful, and it <u>should</u> if there is no straightforward conversion between the measures.

* In this section, 'continuous' is opposed to 'discrete', and has nothing to do with the notion of continuity of a function: a variable is said to vary continuously if it takes its values in an interval of reals.

Note: the reason why conversion may not be straightforward is that a product-entity at the corporate level typically represents the aggregation of several items. Therefore, if such a product-entity groups two items with different ratios of value to volume, then knowing the aggregate production volume of the week does not tell one the value of what was produced, unless an additional information is provided, such as the volume ratio of the two items in the total production. In this case, the 'amounts' of production of the two items should be represented by their value to begin with, and not by their volume.

However, if the resulting error is not significant, one can also define a 'value' for the aggregate product as the weighted average of the values of the items that compose it. A linear conversion then exists between volumes and values of the aggregate product.

In most cases, though, the issue is complicated by the need to consider several characteristics of the product simultaneously: a demand is usually expressed in weight, volume or number of items, whereas a capacity is best expressed in man-hours or machine-hours. The rule is then to aggregate only products having similar ratios of the different characteristics to consider concurrently in the decision process, and different groupings are used for different types of decisions.

The point here is that, since a production or a demand may not be measured by the number of items but by the value of a related characteristic, the discrete nature of these amounts becomes very difficult to retain in the model. Assume for example that the characteristic chosen to measure an aggregate production is the weight. This characteristic varies by steps equal to the unit weight of any item in the aggregate. For it to be measured by integer values, a unit must be found such that the weight of any item is an integer number of this unit.

If there are fifty items in the aggregate and all have different weights, this unit is likely to be very small (smaller than the least difference between two items unit weight) with no physical meaning whatsoever. It is then more convenient to assume a continuous variation and measure the weights in a unit such that the amounts to be measured range in the most 'natural' interval for human manipulation, that is, probably between one and one thousand. For a foundry of aluminum, this unit is the ton…

The conclusion is that it is convenient, in models corresponding to the aggregate levels of a hierarchical planning system and in particular in the flow-control models studied here, to assume that the variables are real-valued, that is, continuous.

## II CONTINUOUS TIME

In the introduction of Zangwill [ZA], the author describes the transition between the work originated by WILSON's economic lot-size formula, in which demand is a continuous function of time and the optimal plan is derived by using calculus, and the work originated by WAGNER and WHITIN [WW], in which time is divided into discrete periods and mathematical programming is used to find the optimal plan. Ever since this transition, both discrete-time and continuous-time models have been used successfully, without any one approach proving to be significantly superior.

In fact, it seems that control theorists have favored continuous-time models, and discrete-time formulations have been more successful among operations researchers. The theory of optimal control actually provides extremely powerful and elegant techniques to solve continuous-time problems of certain types; unfortunately, very few production planning problems have been realistically cast in a format for which the results of control theory were helpful.

Recently, HACKMAN and LEACHMAN [HL], and GERSHWIN and co-workers ([GC]) have made the point that using continuous-time models to represent essentially discrete phenomena could be very insightful. This argument is the major justification for the use, in this work, of continuous time models.

Indeed, the analytical results presented in chapter IV are, for a good part, the continuous-time analogous of some results to be found in GABBAY [GA] with different cost structures. Moreover, the problems specifically considered could be solved by linear programming if formulated in discrete-time. The interest of the continuous-time formulation, however, lies in the fact that the results derived have a simple physical interpretation, that should make them extendable to cases for which there is no simple algorithm, like the case of stochastic demands.

An issue raised by the use of continuous-time, continuous-flow models for production planning is that, at some point, it is necessary to translate the results obtained by using these models into a discrete reality, and there is -a priori- no guarantee of feasibility. In fact, HILLION and PROTH [HP] have recently showed that in a job-shop, provided that a condition on the initial work in process and free resources be satisfied, it is always possible to find a cyclic schedule such that the bottleneck machines are fully utilized in steady-state. In terms of the issue raised, this result means that a production rate decided on a continuous-time model can always be achieved when the system reaches steady-state, as long as this rate is within the capacity of the system.

The models of chapter IV are thus continuous-time models, even though, to simplify the programming, they are implemented in discrete time in chapter V.

# III DETERMINISTIC MODELS

The models studied in this work are deterministic, which means that all the data or parameters they manipulate are assumed known with certainty. However, the industrial world is seldom deterministic itself: such disruptive events as machine failures, fluctuations of the demand volume, and quality problems resulting in variations of the yield do occur, in an unpredictable manner. That does not mean, though, that deterministic models are inapplicable.

## III.1- deterministic models can be used in a stochastic environment

### * modelling uncertainty

Uncertainties existing in production can be modelled in two different ways: either by the events that actually cause them, or by their effect on certain parameters or data of the model. For example, one can actually model the discrete events representing the failures and repairs of a machine, or one can model its capacity as a random variable. In the first case, one considers causes, and in the second, their effect. Also, the second representation implies a loss of resolution with respect to the first one, in the sense that a statistic only gives you a global information.

In fact, knowing that a machine is up 60% of its time does not tell you whether it is up or down at this very moment. If, additionally, the MTBF and MTTR** are on the order of days, the previous information is totally useless in the decision to start an operation that lasts minutes: what is really needed to make that decision is the detailed information concerning the current state of the machine, i.e. whether it is up or down. This type of consideration has been generalized in GERSHWIN [GE3][GE4].

** Mean Time Between Failures, Mean Time To Repair.

* importance of feedback

When the behavior of a system is not deterministic, its controller must be designed so as to limit the negative effect of disruptions. Closed loop controllers are particularly well suited for that purpose. In fact, feedback has been a very important factor in the success of control systems, because it makes it possible to enforce a certain behavior on a system even if it is only partially controllable (i.e. if its output depends both on the controlled variables and on disturbances), or if its 'transfer function' is only roughly approximated by the model (i.e. the effect of input variations on the behavior of the system is not modelled accurately).

Production systems are usually only partially controllable and the models used to represent them are not 'exact' in the sense that they are meant to retain only the dominant characteristics of the system's behavior. Therefore feedback is essential to an efficient control of such systems.

A reproach commomly made to deterministic models is that their output is made obsolete by the first unexpected event. In fact, this defect is not characteristic of deterministic models but of open-loop controllers, and that reproach would hold in the very same manner for stochastic models if they could be used in an open-loop fashion.

* importance of anticipation

However, the controller of a manufacturing system cannot be designed only to react to disruptions: it must also anticipate the future evolution of the environment if the system is supposed to achieve a high performance. Whereas the data corresponding to this future evolution are usually -more or less- subject to some uncertainty, anticipation is possible only if some information exists concerning their variations. Therefore, these data are typically modelled as random variables and their statistics are assumed known (forecasting is the art of evaluating these statistics, based both on

extrapolation of observed occurrences and on information concerning the future). Some of these variables, though, may have a very low variance and will be assumed known with certainty: among them, for example, technical parameters of the process, such as the time required to cast a given type of ingot on a particular casting pit, in an aluminum foundry. This parameter is deterministic for physical reasons, but others that are essentially random can be modelled as deterministic because of the use that is made of them...

* averaging-out of random variables

Consider, for example, a machine for which the data gathered indicate that the time between failures and the time to repair can both be modelled as occurrences of random variables with means $\bar{u} = 28$ mn and $\bar{d} = 4$ mn, and standard deviations $\sigma_u = 4$ mn and $\sigma_d = 1$ mn. Then, after n failures and repairs (n being large), the cumulative amounts of time that the machine has spent either up or down are normally distributed, with means $\bar{U}(n) = n.\bar{u}$ and $\bar{D}(n) = n.\bar{d}$ respectively, and standard deviations $\sigma_U(n) = \sqrt{n}.\sigma_u$ and $\sigma_D(n) = \sqrt{n}.\sigma_d$. Because of the averaging effect, the ratio of standard deviation to mean is divided by a factor $\sqrt{n}$ for the cumulated variables. If the problem considered is to determine the weekly production of this machine over the next two months, the average number of failures and repairs per unit time is $120 \times 60 / 32 = 225$, and the ratios of standard deviation to mean are: $4/28.15 \cong 0.95\,\%$ for the time up, and $1/4.15 \cong 1.67\,\%$ for the time down.

In that context, the error made by assuming a deterministic weekly capacity is insignificant, that is, the average weekly capacity is an accurate representation of the time the machine is operative, and does reflect the occurrence of failures. More generally, this result means that some uncertainties can be taken into account *on average* in a deterministic model; (it also has implications when the models adopted are hierachical, as will be explained in more detail further in this section). Another way to apply deterministic models in an environment subject to uncertainty is to use the concept of rolling horizon.

* rolling horizon

Making certain types of decisions (e.g. production planning) usually requires one to take some forecasts into account, and these forecasts are often more inaccurate as they correspond to a more remote future. In that case, the ratio of the quality of the decision to the complexity of the problem culminates for a certain -finite- length of the forecasting window, called the decision horizon.

Typically the decision problem involves determining the optimal control for a period [t, t+h], where h is the decision horizon (essentially because anticipation means ensuring that the problem remains feasible after any decision, and, in particular, that short-term gains will not cause higher losses on the long run). However, the only decision that really needs to be implemented is that corresponding to time t. Moreover, by the time it is implemented, new forecasts have become available, and the effect on the system of previous decisions can be observed.

It is then advantageous to solve the problem repeatedly and to implement only the immediate decisions, in order that the decision implemented at any time t take into account both the state reached by the system under control and the latest forecasts for the time-window [t, t+h]. This procedure defines the concept of *rolling horizon* (or rolling schedule if the problem is one of scheduling), and its effectiveness is experimentally demonstrated in BAKER [BA]: for a given problem (Dynamic Lot Sizing), it is shown that the gain in quality obtained if the forecasts are perfectly reliable over a significantly longer horizon than the one used in the rolling schedule does not exceed 10%.

To summarize the previous analysis, it may seem, at first sight, that deterministic models cannot adequately represent manufacturing systems, because the data they would manipulate may not be known with certainty, or may be subject to random changes.

But whatever the model chosen for such a non deterministic system, it remains an accurate representation if feedback from the system (or its environment) is used to update it. Furthermore, the necessary anticipation of future changes reaches a degree of difficulty it does not have in a deterministic context.

Deterministic models can nonetheless be used for the control of this type of system, at least in two instances:

* when uncertainty is due to the occurrence of disruptions of significantly higher frequency than that of the decisions to make, their effect can usually be averaged and represented with sufficient accuracy by a constant;

* when uncertainty concerns the future, a periodic reviewing of the decisions, in the light of the most recent -and accurate- forecasts reduces the number of scenari that would be evaluated by a probabilistic model and thus the chances of making 'wrong' decisions by ignoring the uncertainty.

Hierarchical models also have a characteristic that makes them particularly well suited to be applied to non deterministic systems:

III.2- hierarchical systems absorb disturbances at different levels:

One of the advantages of hierarchical systems is that each level needs to consider only the randomness that it can 'absorb'. For example, the destruction by fire of a plant is not the type of random event considered at the detailed scheduling level. On the other hand, the failures of machines that can be repaired in a matter of hours will be considered if, by changing the detailed schedule, it is possible to achieve the production objectives in spite of the failure. If it is not possible, then the next higher level will have to cope with this failure...

In the aluminum industry, for instance, there is no real hedging against the type of technical problem (be it loss of yield or major failure) that would require a significant re-planning at the corporate level. The system is designed so that a reaction to such events remains possible (e.g. the option is always left open to produce high-volume items in more than one plant), and several generic plans are prepared, one of them being finally chosen if needed. However, no actions are systematically taken to limit the impact of such events.

One of the reasons for this attitude is that hedging against a random event is possible only if some information exists about the likeliness of this event. Generally, there is no such information concerning failures or yield-losses of 'corporate' amplitude, simply because, even if they occur once or twice a year, they do not occur in the same circumstances, and deriving statistics would not be sound.Given this buffering effect between levels, the number of disturbances that might need to be modelled at a given level is greatly reduced.

Similarly, the amount and cost of raw material can be highly variable, but large variations of the available supply of raw material are usually the result of an event that is hedged against at the strategic level (major failure or bankrupcy of a supplier, war…), and cost models will give reliable forecasts of the variations that do not result from such events. On the demand side, the randomness comes from the existence of competitors which can attract part of the market by an aggressive price policy. In many cases, however, corporations sign contracts with their major clients and the fluctuations do not affect but a fringe of their demand -except when contracts are revised.

A first conlusion is that deterministic models are well suited for the corporate-level planning, especially since stochastic models, which are usually much more complex, loose some of their efficiency due to the low frequency of the type of random events to be considered.

It also seems, at least for the aluminum industry, that if there was one type of uncertainty to take into account, it would be that affecting the volume of final demand.

## IV BACKLOG VERSUS NO-BACKLOG

Backlogging occurs when a demand cannot be satisfied in time (more precisely, when the cumulative demand exceeds the cumulated production). Although the question of whether or not to allow for backlog in a planning model seems to be essentially a technical point, it raises an issue that has grown into a major concern in industry.

The issue raised is that of due-dates, that is, more generally 'soft constraints' versus 'hard constraints'. Soft constraints are usually modelled as goals to reach, a penalty being incurred if they are not satisfied. There is always a solution to a problem with soft constraints only, and the question is of knowing how good it is. Hard constraints, on the other hand, may result in the problem being infeasible.

There are relatively few situations in which a solution would not be found if cost did not matter, and the concept of soft constraints is appealing because when a problem arises in a practical situation, a bad solution is still better than none. The problem is in defining a penalty cost for not reaching the objective represented by a soft constraint, and in that respect, the penalty to pay for not meeting due dates is particularly hard to evaluate.

Consider the case of an aluminum plant shipping large amounts of metal on barges, and for which the due-dates are expressed in weeks. Not finishing loading a barge on the Friday evening of the contactual shipment week will be accounted for as a big drop in the timeliness index because the delay is rounded up to a week and weighted by the large amount of metal shipped 'late'.

The manager will then resort to overtime on Friday night to finish the loading, although knowing perfectly that doing it on regular hours on Monday morning would have made no difference (except for saving him the cost of overtime), since the barge does not travel on weekends and the metal would, in both cases, be unloaded at the customer's on tuesday morning after a twenty hour 'cruise'.

Note: from a practical point of view this example essentially means that the index chosen is not well suited for its purpose; but defining good indicators of performance is one of the fundamental problems in manufacturing: GOLDRATT's fable [GD] is a very insightful analysis of this shortcoming of traditional accounting procedures.

Where no-backlog models really prove to be superior is in multi-stage systems, for the intermediate inventories. Allowing backlog in such systems means allowing the scheduling of operations on parts that will not be available.

An argument usually given to defend such models is that safety stocks should be kept to physically allow for the otherwise impossible operations on backlogged products. The flaw of this argument is that safety stocks are usually determined to hedge against events whose occurrence will then not be considered in the operation of the system. Hence, assuming that a safety stock will be used to absorb a backlog can mean two things: either the chances of running out of stock because of an event that is not taken into account in the model increase, which may not be acceptable, or the 'safety stock' was in fact intended to absorb the backlog, and then the cost incurred to hold this inventory should be accounted for. It is usually considered a 'good' managerial practice to never let the inventory fall under the safety level on purpose.

Since the work presented here is mainly aimed at multi-stage systems, the models studied are of the no-backlog type, which means that all the inventory levels are bound to be positive.

## V CAPACITY CONSTRAINT

Given the options taken of continuous flows and continuous time, the unknowns of the problem will be time-dependent rates. For example, for a system producing only one type of product, the unknown is the production over time, that is, a function u such that u(t) represents the production rate of the system at time t, i.e. the volume being produced at time t per unit time.

The capacity of such a system is then the maximal value that the production rate can assume, and this value can vary with time: namely the performance of a machine can deteriorate with time, or the machine can be down at some point and its capacity is zero or, if the system consists of several identical machines, its capacity can vary by steps, depending on the set of machines that are up at a given point.

The formulation of a capacity constraint becomes interesting when there is more than one product. In the simplest case, one can assume that the capacity $C(t)$ is the maximum value that the production rate of a reference product can assume. Then, for each of the other products, one determines the ratio $a_i$ of the upper bound on its production rate when produced alone to that of a reference product $i_0$ such that $a_{i_0} = 1$. The capacity constraint becomes:

$$\forall t \in I, \quad \sum_{i=1}^{p} a_i . u_i(t) \leq C(t) \qquad (1)$$

A dual formulation of this constraint, more concise and elegant but less familiar to practitioners, is the one adopted by GERSHWIN and co-workers:

$$\forall t \in I, \quad \sum_{i=1}^{p} \tau_i . u_i(t) \leq \alpha(t) \qquad (2)$$

where $\tau_i$ represents the processing time of product i and $\alpha(t)$ represents the state of the machine: 0 if it is down, and 1 otherwise.

This inequality then means that, when it is up, the machine cannot be utilized more than 100% of the time, and when it is down it cannot be utilized at all. The equivalence between the two formulations casts a new light on the coefficients $a_i = \tau_i . C(t) / \alpha(t)$, which are directly propportional to the processing times.

In general, the capacity of a production system is determined by several inequalities similar to **(1)** or **(2)**, one for each machine in the system, and the set of feasible production rates at a given point in time is a polyhedron. Usually, though, there will be only one of these constraints binding at each point in time: the one corresponding to the 'bottleneck' machine (or limiting resource, if other resources than machines are considered in the capacity constraint). For a given system modelled as a set of resources, the bottleneck depends on the product-mix, but whichever resource it is, its capacity determines the maximal throughput of the entire system. In the models adopted in Chapter IV the assumption is that each system can be approximately modelled as one single limiting resource, regardless of the product-mix.

## VII OBJECTIVE

The objective considered at the flow control level is to coordinate several subsystems in order that products flow as fast as possible between subsystems, or, in different terms, so as to minimize the work in process. In classical models of type, a production cost is usually added to the inventory holding cost (GABBAY [GA], BENSOUSSAN, CROUHY and PROTH [BR]), and sometimes a setup cost is also considered (DZIELINSKI and GOMORY [DG]).

In fact, if the setups seem to be an important issue to take into account, variable production costs can be interpreted only if the upper bound on the production rate (i.e. the capacity) can be raised at some expense (e.g. by overtime). On the other hand, if the capacity

is fixed and the production cost is directly proportional to the volume of production, then the total variable production cost depends only on the demand. Therefore, production costs are not considered in the models of Chapter IV. Moreover, to simplify the solution of the problem, the inventory holding costs of any product i is assumed to be propportional to the volume of inventory $y_i(t)$.

The objective chosen can thus be formulated:

$$\min_{u} \quad \sum_{i} k_i \int_{I} y_i(t)\, dt$$

## VI DELAYS

Even at the corporate level, it may not be adequate to consider that the time spent by products in process is negligible: in fact, certain operations can be long, even at that time-scale. When the flow times are retained by the model, the inventories are defined by the following relations:

$$\forall i \in \{1,..,p\}, \quad \forall t \in I, \qquad y_i(t) = y_i(0) + \int_0^t ( u_i(s-\theta_i) - d_i(s) )\, ds \geq 0$$

where $\theta_i$ represents the duration between the instant when the decision is made to produce product i and the instant when it becomes available, that is, when it enters the final parts inventory.

Note: this modelling of the "lead-time" by a single constant will be more or less accurate, depending on the production system considered. If it is a single and very reliable machine and if the production process is entirely fixed, then the model will be perfect. If it is a whole production plant, it is very unlikely that the time spent by a product in the system will be constant; failures will occur, the idle-time between operations will depend on the scheduling policy, and both these components of the lead-time will be influenced by the load of the system. (It seems that a strong interest has developed recently for models of variable lead times to represent the queuing effect that appears when the load of a system reaches a certain limit).

But this type of restriction applies equally to all models: a model retains only some of the features of the physical system under consideration and its adequacy depends on the relative importance of the ignored features. There are systems for which introducing delays determined as the average lead-times makes it possible to address the most important issues.

Besides this question of adequacy, the introduction of delays raises the question of the definition of the problem to be solved: at time t, the evolution of the inventory of product i is entirely determined on the time interval $[t, t+\theta_i]$ if the demand is known on this period, because the decisions concerning the amounts of production that will be completed during this interval have been made in the past.

Therefore, at least two different objectives could be stated: either determine the amounts of different products to "launch" over a production planning horizon, in order to minimize the resulting inventory holding cost, assuming that the demands are known for the completion times of the decided productions, or assume that demands are known for a certain period (the same for all products) and determine the production that will minimize the inventory holding cost over this demand forecast horizon.

In the first case, demands of products having a long lead-time need to be known earlier than demands of products having a shorter lead-time, and the planning interval is the same for all products, in terms of the starting time. In the second case, the planning interval will be the same for all products in terms of completion time.

The choice between these two possible objectives will depend on the type of capacity constraint: the most practical way to introduce a capacity constraint would be to consider the loading rates of the different products at each point in time. The resulting constraints would be of the type introduced in earlier this chapter:

$$\forall\ t \in I, \quad \sum_{i=1}^{p} a_i \,.\, u_i(t) \leq C(t) \qquad (1)$$

In that case, the first type of objective is the most appropriate. But this is not at all the only possible capacity constraint: if all the products have to undergo a final operation (e.g. packing) for which there is a single facility that actually determines the capacity, it is the completion rates that need be considered in the capacity constraint, which will be stated as:

$$\forall\ t \in I', \quad \sum_{i=1}^{p} a_i \,.\, u_i(t - \theta_i) \leq C(t) \qquad (3)$$

where I' is the interval on which demands are known, the objective being of the second type.

Whatever the objective chosen, the formulation of the problem to be solved will be of the following type:

$$\min_{\underline{u}} \quad \sum_i k_i \int_I y_i(t+\theta_i)\, dt$$

$$\text{s.t.} \quad \forall i, \forall t \in I, \quad y_i(t+\theta_i) = y_i(\theta_i) + \int_I (u_i(s) - d_i(s+\theta_i))\, ds \geq 0$$

$$\forall t \in I, \quad \sum_i u_i(t) \leq C(t)$$

, where the inventories $y_i(\theta_i)$ are known when the decisions are made, and for all i, $d_i$ represents the demand of product i during the planning interval.

Since the demands are assumed known on the interval I, it just takes a variable transformation ($y'_i(t) = y_i(t+\theta_i)$ and $d'_i(t) = d_i(t+\theta_i)$) to show that this formulation is equivalent to that of a no-delay problem in terms of $\underline{y}'$, $\underline{u}$ and $\underline{d}'$.

The problems investigated in this work are thus, without loss of generality, of the no-delay type.

## IX A FLOW-SHOP MODEL

Consider a manufacturing system consisting of a network of n production subsystems feeding intermediate buffers with p different types of products: it is assumed that each product type has to undergo a specific sequence of operations performed in the different subsystems. This means in particular that, if a product can be sold at two different stages of its production process, then it will be split into two different products.

Each product is thus assigned a sequence of subsystems it has to go through in order that its production process be completed. This sequence will be called its route. The production system considered here is of the flow-shop type, which means that the subsystems can be indexed in order that each of the products route is a subsequence of the entire subsystems sequence. In other terms, this restriction means that if a product has to undergo an operation in subsystem i and the following operation in subsystem j, then subsystem j is after subsystem i in the routes of all the products.

It is first assumed that the time spent by the products in process is negligible and, in order to facilitate the formulation, that the demands are the loading rates into a virtual $(n+1)^{th}$ subsystem.

This way, all the routes end with subsystem n+1 and for each product i, its route is: $(m_k)_{k=1,\dots,N_i}$ where $m_{N_i} = n+1$.

The inventory levels are then defined by the equations:

$$\forall\, i \in \{1,..,p\},\ \forall\, k \in \{1,..,N_i-1\},$$

$$\{j = m_k \text{ and } j' = m_{k+1}\} \Rightarrow \quad y_i^j(t) = y_i^j(0) + \int_0^t [\, u_i^j(s) - u_i^{j'}(s)\,]\, ds.$$

where $u_i^j$ is the loading rate of product i into subsystem j, $y_i^j$ the stock of i at subsystem j, and $u_j^{n+1}$ is the demand of product i.

For each subsystem $j \in \{1,..,n\}$, define the set $\Lambda(j)$ of products that have to undergo an operation in subsystem j:

$$\Lambda(j) = \{ i \in \{1..p\} / \exists k \in \{1..N_i-1\} \text{ and } j = m_k \}$$

The problem to solve is then:

$$\underset{(u^j)}{\text{Min}} \quad \sum_{j=1}^{n} \sum_{i=1}^{p} k_i^j \int_0^h y_i^j(t)\, dt$$

$$\text{s.t.} \quad \forall i \in \{1,..,p\}, \forall k \in \{1,..,N_i-1\}, \forall t \in I, \; y_i^{m_k}(t) \geq 0.$$

$$\forall j \in \{1,..,n\}, \forall t \in I, \; \sum_{i=1}^{p} a_i^j . u_i^j(t) \leq C^j(t)$$

This problem can be formulated slightly differently by considering that the products flow through all the subsystems with the restriction that they cannot accumulate but in the inventories belonging to their route, and affect the capacity only for the subsystems in this route. For each product i, define $\Gamma(i)$ as the set of the indices of the subsystems in its route; the formulation of the problem then becomes:

$$\underset{(u^j)}{\text{Min}} \quad \sum_{j=1}^{n} \sum_{i=1}^{p} k_i^j \int_0^h y_i^j(t)\, dt$$

$$\text{s.t.} \quad \forall i \in \{1,..,p\}, \forall j \in \{1,..,n\}, \forall t \in I, \; y_i^j(t) = y_i^j(0) + \int_0^t [u_i^j(s) - u_i^{j+1}(s)]\, ds \geq 0.$$

$$\forall i \in \{1,..,p\}, \forall j \in \{1,..,n\} \setminus \Gamma(i), \; \forall t \in I, \; y_i^j(t) = 0 \qquad (4)$$

$$\forall j \in \{1,..,n\}, \forall t \in I, \; \sum_{i=1}^{p} a_i^j . u_i^j(t) \leq C^j(t).$$

, with the assumption that $\forall j \in \{1,..,n\}, \; i \notin \Lambda(j) \Rightarrow a_i^j = 0.$

This formulation shows that the flow-shop problem could be solved exactly as a multi-stage problem if it were possible to avoid the constraints (4) by introducing corresponding infinite holding costs. In fact, this result follows from those in Chapter IV, and the case of the flow-shop system represented on fig.1 (four subsystems, five products) will be solved in Chapter V.

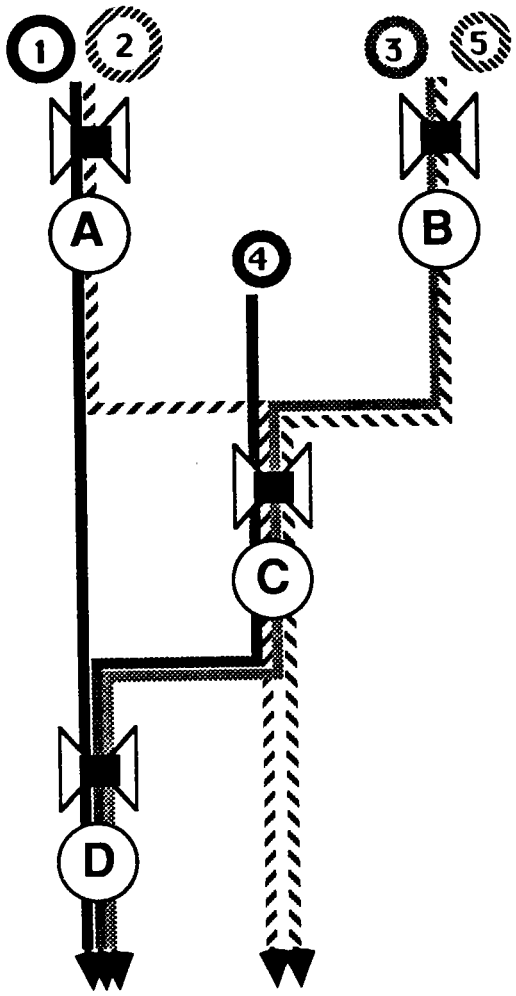



fig.1

# chapter 4:

# ANALYTICAL RESULTS

# THE SINGLE-STAGE, MONO-PRODUCT PROBLEM:

## I FORMULATION:

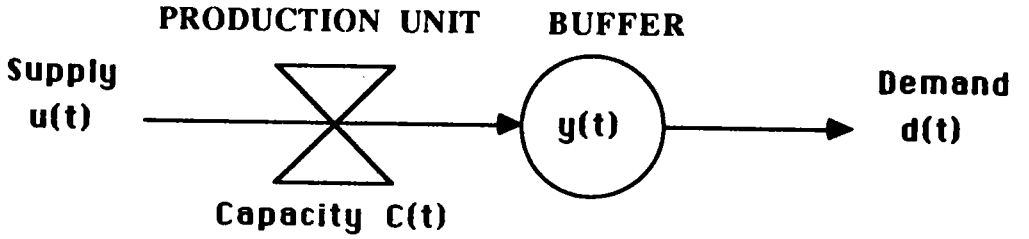



fig. 1.

The system represented on fig.1 consists of a production unit feeding a buffer with a single type of product. This system must satisfy a variable demand, known over the planning interval I=[0,h] and modelled as a positive and piece-wise continuous* function of time.

It is assumed that the time spent by the product in process (also referred to as its flow-time or lead-time) is negligible and that the capacity of the system is limited. Furthermore, backlogging is not allowed and inventory is used to hedge against peaks in demand. The question is then, given that the demand d(t) is known on the planning interval I, how to determine the supply u(t) on this same interval, in order that the demand be satisfied and the total inventory holding cost minimized.

* A function $f$ mapping I into $\mathbb{R}^+$ will be said piece-wise continuous if it has a finite number of discontinuities, all of the first kind, and if it satisfies the condition: $\forall t \in I$, $f(t) \in \{ f(t^-), f(t^+) \}$ , where $f(t^-)$ and $f(t^+)$ are respectively the left-hand and right-hand limits of $f$ at t, which exist since discontinuities are assumed simple. This condition means that $f$ is continuous on some interval $(t-\mu, t]$ , or $[t, t+\mu)$ or on $(t-\mu, t+\mu)$.

This question can be stated in mathematical terms as:

$$SS_1\ (C,d,y_0) \qquad \underset{u \in \Omega}{\text{Min}} \int_0^h y\ (t)\, dt$$

$$\text{s.t.} \quad \forall\, t \in [0,h]\ ,$$

$$y\,(t) = y_0 + \int_0^t [u\,(s) - d\,(s)]\ ds \ \geq 0 \qquad (C_1)$$

$$u\,(t) \leq C\,(t)$$

, where the initial inventory $y_0$ is given and positive, and $\Omega$ is the set of all **positive** and piece-wise continuous functions on I: for the model to remain meaningful, the production rate u must be positive, and it is not restrictive to assume it piece-wise continuous.

note: 1/ for the sake of mathematical rigor, $\Omega$ should be defined as the set of all classes of positive and piece-wise continuous functions on I for the equivalence "is equal to, almost everywhere on I". This simply means that $\Omega$ should not contain two functions that are equal almost everywhere (in short, "a.e."). Indeed, since the objective function is expressed in terms of the integral of the control, two controls equal a.e. on I yield the same value of the objective. Therefore, the optimal control cannot be characterized more precisely than "a.e. on I". However, it was chosen to avoid this sofistication and to repeat the expressions "= a.e. on I" and "modulo the equivalence = a.e. on I" whenever needed, however clumsy it may sound.

2/ a proposition P is said to be satisfied "almost everywhere" on a measurable set E if the subset of E on which P is not satisfied has a zero measure; it is considered here that a real set has a zero measure iff it does not contain any non-empty interval.

## II SIMPLIFICATION:

*The purpose of this section is to prove that problem $SS_1$ can be equivalently formulated with a zero initial inventory.*

**definitions:**

$$\theta : \Omega_x \mathbb{R}^+ \to I$$

$$(f, y_0) \to \tau = \text{Sup} \left\{ t \in [0,h] / \int_0^t f(s)\,ds \leq y_0 \right\}$$

$$\Psi : \Omega_x \mathbb{R}^+ \to \Omega$$

$$(f, y_0) \to g \quad \text{iff} \quad \tau = \theta(f, y_0) \quad \text{and} \quad \begin{cases} \forall t \in [0, \tau] & g(t) = 0 \\ \forall t \in ]\tau, h] & g(t) = f(t) \end{cases}$$

**interpretation:**

If $f$ is a demand function and $y_0$ an initial inventory, then $g = \Psi(f, y_0)$ is the related "net demand", that is, the demand that cannot be satisfied from the initial inventory.

**notation:** $\forall X \in \mathbb{R}$, $X^+$ stands for $\text{Max}(X, 0)$.

**lemma 1.II.1:**

let $f \in \Omega$ and $g = \Psi(f, y_0)$;

**then** $\forall t \in I$, $\displaystyle\int_0^t g(s)\,ds = \left[ -y_0 + \int_0^t f(s)\,ds \right]^+$

**proof:**

let $\tau = \theta(f, y_0)$.

By definition, $\displaystyle\int_0^t g(s)\,ds = \int_\tau^t g(s)\,ds$, whereas $\displaystyle -y_0 + \int_0^t f(s)\,ds = \int_\tau^t f(s)\,ds$.

We thus prove that: $\forall t \in I$, $\displaystyle\int_\tau^t g(s)\,ds = \left[ \int_\tau^t f(s)\,ds \right]^+$

+ if $t \leq \tau$, $\displaystyle\int_\tau^t g(s)\,ds = 0 = \left[ \int_\tau^t f(s)\,ds \right]^+$ because $\displaystyle\int_\tau^t f(s)\,ds \leq 0$

+ if $t > \tau$, $\displaystyle\int_\tau^t g(s)\,ds = \int_\tau^t f(s)\,ds$ by definition of g, and $\displaystyle\int_\tau^t f(s)\,ds \geq 0$ because $f \in \Omega$ is positive; hence, $\displaystyle\int_\tau^t f(s)\,ds = \left[ \int_\tau^t f(s)\,ds \right]^+$.

Consider now the problem:

$$SS_2\ (C,d') \qquad \underset{u \in \Omega}{\text{Min}} \int_0^h y'(t)\,dt$$

$$\text{s.t.} \quad \forall\, t \in [0,h]\ ,$$

$$y'(t) = \int_0^t [u(s) - d'(s)]\,ds \geq 0 \qquad (C'_1)$$

$$u(t) \leq C(t)\ ,$$

where $d' = \Psi(d, y_0)$.

**theorem 1.II.1:**

The problems $(SS_1)$ and $(SS_2)$ stated hereabove are equivalent.

- Or, in other words, u solves $(SS_1)$ iff u solves $(SS_2)$ -

<u>proof:</u>

+ Both objectives are equivalent to: $\underset{u \in \Omega}{\text{Min}} \int_0^h \int_0^t u(s)\,ds\,dt$

(Actually, the other terms of the objectives are constant).

+ It follows from lemma 1.II.1 that the constraints $(C_1)$ and $(C'_1)$ are equivalent:

$$(C_1) \iff \forall\, t \in [0,h]\ , \quad \int_0^t u(s)\,ds \geq -y(0) + \int_0^t d(s)\,ds$$

and, because $u \in \Omega$ is positive, this is equivalent to:

$$\int_0^t u(s)\,ds \geq \left[ -y(0) + \int_0^t d(s)\,ds \right]^+ = \int_0^t d'(s)\,ds$$

(by lemma 1.II.1 )

Hence the simplified formulation of the problem to solve:

$$SS\ (C,d) \qquad \min_{u \in \Omega} \int_0^h \int_0^t u(s)\, ds\, dt$$

$$\text{s.t. } \forall\, t \in [0,h],$$

$$y(t) = \int_0^t [u(s) - d(s)]\, ds \geq 0$$

$$u(t) \leq C(t).$$

Given the previous theorem, it will be staightforward to extend to problem $SS_1\ (C,d,y_0)$ the results obtained for problem $SS\ (C,d)$.

**theorem 1.III.1:**

$$\int_0^h \int_0^t u(s)\, ds\, dt = \int_0^h (h - t)\, u(t)\, dt$$

proof:

The proof is based on an integration by parts:

Let $U(t) = \int_0^t u(s)\, ds$ and $V(t) = t$

then $$\int_0^h \int_0^t u(s)\, ds\, dt = \int_0^h U(t)\, dt = \int_0^h U(t)\, V'(t)\, dt \quad \text{since } V'(t) = 1$$

$$= \Big[ U(t)\, V(t) \Big]_0^h - \int_0^h U'(t)\, V(t)\, dt$$

$$= h\, U(h) - \int_0^h t \,.\, U'(t)\, dt = h \int_0^h u(t)\, dt - \int_0^h t \,.\, u(t)\, dt$$

interpretation:

This result is meant to show how minimizing the inventory holding cost is actually achieved by producing as late as possible.

**corollary 1.III.1:**

The objective of problem SS (C,d) can be rewritten: $\underset{u \in \Omega}{\text{Max}} \int_0^h t.u(t)\,dt.$

## III OPTIMAL FLOW-PLAN:

### definitions:

+ A function u satisfying the constraints of problem SS (C,d) is called an admissible flow plan.

+ Two functions $z = \Phi(C,d)$ and $v = \Xi(C,d)$ can be defined as follows:

$$\forall t \in [0,h], \quad z(t) = \sup_{\tau \in [t,h]} \int_t^\tau [d(s) - C(s)]\,ds \qquad (L)$$

$$v(t) = \begin{cases} \text{Min}(d(t), C(t)) & \text{if } z(t) = 0 \\ C(t) & \text{otherwise} \end{cases}$$

*The purpose of present section is to prove that v is the optimal flow-plan and z the related inventory function.*

**lemma 1.III.1:**

let u be an admissible flow-plan and y the related inventory;

**then** $\forall t \in I, \; y(t) \geq z(t)$.

### interpretation:

z is a lower bound on the admissible inventory functions.

### proof:

let $t \in I$;

$\forall \tau \in [t,h], \; y(t) = y(\tau) - \int_t^\tau [u(s) - d(s)]\,ds \geq - \int_t^\tau [u(s) - d(s)]\,ds$ because $y(\tau) \geq 0$.

Besides, $\forall s \in [t,h], \; u(s) \leq C(s)$ and thus $\int_t^\tau [u(s) - d(s)]\,ds \leq \int_t^\tau [C(s) - d(s)]\,ds$

Therefore, $y(t) \geq \sup_{\tau \in [t,h]} \int_t^\tau [d(s) - C(s)]\,ds = z(t)$

**lemma 1.III.2:**

The functions v and z satisfy the following propositions:

0. $\forall t \in [0,h]$ , $z(t) \geq 0$.

1. z is continuous on $I = [0,h]$.

2. there exists a finite sequence of intervals $([\beta_k, \gamma_{k-1}])_{k=1,...,N}$ such that:

   + $\gamma_0 = h$

   + $\forall k \in \{1,...,N\}$ , $\gamma_k \leq \beta_k \leq \gamma_{k-1}$

   + $\forall t \in [0,h]$ , $z(t) = 0 \Leftrightarrow \exists k \in \{1,..,N\}$ such that $t \in [\beta_k, \gamma_{k-1}]$.

3. v is piece-wise continuous on I, that is, $v \in \Omega$.

4. $\forall t \in [0,h]$ , $z(t) = \int_t^{\beta_t} [\, d(s) - C(s) \,]\, ds$ ,

   where $\beta_t = \text{Inf} \{ \tau \in ]t,h]$ such that $z(\tau) = 0 \}$

<u>interpretation:</u>

The important result of this lemma is that $v \in \Omega$; the other propositions are stepping stones for further results.

<u>proof:</u>

Define the function $\varphi: \tau \rightarrow \int_0^{\tau} [\, d(s) - C(s) \,]\, ds$

Then, $\forall t \in [0,h]$ , $z(t) = \underset{\tau \in [t,h]}{\text{Sup}} [\, \varphi(\tau) - \varphi(t) \,] = \varphi(\tau_t) - \varphi(t)$

, where $\tau_t = \underset{\tau \in [t,h]}{\text{Argsup}}\, \varphi(\tau)$

• <u>Proposition 0</u> results directly from the definition of z (suffice it to consider that for $\tau = t$ , the integral is zero).

• Proposition 1: let $t \in I$ and $\varepsilon > 0$.

d and C being piece-wise continuous on I, $\varphi$ is continuous on I, uniformly because I is compact. Thus $\exists\, \alpha > 0$ such that:

$$\forall\, s,s' \in I,\ |s - s'| < \alpha \Rightarrow |\varphi(s) - \varphi(s')| < \varepsilon/2 \qquad (1)$$

The objective is to prove that: $\forall\, t' \in I,\ |t - t'| < \alpha/2 \Rightarrow |z(t) - z(t')| < \varepsilon$.

let $t' \in\ ]t - \alpha/2, t + \alpha/2[$:

Two cases need be considered:

1- $\tau_t = \tau_{t'}$: then $|z(t) - z(t')| = |\varphi(t') - \varphi(t)| < \varepsilon/2$.

2- $\tau_t \neq \tau_{t'}$: assume that $t < t'$;

Then $\tau_t \in [t,t'[$ and $\varphi(t') \leq \varphi(\tau_{t'}) < \varphi(\tau_t)$: both these properties result from the definition of $\tau_t$ and $\tau_{t'}$.

Moreover, since $\tau_t \in [t,t'[$, $|\tau_t - t'| < \alpha$ and $|\varphi(\tau_t) - \varphi(t')| < \varepsilon/2$: see (1).

Therefore $|z^1(t) - z^1(t')| = |\varphi(t') - \varphi(t) + \varphi(\tau_{t'}) - \varphi(\tau_t)| \leq |\varphi(t') - \varphi(t)| + |\varphi(\tau_{t'}) - \varphi(\tau_t)|$

$$\leq |\varphi(t') - \varphi(t)| + |\varphi(t') - \varphi(\tau_t)| < \varepsilon.$$

Since t and t' have symmetrical roles in the previous development, the same result would be reached under the assumption that $t < t'$.

Hence, $\forall\, t \in I$, $\forall\, \varepsilon > 0$, $\exists\, \alpha > 0$ and $\forall\, t' \in I$, $|t - t'| < \alpha/2 \Rightarrow |z(t) - z(t')| < \varepsilon$, which exactly means that z is continuous on I.

• Proposition 2: z being continuous, the reciprocal image of 0 by z, i.e. the set of all points t in [0,h] such that $z(t) = 0$ is a closed subset of [0,h], i.e. a union of closed intervals. This union is **assumed** finite; the existence and properties of the sequence $([\beta_k, \gamma_{k-1}])_{k=1,\ldots,N}$ asserted in lemma 1.III.2 follow.

• Proposition 3 results from the fact that both d and C are positive and piece-wise continuous.

In fact, if $J = \{ t \in I / z(t) = 0 \}$, then by definition $v = C$ on $I/J$ and $v = \mathrm{Min}(d, C)$ on J; hence v is positive. It can actually be proved that $v = d$ almost everywhere on J, which proves its piece-wise continuity, given that both I and J are finite unions of intervals:

Let $t \in J$ and assume that $d(t) > C(t)$ and that $d - C$ is continuous at t. Then $\exists \varepsilon > 0$ such that: $\forall s \in ]t-\varepsilon, t+\varepsilon[$ , $d(s) > C(s)$ and:

$$z(t) \geq \int_t^{t+\varepsilon} [\, d(s) - C(s) \,] \, ds \; > 0$$ , which contradicts $z(t) = 0$.

Hence $d - C$ is not continuous at the points t of J where $d(t) > C(t)$, that is, the points t of J where $v(t) \neq d(t)$. Since $d - C$ is piece-wise continuous on J, $v = d$ almost everywhere on J.

• Proposition 4:

Let $t \in I$ and $\beta_t = \mathrm{Inf} \{ \tau \in ]t,h] \text{ such that } z(\tau) = 0 \}$.

By definition of $\varphi$ , z and $\beta_t$, it follows that:

$$\sup_{\tau \in [\beta_t, h]} \varphi(\tau) = \varphi(\beta_t) + z(\beta_t) = \varphi(\beta_t) \quad \text{and} \quad \forall s \in ]t, \beta_t] , \; \sup_{\tau \in [s,h]} \varphi(\tau) > \varphi(s)$$

This last inequality implies that:

$$\forall s \in ]t, \beta_t] , \; \sup_{\tau \in [s,h]} \varphi(\tau) = \sup_{\tau \in [\beta_t, h]} \varphi(\tau) = \varphi(\beta_t)$$

and, $\varphi$ being continuous, $\sup_{\tau \in [t,h]} \varphi(\tau) = \varphi(\beta_t)$ , which means that:

$$z(t) = \sup_{\tau \in [t,h]} [\, \varphi(\tau) - \varphi(t) \,] = \varphi(\beta_t) - \varphi(t) = \int_t^{\beta_t} [\, d(s) - C(s) \,] \, ds$$

**lemma 1.III.3:**

$$\forall\, t \in [0,h]\ ,\ z(t) = \int_t^h [\, d(s) - v(s)\,]\, ds$$

proof:

Let $k \in \{1,..,N\}$ and note $w(s) = d(s) - v(s)$;

→ $\forall\, t \in [\beta_k, \gamma_{k-1}]\ ,\ z(t) = 0,$ and $v = d$ a.e on $[\beta_k, \gamma_{k-1}]$;

thus $\forall\, t \in [\beta_k, \gamma_{k-1}]\ ,\ \int_t^{\gamma_{k-1}} w(s)\, ds = 0 = z(t).$

→ $\forall\, t \in\ ]\gamma_k, \beta_k[\ ,\ z(t) > 0$ and $v(t) = C(t)$;

Since $\beta_k = \mathrm{Inf}\,\{\, \tau \in\ ]t,h]\ /\ z(\tau) = 0\,\}$, proposition 4 of lemma 1.II.2 applies and implies that $z(t) = \int_t^{\beta_k} [\, d(s) - C(s)\,]\, ds = \int_t^{\beta_k} w(s)\, ds.$

Hence $\forall\, t \in\ ]\gamma_k, \beta_k[\ ,\ z(t) = \int_t^{\gamma_{k-1}} w(s)\, ds.$

By continuity, this result holds also for $t = \gamma_k$, which means:

$\forall\, k \in \{1,..,N\}\ ,\ \forall\, t \in [\gamma_k, \gamma_{k-1}]\ ,\ z(t) = \int_t^{\gamma_{k-1}} w(s)\, ds$ and $\int_{\gamma_k}^{\gamma_{k-1}} w(s)\, ds = z(\gamma_k) = 0$

A straightforward recursion would yield that $\int_{\gamma_k}^{\gamma_0} w(s)\, ds = 0.$

Since $\gamma_0 = h$, it results that: $\forall\, t \in I\ ,\ z(t) = \int_t^h w(s)\, ds$

**corollary 1.III.3:**

$$\forall\, t \in [0,h]\ ,\ \int_t^h v(s)\, ds = \int_t^{\beta_t} C(s)\, ds + \int_{\beta_t}^h d(s)\, ds$$

proof:

$$z(t) = \int_t^h [d(s) - v(s)]\, ds = \int_t^h [d(s) - C(s)]\, ds \ \Rightarrow\ \int_t^h v(s)\, ds = \int_t^h d(s)\, ds - \int_t^h d(s)\, ds + \int_t^h C(s)\, ds$$

**lemma 1.III.4:**

$$\textbf{if } \forall t \in [0,h],\ \int_0^t C(s)\,ds \geq \int_0^t d(s)\,ds \qquad \textbf{(A)}$$

$$\textbf{then } \quad z(0) = 0 \text{ and } \forall t \in [0,h],\ z(t) = \int_0^t [\,v(s) - d(s)\,]\,ds$$

proof:

If (A) is satisfied, then $z(0) = \sup_{\tau \in [0,h]} \int_0^\tau [\,d(s) - C(s)\,]\,ds \leq 0$ and, since z is positive, $z(0) = 0$.

Hence, $z(t) = \int_t^h [\,d(s) - v(s)\,]\,ds$

$$= -z(0) + \int_t^h [\,d(s) - v(s)\,]\,ds \qquad \text{because } z(0) = 0$$

$$= -\int_0^h [\,d(s) - v(s)\,]\,ds + \int_t^h [\,d(s) - v(s)\,]\,ds$$

$$= \int_0^t [\,d(s) - v(s)\,]\,ds$$

interpretation: *these preliminary lemmas show that, if* (A) *is satisfied, then* v *is an admissible flow-plan: namely, it has been proved that* $v \in \Omega$ *and, by definition,* v *satisfies the capacity constraint; moreover, it was just proved that* z *is the inventory function related to* v *and, previously, that* z *is positive.*

**corollary 1.III.4:**

There exist admissible flow plans **iff** (A) is satisfied.

**proof:**

The "if" part of the statement has just been proved: if (A) is satisfied, then v is an admissible flow-plan. Reciprocally, if there exists an admissible flow-plan u, it follows that:

$\forall t \in I$, $\int_0^t u(s)\,ds \geq \int_0^t d(s)\,ds$ because $y(t) \geq 0$.

Moreover, $\forall s \in [0,t]$, $u(s) \leq C(s)$ and thus $\int_0^t C(s)\,ds \geq \int_0^t u(s)\,ds$

Hence (A) is satisfied.

**theorem 1.III.1:**

**If** SS (C,d) has a solution, **then** $v = \Xi(C,d)$ is this solution, and it is unique modulo the equivalence "= a.e. on I".

In other words:

**If** u is an optimal flow-plan ( i.e a solution to SS (C,d) ), **then** u = v almost everywhere on I.

**proof:**

If $\exists J = ]t_0, t_1[$, $t_0 \neq t_1$ and $\forall s \in J$, $u(s) \neq v(s)$ then $y \neq z$ a.e. on J and it follows from lemma 1.III.1 that $y > z$ a.e. on J.

Thus $\int_0^h y(t)\,dt > \int_0^h z(t)\,dt$ and u would not be optimal.

Therefore u = v a.e. on I.

**Conclusion:**

Since it was proved in theorem 1.II.1 that problem $SS_1(C,d,y_0)$ is equivalent to $SS(C, \Psi(d,y_0))$, the initial problem $SS_1(C,d,y_0)$ is solved by $v = \Xi(C, \Psi(d,y_0))$.

## IV PHYSICAL INTERPRETATION: BACKWARD SMOOTHING

Fig. 2. illustrates better than any comment how the optimal flow-plan is derived from the demand and capacity functions:

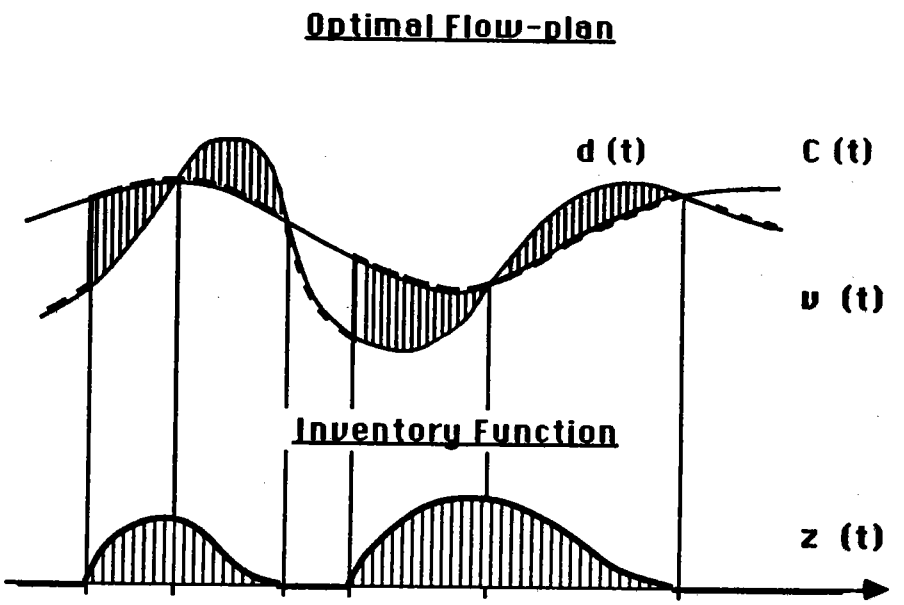



fig. 2.

The rationale for this solution is to produce in advance when the future demand is expected to exceed the capacity, and to do so "as late as possible". Consequently, the production is either equal to the demand or to the capacity. In fact, the production will be equal to the capacity during the peak periods (when demand exceeds capacity) and during an immediately preceding slack period (during which inventory is built up). The rest of the time, the production is equal to the demand and the inventory kept at zero. These results are summarized in the following equation: $\mathbf{y \cdot (u - C) = 0}$ a.e. on I.

note: the procedure yielding $v = \Xi(C,d)$ will be referred to as "backward smoothing" in the remainder of this work. Insofar that the optimal production plan is derived from the demand, a system based on this procedure would belong to the class of "pull systems".

## V PRELIMINARY RESULTS:

The purpose of this section is to present some results which are not really insightfull by themselves but will be useful in the remainder of the chapter. The first subsection concerns the backward smoothing procedure. The second one introduces the "opposite" procedure, which consists of pushing (instead of pulling) material through a capacitated system.

### -A- Properties of the function $\Xi$

Let $C,d \in \Omega$ and consider the problems $MxF_t(C,d)$ stated hereunder:

$$MxF_t(C,d) \qquad \max_{u \in \Omega} \int_t^h u(s)\,ds$$

$$\text{s.t. } \forall s \in [t,h], \quad \int_s^h u(r)\,dr \le \int_s^h d(r)\,dr \qquad (C_1)$$

$$u(s) \le C(s) \qquad (C_2)$$

**notation:**

For any function $f$ in $\Omega$, $f_t$ is the restriction of $f$ to $[t,h]$

**lemma 1.V.1:**

**if** $v = \Xi(C,d)$ , **then** $\forall t \in I$, $v_t$ solves $MxF_t(C,d)$.

**proof:**

The proof is essentially the same as for lemma 1.III.1: let u be a solution to $MxF_t(C,d)$ and $\tau \in [t,h]$;

$$\int_t^h (d(s)-u(s))\,ds \ge \int_t^\tau (d(s)-u(s))\,ds \quad \text{because} \quad \int_\tau^h (d(s)-u(s))\,ds \ge 0 \text{ by } (C_1)$$

$$\ge \int_t^\tau (d(s)-C(s))\,ds \quad \text{because of } (C_2).$$

Hence $\int_t^h (d(s)-u(s))ds \geq \sup_{\tau \in [t,h]} \int_t^\tau (d(s)-C(s))ds = z(t) = \int_t^h (d(s)-v_t(s))ds$

Therefore $\int_t^h u(s)ds \leq \int_t^h v_t(s)ds$

Since $v_t$ satisfies the constraints of problem $MxF_t(C,d)$, it is one of the solutions, although generally not the only one.

**definition:**

Two partial orders $\ll^h$ and $\ll^0$ can be defined on $\Omega$ as follows:

$\forall f_1, f_2 \in \Omega$, $f_1 \ll^h f_2$ **iff** $\forall t \in I$, $\int_t^h f_1(s)ds \leq \int_t^h f_2(s)ds$

$f_1 \ll^0 f_2$ **iff** $\int_0^h f_1(s)ds = \int_0^h f_2(s)ds$ and $f_2 \ll^h f_1$

which means: $\forall t \in I$, $\int_0^t f_1(s)ds \leq \int_0^t f_2(s)ds$.

Both of these binary relations satisfy the conditions to be orders on $\Omega$ : reflexivity and transitivity follow from that of $\leq$ on $\mathbb{R}$, and antisymmetry results from the proposition:

$$\forall x \in \{0,h\}, \quad (f_1 \ll^x f_2 \text{ and } f_2 \ll^x f_1) \Rightarrow f_1 = f_2 \text{ a.e. on I,}$$

and the fact that by definition, $\Omega$ is the set of all positive and piece-wise continuous functions on I, modulo the equivalence "equal almost everywhere on I" (see section 1.I).

**theorem 1.V.1:**

Let C, d, d' $\in \Omega$ and such that both SS (C,d) and SS (C,d') are solvable;

**if** $d \ll^h d'$, **then** $\Xi(C, d) \ll^h \Xi(C, d')$.

**interpretation:** $\Xi$ preserves the order $\ll^h$.

**proof:**

Let $u = \Xi(C, d)$, $u' = \Xi(C, d')$ and $t \in [0,h]$;

$u_t$ is admissible for $MxF_t(C,d')$ because $\forall r \in I$, $\int_r^h d(s)ds \leq \int_r^h d'(s)ds$.

Therefore, since $u'_t$ is a solution of $MxF_t(C,d')$, $u_t$ and $u'_t$ satisfy:

$$\int_t^h u'_t(s)ds \geq \int_t^h u_t(s)ds \quad \text{that is,} \quad \int_t^h u'(s)ds \geq \int_t^h u(s)ds.$$

As this inequality holds for all $t \in [0,h]$, it results that $u \ll^h u'$.

**corollary 1.V.1:**

Let $C, d, d' \in \Omega$ and such that $SS(C,d)$ is solvable;

**if** $d' \ll^0 d$, **then** $SS(C,d')$ is solvable,

**and** $\Xi(C, d') \ll^0 \Xi(C, d)$.

**proof:**

$\rightarrow$ $d' \ll^0 d$ implies that $\forall t \in I$, $\int_0^t d'(s)ds \leq \int_0^t d(s)ds$.

Also, since $SS(C,d)$ is solvable, it follows from the "necessary part" of corollary 1.III.4 that $\forall t \in I$, $\int_0^t d(s)ds \leq \int_0^t C(s)ds$.

Corollary 1.III.4 states that this condition is also sufficient and thus, by combining the two previous inequalities, it results that $SS(C,d')$ is solvable.

$\rightarrow$ $\forall C, f \in \Omega$ such that $SS(C,d)$ is solvable, $\int_0^h f(s)ds = \int_0^h \Xi(C, d)(s)ds$

results from the fact that if $z = \Phi(C, d)$ is the inventory function associated with the control $\Xi(C, d)$ (see definitions in section III), then, by definition $z(h) = 0$, and results also from lemma 1.III.3, which states: $z(h) = \int_0^h f(s)ds - \int_0^h \Xi(C, d)(s)ds$.

$\rightarrow d' \ll^0 d \Rightarrow \int_0^h d'(s)ds = \int_0^h d(s)ds$ and thus: $\int_0^h \Xi(C,d)(s)ds = \int_0^h \Xi(C,d')(s)ds.$

Also, $d' \ll^0 d \Rightarrow d \ll^h d'$ and, after theorem 1.V.1, $\Xi(C,d) \ll^h \Xi(C,d')$. Combining these two results ends the proof of the corollary.

**theorem 1.V.2:**

Let $C, d, \delta \in \Omega$; assuming these functions are such that the notations make sense, $\Xi$ satisfies the following properties:

$P_1$ $\Xi(C+\delta, d+\delta) = \Xi(C,d) + \delta$ a.e. on I,

$P_2$ $\Xi(C+\delta, \Xi(C,d)) = \Xi(C,d)$ a.e. on I,

$P_3$ $\Xi(C-\delta, \Xi(C,d)) = \Xi(C-\delta,d)$ a.e. on I.

<u>interpretation:</u> these properties become handy in the study of multi-stage systems.

<u>proof:</u>

$P_1$: let $u = \Xi(C, d)$, $u' = \Xi(C+\delta, d+\delta)$ and $y, y'$ the related inventories.

Then $\forall t \in I$, $y'(t) = \underset{\tau \in [t,h]}{\text{Sup}} \int_t^\tau [(d+\delta)(s) - (C+\delta)(s)]ds = y(t)$,

that is, $\forall t \in I$, $\int_0^t [u'(s) - (d+\delta)(s)]ds = \int_0^t [u(s) - d(s)]ds$,

and $\forall t \in I$, $\int_0^t [u'(s) - (u+\delta)(s)]ds = 0$, which proves $P_1$.

$P_2$: let $u^\circ = \Xi(C+\delta, u)$ and $y^\circ$ the related inventory.

Then $\forall t \in I$, $y^\circ(t) = \underset{\tau \in [t,h]}{\text{Sup}} \int_t^\tau [u(s) - (C+\delta)(s)]ds = 0$ because $u$ satisfies the constraints of SS(C,d) and thus $\forall s \in I$, $u(s) \leq C(s) \leq (C+\delta)(s)$.

Therefore, $u^\circ = u$ a.e. on I.

$\mathbf{P_3}$: let $u_1 = \Xi(C-\delta, u)$, $u_2 = \Xi(C-\delta, d)$ and $y_1$, $y_2$ the related inventories.

**lemma 1:** $\forall\, t \in I,\ y(t) \le y_2(t)$.

Namely, $\delta \ge 0 \Rightarrow \sup_{\tau \in [t,h]} \int_t^\tau [u(s) - (C-\delta)(s)]\, ds \ge \sup_{\tau \in [t,h]} \int_t^\tau [u(s) - C(s)]\, ds$

Hence the result, and a direct consequence: $y_2(t) = 0 \Rightarrow y(t) = 0$.

**lemma 2:** $\forall\, t \in I,\ y_2(t) \le y(t) + y_1(t)$.

$u = \Xi(C, d) \Rightarrow \forall\, t \in I,\ \int_t^h u(s)ds \le \int_t^h d(s)ds$ , that is, $u \ll^h d$.

Thus, by theorem 1.V.1, $u_1 = \Xi(C-\delta, u) \ll^h \Xi(C-\delta, d) = u_2$ , that is:

$\forall\, t \in I,\ \int_t^h u_1(s)ds \le \int_t^h u_2(s)ds$ , which is equivalent to:

$\forall\, t \in I,\ \int_t^h [d(s) - u_2(s)]ds \le \int_t^h d(s)ds - \int_t^h u_1(s)ds$ , and also to:

$\forall\, t \in I,\ \int_t^h [d(s) - u_2(s)]\, ds \le \int_t^h [d(s) - u(s)]\, ds + \int_t^h [u(s) - u_1(s)]\, ds$ Q.E.D.

**lemma 3:** $\forall\, t \in I,\ y_1(t) \le y_2(t)$.

The proof is by contradiction: $y_1(t) > y_2(t) \Rightarrow \exists\, \tau_0 \in [t,h]$ such that:

$$\forall\, \tau \in [t,h],\ \int_t^{\tau_0} [u(s) - (C-\delta)(s)]\, ds > \int_t^\tau [d(s) - (C-\delta)(s)]\, ds$$

$$\text{i.e.}\quad \forall\, \tau \in [t,h],\ \int_t^{\tau_0} [u(s) - d(s)]\, ds > \int_{\tau_0}^\tau d(s)\, ds - \int_t^\tau (C-\delta)(s)\, ds + \int_t^{\tau_0} (C-\delta)(s)\, ds$$

$$\text{or}\quad \forall\, \tau \in [t,h],\ y(\tau_0) - y(t) > \int_{\tau_0}^\tau [d(s) - (C-\delta)(s)]\, ds$$

In that case, $y(\tau_0) \ge y(\tau_0) - y(t) > \sup_{\tau \in [t,h]} \int_{\tau_0}^\tau [u(s) - (C-\delta)(s)]\, ds$

$$\ge \sup_{\tau \in [\tau_0,h]} \int_{\tau_0}^\tau [u(s) - (C-\delta)(s)]\, ds = y_2(\tau_0)$$

It then results that $y(\tau_0) > y_2(\tau_0)$, which contradicts lemma 1.

Q.E.D.

corollary: $\forall t \in I,\ y_1(t) = 0 \Rightarrow y_2(t) = y(t)$.

Namely, by lemma 2, $y_1(t) = 0 \Rightarrow y_2(t) \leq y(t)$ and, by lemma 1, $y(t) \leq y_2(t)$.

proof ($P_3$):

For $j = 1, 2$ and $t \in I$, let $\beta^j_t = \text{Inf}\ \{ \tau \in ]t,h]$ such that $y_j(\tau) = 0 \}$;

From corollary 1.III.3 it follows that:

$$\forall t \in [0,h],\ \int_t^h u_1(s)\,ds = \int_t^{\beta^1_t} (C-\delta)(s)\,ds + \int_{\beta^1_t}^h u(s)\,ds$$

$$\text{and} \quad \int_t^h u_2(s)\,ds = \int_t^{\beta_t^2} (C-\delta)(s)\,ds + \int_{\beta_t^2}^h d(s)\,ds$$

Since $y_2 \geq y$ and $y_2(\beta_t^2) = 0$, $y(\beta_t^2) = 0$ and $\int_{\beta_t^2}^h u(s)\,ds = \int_{\beta_t^2}^h d(s)\,ds$.

Therefore, $\int_t^h [u_2(s) - u_1(s)]\,ds = \int_{\beta^1_t}^{\beta_t^2} [(C-\delta)(s) - u(s)]\,ds$

$$= -\int_{\beta^1_t}^{\beta_t^2} [d(s) - (C-\delta)(s)]\,ds + \int_{\beta^1_t}^{\beta_t^2} [u(s) - d(s)]\,ds.$$

$$= -y_2(\beta^1_t) + y(\beta^1_t) - y(\beta_t^2)$$

$$= -y_2(\beta^1_t) + y(\beta^1_t) \text{ because } y(\beta_t^2) \leq y_2(\beta_t^2) = 0$$

$$= 0 \text{ because } y_1(\beta^1_t) = 0 \Rightarrow y_2(\beta^1_t) = y(\beta^1_t)$$

Push systems are multi-stage systems in which the flows are dictated by the loading rates into the system, as opposed to pull systems, in which the flows are driven by the demand of end-product. The results presented in the following sub-section are applicable to such systems.

## -B- "Push systems"

Consider the problem:

$$MxS\,(C^1, C^2, y_0) \qquad \max_{u \in \Omega} \int_0^h \int_0^t u(s)\,ds\,dt$$

$$\text{s.t. } \forall t \in [0,h],$$

$$\int_0^t u(s)\,ds \le \int_0^t C^1(s)\,ds + y_0$$

$$u(t) \le C^2(t)$$

This problem corresponds to that of maximizing the inventory holding cost -by producing as early as possible on the downstream machine- in a system like the one represented on fig. 1, in which inventory holding costs are incurred only for the end-product.

(While such an objective may seem sacrilegious in this piece of work, the usefulness of the following results will become clear in the forthcoming chapter).

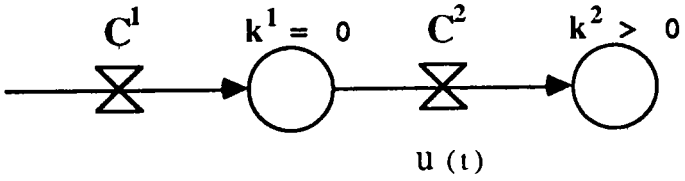



**fig. 1**

Following from theorem 1.III.1 the objective of $MxS\,(C^1, C^2, y_0)$ can be rewritten:

$$\min_{u \in \Omega} \int_0^h t \cdot u(t)\,dt$$

Also, $MxS(C^1, C^2, y_0)$ can be equivalently rewritten with a zero initial inventory if $C^1$ is replaced by, for example, $C'^1 = \Gamma(C^1, C^2, y_0)$ defined by:

$$\begin{cases} \forall t \in [0, \tau], \quad C'^1(t) = C^2(t) \\ \forall t \in ]\tau, h], \quad C'^1(t) = C^1(t) \end{cases}$$

$$\text{where } \tau = \theta(C^2 - C^1, y_0) = \text{Sup}\left\{ t \in [0,h] \ / \ \int_0^t [C^2(s) - C^1(s)]\, ds \right\}$$

The solutions to $MxS(C^1, C^2, y_0)$ and $MxS(C^1, C^2)$ will then be equal.

**definitions:**

$$\forall t \in [0,h], \quad \zeta(t) = \sup_{\tau \in [t,h]} \int_t^\tau [C^1(s) - C^2(s)]\, ds,$$

$$\varpi(t) = \begin{cases} \text{Min}(C^1(t), C^2(t)) \text{ if } \zeta(t) = 0 \quad (\text{i.e. } C^1(t) \text{ a.e. on } \zeta^{-1}(0)) \\ C^2(t) \text{ otherwise} \end{cases}$$

The following results are given without proof because they are in all respects similar to those stated in section 1.III.

0. $\forall t \in [0,h], \ \zeta(t) \geq 0.$

1. $\zeta$ is continuous on $I = [0,h]$.

2. $\varpi$ is piece-wise continuous on I, that is, $\varpi \in \Omega^*$.

3. $\forall t \in [0,h], \ \zeta(t) = \int_{\alpha_t}^t [C^1(s) - C^2(s)]\, ds$,

where $\alpha_t = \text{Sup}\ \{ \tau \in [0,t[ \text{ such that } \zeta(\tau) = 0 \}$

4. $\forall t \in [0,h], \ \zeta(t) = \int_0^t [C^1(s) - \varpi(s)]\, ds$

5. if u is admissible for $MxS(C^1, C^2)$, then $\forall t \in [0,h], \ \int_0^t u(s)\, ds \leq \int_0^t \varpi(s)\, ds$

6. $\varpi$ solves $MxS(C^1, C^2)$ and it is the only solution, modulo the equivalence "= a.e. on I".

* with the same assumption as in the proof of prop. 2 of lemma 1.III.2.

**notation:** $\varpi \equiv \Pi(C^1, C^2)$

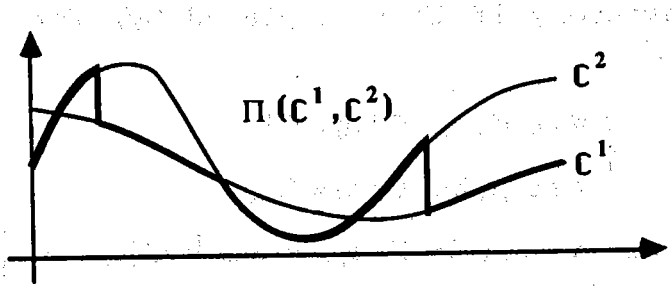



**fig. 2**

**definition:** $\forall f \in \Omega$, $f^t$ is the restriction of $f$ to $[0,t]$.

$\forall t \in [0,h]$, consider the problem defined as:

$$MxF^t(C^1, C^2) \qquad \max_{u \in \Omega} \int_0^t u(s)\,ds$$

$$\text{s.t. } \forall s \in [0,t], \quad \int_0^s u(r)\,dr \leq \int_0^s C^1(r)\,dr$$

$$u(s) \leq C^2(s)$$

The following result is given without proof:

**7.** if $\varpi = \Pi(C^1, C^2)$, then $\forall t \in [0,h]$, $\varpi^t$ solves $MxF^t(C^1, C^2)$.

# THE MULTI-STAGE, MONO-PRODUCT PROBLEM:

## I FORMULATION:

The system considered -in fact, a production line- consists of n single-stage subsystems in series ; each production unit transfers material from a buffer to the next downstream one (the first production unit is externally supplied in raw material and the supply is assumed illimited). The buffers are accumulation points characterized by an inventory holding cost, whereas the production units are flow-limiting devices, characterized by a time-varying capacity. The objective is, as in the single-stage case, to determine the flow-plan (here a sequence of functions $(u^l)_{l=1,..n}$ ) that minimizes the total inventory holding cost.

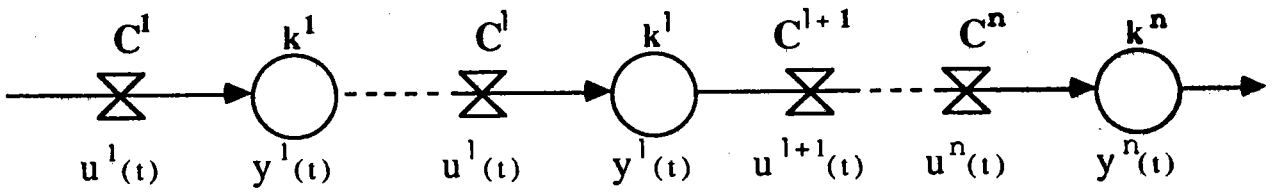



**fig.1.**

The problem to solve can be formulated:

$$MS_0\ ((C^l), d, (y_0^l), (k^l)) \quad \underset{(u^l) \in \Omega^n}{\text{Min}} \sum_{l=1}^{n} k^l \int_0^h y^l(t)\,dt$$

$$\text{s.t.} \quad u^{n+1} \equiv d$$

$$\forall\, l \in \{1,..,n\}\ ,\ \forall\, t \in [0,h]\ ,$$

$$y^l(t) = y_0^l + \int_0^t [u^l(s) - u^{l+1}(s)]\,ds \geq 0$$

$$u^l(t) \leq C^l(t)$$

Or, equivalently:

$$MS_1\ ((C^l), d, (y_0^l), (\kappa^l)) \quad \underset{(u^l) \in \Omega^n}{\text{Min}} \sum_{l=1}^{n} \kappa^l \int_0^h \int_0^t u^l(s)\, ds\, dt$$

$$\text{s.t.} \quad u^{n+1} \equiv d$$

$$\forall\, l \in \{1,..,n\}\ ,\ \forall\, t \in [0,h]\ ,$$

$$y_0^l + \int_0^t u^l(s)\, ds \ \geq \ \int_0^t u^{l+1}(s)\, ds$$

$$u^l(t) \leq C^l(t)$$

$$\text{with:}\ \ \kappa^1 = k^1 \ \ \text{and}\ \ \forall\, l \in \{2,..,n\}\ ,\ \ \kappa^l = k^l - k^{l-1}$$

**note:** this last formulation is directly derived from the definition of the inventory functions $y^l(t)$ and both formulations implicitly rely on the assumption that the measurement unit at any stage $l$ is the quantity required to produce one unit of end product.

## II OPTIMAL FLOW-PLAN:

**theorem 2.II.1:**

Let $(u^l)_{l=1,..,n}$ be a solution to $MS_1\ ((C^l), d, (y_0^l), (\kappa^l))$.

**if** $\exists\, j \in \{1,..,n\}$ and $\kappa^j > 0$ **then** $u^j = \Xi\,(C^j, \Psi\,(u^{j+1}, y_0^j))$ a.e. on I.

which means that $u^j$ will be derived from $u^{j+1}$ through the "backward smoothing" procedure described in the single-stage case.

**interpretation:**

*If the inventory holding costs increase along the production line, the optimal flow plan can be determined by iterating the single-stage "backward smoothing" procedure from the last production stage up.*

proof:

Let $v = \Xi(C^j, \Psi(u^{j+1}, y_0^j))$ and consider the flow-plan $(u'^l)$ defined by:

$$\forall l \in \{1,..,n\},\ l \neq j,\ u'^l = u^l \text{ and } u'^j = v.$$

Since $u^j$ is admissible for the single-stage problem $SS_1(C^j, u^{j+1}, y_0^j)$, lemma 1.III.1 applies and $\int_0^t u^j(s)\,ds \geq \int_0^t v(s)\,ds \geq \int_0^t u^{j+1}(s)\,ds$. Moreover, $(u^l)$ being admissible for $MS_1((C^l), d, (y_0^l))$, $\int_0^t u^{j-1}(s)\,ds \geq \int_0^t u^j(s)\,ds$. Thus $\int_0^t u'^{j-1}(s)\,ds \geq \int_0^t u'^j(s)\,ds \geq \int_0^t u'^{j+1}(s)\,ds$ and $(u'^l)$ is admissible.

Now, the same argument as in the proof of theorem 1.III.1 can be used to prove that either $u^j = v$ a.e. on I or the cost resulting from $(u'^l)$ will be less than that resulting from $(u^l)$:

If $\exists J = ]t_0, t_1[$ , $t_0 \neq t_1$ and $\forall s \in J$ , $u^j(s) \neq v(s)$ then $y \neq z$ a.e. on J and $\int_0^t u^j(s)\,ds > \int_0^t v(s)\,ds$ for almost all t in J. Hence, since $\kappa^j > 0$,

$$\sum_{l=1}^{n} \kappa^l \int_0^h \int_0^t u^l(s)\,ds\,dt > \sum_{l=1}^{n} \kappa^l \int_0^h \int_0^t u'^l(s)\,ds\,dt$$

**theorem 2.II.2:**

**if** $\exists j \in \{1,..,n\}$ and $\kappa^j < 0$ **then** $u^j = \Pi(\Gamma(u^{j-1}, C^j, y_0^j))$ a.e. on I.

interpretation:

*The optimal control for a multi-stage system implies "draining" a buffer at the highest possible rate whenever the next one down the line has a lower cost. Unfortunately, this result is not as directly usable as the previous one, since it only allows to determine the optimal control downwards, whereas the data input to the problem is the final demand.*

proof:

Let $v = \Pi(\Gamma(u^{j-1}, C^j, y_0^j), C^j)$;

→ $u^j$ is admissible for the problem MxS $(u^{j-1}, C^j, y_0^j)$ and, since v solves MxS (see section 1.V.2 ), it follows that $\int_0^h \int_0^t u^j(s)\,ds\,dt \leq \int_0^h \int_0^t v(s)\,ds\,dt$.

→ Moreover, $\forall t \in [0,h]$, both $v^t$ and $u^{jt}$ are admissible for the problem MxF$^t$ $(\Gamma(u^{j-1}, C^j, y_0^j), C^j)$ and $\forall t \in [0,h]$, $\int_0^t v(s)\,ds \geq \int_0^t u^j(s)\,ds$ because $v^t$ is one possible solution to MxF$^t$ ( section 1.V.2 ). Therefore, v will satisfy the equation of conservation of flow in $MS_1$ as $u^j$ does.

The rest of the argument is the same as in the proof of the previous theorem: a modified control $(u'^1)$ would be admissible for $MS_1$ and would yield a stricly lower value of the objective if $\kappa^j < 0$ and $u^j \neq v$. Hence the result.

The direct consequence of the limitation pointed out in the interpretation of this theorem is that there is no analytical solution for systems in which inventory holding costs can decrease along the line. This is illustrated in the next section.

## III THE TWO-STAGE SYSTEM WITH DECREASING COSTS

2S $(C^{1,2}, d, k^{1,2})$

$$\min_{u^1, u^2 \in \Omega} J(u^1, u^2) = k^1 \int_0^h \int_0^t u^1(s)\,ds\,dt - (k^1 - k^2) \int_0^h \int_0^t u^2(s)\,ds\,dt$$

s.t. $\forall t \in [0,h]$,

$$\int_0^t u^1(s)\,ds \geq \int_0^t u^2(s)\,ds \geq \int_0^t d(s)\,ds$$

$$u^1(t) \leq C^1(t) \quad \text{and} \quad u^2(t) \leq C^2(t)$$

One difference with the single stage problem (or the two-stage problem with increasing costs) is that, since $k^1 > k^2$, the coefficient of the second term of the objective is negative. That means that the objective is not simply to load the upstream system so as to keep the least amount of inventory, but rather to combine this objective with that of maximizing the flow of material into the downstream buffer.

Another difference with the single-stage problem is in the expression of the capacity constraint for the whole system. Two particular cases are insightfull in that respect:

**theorem 2.II.1:**

Let $(u^1, u^2)$ be a solution to problem $2S(C^{1,2}, d, k^{1,2})$;

**1. if** $C^1 \geq \Xi(C^2, d)$ **then** $u^1 = u^2 = \Xi(C^2, d)$ a.e. on I.

**2. if** $C^2 \geq \Xi(C^1, d)$ **then** $u^1 = u^2 = \Xi(C^1, d)$ a.e. on I.

proof:

lemma 1:

Let $v^2 = \Xi(C^2, d)$ and $\forall t \in [0,h]$, $z^2(t) = \int_0^t (v^2(s) - d(s))\, ds$,

$$y^2(t) = \int_0^t (u^2(s) - d(s))\, ds.$$

**then** $\forall t \in [0,h]$, $y^2(t) \geq z^2(t)$.

interpretation: the end-product optimal inventory is higher for the two-stage system than it would be for its lower (single-stage) portion with unconstrained supply.

The rationale for this result is that if there existed a control $(u^1, u^2)$ admissible for $2S(C^{1,2}, d, k^{1,2})$ and such that the inequality $y^2 \geq z^2$ were not satisfied, $u^2$ would a fortiori be admissible for $SS(C^2, d)$ and lemma 1.III.1 would be contradicted; (this lemma states that $z^2$ is a lower bound on the inventory functions admissible for the single-stage problem).

This lemma implies that $(k^2 - k^1) \int_0^h \int_0^t v^2(s)\,ds\,dt$ is a lower bound on the objective of $2S(C^{1,2}, d, k^{1,2})$, and that it is achieved by a control $(u^1, u^2)$ if and only if $u^1 = u^2 = v^2$ a.e. on I. If $C^1 \geq v^2$ -and in particular if $C^1 \geq C^2$- this control is clearly admissible and thus it is the only solution to $2S(C^{1,2}, d, k^{1,2})$.

<u>lemma 2:</u> Let $v^1 = \Xi(C^1, d)$ and $\forall t \in [0,h]$, $z^1(t) = \int_0^t (v^1(s) - d(s))\,ds$,

$$y^1(t) = \int_0^t (u^1(s) - u^2(s))\,ds.$$

**then** $\forall t \in [0,h]$, $y^1(t) + y^2(t) \geq z^1(t)$.

<u>interpretation:</u> the work in process in a two-stage system facing a given demand is higher than the inventory there would be in its upper portion facing the same demand (again, the effect of an additional capacity constraint: lemmas 1 and 2 illustrate how inventories are related to capacities).

The explanation of this result is the same as that of lemma 1: in essence, $u^1$ is admissible for $SS(C^1, d)$ if it is for $2S(C^{1,2}, d, k^{1,2})$ and lemma 1.III.1 applies.

As for the second half of the theorem, lemma 2 means that if $k^1 > k^2$, $k^2 \int_0^h \int_0^t v^1(s)\,ds\,dt$ is a lower bound on the objective of $2S(C^{1,2}, d, k^{1,2})$ that can be achieved if and only if $u^1 = u^2 = v^1$ a.e. on I, which is an admissible solution when $C^2 \geq v^1$ -and in particular if $C^2 \geq C^1$-.

Considering these results, one way to determine the loading rate $u^1$ could be to assume that it is the capacity of the upstream subsystem that limits the whole system; in other words, $u^1$ could be taken equal to $\Xi(C^1,d)$ and $u^2$ equal to $\Pi(u^1,C^2)$. Unfortunately, there is no guarantee that $u^2$ will actually meet the demand $d$.

Another way to approach the problem is to use the same loading policy as for a single stage system with a capacity function equal to the inferior envelope $\underline{C}$ of $C^1$ and $C^2$. But this is clearly a very coarse approximation, since it consists of not making use of the intermediate buffer to increase the throughput of the system.

These two solutions are compared in the following example:

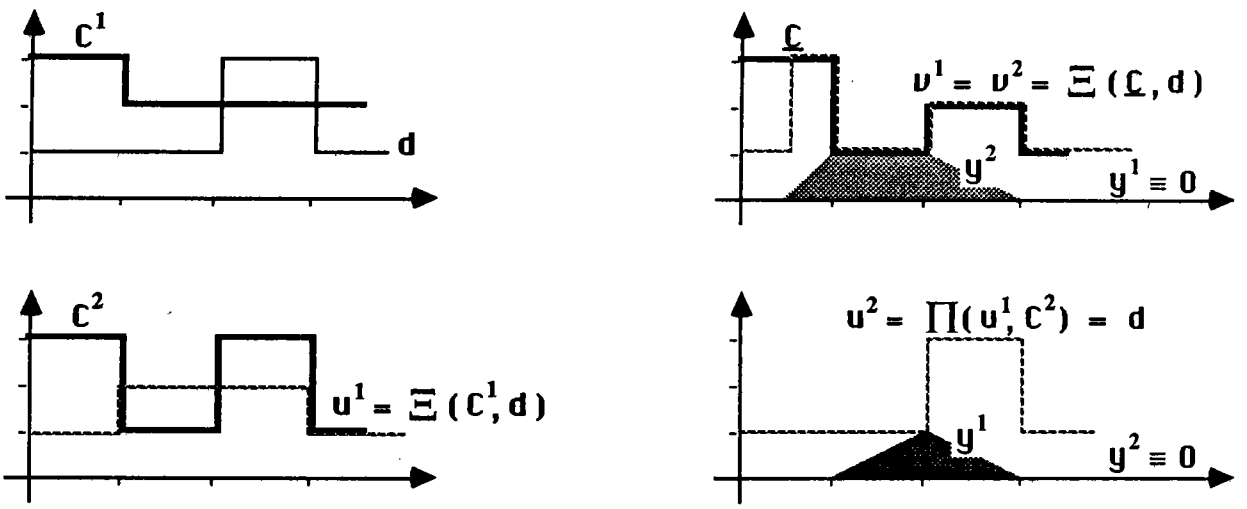



**fig. 3**

This example reveals that, whereas in the case of increasing costs, the relative importance of the cost coefficients had no influence on the optimal policy, this property does not hold when costs are decreasing.

In fact, in the example, the optimal control is $u^1 = u^2 = \Xi(C,d)$ if $k^1 > 1.75\,k^2$, whereas if $k^1 < 1.75\,k^2$, the optimal control is $u^1 = \Xi(C^1,d)$ and $u^2 = \Pi(u^1,C^2) = d$; in this outset, the critical value 1.75 of the ratio $r = k^1/k^2$ is dictated by the characteristics of the demand and capacity functions. It is then easy to imagine that, by cumulating over the horizon of a problem a series of scenari like that studied in this example, one would end up with several critical values of the ratio r and different optimal controls in the different ranges of r.

The conclusion is that there is no analytical solution to multi-stage problems in which costs do not increase with the stage, except in special cases. A numerical method -dynamic programming- providing good solutions at the expense of heavy computations is described in Chapter 5.

## IV NOTE ON THE INVENTORY HOLDING COSTS

Assuming that inventory holding costs increase as the product visits successive machines or production systems can be justified by its increasing added value. This is satisfying if the inventory holding cost is predominantly due the to the capital value of the products. In some cases, however, this assumption cannot be made.

As explained in the formulation of the problem, the unit in which the flows are measured at each stage are not the same physical units. In fact, different units are defined at each stage l, so that producing one "l unit" of product at stage l requires requires exactly one "l-1 unit" of product at stage l-1.

For example, in a single-stage chemical process, if 1 litre of end-product requires 3.6 litres of raw material, the demand will be measured in litres, but the supply will be measured in a unit equivalent to 3.6 litres (call it a gallon), in order to simplify the formulation of the mass-balance equation.

In such cases, it is clear that if the inventory holding cost is mainly due to the cost of the storing equipment or storage space, the inventory holding cost coefficient may be higher for the raw material than for the end product. This is also true in all industries where the raw material is perishable -i.e. requires special equipment- and the end-product is not: canneries or creameries belong to this category.

*The assumption in this work is that the class of multi-stage production systems in which the inventory holding costs increase with the stage is sufficiently large to justify the research conducted.*

# THE SINGLE-STAGE, MULTI-PRODUCT PROBLEM:

## I FORMULATION:

The problem considered is the same as in Part 1, except that there are p different types of goods to produce in order that the expected demands of each of them be satisfied. These goods are produced in a single system of limited capacity, which is expressed in terms of their flow rates through the following inequality:

$$\forall t \in I, \quad \sum_{i=1}^{p} a_i . u_i(t) \leq C(t) \qquad (1)$$

The objective sought at the level for which this model is intended is to minimize the total inventory holding cost: for each product, the inventory holding cost is assumed linear and, if $y_i(t)$ stands for the inventory level of product i at time t and $k_i$ is its inventory holding cost per time unit, the problem to solve can be stated:

$SM_1\ (C, \underline{d}, \underline{k}, \underline{a}, \underline{y}^0)$:

$$\min_{\underline{u} \in \Omega^p} \sum_{i=1}^{p} k_i \int_0^h y_i(t)\, dt$$

s.t. $\forall t \in I,$

$$\forall i \in \{1,..,p\}, \quad y_i(t) = y_i^0 + \int_0^t [u_i(s) - d_i(s)]\, ds \geq 0 \qquad (C_1)$$

$$\sum_{i=1}^{p} a_i . u_i(t) \leq C(t) \qquad (C_2)$$

where the initial inventories $y_i^0$ are given.

**notation:** underscored variables as $\underline{u}$ represent vectors of reals or vectors of functions.

## II SIMPLIFICATION OF THE PROBLEM:

It follows directly from theorem 1.II.1 that problem $SM_1$ can be reformulated with zero initial inventories if the demand functions $d_i$ are replaced by the net-demand functions $\Psi(d_i, y_i^0)$. The purpose of this section is to further simplify the formulation in order that the capacity constraint become:

$$\forall t \in I, \quad \sum_{i=1}^{p} u_i(t) \leq C(t)$$

Consider the problem stated hereunder:

$SM_2(C, \underline{d}', \underline{\kappa})$:

$$\min_{\underline{u}' \in \Omega^p} \sum_{i=1}^{p} \kappa_i \int_0^h y'_i(t)\, dt$$

$$\text{s.t. } \forall t \in I, \quad \forall i \in \{1,..,p\}, \; y'_i(t) = \int_0^t [u'_i(s) - d'_i(s)]\, ds \geq 0 \qquad (C'_1)$$

$$\sum_{i=1}^{p} u'_i(t) \leq C(t) \qquad (C'_2)$$

**theorem 3.II.1:**

$\underline{u}$ solves $SM_1(C, \underline{d}, \underline{k}, \underline{a}, \underline{y}^0)$ **iff** $\underline{a}.\underline{u}$ solves $SM_2(C, \underline{d}', \underline{k}/\underline{a})$,

where $\underline{d}'$ is defined by $\forall i \in \{1,..,p\}$, $d'_i = a_i.\Psi(d_i, y_i^0)$, $\underline{a}.\underline{u} = (a_i.d_i)_{i=1,..,p}$ and $\underline{k}/\underline{a} = (k_i/a_i)_{i=1,..,p}$.

**interpretation:** The capacity constraint becomes simpler if the units in terms of which the flow-rates are measured are modified so that a unit of any product requires the same amount of "capacity"; the only additional requirement is that the cost coefficients be considered in these new units.

proof:

$SM_2(C, \underline{d}', \underline{k}/\underline{a})$ can be stated as:

$$\underset{\underline{u}' \in \Omega^p}{\text{Min}} \sum_{i=1}^{p} k_i / a_i \int_0^h \int_0^t u'_i(s)\, ds\, dt$$

$$\text{s.t. } \forall t \in I, \quad \forall i \in \{1,..,p\}, \quad \int_0^t u'_i(s)\, ds \geq a_i \cdot \int_0^t \Psi(d_i, y_i^0)(s)\, ds \qquad (C'_1)$$

$$\sum_{i=1}^{p} u'_i(t) \leq C(t) \qquad (C'_2)$$

Since, by theorem 1.II.1, $(C'1) \Leftrightarrow \int_0^t u'_i(s)\, ds \geq a_i \cdot \left[ - y_i^0 + \int_0^t d_i(s)\, ds \right]$,

it becomes clear that $\{ SM_2(C, \underline{d}', \underline{k}/\underline{a})$ and $\underline{u}' = \underline{a} \cdot \underline{u} \} \Leftrightarrow SM_1(C, \underline{d}, \underline{k}, \underline{a}, \underline{y}^0)$.

## III PROPERTIES OF THE PRODUCTION PLANS:

definition: a vector of functions $\underline{u} = (u_1,.., u_p) \in \Omega^p$ is called a production plan, flow-plan (or simply plan) and termed admissible if it satisfies the constraints of problem $SM(C, \underline{d}, \underline{k})$ stated hereunder:

$SM(C, \underline{d}, \underline{k})$:

$$\underset{\underline{u} \in \Omega^p}{\text{Min}} \sum_{i=1}^{p} k_i \int_0^h \int_0^t u_i(s)\, ds\, dt$$

$$\text{s.t. } \forall t \in I, \quad \forall i \in \{1,..,p\}, \quad \int_0^t u_i(s)\, ds \geq \int_0^t d_i(s)\, ds \qquad (C_1)$$

$$\sum_{i=1}^{p} u_i(t) \leq C(t) \qquad (C_2)$$

In current section, it is assumed that $SM(C, \underline{d}, \underline{k})$ has a solution $\underline{u}$ and some of its properties are presented.

**theorem 3.III.1:**

**If** $\exists\, t \in I$ and $\exists\, i \in \{1,..,p\}$ such that $y_i(t) > 0$,

**then** $\exists\, \eta > 0$ and $\forall\, j \in \{1,..,p\}$, $k_j < k_i \Rightarrow \forall\, s \in [t-\eta, t+\eta]$, $u_j(s) = 0$.

interpretation: when a product is kept in inventory, there is no production of the ones that have a lower inventory holding cost: these latter products are supplied from inventory.

definition:

$$\forall\, i \in \{1,..,p\},\ \forall\, t \in I,\ \delta_i(t) = u_i(t) - d_i(t) = \dot{y}_i(t).$$

proof:

Since $\delta_i$ is piece-wise continuous, $y_i$ is continuous; therefore, $y_i(t) > 0$ entails that $\exists\, \eta > 0$ and $\forall\, s \in [t-\eta, t+\eta]$, $y_i(s) > 0$.

→ Assume that $\exists\, j \in \{1,..,p\}$, $k_j < k_i$ and $\exists\, s \in ]t-\eta, t+\eta[$ such that $u_j(s) > 0$;

Then, $u_j$ being piece-wise continuous, there exists an interval $L = [z, z+\mu] \subset [t-\eta, t+\eta]$ such that: $\int_z^{z+\mu} u_j(r)\,dr > 0$.

Since $y_i(z) > 0$, there has been overproduction of product i before time z and it is possible to belate this overproduction and instead produce product j in advance: see fig. 2.

→ Define $T = \{ s \in [0, z] \,/\, \exists\, \alpha \in ]0, s]$ and $\forall\, r \in [s-\alpha, s]$, $\delta_i(r) > 0 \}$. T is not empty because $y_i(0) = 0$ and $y_i(z) > 0$. One can thus define:

$x = \mathrm{Sup}(T)$, $\varepsilon = \mathrm{Sup}\{ \alpha \in [0, x] \,/\, \forall\, r \in [x-\alpha, x],\ \delta_i(r) > 0 \}$ and $K = [x-\varepsilon, x]$.

$\varepsilon$ can be reduced in order that: $\int_{x-\varepsilon}^{x} \delta_i(t)\,dt \le \min_{s \in [z, z+\mu]} y_i(s)$ and then,

$\varepsilon$ and $\mu$ can be adjusted so that: $\Delta = \int_{x-\varepsilon}^{x} \delta_i(t)\,dt = \int_z^{z+\mu} u_j(t)\,dt > 0$.

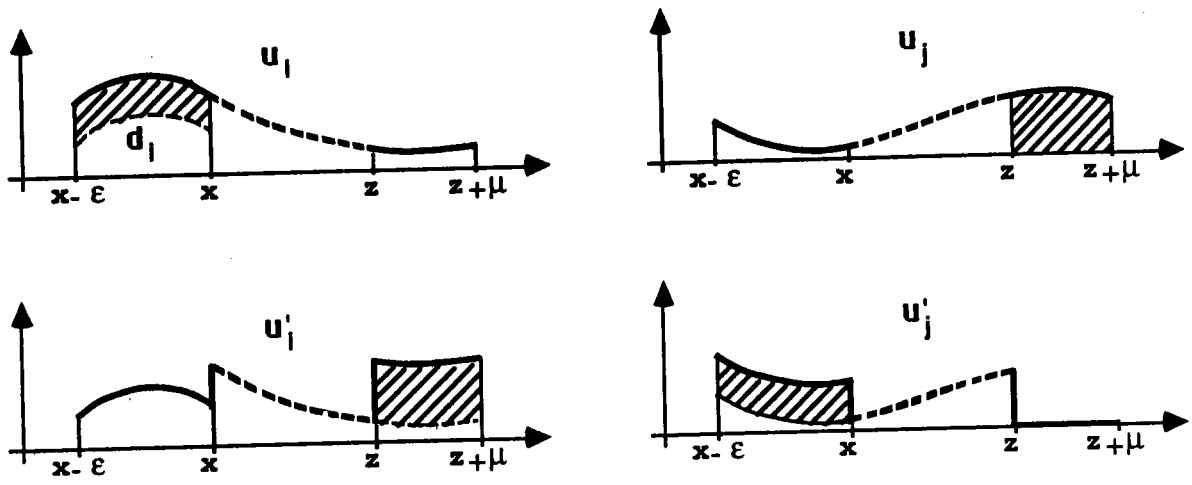



**fig. 2.**

→ Define then the plan $\underline{u}'$ by:

. $\forall k \in \{1,..,p\} \setminus \{i,j\}$ , $u'_k = u_k$ .

. $\forall s \in I \setminus (K \cup L)$ , $\forall k \in \{i,j\}$ , $u'_k(s) = u_k(s)$ .

. $\forall s \in [x-\varepsilon, x]$ , $u'_i(s) = u_i(s) - \delta_i(s)$ and $u'_j(s) = u_j(s) + \delta_i(s)$ .

. $\forall s \in [z, z+\mu]$ , $u'_i(s) = u_i(s) + u_j(s)$ and $u'_j(s) = 0$.

→ This new plan $\underline{u}'$ is admissible:

+ Since $\forall s \in [x-\varepsilon, z+\mu]$ , $u'_i(s) + u'_j(s) = u_i(s) + u_j(s)$ and the other products flows are not modified, the capacity constraint is satisfied.

+ The inventories of products i and j are modified on $[x-\varepsilon, z+\mu]$ only, whereas the other inventories are not modified at all:

. $\forall s \in [x-\varepsilon, x]$ , $y'_i(s) - y_i(s) = -\int_{x-\varepsilon}^{s} \delta_i(r)\,dr$ and $y'_j(s) - y_j(s) = \int_{x-\varepsilon}^{s} \delta_i(r)\,dr$

. $\forall s \in [x, z]$ , $y'_i(s) - y_i(s) = -\Delta$ and $y'_j(s) - y_j(s) = \Delta$

. $\forall s \in [z, z+\mu]$ , $y'_i(s) - y_i(s) = -\int_{s}^{z+\mu} u_j(r)\,dr$ and $y'_j(s) - y_j(s) = \int_{s}^{z+\mu} u_j(r)\,dr$

+ These inventories are still positive: this is obvious for product j since $\forall s \in I$, $y'_j(s) \geq y_j(s)$.

  For product i, $\forall s \in [x-\varepsilon, x]$, $y'_i(s) = y_i(x-\varepsilon) \geq 0$;
  moreover, $x = \mathrm{Sup}(T)$ means that $\forall s \in [x, z]$, $y_i(s) \geq y_i(z)$. Therefore $\mathrm{Min}\{ y_i(s) / s \in [x, z+\mu] \} = \mathrm{Min}\{ y_i(s) / s \in [z, z+\mu] \}$ and it follows from the expressions of $y'_i - y_i$ that $\forall s \in [x, z+\mu]$, $y'_i(s) \geq y_i(s) - \Delta \geq 0$ because $\Delta \leq \mathrm{Min}\{ y_i(s) / s \in [z, z+\mu] \}$.

  Hence, $\forall s \in [x-\varepsilon, z+\mu]$, $y'_i(s) \geq 0$.

→ The difference between the cost K' resulting from plan $\underline{u}'$ and the cost K resulting from $\underline{u}$ is:

$$K' - K = \sum_{l=i,j} k_l \int_{x-\varepsilon}^{z+\mu} [\, y'_l(s) - y_l(s) \,]\, ds$$

$$= (k_j - k_i) \left[ \int_{x-\varepsilon}^{x} \int_{x-\varepsilon}^{t} \delta_i(s)\, ds\, dt + \Delta . (z-x) + \int_{z}^{z+\mu} \int_{t}^{z+\mu} u_j(s)\, ds\, dt \right]$$

Since $k_j < k_i$, this difference is strictly negative and the contradiction that proves theorem 3.III.1 has been found.

<u>assumption:</u>

In the formulation of problem SM (C, $\underline{d}$, $\underline{k}$), the products differ only by their inventory holding costs; hence products with the same costs can be aggregated (i.e. the related variables can be added without loss of information) and it is assumed hereunder that all products have different costs. Moreover, to facilitate the formulation of the following results, the products are ranked by <u>decreasing</u> inventory holding costs: $k_1 > k_2 > \ldots > k_p$.

<u>definition:</u>

$$\forall t \in I\ ,\ \Delta_0(t) = C(t) \text{ and } \forall i \in \{1,..,p\}\ ,\ \Delta_i(t) = C(t) - \sum_{j=1}^{i} u_j(t)\ .$$

**corollary 3.III.1:**

**If** $\exists t \in I$ and $\exists i \in \{1,..,p\}$ such that $y_i(t) > 0$,

**then** $\exists \eta > 0$ and $\forall s \in [t-\eta, t+\eta]$, $\Delta_i(s) = 0$,

which means $\sum_{j=1}^{i} u_j(s) = C(s)$ or, in other terms, $u_i(s) = \Delta_{i-1}(s)$.

<u>interpretation:</u> whenever a product is kept in inventory, the system is working at capacity.

<u>proof:</u> from theorem 3.III.1, it follows that $\exists \eta > 0$ and $\forall j \in \{i+1,..,p\}$, $\forall s \in [t-\eta, t+\eta]$, $u_j(s) = 0$. Hence, $\Delta_i(s) \neq 0$ means that the system is not operated at capacity. If this were true, the passed overproduction of product i (made necessary by the positive inventory at time t) could be delayed and the inventory holding cost reduced: the proof would be identical to that of theorem 3.III.1. Thus, $\forall s \in [t-\eta, t+\eta]$, $\Delta_i(s) = 0$.

Note that the proof of this result could have been lumped with that of theorem 3.III.1, without any significant difficulty or advantage.

**corollary 3.III.2:**

**If** $\exists J = ]t_0, t_1[ \subset I$, $t_0 \neq t_1$ and $\exists i \in \{1,..,p\}$ such that $u_i(t) \neq d_i(t)$ a.e. on J,

**then :** $\mathbf{P_1}$ $\forall t \in J$, $u_i(t) = \Delta_{i-1}(t)$ and thus $\forall j \in \{i+1,..,p\}$, $u_j(t) = 0$.

**If**, additionally, $u_i(t) \neq 0$ a.e. on J,

**then :** $\mathbf{P_2}$ $\forall t \in J$, $\forall j \in \{1,..,i-1\}$, $u_j(t) = d_j(t)$.

<u>interpretation:</u> if an item is produced at a rate that is different from its demand rate, then cheaper items are not produced at all, more expensive items are produced in order to satisfy their demand exactly, and the system is working at maximum rate.

That means in particular that, almost everywhere on I, $\underline{u}$ assumes one of -at most- p+1 values, called *elementary production vectors* and defined hereunder.

**proof:**

$P_1$: based on the fact that: $u_i(t) \neq d_i(t)$ a.e. on $J \Rightarrow y_i(t) > 0$ a.e. on $J$, $P_1$ results directly from theorem 3.III.1.

$P_2$: assume $P_2$ does not hold; then $\exists j \in \{1,..,p\}$, $k_j > k_i$ and $\exists t \in J$ such that $u_j(t) \neq d_j(t)$; $u_j(t)$ being piece-wise continuous, this relation holds on an interval $K \subset J$ and, consequently, $\exists K' \subset K$ and $\forall t \in K'$, $y_j(t) \neq 0$. Application of theorem 3.III.1 yields that $\forall t \in K'$, $u_i(t) = 0$, which contradicts the additional condition $u_i(t) \neq 0$ a.e. on J.

**definitions :**

+ $\forall t \in I$, $\pi(t) = \text{Max}\left\{ i \in \{1,..,p+1\} / \sum_{j=1}^{i-1} d_j(t) \leq C(t) \right\}$ is the number of elementary production vectors at time t.
+ These vectors $\underline{v}^1(t), .., \underline{v}^{\pi(t)}(t)$ can be defined as follows :

$$\forall i \in \{1,..,\pi(t)\}, \quad \begin{cases} \forall j \in \{1,..,i-1\}, & v_j^i(t) = d_j(t) \\ \forall j \in \{i,..,p\}, & v_j^i(t) = \Delta_{j-1}(t) \end{cases}$$

+ Let $\vartheta(I)$ be defined by: $P \in \vartheta(I)$ iff $\exists N \in \mathbb{N}$ and a sequence of sub-intervals $(I_j)_{j=1,..,N}$ such that $P = \{ I_1,.., I_N \}$, and also:

and
1. $\forall j \in \{1,..,N\}$, $I_j \subset I$, $\bigcap_{j=1}^{N} I_j = \emptyset$ and $\bigcup_{j=1}^{N} I_j = I$.
2. $\forall j \in \{1,..,N\}$, $\exists \pi_j \in \{1,..,p+1\}$ such that $\forall t \in I_j$, $\pi(t) = \pi_j$.

$\vartheta(I)$ is the set of partitions of I such that the number of elementary production vectors is constant on each sub-interval.

+ $\Gamma = \left\{ f \in \Omega^p / \exists P_f = \{ I_1,.., I_N \} \in \vartheta(I) \text{ such that:} \right.$
$\left. \forall k \in \{1,..,N\}, \exists i \in \{1,.., \pi_k\} \text{ and } \forall t \in I_k, f(t) = \underline{v}^i(t) \right\}$
is the set of all combinations of the elementary plans.

**summary :** **if** $\underline{u}$ is optimal, **then** $\exists! \underline{v} \in \Gamma$ such that $\underline{u} = \underline{v}$ a.e. on I.

The main result of this part is introduced and proved in the coming section. It means that the optimal plan can be found by solving mono-product problems in sequence. In this process, the first product whose flow is determined is the one with the highest inventory cost and all the capacity of the system is available for it. Then the flow of the second higher-cost product is determined, given that it is limited by the remaining capacity, and the proces goes on.

## IV OPTIMAL PRODUCTION PLAN:

**lemma 3.IV.1:**

if $\underline{y}$ represents the inventory resulting from an optimal production plan $\underline{u}$, then $\underline{y}(h) = \underline{0}$: the final inventory is zero.

proof:

This result is quite obvious and a formal proof would be similar to the proofs presented in the previous section: since the initial inventories are zero, if the final inventory of a product i is not zero, then there must have been overproduction of i during the planning interval and the production plan can be modified to avoid this useless overproduction and yield a lower inventory holding cost.

**theorem 3.IV.1:**

Let $\underline{y}$ be an optimal inventory vector and $i \in \{1,..,p\}$; **if** $\Delta_{i-1}$ is known, **then** $y_i$ is uniquely determined by the following relation:

$$\forall\, t \in I\ ,\ y_i(t) = \int_t^{\beta_t} [\, d_i(s) - \Delta_{i-1}(s)\,]\, ds \quad (1),$$

where $\beta_t = \mathrm{Inf}\,\{\, \tau \in ]t,h]\ /\ y_i(\tau) = 0\,\}$ .

and $u_i = \Xi\,(\Delta_{i-1}, d_i)$ a.e. on $I$ , with the notations of part 1.

**proof:**

$\beta_t$ exists for all t in I because $\underline{y}(h) = \underline{0}$ and it follows from its definition that: $y_i(t) = \int_t^{\beta_t} [\, d_i(s) - u_i(s) \,]\, ds$. Two cases need then be considered:

1- $\beta_t = t$ : then $y_i(t) = 0$ by definition of $\beta_t$ and (1) is satisfied.

2- $\beta_t \neq t$ : then $y_i(t) > 0$ and it follows from corollary 3.III.1 that $u_i(s) = \Delta_{i-1}(s)$ a.e. on $[\, t, \beta_t \,]$ , which proves (1) completely.

The second statement of the theorem follows directly from proposition 4 of lemma 1.III.2.

**interpretation:**

This theorem means that $u_i$ and $y_i$ can be determined by solving the mono-product problem $SS(\Delta_{i-1}, d_i)$ with $\Delta_{i-1}$ as capacity function, that is, with the remaining capacity after the productions of items 1 to i-1 - i.e. items with higher inventory holding cost- have been planned.

**corollary 3.IV.1:**

Since $\Delta_0 = C$ is known, 'the' optimal plan $\underline{u}$ will be determined by solving the monoproduct problems $SS(\Delta_{i-1}, d_i)$ for $i = 1$ to $p$. It will therefore be defined by: $\forall i \in \{1,..,p\}$, $u_i = \Xi(\Delta_{i-1}, d_i)$.

**note:**

Only the ranking of the products inventory holding costs and not their values will affect the optimal flow plan.

**notation:**

This optimal flow-plan will thus be noted either $\underline{u} = \Xi^{\times}(C,\underline{d})$ or $\underline{u} = \Xi^{\times}(C, \underline{d}, \underline{k})$, depending on whether or not the products are assumed ranked by decreasing inventory holding cost.

## V EXISTENCE OF AN OPTIMAL FLOW-PLAN :

The condition for the existence of an optimal flow-plan is the same as for a monoproduct system of equivalent capacity facing a demand equal to the sum of the different products demands.

**theorem 3.V.1:**

Let $j \in \{1,..,p\}$ and assume the functions $u_i = \Xi(\Delta_{i-1}, d_i)_{i=1,..,j-1}$ have been defined;

$$\textbf{if } \forall t \in I\ ,\ \int_0^t C(s)\,ds \geq \sum_{i=1}^{p} \int_0^t u_i(s)\,ds \qquad (A)$$

$$\textbf{then } \forall t \in I\ ,\ \int_0^t \Delta_{j-1}(s)\,ds \geq \int_0^t d_j(s)\,ds$$

interpretation:

If the problems $SS(\Delta_{i-1}, d_i)_{i=1,..,j-1}$ have been solved and C and $\underline{d}$ satisfy condition (A), then, whatever the cost vector $\underline{k}$, the problem $SS(\Delta_{j-1}, d_j)$ is solvable.

proof:

What needs to be proved is: $\forall t \in I\ ,\ \int_0^t d_j(s)\,ds + \sum_{i=1}^{j-1} \int_0^t u_i(s)\,ds \leq \int_0^t C(s)\,ds$

Let $t \in I$ and $\beta_t = \text{Inf}\{\tau \in ]t,h] \ /\ \forall i \in \{1,..,j-1\}\ y_i(\tau) = 0\}$.

then $\forall i \in \{1,..,j-1\}\ ,\ \int_0^t u_i(s)\,ds = \int_0^t d_i(s)\,ds$ (1)

and $\forall s \in [t, \beta_t]$, since $\exists i \in \{1,..,j-1\}$ such that $y_i(s) > 0$, $\sum_{i=1}^{j-1} u_i(s) = C(s)$ (2)

(this follows from corollary 3.III.1.)

Hence, $\int_0^t C(s)\,ds = \int_0^{\beta_t} C(s)\,ds - \int_0^{\beta_t} \left[ \sum_{i=1}^{j-1} u_i(s) \right] ds$ because of (2).

$$\geq \sum_{i=1}^{p} \int_0^{\beta_t} d_i(s)\,ds - \sum_{i=1}^{j-1} \int_t^{\beta_t} u_i(s)\,ds$$ follows from (A).

$$\geq \sum_{i=1}^{j} \int_0^{\beta_t} d_i(s)\,ds - \sum_{i=1}^{j-1} \int_t^{\beta_t} u_i(s)\,ds$$ because the functions $d_i$ are positive.

following from (1), this last term is equal to: $\int_0^{\beta_t} d_i(s)\,ds + \sum_{i=1}^{j-1} \int_0^t u_i(s)\,ds$

which implies, (again, because $d_i \geq 0$): $\int_0^t C(s)\,ds \geq \int_0^t d_i(s)\,ds + \sum_{i=1}^{j-1} \int_0^t u_i(s)\,ds$

**corollary 3.V.1:**

Problem $SM(C, \underline{d}, \underline{k})$ is solvable **iff** $\forall t \in I$, $\int_0^t C(s)\,ds \geq \sum_{i=1}^{p} \int_0^t u_i(s)\,ds$ (A)

<u>proof:</u>

+ (A) is clearly necessary, since it follows directly from the constraints $(C_1)$ and $(C_2)$ of problem $SM(C, \underline{d}, \underline{k})$:

$(C_1) \Rightarrow \forall t \in I,\ \sum_{i=1}^{p} \int_0^t u_i(s)\,ds \geq \sum_{i=1}^{p} \int_0^t d_i(s)\,ds$

$(C_2) \Rightarrow \forall t \in I,\ \int_0^t \left[ \sum_{i=1}^{p} u_i(s) \right] ds \leq \int_0^t C(s)\,ds$

Hence $(C_1)$ and $(C_2) \Rightarrow$ (A).

+ (A) is also sufficient: namely, it was proved in part 1 that problem $SS(C, d)$ has a solution if and only if:

$$\forall t \in I\ ,\ \int_0^t C(s)\,ds \geq \int_0^t d(s)\,ds$$

Moreover, theorem 3.IV.1 states that if the problems $SS(\Delta_{i-1}, d_i)$ can be solved in sequence for $i = 1$ to $p$, then $SM(C, \underline{d}, \underline{k})$ is solved.

Since $\Delta_0 = C$, (A) implies that $\forall t \in I\ ,\ \int_0^t \Delta_0(s)\,ds \geq \int_0^t d_1(s)\,ds$ and hence that $SS(\Delta_0, d_1)$ is solvable; therefore, a simple induction on theorem 3.V.1 proves that problems $SS(\Delta_{i-1}, d_i)_{i=1,\ldots,p}$ are solvable.

## VI PROPERTIES OF $\Xi^x$:

### notation:

let $j \in \{1,..,p\}$, and define $\Sigma_j$ and $\Sigma^j$ as follows:

$$\forall \underline{v} \in \Omega^p,\ \forall t \in I,\ \Sigma^j(\underline{v})(t) = \sum_{i=1}^{j} v_i(t) \quad \text{and} \quad \Sigma_j(\underline{v})(t) = \sum_{i=j}^{p} v_i(t)$$

moreover, $\forall i,j \in \{1,..,p\}, j<i$, $\Sigma_j^i$ will denote $\Sigma_j - \Sigma_{i+1} = \Sigma^i - \Sigma^{j-1}$

(with the assumption that $\Sigma_{p+1} = \Sigma^0 = 0$).

**theorem 3.VI.1:**

**if** $\underline{u} = \Xi^x(C,\underline{d})$ , **then** $\forall\ j \in \{1,..,p\}$, $\Sigma^j(\underline{u}) = \Xi(C, \Sigma^j(\underline{d}))$ a.e. on I.

### interpretation:

If one considers the mono-product problem resulting of the aggregation of several products (of successive rank if ranked by decreasing inventory holding cost), the optimal flow-plan is the aggregation of the corresponding flows determined by solving the multi-product problem.

proof: by induction.

+ $\Sigma^1(\underline{u}) = u_1 = \Xi(\Delta_0, d_1) = \Xi(C, d_1) = \Xi(C, \Sigma^1(\underline{d}))$.

+ let $j \in \{1,..,p\}$ and assume $\Sigma^j(\underline{u}) = \Xi(C, \Sigma^j(\underline{d}))$ a.e. on I;

$u_{j+1} = \Xi(\Delta_j, d_{j+1}) = \Xi(C - \Sigma^j(\underline{u}), d_{j+1})$ by definition of $\Delta_j$.

Thus, using proposition 1 of theorem 1.V.2:

$$\begin{aligned} u_{j+1} + \Sigma^j(\underline{u}) - d_{j+1} &= \Xi(C - d_{j+1}, \Sigma^j(\underline{u})) \\ &= \Xi(C - d_{j+1}, \Xi(C, \Sigma^j(\underline{d}))) \\ &= \Xi(C - d_{j+1}, \Sigma^j(\underline{d})) \quad \text{by proposition 3.} \end{aligned}$$

Therefore, using proposition 1 again:

$$\Sigma^{j+1}(\underline{u}) = \Xi(C, \Sigma^j(\underline{d}) + d_{j+1}) = \Xi(C, \Sigma^{j+1}(\underline{d})).$$

**note:** this result, together with corollary 1.III.4 (necessary and sufficient condition on the difference between capacity and demand functions for the single stage, monoproduct problem to have a solution) would provide a staightforward proof for corollary 3.V.1.

**corollary 3.VI.1:**

let $\underline{u} = \Xi^{x}(C, \underline{d})$ and $\underline{d}' \in \Omega^{p}$ such that:

$$\exists\, i, j \in \{1,..,p\},\ j<i \quad \text{and} \quad \begin{cases} \forall\, k \in \{1,..,p\}\backslash\{j,..,i\}\ ,\ d'_k = d_k \\ \Sigma_j^i(\underline{d}') = \Sigma_j^i(\underline{d}) \end{cases}$$

$$\textbf{if } u' = \Xi^{x}(C, d'), \textbf{ then} \begin{cases} \forall\, k \in \{1,..,p\}\backslash\{j,..,i\}\ ,\ u'_k = u_k \\ \Sigma_j^i(\underline{u}') = \Sigma_j^i(\underline{u}) \end{cases}$$

interpretation:

If two demand vectors are such that there is a group of products whose cumulated demand is the same in both vectors, and if the individual demands of more expensive products are equal, then the optimal flow-plan has the same characteristics.

proof:

+ $\forall\, k \in \{1,.., j-1\}$, $\Sigma^k(\underline{d}') = \Sigma^k(\underline{d})$ and, following from theorem 3.VI.1, $\Sigma^k(\underline{u}') = \Sigma^k(\underline{u})$ , which yields that $u'_k = u_k$ (because $\Sigma^1(\underline{u}) = u_1$).

+ by definition, $\Sigma^i(\underline{d}') = \Sigma^i(\underline{d})$, and thus $\Sigma^i(\underline{u}') = \Sigma^i(\underline{u})$; hence, by difference, $\Sigma_j^i(\underline{u}') = \Sigma^i(\underline{u}') - \Sigma^{j-1}(\underline{u}') = \Sigma^i(\underline{u}) - \Sigma^{j-1}(\underline{u}) = \Sigma_j^i(\underline{u})$.

+ $\forall\, k \in \{i+1,.., p\}$, $\Sigma^k(\underline{d}') = \Sigma^k(\underline{d})$ and $\Sigma^k(\underline{u}') = \Sigma^k(\underline{u})$; therefore, since $\Sigma^i(\underline{u}') = \Sigma^i(\underline{u})$, $u'_k = u_k$.

# THE MULTI-STAGE, MULTI-PRODUCT PROBLEM

## I FORMULATION

The system modelled here consists of a sequence of single-stage, multi-product sub-systems of the type considered in the previous part. The problem to solve can be stated, with self-explanatory notations:

$MM_1((C^l), \underline{d}, (\underline{k}^l), (\underline{a}^l), (\underline{y}^{l\circ}))$:

$$\underset{(\underline{u}^l) \in \Omega^{p \times n}}{\text{Min}} \sum_{l=1}^{n} \sum_{i=1}^{p} k^l_i \int_0^h y^l_i(t)\, dt \qquad (O^1)$$

s.t. $\forall i \in \{1,..,p\}$, $u_i^{n+1} \equiv d_i$

$\forall i \in \{1,..,p\}$, $\forall l \in \{1,..,n\}$, $\forall t \in I$, $y^l_i(t) = y^{l\circ}_i + \int_0^t [u^l_i(s) - u_i^{l+1}(s)]\, ds \geq 0$

$\forall l \in \{1,..,n\}$, $\forall t \in I$, $\sum_{i=1}^{p} a^l_i . u^l_i(t) \leq C^l(t)$

where the initial inventories $y^{l\circ}_i$ are given.

<u>note:</u> the objective of $MM_1((C^l), \underline{d}, (\underline{k}^l), (\underline{a}^l), (\underline{y}^{l\circ}))$ can be rewritten:

$$\underset{(\underline{u}^l) \in \Omega^{p \times n}}{\text{Min}} \sum_{l=1}^{n} \sum_{i=1}^{p} \kappa^l_i \int_0^h \int_0^t u^l_i(s)\, ds\, dt \qquad (O^2)$$

where $\forall i \in \{1,..,p\}$, $\kappa^1_i = k^1_i$ and $\forall j \in \{2,..,n\}$, $\kappa^j_i = k^j_i - k^{j-1}_i$.

## II PROPERTIES OF THE OPTIMAL CONTROL

**theorem 4.II.1**

Let $(\underline{u}^l)$ be a solution to $MM_1((C^l), \underline{d}, (\underline{k}^l), (\underline{a}^l), (\underline{y}^{l\circ}))$ and $j \in \{1,..,n\}$;

**if** there exists an interval $L \subset I$, $L \neq \emptyset$, and a product $i \in \{1,..,p\}$ such that $\forall t \in L$, $u^j_i(t) \notin \{0, u_i^{j+1}(t)\}$,

**then 1-** $\sum_{k=1}^{p} a^j_k . u^j_k(t) = C^j(t)$ a.e. on L.

**and 2-** $\forall k \in \{1,..,p\}$, $\{ u^j_k(t) \in \{0, u_k^{j+1}(t)\}$ a.e. on L $\}$ or $\{ \kappa^j_k / a^j_k = \kappa^j_i / a^j_i \}$.

**interpretation**

If at a given stage, the products have all different ratios of cost increment to capacity coefficient then if there is a product whose flow-rate is neither equal to zero nor to its demand at that stage, then there is only one, and the stage is working at capacity.

**notations** directly derived from those of part 3:

$$\forall j \in \{1,..,n\}, \forall i \in \{1,..,p\}, \forall t \in I, \quad \delta_i^j(t) = u_i^j(t) - d_i^j(t) = y_i^j(t).$$

$$\Delta_i^j(t) = C^j(t) - \sum_{k=1}^{i} a_k^j \cdot u_k^j(t).$$

**proof**

Both propositions can be proved by contradiction: they are assumed false and the optimality of $(\underline{u}^1)$ is contradicted by the existence of a plan $(\underline{u}'^1)$ resulting in a lower value of the objective function.

**Proposition 1:**

$$\sum_{k=1}^{p} a_k^j \cdot u_k^j(t) = C^j(t) \quad \text{is equivalent to} \quad \Delta^j{}_p(t) = 0.$$

$\longrightarrow$ Assume there exists an interval $K \subset L$ such that $\Delta^j{}_p(t) > 0$ on $K \neq \emptyset$. $\longleftarrow$

1- Since $\forall t \in L$ -and thus $\forall t \in K$-, $u_i^j(t) \notin \{0, u_i^{j+1}(t)\}$, it follows from the piece-wise continuity of these functions that:

$$\exists J \subset K,\ J \neq \emptyset, \text{ such that } \forall t \in J,\ u_i^j(t) > u_i^{j+1}(t) \geq 0,$$

or $\exists J \subset K,\ J \neq \emptyset$, such that $\forall t \in J,\ 0 < u_i^j(t) < u_i^{j+1}(t)$.

In both cases, $\exists [z, z+\mu] \subset J$, $\mu > 0$, such that $\forall t \in [z, z+\mu]$, $u_i^j(t) > 0$ and

$$0 < \int_z^{z+\mu} u_i^j(s)\, ds < y_i^j(z).$$

In fact, in the first case, inventory is built up during the interval J and, in the second case, it is depleted; therefore, in both cases it must be positive at some point in J. It can be assumed without loss of generality that $J = [z, z+\mu]$.

2- Then $\exists \varepsilon \in ]0,\mu[$ such that $0 < \int_z^{z+\varepsilon} u_i^j(s)\,ds = 1/a_i^j \cdot \int_{z+\varepsilon}^{z+\mu} \Delta j_p(s)\,ds < y_i^j(z)$.

3- It is thus possible to define a new control $(\underline{u}'^l)$ such that:

- $\forall l \neq j,\ \underline{u}'^l = \underline{u}^l$.
- $\forall k \in \{1,\dots,p\}\backslash\{i\},\ u'^j_k = u^j_k$.
- $\forall t \in I\backslash[z, z+\mu],\ u'^j_i(t) = u^j_i(t)$.
- $\forall t \in [z, z+\varepsilon],\ u'^j_i(t) = 0$.
- $\forall t \in [z+\varepsilon, z+\mu],\ u'^j_i(t) = u^j_i(t) + \Delta j_p(t) / a^j_i$.

4- This control is admissible for problem $MM_1((C^l), \underline{d}, (\underline{k}^l), (\underline{a}^l), (\underline{y}^{l\circ}))$. In fact, on $[z+\varepsilon, z+\mu]$, $u'^j_i(t)$ is defined so that exactly all the capacity is utilized; on the rest of I, the flows $\underline{u}'^l$ are less or equal than the flows $\underline{u}^l$.

Also, $(\underline{u}'^l)$ satisfies the constraint of positive inventories, because the inventory $y^j_i(z)$ of product i is sufficient to allow for the reduction of production over $[z, z+\varepsilon]$, and because the inventory trajectories are not modified outside of J (this results from the fact that $y'^j_i(z+\mu) = y^j_i(z+\mu)$) .

5- The difference in the costs resulting from $(\underline{u}'^l)$ and $(\underline{u}^l)$ is:

$$K - K' = \int_z^{z+\mu} \int_z^t \kappa^j_i (u^j_i(s) - u'^j_i(s))\,ds\,dt$$

$$= \int_z^{z+\varepsilon} \int_z^t \kappa^j_i (u^j_i(s) - u'^j_i(s))\,ds\,dt - \int_{z+\varepsilon}^{z+\mu} \int_t^{z+\mu} \kappa^j_i (u^j_i(s) - u'^j_i(s))\,ds\,dt$$

$$= \kappa^j_i \cdot \int_z^{z+\varepsilon} \int_z^t u^j_i(s)\,ds\,dt + \kappa^j_i / a^j_i \cdot \int_{z+\varepsilon}^{z+\mu} \int_t^{z+\mu} \Delta j_p(s)\,ds\,dt$$

Given that both $u^j_i$ and $\Delta j_p$ are strictly positive on J, this difference is strictly positive, which contradicts the optimality of the plan $(\underline{u}^l)$. Q.E.D.

**Proposition 2:**

→ Assume $\exists k \in \{1,..,p\}$ such that $\{ u_k^j(t) \in \{0, u_k^{j+1}(t)\}$ a.e. on $L\}$ is false. ←

1- This means there exists an interval $K \subset L$, $K \neq \emptyset$, such that $\forall t \in K$, $u_k^j(t) \notin \{0, u_k^{j+1}(t)\}$. The same conclusions as for $u_i^j$ can be drawn, and in particular: $\exists J \subset K$, $J \neq \emptyset$, such that $\forall t \in J$, $u_k^j(t) > 0$ and $y_k^j(t) > 0$.

2- Since $\forall t \in J$, $u_i^j(t) \notin \{0, u_i^{j+1}(t)\}$, $\exists [z, z+\mu] \subset J$, $\mu > 0$, such that:

$$\forall t \in [z, z+\mu],\ u_i^j(t) > 0 \text{ and } \int_z^{z+\mu} u_i^j(s)\,ds < y_i^j(z).$$

And, since $z \in J$, $y_k^j(z) > 0$ and one can reduce $\mu$ in order that:

$$\int_z^{z+\mu} u_k^j(s)\,ds < y_k^j(z) \quad \text{also.}$$

3- Then $\exists \varepsilon \in ]0,\mu[$ such that $0 < \int_z^{z+\varepsilon} u_i^j(s)\,ds = \int_{z+\varepsilon}^{z+\mu} a_k^j/a_i^j \cdot u_k^j(s)\,ds < y_i^j(z)$.

4- It is thus possible to define a new control $(\underline{u}'^l)$ such that:

- $\forall l \neq j$, $\underline{u}'^l = \underline{u}^l$.
- $\forall \iota \in \{1,..., p\}\backslash\{i,k\}$, $u'^j_\iota = u^j_\iota$.
- $\forall \iota \in \{i,k\}$, $\forall t \in I \backslash [z, z+\mu]$, $u'^j_\iota(t) = u^j_\iota(t)$.
- $\forall t \in [z, z+\varepsilon]$, $u'^j_i(t) = 0$ ; $u'^j_k(t) = u^j_k(t) + a_i^j/a_k^j\ u_i^j(t)$.
- $\forall t \in [z+\varepsilon, z+\mu]$, $u'^j_k(t) = 0$ ; $u'^j_i(t) = u_i^j(t) + a_k^j/a_i^j\ u_k^j(t)$.

5- Since $\forall t \in I$, $a_i^j u'^j_i + a_k^j u'^j_k = a_i^j u_i^j + a_k^j u_k^j$, $(\underline{u}'^l)$ satisfies the capacity constraint.

Similarly, since $\forall \iota \in \{i,k\}$, $\int_z^{z+\mu} u'^j_\iota(s)\,ds = \int_z^{z+\mu} u_\iota^j(s)\,ds$, it follows that $\forall \iota \in \{i,k\}$, $\forall t \in I \setminus [z, z+\mu]$, $y'^j_\iota(t) = y_\iota^{j+1}(t)$. As the inventory of product i at time z is sufficient to make for the subsequent reduction in production, $(\underline{u}'^l)$ satisfies the inventory constraint.

6- The difference in the costs resulting from $(\underline{u}'^{j})$ and $(\underline{u}^{j})$ is:

$$K - K' = \int_z^{z+\mu} \int_z^t [\, \kappa_i^j (u_i^j(s) - u'^j_i(s)) + \kappa_k^j (u_k^j(s) - u'^j_k(s)) \,] \, ds \, dt$$

$$= \int_z^{z+\varepsilon} \int_z^t [\, \kappa_i^j . u_i^j(s) - \frac{a_i^j}{a_k^j} \kappa_k^j . u_i^j(s) \,] \, ds \, dt - \int_{z+\varepsilon}^{z+\mu} \int_t^{z+\mu} [\, \kappa_k^j . u_k^j(s) - \frac{a_k^j}{a_i^j} \kappa_i^j . u_k^j(s) \,] \, ds \, dt$$

$$= \left[ \frac{\kappa_i^j}{a_i^j} - \frac{\kappa_k^j}{a_k^j} \right] \left[ a_i^j \int_z^{z+\varepsilon} \int_z^t u_i^j(s) \, ds \, dt + a_k^j \int_{z+\varepsilon}^{z+\mu} \int_t^{z+\mu} u_k^j(s) \, ds \, dt \right] \equiv \Delta K_i$$

It follows from the symmetry of the expression in terms of i and k that, had it been chosen to decrease the flow of k over $[z, z+\varepsilon]$ instead of decreasing that of i, the difference in costs would have been:

$$K - K' = \left[ \frac{\kappa_k^j}{a_k^j} - \frac{\kappa_i^j}{a_i^j} \right] \left[ a_k^j \int_z^{z+\varepsilon} \int_z^t u_k^j(s) \, ds \, dt + a_i^j \int_{z+\varepsilon}^{z+\mu} \int_t^{z+\mu} u_i^j(s) \, ds \, dt \right] \equiv \Delta K_k$$

The control $(\underline{u}^{j})$ being optimal, both $(\Delta K_i)$ and $(\Delta K_k)$ are negative. Since the second factor in both expressions is strictly positive, $\Delta K_i \leq 0$ and $\Delta K_k \leq 0$ implies that the first factor is also negative, which means:

$$\frac{\kappa_i^j}{a_i^j} - \frac{\kappa_k^j}{a_k^j} \leq 0 \quad \text{and} \quad \frac{\kappa_k^j}{a_k^j} - \frac{\kappa_i^j}{a_i^j} \leq 0 \text{ , that is, } \frac{\kappa_k^j}{a_k^j} = \frac{\kappa_i^j}{a_i^j}.$$

<u>Conclusion:</u> for a given stage j, if the demand $\underline{u}^{j+1}$ defined by the flows at the next stage is known and if the products have all different ratios $\kappa_i^j/a_i^j$, the control $\underline{u}^j$ can take only a finite number of values at any given time.

The number of values it can take is at most $(p+2).2^{p-1}$ because there are at most $2^p$ controls in which the flow of every product i takes its values in $\{0, u_i^{j+1}\}$, and $p.2^{p-1}$ other controls.

In fact, once the only 'particular' product whose flow is not in $\{0, u_i^{j+1}\}$ is chosen, the value of its flow is entirely determined by the values of the other flows, since they must add up to the capacity; there are p candidates for the position of 'particular' product, and at most $2^{p-1}$ values for the rest of the control, hence $p.2^{p-1}$ controls.

Comparison with the single-stage case:

For a single-stage system, it has been proved that the 'particular' product at a given time is also the most 'expensive' product whose inventory is strictly positive, 'expensive' meaning 'with high ratio $\kappa_i^j/a_i^j$ of cost increment to capacity coefficient'. It has also been proved that the flows of cheaper products are zero and the flows of more expensive products are equal to their demand. Given this result, the number of possible controls at any given time is reduced to p+1.

All the results leading to this tighter characterization of the optimal control were derived by considering the impact on the objective of a local modification of the control. In a single-stage system, the supply to the system being unconstrained, the only concern when modifying the flow of a product -within the limits defined by the capacity- was to ensure that demand would still be met. In the case of a multi-stage system, the supply is bounded by the flows of upstream stages. One must therefore ensure that either the modification of a flow can be 'absorbed' by the next upstream stage, or that the impact on the objective due to the propagation of the modification does not counterbalance the effect of the local modification.

The strong results of the single-stage case were obtained by advancing the production of an item of low inventory holding cost in order to belate that of a more 'expensive' product. Such a modification cannot be achieved locally in a multi-stage system. Therefore, these strong results will be derived only for a particular class of systems, for which it is possible to measure the effect of propagating a modification.

## III SIMPLIFIED PROBLEM

There is no straightforward method to simplify the formulation of $MM_1((C^l), \underline{d}, (\underline{k}^l), (\underline{a}^l), (\underline{y}^{l^\bullet}))$ so as to avoid initial inventories and capacity coefficients, as there was in the single-stage case.

In fact, it is not possible in general* to find units to measure the different flows so as to simplify the capacity constraints, and there is no equivalent to the 'net' demand: the initial inventories existing in all but the last stage cannot be made available instantaneously for end-product demand because of the intermediate sub-systems limit the flow, and determining the most cost-effective way to 'transfer' them is of the same order of difficulty as solving the problem $MM_1$ with zero inventories.

The scope of this section is therefore limited to problem $MM((C^l), d, (k^l))$, that is, the simplified version of $MM_1$ with zero initial inventories and no capacity coefficients. The system considered will also be particularized by the cost assumptions required to derive the same type of results as in part 3.

*<u>note:</u> the capacity constraint can be simplified in two particular cases: when the coefficients $a^l_i$ are independent of l, which means that the ratios of the processing times of different products are the same at all stages, and when the coefficients $a^l_i$ are independent of i, which means that the processing times at a given stage are the same for all the products.

In the first case, the flow of every product i will be measured in a unit equal to $a^l_i = a_i$ times the base unit; in the second case, the capacity functions $C^l$ will be divided by $a^l_i = a^l$, that is, the capacity of stage l will be measured in a unit equal to $a^l$ base units.

cost assumptions:

1- the inventory holding cost of a product increases with the stage at which it is considered: the further down in the line, the higher the cost:

$$\forall i \in \{1,..,p\} \ , \ k^1_i < k^2_i < \cdots < k^n_i$$

2- the products can still be ranked by increasing inventory holding cost, regardless of the stage at which they are considered:

$$\forall l \in \{1,..,n\} \ , \ k^l_1 > k^l_2 > \cdots > k^l_p$$

3- the increment in cost from stage to stage is higher for more expensive products:

$$\forall l \in \{1,..,n\} \ , \ \kappa^l_1 > \kappa^l_2 > \cdots > \kappa^l_p$$

A particular cost structure satisfying these conditions (and previously introduced in GABBAY [GA]) would be one in which each stage l and each product i can be associated respectively cost coefficients $\beta^l$ and $\alpha_i$ such that:

$$\alpha_1 > \cdots > \alpha_p \ , \ \beta^1 < \cdots < \beta^n \text{ and } \forall l \in \{1,..,n\}, \forall i \in \{1,..,p\} \ , \ k^l_i = \alpha_i . \beta^l . \quad \textbf{(CS)}$$

note: From a practical point of view, the second and third cost assumptions can appear too restrictive: if the cost coefficients are inventory holding costs "stricto sensu" and are assumed proportional to the value of the product at a given stage, then the value added at a given stage could perfectly alter the ranking of the costs and the second assumption would not hold.

On the other hand, if the cost coefficients just represent some priorities for holding inventory, then the cost structure described by (CS) is perfectly adequate since it combines these priorities with the first cost assumption, which is realistic in a wide class of problems.

**theorem 4.III.1:**

Let $(\underline{u}^j)$ be an optimal flow-plan;

**If** $\exists i \in \{1,..,p\}$ and $\exists t \in I$ such that $y_i^l(t) > 0$,

**then** $\exists \eta > 0$ such that $\forall s \in [t-\eta, t+\eta]$, $\Delta^l_p(s) = 0$.

<u>interpretation:</u> if there is some inventory kept at a stage of the system, then this stage is working at capacity.

<u>proof:</u>

This proof is, in essence, very similar to that of theorem 3.III.1:

$y_i^l$ being continuous, $y_i^l(t) > 0 \Rightarrow \exists \eta > 0$ and $\forall s \in [t-\eta, t+\eta]$, $y_i^l(s) > 0$.

**(H)** assume $\exists z \in ]t-\eta, t+\eta[$, $\exists \mu > 0$ such that $L = [z, z+\mu] \subset [t-\eta, t+\eta]$ and that $\int_z^{z+\mu} \Delta^l_p(r)\,dr > 0$. Since $y_i^l(z) > 0$ and $y_i^l(0) = 0$ the following definitions are valid: $x = \text{Sup}\left\{ r \in [0,z] \,/\, \exists \alpha \in ]0,r] \text{ and } \forall s \in [r-\alpha, r],\ \delta^l_i(s) > 0 \right\} \neq \emptyset$,

$$\varepsilon = \text{Sup}\left\{ \alpha \in ]0,x] \,/\, \forall s \in [x-\alpha, x],\ \delta^l_i(s) > 0 \right\} \quad \text{and} \quad K = [x-\varepsilon, x].$$

$\varepsilon$ can be reduced in order that: $\int_{x-\varepsilon}^{x} \delta^l_i(t)\,dt \leq \underset{s \in [z, z+\mu]}{\text{Min}} y^l_i(s)$ and then,

$\varepsilon$ and $\mu$ can be adjusted so that: $\int_{x-\varepsilon}^{x} \delta^l_i(t)\,dt = \int_z^{z+\mu} \Delta^l_p(t)\,dt = Q > 0.$

A new plan $(\underline{u}'^l)$ can thereafter be constructed, that will yield a lower value of the objective function:

- $\forall j \in \{1,.., n\} \setminus \{l\}$, $\underline{u}'^j = \underline{u}^j$.
- $\forall k \in \{1,.., p\} \setminus \{i\}$, $u'^l_k = u^l_k$.
- $\forall s \in I \setminus K \cup L$, $u'^l_i(s) = u^l_i(s)$.
- $\forall s \in K$, $u'^l_i(s) = u^l_i(s) - \delta^l_i(s)$.
- $\forall s \in L$, $u'^l_i(s) = u^l_i(s) + \Delta^l_p(s)$.

This plan satisfies the constraints of the problem, and the difference in inventory holding cost with respect to $(\underline{u}^l)$ is:

$$K - K' = \sum_{j=1}^{n} \sum_{k=i}^{p} \kappa_k^j \int_0^h \int_0^t (u_k^j - u'^j_k)(s)\, ds\, dt$$

$$= \kappa^l_i \left[ \int_{x-\varepsilon}^{x} \int_{x-\varepsilon}^{t} \delta^l_i(s)\, ds\, dt + Q(z-x) + \int_z^{z+\mu} \int_t^{z+\mu} \Delta^l_p(s)\, ds\, dt \right] > 0.$$

This would contradict the optimality of $(\underline{u}^l)$; therefore the initial assumption is false and $\forall s \in [t-\eta, t+\eta]$, $\Delta^l_p(s) = 0$.

**theorem 4.III.2:**

Let $(\underline{u}^j)$ be an optimal flow-plan such that:

$$\exists l \in \{1,..,n\} \text{ and } \forall j \in \{1,..,l-1\},\ \underline{u}^j = \Xi^{\times}(C^j, \underline{u}^{j+1}). \qquad \textbf{(HR}^l\textbf{)}$$

**If** $\exists i \in \{1,..,p\}$ and $\exists t \in I$ such that $y_i^l(t) > 0$,

**then** $\exists \eta > 0$ such that $\forall s \in [t-\eta, t+\eta]$, $\forall k \in \{i+1,..,p\}$, $u^l_k(s) = 0$.

interpretation: under the hypothesis **(HR$^l$)**, if there is a positive inventory of product i at stage l, then cheaper products are supplied from inventory to the downstream stages and not produced.

proof:

$y_i^l$ being continuous, $y_i^l(t) > 0 \Rightarrow \exists \eta > 0$ and $\forall s \in [t-\eta, t+\eta]$, $y_i^l(s) > 0$.

→ The remainder of the proof is by contradiction: ←

**(H)** assume $\exists z \in ]t-\eta, t+\eta[$, $\exists \mu > 0$ such that $L = [z, z+\mu] \subset [t-\eta, t+\eta]$ and:

$$\int_z^{z+\mu} \Sigma_{i+1}(\underline{u}^l)(t)\, dt > 0, \text{ that is, } \int_z^{z+\mu} \sum_{k=i+1}^{p} u^l_k(t)\, dt > 0.$$

Since $y_i^l(z) > 0$ and $y_i^l(0) = 0$ the following definitions are valid:

$$x = \text{Sup}\ \left\{ r \in [0,z] \ / \ \exists\, \alpha \in ]0,r] \text{ and } \forall\, s \in [r-\alpha, r],\ \delta^l_i(s) > 0 \right\} \neq \emptyset \qquad (1$$

$$\varepsilon = \text{Sup}\ \left\{ \alpha \in ]0,x] \ / \ \forall\, s \in [x-\alpha, x],\ \delta^l_i(s) > 0 \right\} \quad \text{and} \quad K = [x-\varepsilon, x].$$

$\varepsilon$ can be reduced in order that: $\displaystyle\int_{x-\varepsilon}^{x} \delta^l_i(t)\,dt \leq \min_{s \in [z, z+\mu]} y^l_i(s)$ and then

$\varepsilon$ and $\mu$ can be adjusted so that: $\displaystyle\int_{x-\varepsilon}^{x} \delta^l_i(t)\,dt = \int_{z}^{z+\mu} \Sigma_{i+1}(\underline{u}^l)(t)\,dt > 0.$ (2

A new plan $(u'^j_k)$ can thereafter be constructed, that will yield lower value of the objective function:

Let $\varepsilon_i = \varepsilon$;

for $k = i+1,..,p$ $\exists\ \varepsilon_k \in [0, \varepsilon_{k-1}]$ such that $\displaystyle\int_{x-\varepsilon_{k-1}}^{x-\varepsilon_k} \delta^l_i(t)\,dt = \int_{z}^{z+\mu} u^l_k(t)\,dt$

(The existence of the $\varepsilon_k$ follows from (2))

* $\underline{u}'^l$ can then be defined as follows:

. $\forall\, k \in \{1,..,i-1\},\ u'^l_k = u^l_k$.

. $\forall\, k \in \{i,..,p\},\ \forall\, s \in I \setminus (K \cup L),\ u'^l_k(s) = u^l_k(s)$.

. $\forall\, s \in K,\ u'^l_i(s) = u^l_i(s) - \delta^l_i(s)$ ; $\forall\, s \in L,\ u'^l_i(s) = u^l_i(s) + \Sigma_{i+1}(\underline{u}^l)(s)$.

. $\forall\, k \in \{i+1,..,p\},\ \forall\, s \in [x-\varepsilon_{k-1}, x-\varepsilon_k],\ u'^l_k(s) = u^l_k(s) + \delta^l_i(s)$

$\forall\, s \in K \setminus [x-\varepsilon_{k-1}, x-\varepsilon_k],\ u'^l_k(s) = u^l_k(s)$

$\forall\, s \in L,\ u'^l_k(s) = 0$

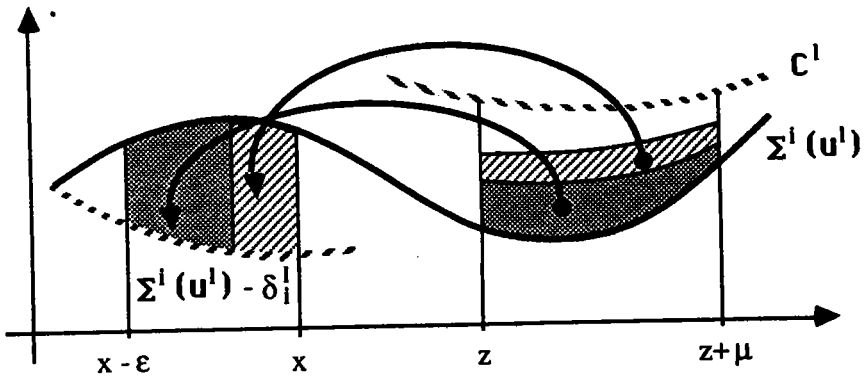



**fig. 1**

* the flow-plans relative to the other stages are defined by:

. $\forall j \in \{1,..,l-1\}$, $\underline{u}'^j = \Xi^{\times}(C^j, \underline{u}'^{j+1})$.

. $\forall j \in \{l+1,..,n\}$, $\underline{u}'^j = \underline{u}^j$.

It was stated in corollary 3.V.1 that the existence of a solution to a multiproduct problem depends only on the total demand with respect to the capacity. By construction, $\sum^p(\underline{u}'^l) = \sum^p(\underline{u}^l)$, and thus, since it was assumed that $\Xi^{\times}(C^{l-1},\underline{u}^l)$ exists, so does $\Xi^{\times}(C^{l-1},\underline{u}'^l)$. Moreover, it results from corollary 3.VI.1 that the property $\sum^p(\underline{d}') = \sum^p(\underline{d})$ is preserved by the function $\Xi^{\times}$. It could then be proved more formally by induction that the definition of the plans $\underline{u}'^j$, $j \in \{1,..,l-1\}$ is valid.

In order to show that this new plan satisfies the constraints of problem MM, one only needs to prove that $\underline{u}'^l$ satisfies the capacity constraint of stage l (but this is easy to verify by considering the definition of $\underline{u}'^l$) and that no stockout will result at stage l from replacing $\underline{u}^l$ by $\underline{u}'^l$, given that $\underline{u}'^{l+1} = \underline{u}^{l+1}$.

Qualitatively, the argument is that items i+1 to p are produced in advance, which gives them a higher inventory function, whereas for item i, equations (1) and (2) ensure that postponing the production will not result in a stockout.

The following propositions result from the definition of $\underline{u}^l$ and the properties of $\Xi$ and $\Xi^{\times}$; they are used to prove that $(\underline{u}'^j)$ yields a lower value of the total cost than $(\underline{u}^j)$.

$$\begin{cases} . \forall k \in \{1,.., i-1\},\ u'_k{}^l = u_k{}^l \\ . \Sigma_i(\underline{u}'^l) = \Sigma_i(\underline{u}^l) \\ . u'_i{}^l \ll^0 u_i{}^l \\ . \forall k \in \{i+1,.., p\},\ u_k{}^l \ll^0 u'_k{}^l \end{cases} \Rightarrow \forall j \in \{1,..,l-1\}, \begin{cases} .\forall k \in \{1,.., i-1\},\ u'_k{}^j = u_k{}^j \\ .\Sigma_i(\underline{u}'^j) = \Sigma_i(\underline{u}^j) \\ .u'_i{}^j \ll^0 u_i{}^j \\ .\forall k \in \{i+1,.., p\},\ u_k{}^j \ll^0 u'_k{}^j \end{cases}$$

The first two propositions follow directly from corollary 3.VI.1, and the last two follow from corollary 1.V.1.

The difference between the costs resulting from the two flow-plans is:

$$K - K' = \sum_{j=1}^{l} \sum_{k=i}^{p} \kappa^j_k \int_0^h \int_0^t (u^j_k - u'^j_k)(s)\, ds\, dt \quad = \quad X + Y$$

$$\text{where } X = \sum_{j=1}^{l-1} \sum_{k=i}^{p} \kappa^j_k \int_0^h \int_0^t (u^j_k - u'^j_k)(s)\, ds\, dt$$

$$\text{and } \quad Y = \sum_{k=i}^{p} \kappa^l_k \int_0^h \int_0^t (u^l_k - u'^l_k)(s)\, ds\, dt$$

$\forall j \in \{1,..,l-1\}$, using the equation $\Sigma_i(\underline{u}'^j) = \Sigma_i(\underline{u}^j)$, which is equivalent to $u^j_i - u'^j_i = \Sigma_{i+1}(\underline{u}'^j) - \Sigma_{i+1}(\underline{u}^j)$, the expression of X becomes:

$$X = \sum_{j=1}^{l-1} \left[ \kappa^j_i \sum_{k=i+1}^{p} \int_0^h \int_0^t (u'^j_k - u^j_k)(s)\, ds\, dt \; - \sum_{k=i+1}^{p} \kappa^j_k \int_0^h \int_0^t (u'^j_k - u^j_k)(s)\, ds\, dt \right]$$

Since, by construction, $\forall j \in \{1,..,l-1\}$, $\forall k \in \{i+1,..,p\}$, $u^j_k \ll^0 u'^j_k$ and it was assumed that $\kappa^j_i > \kappa^j_{i+1} > \cdots > \kappa^j_p$, $\underline{X \geq 0}$. (This may seem an easy result, but the major difficulty in this theorem was to bound the effect of a local modification of the control on the inventory cost of upstream stages, and that is exactly what the previous inequality achieves...)

As regards Y, $\underline{u}'^l$ has been defined in order that $Y > 0$, in the very same manner as in the single-stage case:

<u>definition:</u> $\forall k \in \{i,..,p\}, \forall t \in I$, let $\varphi_k(t) = \int_0^t (u^l_k - u'^l_k)(s)\, ds$

The equations stated hereunder follow:

. $\forall t \in K$, $\varphi_i(t) = \int_{x-\varepsilon}^{t} \delta^l_i(s)\, ds$

. $\forall k \in \{i+1,..,p\}$, $\forall t \in [x - \varepsilon_{k-1}, x - \varepsilon_k]$, $\varphi_k(t) = -\int_{x-\varepsilon_{k-1}}^{t} \delta^l_i(t)\, dt$

. $\forall t \in [x,z]$, $\varphi_i(t) = \Delta = \int_{x-\varepsilon}^{x} \delta^l_i(s)\,ds = \int_{z}^{z+\mu} \Sigma_{i+1}(\underline{u}^l)(s)\,ds$

. $\forall t \in [x,z]$, $\forall k \in \{i+1,..,p\}$, $\varphi_k(t) = -\Delta_k = -\int_{z}^{z+\mu} u^l_k(s)\,ds = -\int_{x-\varepsilon_{k-1}}^{x-\varepsilon_k} \delta^l_i(t)\,dt$

. $\forall t \in L$, $\varphi_i(t) = \int_{t}^{z+\mu} \Sigma_{i+1}(\underline{u}^l)(s)\,ds$ and $\forall k \in \{i+1,..,p\}$, $\varphi_k(t) = -\int_{t}^{z+\mu} u^l_k(s)\,ds$

On the rest of I, the functions $\varphi_k$, $k \in \{i,..,p\}$, are null. Thus the expression of Y is:

$$Y = \kappa^l_i \left[ \int_{x-\varepsilon}^{x} \int_{x-\varepsilon}^{t} \delta^l_i(s)\,ds\,dt + \Delta(z-x) + \int_{z}^{z+\mu} \int_{t}^{z+\mu} \Delta^l_i(s)\,ds\,dt \right]$$

$$- \sum_{k=i+1}^{p} \kappa^l_k \left[ \int_{x-\varepsilon_{k-1}}^{x-\varepsilon_k} \int_{x-\varepsilon_{k-1}}^{t} \delta^l_i(s)\,ds\,dt + \Delta_k(z-x+\varepsilon_k) + \int_{z}^{z+\mu} \int_{t}^{z+\mu} u^l_k(s)\,ds \right]$$

$$= \kappa^l_i \; Z_i - \sum_{k=i+1}^{p} \kappa^l_k \; Z_k$$

where the $Z_j, j \in \{i,...,p\}$ represent the expressions in brackets.

By construction, $Z_i = \sum_{k=i+1}^{p} Z_k > 0$ and it was assumed that $\kappa^j_i > \cdots > \kappa^j_p$.

Therefore, $\underline{Y \geq 0}$, and thus $K > K'$, which contradicts the optimality of the plan and ends the proof. In fact, the initial hypothesis **(H)** must be false, which means:

$\forall z, \mu > 0$ such that $L = [z, z+\mu] \subset [t-\eta, t+\eta]$, $\int_{z}^{z+\mu} \sum_{k=i+1}^{p} u^l_k(r)\,dr \leq 0$.

**corollary 4.III.1:**

Under the cost assumptions stated at the beginning of this section, **if** $\exists l \in \{1,..,n\}$ and $\forall j \in \{1,..,l-1\}$, $\underline{u}^j = \Xi^x(C^j, \underline{u}^{j+1})$, **then** $\underline{u}^j = \Xi^x(C^j, \underline{u}^{j+1})$.

proof:

The procedure to derive this result from that of theorems 4.III.1 and 4.III.2 would be the same as in part 3.

**corollary 4.III.2:**

Under the cost assumptions stated previously in this section, the solution to $MM((C^l), \underline{d}, (\underline{k}^l))$ can be determined backwards from the demand of the final product by the single-stage procedure symbolized by the function $\Xi^x$.

In mathematical terms, with the notation $u_{n+1} \equiv d$:

**if** $(\underline{u}^j)$ solves $MM((C^l), \underline{d}, (\underline{k}^l))$ **then** $\forall l \in \{1,..,n\}$, $\underline{u}^l = \Xi^x(C^j, \underline{u}^{l+1})$ a.e. on I

proof:

Given the result of corollary 4.II.1, the only point that remains to be proved in order to derive corollary 4.III.2 by induction is that $\underline{u}^1 = \Xi^x(C^j, \underline{u}^2)$. This point becomes obvious once one notices that if $\underline{u}$ is given, the problem of determining $\underline{u}^1$ is a single-stage problem of the type considered in part 3.

interpretation:

If each of the subsystems is controlled individually along the optimal policy defined in Part 3 for single-stage systems, then the resulting control is **globally optimal**, provided that the inventory holding costs satisfy the conditions stated at the beginnng of this section.

# chapter 5:

# APPLICATIONS

# THE MONO-PRODUCT, TWO-STAGE PROBLEM WITH DECREASING COSTS

It was claimed in the previous chapter that no analytical solution to the two-stage problem with decreasing costs could be found. BENSOUSSAN and PROTH [BP] consider a similar -single stage- problem with a more complex cost structure: the instantaneous cost is the sum of a concave function of the inventory and a concave function of the flow-rate, both functions being also time-dependent. The authors use the same approach as WAGNER and WHITIN [WW] in their famous dynamic lot-sizing algorithm: they characterize the optimal control to reduce the dimension of the search space in a discretized version of the problem that they solve thereafter by dynamic programming.

This approach is applied here to the two-stage problem.

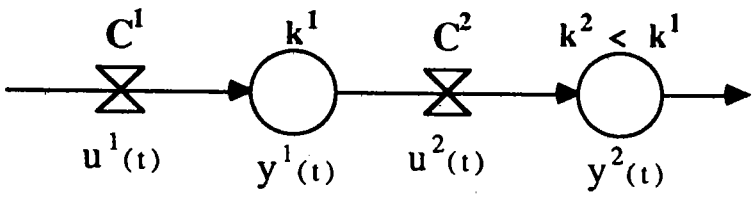



**fig. 1**

## I PROPERTIES OF THE OPTIMAL CONTROL

theorem 5.I.1:

Let $(u^1,u^2)$ be a solution to $2S(C^{1,2},d,k^{1,2})$ and $(y^1,y^2)$ the resulting inventories;

1. **if** $y^1(t)>0$, **then** $u^1=C^1$ and $u^2=C^2$ on a neighborhood of $t$.
2. **if** $y^1=0$ and $y^2>0$ on $J\subset I$, $J\neq\emptyset$, **then** $u^1=u^2=\underline{C}$ a.e. on $J$.
3. **if** $y^1=y^2=0$ on $J\subset I$, $J\neq\emptyset$, **then** $u^1=u^2=d$ a.e. on $J$.

**proof:**

1. The inequality $y^1(t) > 0$ holds on a neighborhood J of t because $y^1$ is a continuous function, and means that $\underset{\tau \in [t,h]}{\text{Sup}} \int_t^\tau [u^1(s) - C^2(s)]\, ds > 0$, because $u^2 = \Pi(u^1, C^2)$. Thus, by definition of $\Pi$, $u^2 = C^2$ on J.

Also, the first half of the proposition can be proved just as in the single-stage case (theorem 3.III.1) : $y^1(t) > 0$ implies that $u^1$ has been positive at some earlier point in time s (because $y^1(t) = 0$) and if $u^1$ were not equal to $C^1$ in the neighborhood of t, it would be possible to cut down the inventory holding cost by delaying the flow $u^1$.

2. The argument is the same to show that if $y^1 = 0$ and $y^2 > 0$ on J, then $u^1 = u^2 = \underline{C}$: $y^1 = 0$ implies that $u^1 = u^2$ and the capacity constraint then becomes $u^1 = u^2 \leq \underline{C}$. If this inequality were strict, it would be possible to postpone the previous overproduction that resulted in $y^2 > 0$ and thus reduce the inventory cost. Again, this would contradict the optimality of $(u^1, u^2)$.

3. The last proposition does not really need a proof and is only stated for the sake of completeness.

note: Propositions 1 and 2 can be interpreted as follows: if there is some inventory kept in the system, (that is, if $y^1 + y^2 > 0$,) the system is operated "at capacity". The only reason for two different outcomes is that if $y^1 = 0$, the two-stage system behaves as a single stage system whose capacity is that of the less performant stage; on the contrary, if $y^1 > 0$, the work-in-process in the intermediate buffer decouples the two stages and allows them to be operated up to their respective capacities. Nothing really surprising in these results...

**theorem 5.I.2:**

Let $(u^1,u^2)$ be a solution to $2S(C^{1,2},d,k^{1,2})$, $(y^1,y^2)$ the resulting inventories and $v^1=\Xi(C^1,d)$, $v^2=\Xi(C^2,d)$; then it can be proved that:

1. $y^1(h) = y^2(h) = 0$.

2. $\forall t\in I,\ 0 \le y^1(t) \le \text{Min}\left\{ \int_0^t (C^1(s) - v^2(s))\,ds,\ \int_0^h d(s)\,ds - \int_0^t v^2(s)\,ds \right\}$

3. $\forall t\in I,\ \int_0^t (v^2(s) - d(s))\,ds \le y^2(t) \le \text{Min}\left\{ \int_0^t (C^2(s) - d(s))\,ds,\ \int_t^h d(s)\,ds \right\}$

4. $\forall t\in I,\ y^1(t) + y^2(t) \ge \int_0^t (v^1(s) - d(s))\,ds$

<u>proof:</u>

1. Because $y^1(0) = y^2(0) = 0$ and $y^1(h) \ge 0$, $y^2(h) \ge 0$, one can state:

$\exists\, t^1, t^2 \in I$ such that $\int_{t^1}^h u^1(s)\,ds = y^1(h) + y^2(h)$ and $\int_{t^2}^h u^2(s)\,ds = y^2(h)$;

It is then easy to show that $y^1(t^1) = \int_{t^1}^{t^2} u^2(s)\,ds$ and $y^2(t^2) = \int_{t^2}^h d(s)\,ds$.

This implies in particular that $t^1 \le t^2$.

Define the control u' so that for $k = 1,2$, $u'^k = u^k$ on $[0, t^k]$ and $u'^k = 0$ on $[t^k, h]$; u' is admissible for $2S(C^{1,2},d,k^{1,2})$ if u is:

. for $k = 1,2$, $y'^k = y^k \ge 0$ on $[0, t^k]$

. $\forall t \in [t^1, t^2]$, $y'^1(t) = y^1(t^1) - \int_{t^1}^t u^2(s)\,ds \ge y^1(t^1) - \int_{t^1}^{t^2} u^2(s)\,ds = 0$

. $\forall t \in [t^2, h]$, $y'^1(t) = 0$ and $y'^2(t) = y^2(t^2) - \int_{t^2}^t d(s)\,ds \ge y^2(t^2) - \int_{t^2}^h d(s)\,ds = 0$

The difference between the costs resulting from the two controls,

$$K - K' = k^1 \int_{t^1}^{t^2} (t^2 - t)\,u^2(t)\,dt + k^1 \int_{t^2}^h y^1(t)\,dt + k^2 \int_{t^2}^h (h - t)\,d(t)\,dt$$

clearly contradicts the optimality of u if $y^1(h) + y^2(h) > 0$.

2. Following from the constraints of $2S(C^{1,2},d,k^{1,2})$, $y^1 \geq 0$ and $u^1 \leq C^1$; also, it was stated in lemma 2.II.1-1 that $\forall t \in I$, $\int_0^t u^2(s)\,ds \geq \int_0^t v^2(s)\,ds$. Since $y^1(t) = \int_0^t u^1(s)\,ds - \int_0^t u^2(s)\,ds$, it follows that $y^1(t) \leq \int_0^t (C^1(s) - v^2(s))\,ds$. Additionally, $y^1(h) = y^2(h) = 0$ implies that $\int_0^h u^1(s)\,ds = \int_0^h d(s)\,ds$ and thus: $\int_0^t (u^1(s) - u^2(s))\,ds \leq \int_0^h d(s)\,ds - \int_0^t v^2(s)\,ds$, which proves proposition 2.

3. Similarly, $y^2(t) = \int_0^t u^2(s)\,ds - \int_0^t d(s)\,ds$, and hence, the first inequality in proposition 3 also follows from lemma 2.II.1-1; the second one follows in part from the constraint $u^2 \leq C^2$ and in part from the fact that $y^2(t) = y^2(h) - \int_t^h u^2(s)\,ds + \int_t^h d(s)\,ds \leq y^2(h) + \int_t^h d(s)\,ds$, and $y^2(h) = 0$.

4. Proposition 4 merely restates lemma 2.II.1-2.

## IV DYNAMIC PROGRAMMING

Dynamic programming is an optimization technique developed by BELLMAN and based on the Principle of Optimality. Roughly speaking, this principle states that any portion of an optimal path is optimal (for a more in-depth explanation, see BERTSEKAS [BE], FLEMING and RISHEL [FR] or KAMIEN and SHWARTZ [KS]). The application of this technique to the two-stage problem is described here.

### notations:

→ $u = (u^1, u^2)$ is the control vector.

→ $x = (x^1, x^2)$ , where $\forall l \in \{1,2\}$, $x^l(t) = \int_0^t u^l(s)\,ds$ is the state vector.

These two vectors are linked by the equation : $\dot{x} = u$.

Note that the inventories could have been chosen as state variables but then, writing this last equation would have required to use an augmented state vector.

→ $J((z,t);u)$ is the "value function" or "cost-to-go" function which represents the cost incurred over the time interval $[t,h]$ if the state at time t is z and the control u is applied:

$$J((z,t);u) = \int_t^h \kappa' x(s)\, ds \quad , \quad \text{where} \quad \forall s \in [t,h],\ x(s) = z + \int_t^s u(r)\, dr$$

and ' represents the inner product.

Thus $J((z,t);u) = \sum_{l=1}^{2} \kappa^l \left[ z^l + \int_t^h (h-s)\, u^l(s)\, ds \right]$

**definitions:**

→ $\forall t \in [0,h],\ \Delta_t = \{ u \in \Omega^2 / \forall s \in [t,h],\ \forall l \in \{1,2\},\ u^l(s) \le C^l(s) \}$

→ For given initial conditions $(z,t)$, a control u is said feasible if it results in positive inventories, that is, if

$$\forall s \in [t,h],\quad z^1 + \int_t^s u^1(r)\, dr \ \ge\ z^2 + \int_t^s u^2(r)\, dr \ \ge\ \int_t^s d(r)\, dr$$

→ $J((z,t);u) \equiv +\infty$ if the control u is not feasible for the initial conditions $(z,t)$.

→ $J^*(z,t) = \underset{u}{\text{Min}}\ J((z,t);u)$, that is, $J^*$ is the optimal value function.

## 1- Differential Equation of Dynamic Programming:

(FLEMING and RISHEL [FR])

Let $(z,t)$ be feasible initial conditions, ( i.e. such that there exist $t' > t$ and a feasible control defined on $[t,t']$ ), such that $J^*$ is differentiable at $(z,t)$. If there exists an optimal control $u^*$, then the partial differential equation:

$$\underset{u \in \Delta_t}{\text{Min}} \left\{ \frac{\partial J^*}{\partial t} + \frac{\partial J^*}{\partial z} . u \right\} = 0$$

is satisfied, and the minimum is achieved by $u^*(t^+)$ (that is, the right limit of $u^*$ at t).

Hence $u^*$ solves both a differential equation:

$$\frac{\partial J^*}{\partial t} + \frac{\partial J^*}{\partial z} . u = 0$$

and a linear optimization problem:

$$\underset{u \in \Delta_t}{\text{Min}} \ \frac{\partial J^*}{\partial z} . u$$

This last result is central in the work of KIMEMIA and GERSHWIN [KG] on the optimal control of a Flexible Manufacturing System subject to machine failures. In order to solve the stochastic control problem resulting from the formulation of their model, the authors suggest that the cost-to-go function J be evaluated off-line and the loading rates then determined on-line by solving the stochastic equivalent of the previous linear program.

In subsequent work, GERSHWIN, AKELLA and CHOONG [GC] point out that the exact shape of the J function has little or no influence on the optimal control and they choose to approximate it by a quadratic function, which considerably simplifies the computation of the sub-optimal control.

In the current problem, the demand varies over the horizon and evaluating the J function is of the same order of difficulty as solving the entire problem. Therefore, the optimal control is not derived from the value function J but generated in the very same way as the shortest path in a graph.

## 2- Discrete Formulation of Backward Dynamic Programming:

Let $\delta t = h/N$ be a given time-step; the Principle of Optimality can be directly formulated as follows:

$$\textbf{(BDP)} \quad \forall t \in I, \ \forall z \geq 0, \ J^*(z,t) = \underset{u \in \Delta(z,t)}{\text{Min}} \left[ \int_t^{t+\delta t} \kappa' x(s) \, ds + J^*(x(t+\delta t), t+\delta t) \right]$$

where (1) $J^*$ is the optimal value function, i.e. the one resulting from the optimal control,

(2) $\forall s \in [t,t+\delta t],\ x(s) = z + \int_t^s u(r)\,dr$,

(3) $\Delta(z,t)$ is the set of feasible controls for the initial conditions $(z,t)$, that is:

$$\Delta(z,t) = \left\{ u \in \Delta_t / \forall s \in [t,t+\delta t],\ z^1 + \int_t^s u^1(r)\,dr \geq z^2 + \int_t^s u^2(r)\,dr \geq \int_t^s d(r)\,dr \right\}$$

In order to derive an algorithm from this equation, it is necessary to limit the computation of $J^*(z,t)$ to a finite number of pairs $(z,t)$ and, for each of them, reduce $\Delta(z,t)$ to a finite set.

1- time discretization:

$J^*(z,t)$ is determined only for $t = i\delta t$, $i \in \{1,..,N\}$.

2- state discretization:

Theorem 5.I.2 implies that $\forall t \in I$, $x(t) \in \Sigma(t) = [\,m^1(t), M^1(t)] \times [\,m^2(t), M^2(t)]$, where the bounding functions are:

$$m^2(t) = \int_0^t v^2(s)\,ds \qquad \text{and} \qquad M^2(t) = \text{Min}\left\{ \int_0^t C^2(s)\,ds,\ \int_0^h d(s)\,ds \right\}$$

$$m^1(t) = \text{Max}\left\{ m^2(t),\ \int_0^t v^1(s)\,ds \right\} \quad \text{and} \quad M^1(t) = \text{Min}\left\{ \int_0^t C^1(s)\,ds,\ \int_0^h d(s)\,ds \right\}$$

Note that the interpretation of these bounds is easier in terms of the cumulative productions, than it is in terms of the inventories.

Given a state-step $\delta x = (\delta x^1, \delta x^2)$, $J^*(z,t)$ is determined at any time $t$ only for $z \in \Sigma'(t) = \{ x \in \Sigma(t) / x = (j.\delta x^1, k.\delta x^2),\ j,k \in \mathbb{N} \}$, assuming that $\delta x$ is chosen so that the extreme points of $\Sigma(t)$ are in $\Sigma'(t)$.

## 3- control discretization:

Let $U^1$, $U^2$ and $U^3$ represent the vectors (d,d), $(C^1, C^2)$ and $(\underline{C}, \underline{C})$, and let $u^*$ denote the optimal control; it was shown in Section I that for almost every t in I, $u^*(t) \in \{ U^1(t), U^2(t), U^3(t) \}$. In order to use this result in the discrete case, an assumption is required:

### Assumption:

$\forall i \in \{1,..,N\}$, **if** $\exists t \in [(i-1)\delta t, i\delta t[$ and $\exists k \in \{1,2,3\}$ such that $u^*(t) = U^k(t)$,

**then** $\forall s \in [(i-1)\delta t, i\delta t[$, $u^*(s) = U^k(s)$. **(As)**

**(As)** means that the optimal control $u^*$ is sought such that $u^*(t)$ switches from $U^j(t)$ to $U^k(t)$ only at times $i\delta t$, $i \in \{1,..,N\}$. Under this assumption, the number of elements in $\Delta(z,t)$ is reduced to at most three. In fact, if $d(t) > \underline{C}(t)$, $U^1(t) \notin \Delta(z,t)$, and if $C^1(t) = C^2(t)$, $U^2(t) = U^3(t)$, and also, $U^1(t) = U^2(t) = U^3(t)$ if $C^1(t) = C^2(t) = d(t)$.

In order that the assumption **(As)** does not preclude finding the optimal control, the time-step $\delta t$ must be chosen very small with respect to the variations of the functions d, $C^1$ and $C^2$. In fact, if $\delta$ is not small enough, the continuous-time optimal control can "switch" between $U^1$, $U^2$ and $U^3$ during an interval $[(i-1)\delta t, i\delta t[$, and the value of $u^*$ that should be found by a discrete-time algorithm is a weighted combination of $U^1$, $U^2$ and $U^3$ (as opposed to simply one of these vectors).

The trade-off in the algorithm is thus between chosing a very small time-step or forgetting the "bang-bang" characteristics of the optimal control. Both alternatives result in an increased dimensionality of the problem and longer computations. In order to choose between the two approaches a criterion is required: the ratio of the complexity of the algorithm to the quality of the solution it yields seems to be a reasonable one.

4- algorithm:

**1- Backward equations:** determination of $J^*(x_0,0)$.

- $\Sigma'(h) := \{(0,0)\}$ ;
- $J^*((0,0),h) := 0$ ;
- For i := N -1 downto 0 do
  - For all $z \in \Sigma'(i.\delta t)$ do
    - $J^*(z, i.\delta t) := +\infty$ ;
    - $v^*(z, i.\delta t) := 0$ ;
    - For all $u \in \Delta(z, i.\delta t)$ do
      - $x := z + u.\delta t$ ;
      - if $x \in \Sigma'((i+1).\delta t)$ then
        $J := (k'z) + J^*(x, (i+1).\delta t)$ {interpolated if needed}
        else $J := +\infty$ ;
      - if $J < J^*(z, i.\delta t)$ then
        $J^*(z, i.\delta t) := J$ ;
        $v^*(z, i.\delta t) := u$
- if $J^*(x_0,0) = +\infty$ then no solution
  else:

**2- Forward equations:** determination of $u^*$.

- $x^*(0) := x_0$ ;
- For i := 1 to N -1 do
  - $u^*(i.\delta t) := v^*(x^*((i-1).\delta t), i.\delta t)$ ; {interpolated if needed}
  - $x^*(i.\delta t) := x^*((i-1).\delta t) + u^*(i.\delta t).\delta t$

The complexity of the algorithm is :

in the order of $\sum_{i=0}^{N} \sum_{z \in \Sigma(t)} \text{card}(\Delta(z, i.\delta t))$ for the first phase,

and in the order of N for the second one (this term is thus negligible compared to the previous one).

Assuming that the quality of the solution is measured by its "time resolution" that is, by the number N of time-steps in the horizon, and disregarding the fact that the chances of success of the algorithm also depend on this parameter, the ratio complexity/quality of the algorithm does not depend on N. In fact, if the average number of states in $\Sigma'(i.\delta t)$ is $nx_1 \times nx_2$, and the average number of controls in $\Delta(z,i.\delta t)$ is $nu_1 \times nu_2$, then the ratio complexity/quality is:

$$C^x / Q^y \approx nx_1 \times nx_2 \times nu_1 \times nu_2$$

It is thus always beneficial to reduce the size of $\Delta(z,i.\delta t)$. One way to achieve this objective is to go further in the use of the properties of the optimal control. In fact, the state variables and the inventories are linked by the equations: $\forall t \in I$, $y^1(t) = x^1(t) - x^2(t)$ and $y^2(t) = x^2(t) - D(t)$, where $D(t)$ represents the cumulative demand at time t. Hence, theorem 5.I.2 could be rewritten, under the assumption **(As)**:

(1) * if $z^1 > z^2$, $\Delta(z,t) = \begin{cases} \{(C^1(t), C^2(t))\} & \text{if } z^1 + \int_t^{t+\delta t} C^1(r)\,dr \geq z^2 + \int_t^{t+\delta t} C^2(r)\,dr \\ & \text{and } z^2 + \int_t^{t+\delta t} C^2(r)\,dr \geq D(t+\delta t) \\ \emptyset & \text{otherwise} \end{cases}$

(2) * if $z^1 = 0$ and $z^2 > D(t)$, $\Delta(z,t) = \begin{cases} \{(\underline{C}(t), \underline{C}(t))\} & \text{if } z^2 + \int_t^{t+\delta t} \underline{C}(r)\,dr \geq D(t+\delta t) \\ \emptyset & \text{otherwise} \end{cases}$

(3) * if $z^1 = z^2 = D(t)$, $\Delta(z,t) = \begin{cases} \{(d(t), d(t))\} & \text{if } d(t) \leq \underline{C}(t) \\ \emptyset & \text{otherwise} \end{cases}$

<u>note:</u> to be entirely rigorous, one should require that the previous inequalities hold on the entire interval $[t,t+\delta t]$; for instance, for the admissibility of $(\underline{C}(t), \underline{C}(t))$, the inequality should be:

$$\forall s \in [t,t+\delta t],\ z^2 + \int_t^s \underline{C}(r)\,dr \geq \int_t^s d(r)\,dr.$$

In fact, the conditions stated previously merely require that the inventory levels be positive at the instants $i\delta t$, $i \in \{1,..,N\}$. In the implementation, the values of the functions are known only at these instants and thus, even the previous inequalities must be approximated. For example, the first condition will be checked in the form: $z^1 + C^1(t)\,\delta t \geq z^2 + C^2(t)\,\delta t$, resulting in a rounding error that will be all the more important as the time step is large.

However, this is not the main issue raised by these results: indeed, if the propositions (1), (2) and (3) were true, the optimal control would be entirely determined by the initial state. In particular, if the initial state is zero, an algorithm based on the previous equations will either find that the optimal control is $U^1$, that is, $u^1 = u^2 = d$, or it will fail to find a solution if $U^1$ is not admissible.

This outcome clearly shows that theorem 5.I.2 cannot be translated straightforwardly to be used in the discrete-time case. In particular, a distinction must be made between the inequalities like $x^2(t) > D(t)$ and the equations like $x^1(t) = x^2(t)$. Namely, the former ones must hold on a neighborhood of t because of the continuity of the functions considered, and they would hold on a whole interval $[(i-1)\delta t, i\delta t[$ if $\delta t$ were small enough, but the latter ones may hold only locally.

Therefore the implemented algorithm uses only propositions (1) and (2) to reduce the size of the sets $\Delta(z,t)$.

## 5- numerical results:

The algorithm described in the previous section was implemented in Turbo Pascal™ on an Apple Macintosh™ and run on the example firs presented in Chapter IV, which motivated this study. The solution was found in less than ten seconds including input/ouput operations to a floppy-disc drive.

The results obtained are described hereunder: they show in particular how, for a cost ratio $k_1/k_2$ lower than 1.75, the optimal control consists of storing in the upstream buffer only:

horizon : 4.0
number of time steps : 8
upstr. stock step : 0.50
dnstr. stock step : 0.50
upstream cost : **1.74**
downstream cost : 1.00

| time : | 0.0 | 0.5 | 1.0 | 1.5 | 2.0 | 2.5 | 3.0 | 3.5 | 4.0 |
|---|---|---|---|---|---|---|---|---|---|
| demands : | | 1.0 | 1.0 | 1.0 | 1.0 | 3.0 | 3.0 | 1.0 | 1.0 |
| cpct up : | 3.0 | 3.0 | 2.0 | 2.0 | 2.0 | 2.0 | 2.0 | 2.0 | |
| cpct dn : | 3.0 | 3.0 | 1.0 | 1.0 | 3.0 | 3.0 | 1.0 | 1.0 | |
| opt u1 : | 1.0 | 1.0 | 2.0 | 2.0 | 2.0 | 2.0 | 1.0 | 1.0 | |
| opt u2 : | 1.0 | 1.0 | 1.0 | 1.0 | 3.0 | 3.0 | 1.0 | 1.0 | |
| opt x1 : | 0.0 | 0.0 | 0.0 | **0.5** | **1.0** | **0.5** | 0.0 | 0.0 | 0.0 |
| opt x2 : | **0.0** | **0.0** | **0.0** | **0.0** | **0.0** | **0.0** | **0.0** | **0.0** | **0.0** |

**---> optimal cost = 1.74 <---**

**tables containing the values of** $J^*((x_1,x_2), i.\delta t)$ :

time: 3.5 ; bounds: 0.50 , 0.50 ; d: 1.0 ; C1: 2.0 ; C2: 1.0

| | |
|---|---|
| -1.00 | -1.00 |
| **0.00** | -1.00 |

time: 3.0 ; bounds: 1.00 , 1.00 ; d: 1.0 ; C1: 2.0 ; C2: 1.0

| | | |
|---|---|---|
| -1.00 | -1.00 | -1.00 |
| -1.00 | -1.00 | -1.00 |
| **0.00** | -1.00 | -1.00 |

time: 2.5 ; bounds: 2.50 , 2.00 ; d: 3.0 ; C1: 2.0 ; C2: 3.0

| | | | | | |
|---|---|---|---|---|---|
| -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| **0.25** | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| -1.00 | **0.44** | -1.00 | -1.00 | -1.00 | -1.00 |

time: 2.0 ; bounds: 3.00 , 2.00 ; d: 3.0 ; C1: 2.0 ; C2: 3.0

| | | | | | | |
|---|---|---|---|---|---|---|
| -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| **0.75** | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| -1.00 | **0.94** | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| -1.00 | -1.00 | **1.31** | -1.00 | -1.00 | -1.00 | -1.00 |

time: 1.5 ; bounds: 2.50 , 2.00 ; d: 1.0 ; C1: 2.0 ; C2: 1.0

| | | | | | |
|---|---|---|---|---|---|
| -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| **1.25** | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| -1.00 | **1.74** | -1.00 | -1.00 | -1.00 | -1.00 |

time: 1.0 ; bounds: 2.00 , 2.00 ; d: 1.0 ; C1: 2.0 ; C2: 1.0

| | | | | |
|---|---|---|---|---|
| -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| **1.75** | -1.00 | -1.00 | -1.00 | -1.00 |
| -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| **1.74** | -1.00 | -1.00 | -1.00 | -1.00 |

time: 0.5 ; bounds: 1.00 , 1.00 ; d: 1.0 ; C1: 3.0 ; C2: 3.0

| | | |
|---|---|---|
| -1.00 | -1.00 | -1.00 |
| -1.00 | -1.00 | -1.00 |
| **1.74** | -1.00 | -1.00 |

time: 0.0 ; bounds: 0.00 , 0.00 ; d: 1.0 ; C1: 3.0 ; C2: 3.0

**1.74**

These tables show how the cost-to-go function is determined for the values of the the state-vector that correspond to grid-points, so that only potential portions of an optimal path are kept. The next listing illustrates the effect of a slight change in the cost ratio and how a path that was dominated hereabove becomes optimal for $k_1 / k_2 > 1.75$: over that value of the ratio, it becomes cheaper to store downstream.

horizon : 4.0
number of time steps : 8
upstr. stock step : 0.50
dnstr. stock step : 0.50
upstream cost : **1.76**
downstream cost : 1.00

| time : | 0.0 | 0.5 | 1.0 | 1.5 | 2.0 | 2.5 | 3.0 | 3.5 | 4.0 |
|---|---|---|---|---|---|---|---|---|---|
| opt u1 : | 1.0 | 3.0 | 1.0 | 1.0 | 2.0 | 2.0 | 1.0 | 1.0 | |
| opt u2 : | 1.0 | 3.0 | 1.0 | 1.0 | 2.0 | 2.0 | 1.0 | 1.0 | |
| opt x1 : | **0.0** | **0.0** | **0.0** | **0.0** | **0.0** | **0.0** | **0.0** | **0.0** | **0.0** |
| opt x2 : | 0.0 | 0.0 | **1.0** | **1.0** | **1.0** | **0.5** | 0.0 | 0.0 | 0.0 |

**---> optimal cost = 1.75 <---**

**tables containing the values of** $J^*((x_1,x_2),i.\delta t)$ :

time: 3.5 ; bounds: 0.50 , 0.50 ; d: 1.0 ; C1: 2.0 ; C2: 1.0

| | |
|---|---|
| -1.00 | -1.00 |
| **0.00** | -1.00 |

time: 3.0 ; bounds: 1.00 , 1.00 ; d: 1.0 ; C1: 2.0 ; C2: 1.0

| | | |
|---|---|---|
| -1.00 | -1.00 | -1.00 |
| -1.00 | -1.00 | -1.00 |
| **0.00** | -1.00 | -1.00 |

time: 2.5 ; bounds: 2.50 , 2.00 ; d: 3.0 ; C1: 2.0 ; C2: 3.0

| | | | | | |
|---|---|---|---|---|---|
| -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| **0.25** | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| -1.00 | **0.44** | -1.00 | -1.00 | -1.00 | -1.00 |

time: 2.0 ; bounds: 3.00 , 2.00 ; d: 3.0 ; C1: 2.0 ; C2: 3.0

| | | | | | | |
|---|---|---|---|---|---|---|
| -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| **0.75** | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| -1.00 | **0.94** | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| -1.00 | -1.00 | **1.32** | -1.00 | -1.00 | -1.00 | -1.00 |

time: 1.5 ; bounds: 2.50 , 2.00 ; d: 1.0 ; C1: 2.0 ; C2: 1.0

| | | | | | |
|---|---|---|---|---|---|
| -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| **1.25** | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| -1.00 | **1.76** | -1.00 | -1.00 | -1.00 | -1.00 |

time: 1.0 ; bounds: 2.00 , 2.00 ; d: 1.0 ; C1: 2.0 ; C2: 1.0

| | | | | |
|---|---|---|---|---|
| -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| **1.75** | -1.00 | -1.00 | -1.00 | -1.00 |
| -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| **1.76** | -1.00 | -1.00 | -1.00 | -1.00 |

time: 0.5 ; bounds: 1.00 , 1.00 ; d: 1.0 ; C1: 3.0 ; C2: 3.0

| | | |
|---|---|---|
| -1.00 | -1.00 | -1.00 |
| -1.00 | -1.00 | -1.00 |
| **1.75** | -1.00 | -1.00 |

time: 0.0 ; bounds: 0.00 , 0.00 ; d: 1.0 ; C1: 3.0 ; C2: 3.0

**1.75**

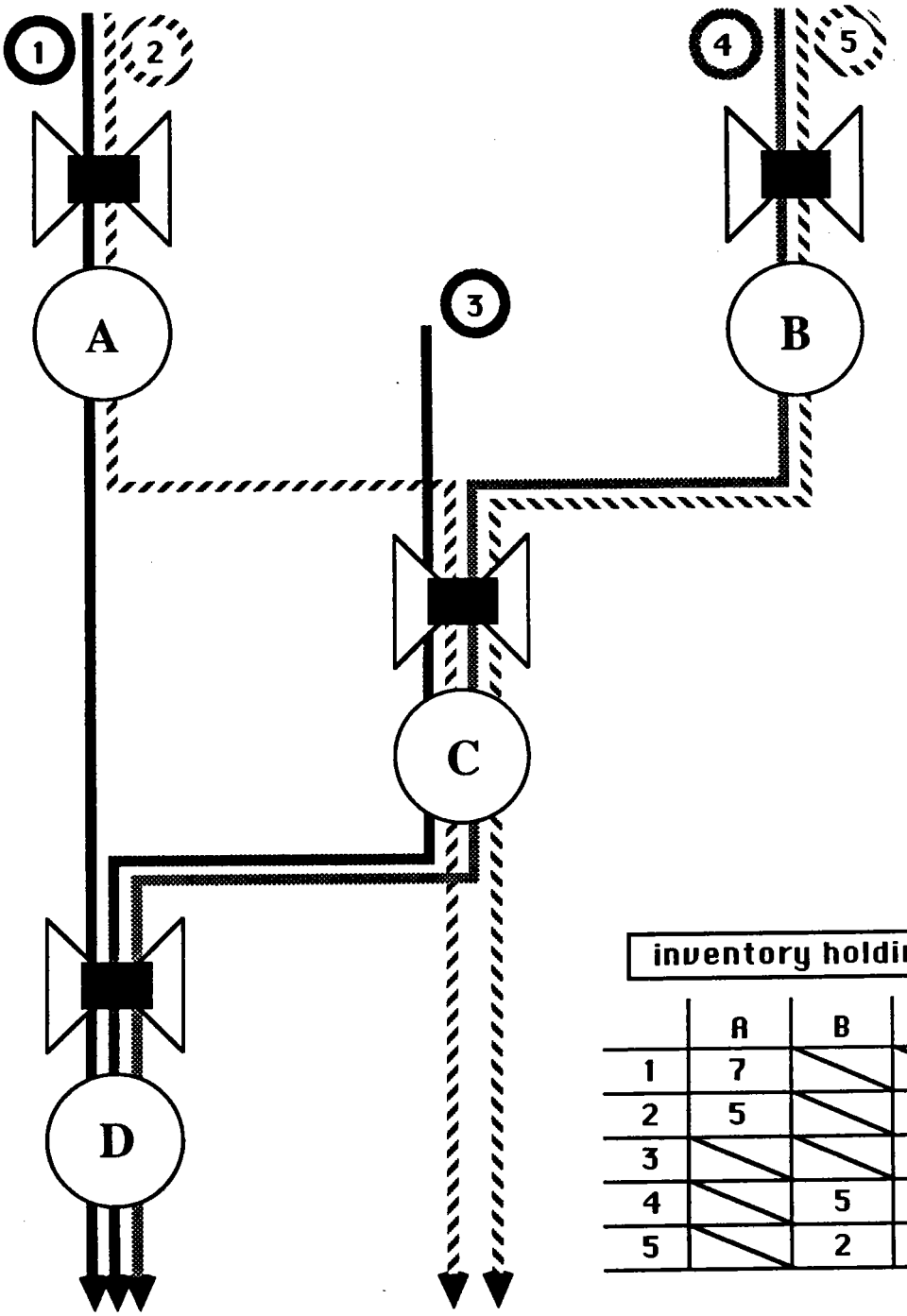
## Conclusion:

Although this application proves the dynamic programming approach to be realistic in the sense that it has generated a succesful numerical method, it also shows that the issue of the convergence of the algorithm with respect to the length of the time step must be raised, and, which is worse, that extensions to larger scale systems will necessarily be at the expense of an exponentially growing computational burden. In fact, the system pictured at the end of Chapter III (five product types, four subsystems) was inially intended for a Dynamic Programming-based application, in the most general case of cost structure. However, when the problem is considered more closely, its dimensionality becomes appearant.

In fact, the state space corresponding to this system is of dimension 11 because the inventory vector has only 11 significant components. the other ones being bound to be zero. Henceforth, if the bounds on the optimal state trajectory are found and the region they delimit is discretized, the size of the state space will be of the same order of magnitude as the average number of discretization steps along the different components *to the eleventh!* Moreover, if no cost assumptions are made and only the weaker results of Chapter IV can be used to characterize the optimal control, then the number of control candidates for each state and time step will also be large, in the order of $(p+2).2^{p-1}$, which, for $p=5$ products, means 112 controls, but would also mean 6144 for $p=10$...

The bottom line of this section is that the numerical method just developed would require an excessive amount of computations for most problems of realistic size.

## APPLICATION TO A FLOW-SHOP.




**inventory holding costs**

| | A | B | C | D |
|---|---|---|---|---|
| 1 | 7 | | | 10 |
| 2 | 5 | | 9 | |
| 3 | | | 8 | 9 |
| 4 | | 5 | 6 | 7 |
| 5 | | 2 | 3 | |

fig.1

The purpose of this short section is to contrast the difficulty of obtaining results by the dynamic programming approach, even for a small scale problem, with the extreme simplicity of implementing the results found in Chapter IV. In fact, the flow control problem for the system reprented on figure 1, in which five different part types visit four different workcenters is very easily solved using a spreadsheet on a microcomputer.

The first step of the resolution is to discretize the final demand functions; then, the flow-plans are determined for all the subsystems sequentially by solving "single-stage" problems. The only constraint on the sequence in which these plans are determined is that, in order to determine the optimal production of a subsystem, the demand it faces (i.e. the production of all downstream subsystems) must be known. Then the algorithm to apply at each stage is the following:

algorithm (H) :

$\Delta_0 \equiv C.$

For $i = 1,..,p$ do :

  $y_{i,h} := 0$

  For $t = h,..,1$ do :

    . $u_{i,t} := \min \{ \Delta_{i-1,t} , d_{i,t} + y_{i,t} \}$

    . $y_{i,t-1} := y_{i,t} + d_{i,t} - u_{i,t}$

    . $\Delta_{i,t} := \Delta_{i-1,t} - u_{i,t}$

, where C is the capacity of the subsystem considered and $\underline{d}$ is the vector of the demands it faces, the products being ranked by decreasing inventory holding costs.

Note that when there does not exist any admissible plan, the algorithm (H) yields non-zero initial inventories, the amount of which represents the extra capacity required for the problem to have a solution. Numerical results are presented in the remainder of the section for illustration purposes.

## End Product Demands

| period: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| prod. 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| prod. 2 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 0 | 0 | 0 | 0 |
| prod. 3 | 1 | 1.2 | 1.4 | 1.6 | 1.8 | 2.0 | 2.2 | 2.1 | 1.8 | 1.5 |
| prod. 4 | 2 | 1.9 | 1.8 | 1.7 | 1.6 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 |
| prod. 5 | 0.8 | 0.8 | 0.8 | 1.2 | 0 | 0 | 1.3 | 1.4 | 1.3 | 1.5 |

## Numerical Results

| period | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SUBSYSTEM D | | | | | | | | | | | | | |
| product cost | 10 | | | | | | | | | | | | |
| demands | | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | | |
| capacity | | 8.5 | 8.2 | 7.9 | 7.6 | 0 | 0 | 5 | 5.5 | 6 | 6.5 | | |
| production | | 2 | 2 | 2 | 6 | 0 | 0 | 2 | 2 | 2 | 2 | | |
| inventory | 0 | 0 | 0 | 0 | 4 | 2 | 0 | 0 | 0 | 0 | 0 | inventory cost: | 60 |
| product cost | 9 | | | | | | | | | | | | |
| demands | | 1 | 1.2 | 1.4 | 1.6 | 1.8 | 2 | 2.2 | 2.1 | 1.8 | 1.5 | | |
| rmg capacity | | 6.5 | 6.2 | 5.9 | 1.6 | 0 | 0 | 3 | 3.5 | 4 | 4.5 | | |
| production | | 1 | 1.2 | 5.2 | 1.6 | 0 | 0 | 2.2 | 2.1 | 1.8 | 1.5 | | |
| inventory | 0 | 0 | 0 | 3.8 | 3.8 | 2 | 0 | 0 | 0 | 0 | 0 | inventory cost: | 86.4 |
| product cost | 7 | | | | | | | | | | | | |
| demands | | 2 | 1.9 | 1.8 | 1.7 | 1.6 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | | |
| rmg capacity | | 5.5 | 5 | 0.7 | 0 | 0 | 0 | 0.8 | 1.4 | 2.2 | 3 | | |
| production | | 5.4 | 5 | 0.7 | 0 | 0 | 0 | 0.8 | 1.3 | 1.2 | 1.1 | | |
| inventory | 0 | 3.4 | 6.5 | 5.4 | 3.7 | 2.1 | 0.6 | 0 | 0 | 0 | 0 | inventory cost: | 151.9 |
| SUBSYSTEM C | | | | | | | | | | | | | |
| product cost | 9 | | | | | | | | | | | | |
| demands | | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 0 | 0 | 0 | 0 | | |
| capacity | | 9 | 9 | 9 | 5 | 1 | 1 | 4 | 4 | 6 | 6 | | |
| production | | 1.5 | 1.5 | 1.5 | 2.5 | 1 | 1 | 0 | 0 | 0 | 0 | | |
| inventory | 0 | 0 | 0 | 0 | 1 | 0.5 | 0 | 0 | 0 | 0 | 0 | inventory cost: | 13.5 |

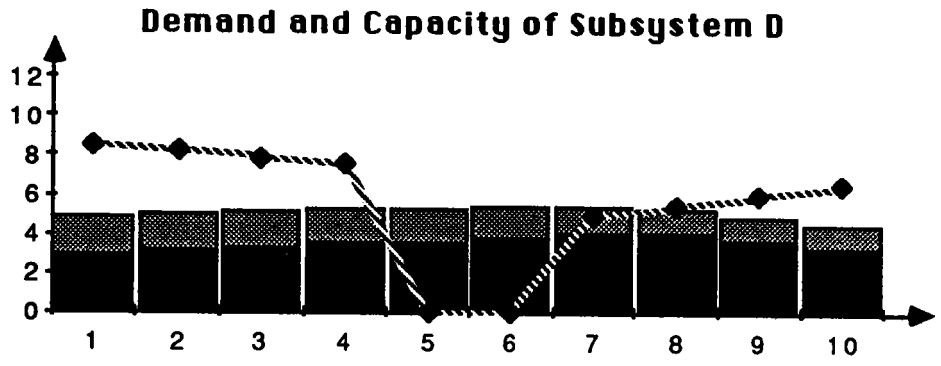
| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| product cost | 8 | | | | | | | | | | | | |
| demands | | 1 | 1.2 | 5.2 | 1.6 | 0 | 0 | 2.2 | 2.1 | 1.8 | 1.5 | | |
| rmg capacity | | 7.5 | 7.5 | 7.5 | 2.5 | 0 | 0 | 4 | 4 | 6 | 6 | | |
| production | | 1 | 1.2 | 5.2 | 1.6 | 0 | 0 | 2.2 | 2.1 | 1.8 | 1.5 | | |
| inventory | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | inventory cost: | 0 |
| product cost | 6 | | | | | | | | | | | | |
| demands | | 5.4 | 5 | 0.7 | 0 | 0 | 0 | 0.8 | 1.3 | 1.2 | 1.1 | | |
| rmg capacity | | 6.5 | 6.3 | 2.3 | 0.9 | 0 | 0 | 1.8 | 1.9 | 4.2 | 4.5 | | |
| production | | 5.4 | 5 | 0.7 | 0 | 0 | 0 | 0.8 | 1.3 | 1.2 | 1.1 | | |
| inventory | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | inventory cost: | 0 |
| product cost | 3 | | | | | | | | | | | | |
| demands | | 0.8 | 0.8 | 0.8 | 1.2 | 0 | 0 | 1.3 | 1.4 | 1.3 | 1.5 | | |
| rmg capacity | | 1.1 | 1.3 | 1.6 | 0.9 | 0 | 0 | 1 | 0.6 | 3 | 3.4 | | |
| production | | 0.9 | 1.3 | 1.6 | 0.9 | 0 | 0 | 1 | 0.6 | 1.3 | 1.5 | | |
| inventory | 0 | 0.1 | 0.6 | 1.4 | 1.1 | 1.1 | 1.1 | 0.8 | 0 | 0 | 0 | inventory cost: | 18.6 |

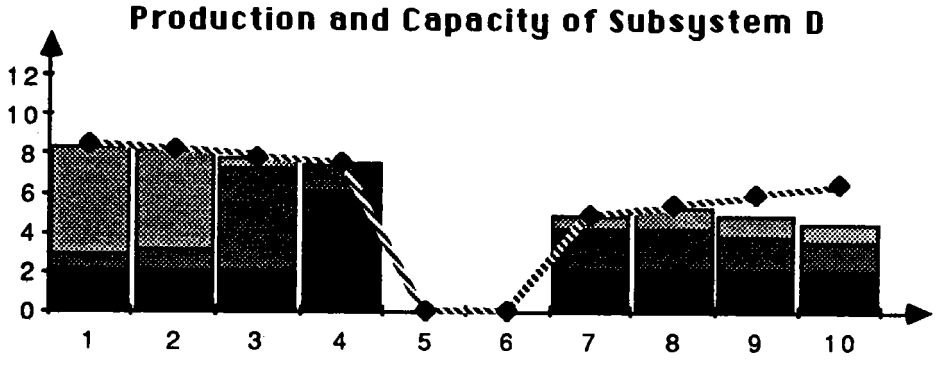
SUBSYSTEM B

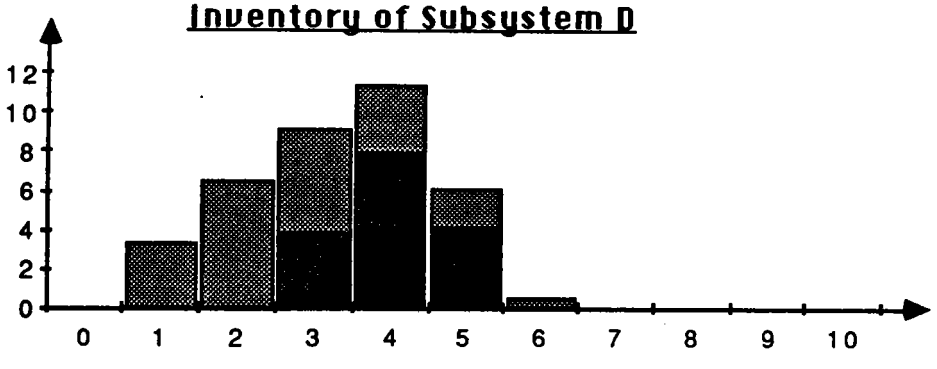
| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| product cost | 5 | | | | | | | | | | | | |
| demands | | 5.4 | 5 | 0.7 | 0 | 0 | 0 | 0.8 | 1.3 | 1.2 | 1.1 | | |
| capacity | | 7 | 7 | 3 | 4 | 4 | 0 | 0 | 1 | 0 | 1 | | |
| production | | 5.4 | 5 | 0.7 | 0 | 2.4 | 0 | 0 | 1 | 0 | 1 | | |
| inventory | 0 | 0 | 0 | 0 | 0 | 2.4 | 2.4 | 1.6 | 1.3 | 0.1 | 0 | inventory cost: | 39 |
| product cost | 2 | | | | | | | | | | | | |
| demands | | 0.9 | 1.3 | 1.6 | 0.9 | 0 | 0 | 1 | 0.6 | 1.3 | 1.5 | | |
| rmg capacity | | 1.6 | 2 | 2.3 | 4 | 1.6 | 0 | 0 | 0 | 0 | 0 | | |
| production | | 0.9 | 1.3 | 1.6 | 3.7 | 1.6 | 0 | 0 | 0 | 0 | 0 | | |
| inventory | 0 | 0 | 0 | 0 | 2.8 | 4.4 | 4.4 | 3.4 | 2.8 | 1.5 | 0 | inventory cost: | 38.6 |

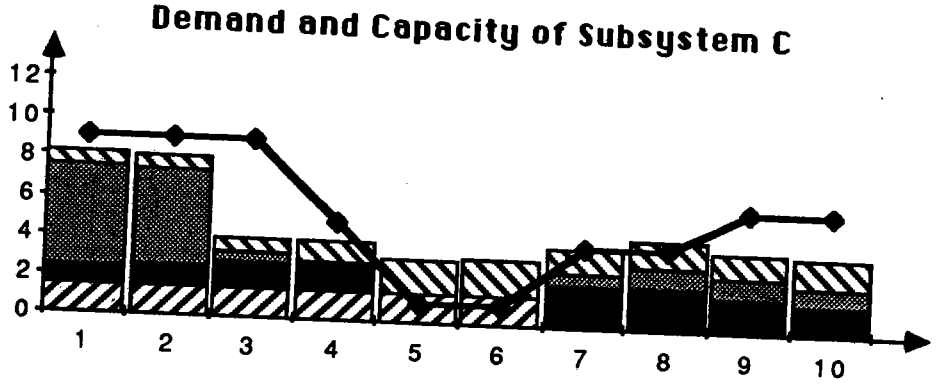
SUBSYSTEM A

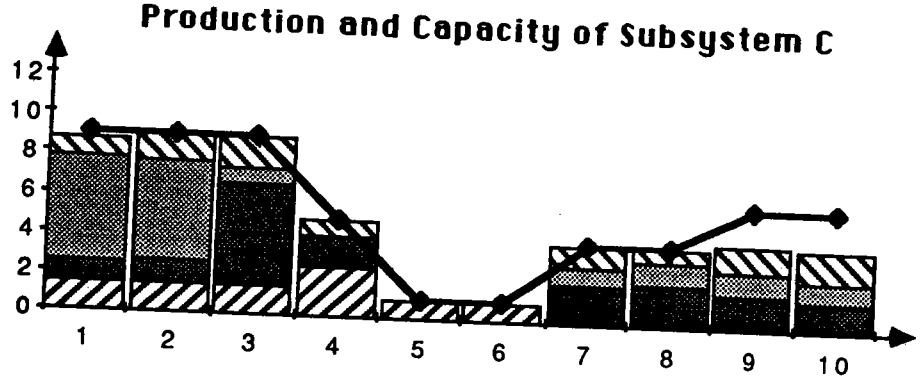
| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| product cost | 7 | | | | | | | | | | | | |
| demands | | 2 | 2 | 2 | 6 | 0 | 0 | 2 | 2 | 2 | 2 | | |
| capacity | | 4 | 4 | 6 | 6 | 3 | 0 | 0 | 3 | 3 | 3 | | |
| production | | 2 | 2 | 2 | 6 | 2 | 0 | 0 | 2 | 2 | 2 | | |
| inventory | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | inventory cost: | 28 |
| product cost | 5 | | | | | | | | | | | | |
| demands | | 1.5 | 1.5 | 1.5 | 2.5 | 1 | 1 | 0 | 0 | 0 | 0 | | |
| rmg capacity | | 2 | 2 | 4 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | | |
| production | | 2 | 2 | 4 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | | |
| inventory | 0 | 0.5 | 1 | 3.5 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | inventory cost: | 35 |

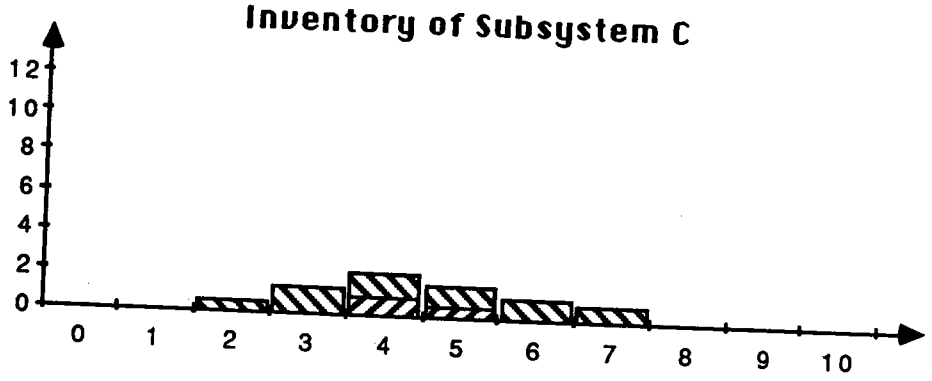

Demand and Capacity of Subsystem D
12
10
8
6
4
2
0
1
2
3
4
5
6
7
8
9
10

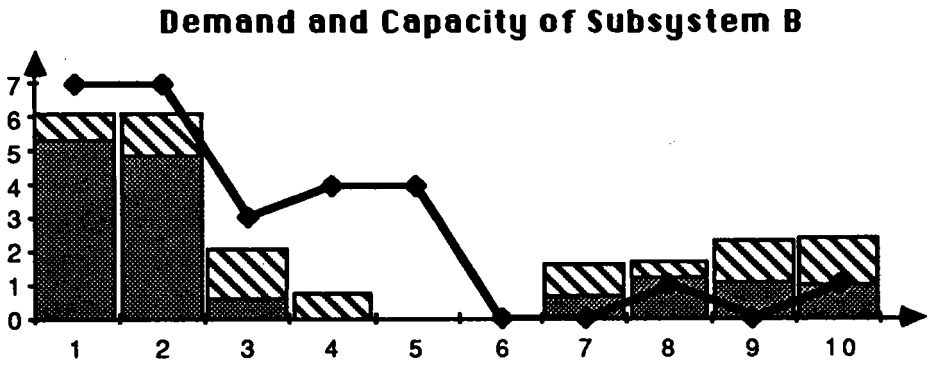

Production and Capacity of Subsystem D
12
10
8
6
4
2
0
1
2
3
4
5
6
7
8
9
10

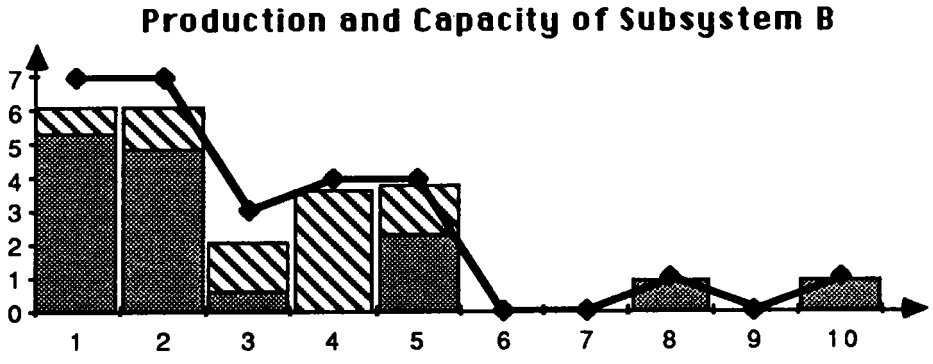

Inventory of Subsystem D
12
10
8
6
4
2
0
0
1
2
3
4
5
6
7
8
9
10

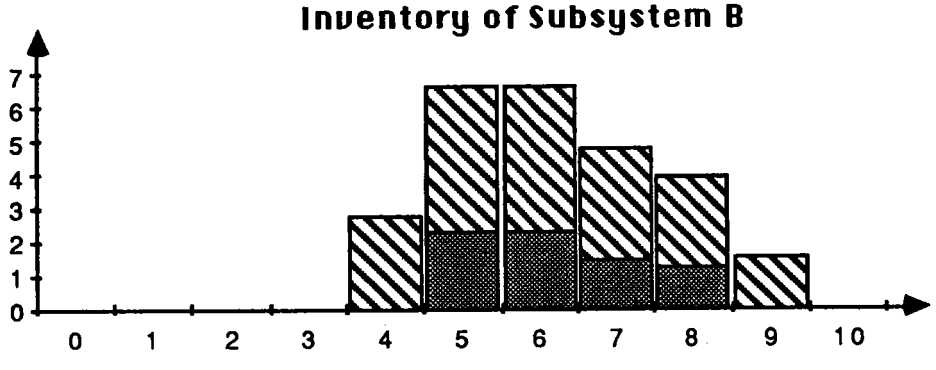

Demand and Capacity of Subsystem C
12
10
8
6
4
2
0
1
2
3
4
5
6
7
8
9
10

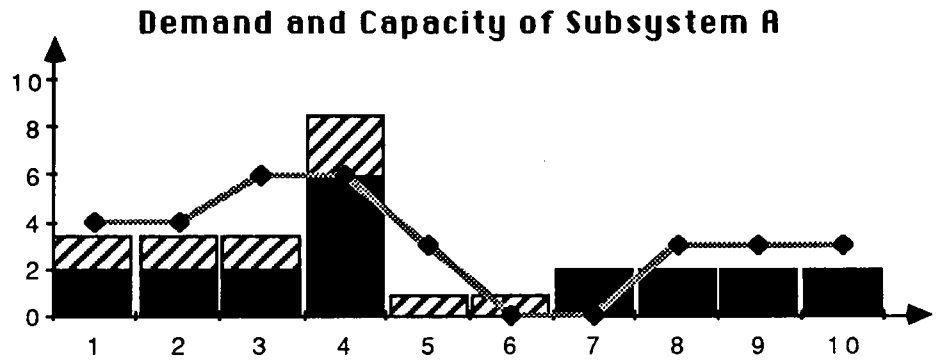

Production and Capacity of Subsystem C
12
10
8
6
4
2
0
1
2
3
4
5
6
7
8
9
10

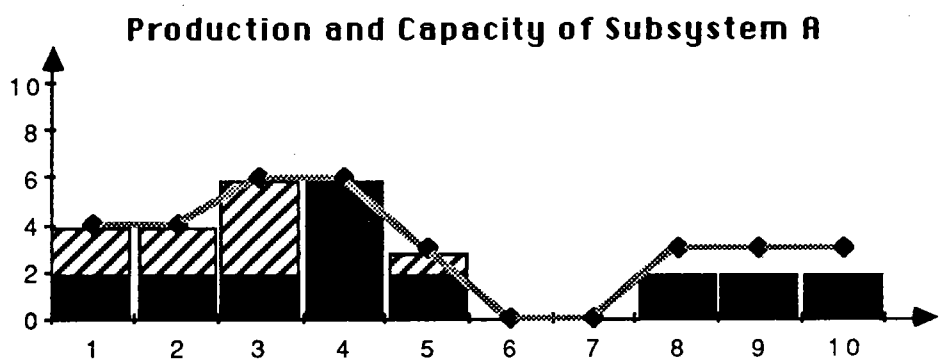

Inventory of Subsystem C
12
10
8
6
4
2
0
0
1
2
3
4
5
6
7
8
9
10

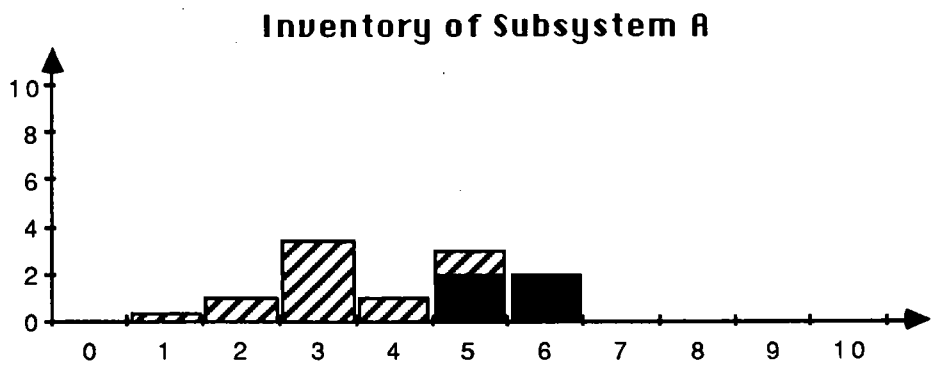

Demand and Capacity of Subsystem B
7
6
5
4
3
2
1
0
1 2 3 4 5 6 7 8 9 10


Production and Capacity of Subsystem B
7
6
5
4
3
2
1
0
1 2 3 4 5 6 7 8 9 10


Inventory of Subsystem B
7
6
5
4
3
2
1
0
0 1 2 3 4 5 6 7 8 9 10


Demand and Capacity of Subsystem A
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 10


Production and Capacity of Subsystem A
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 10


Inventory of Subsystem A
10
8
6
4
2
0
0 1 2 3 4 5 6 7 8 9 10

# CONCLUSIONS

There are many ways to improve the performance of a firm, and improving its operation by a better planning is only one of them, although not necessarily the easiest. In fact, finding ways to a better planning requires a good understanding of the interactions between different decisions and applying them supposes a commitment of all the personnel.

Up to now, the understanding gained from research has more often than not remained unapplied because the models used to conduct this research had to be focussed on a limited number of issues, and legitimately so: considering the most general model is rarely insightful, because the problems to solve are then totally intractable. The consequence of this necessary oversimplification of the models used is that the results obtained could not be directly translated into applications, except in the industries in which the issues considered would account for most of the system's behavior.

Hence the frustration of researchers who saw their understanding fade away when they tried to apply it to a real situation complicated by phenomena that they did not take into account, and the frustration of practitioners observing how the gains they could expect from research would not materialize.

The outcome of this frustration was a relative 'isolationism' on both parts: the plants in western countries were operated by managers who would acquire a local understanding of their behavior, and would use their 'common sense' and some slack to ensure a <u>smooth</u> operation; researchers, on the other hand, coined a number of mathematical problems deemed relevant to the field of production planning ('lot sizing', 'discrete scheduling',…), and worked hard to solve them. Meanwhile, the Japanese experimented with the concept of 'pushing' a production system to its limits in order to determine its weak points and improve them. Different aspects of this concept (Just In Time, Total Quality…) were successfully implemented in several companies.

The competition has since grown fiercer in many traditional fields and the decline of some major U.S. manufacturers due to the competition of better operated Japanese or Korean firms has illustrated painfully how a better production management can make the difference between profits and losses. The aluminum industry might become one more illustration of this statement: it had its best years in the past decades, when the electrolytic process was new and casting an ingot was technologically challenging. Now, technology alone cannot guarantee the profitability of a company.

The production cost of an aluminum ingot depends primarily on the cost of electricity. Therefore, the major aluminum producers have installed "state of the art" smelters in remote regions of the world with low electricity cost, thereby complicating their logistics. Also, small companies have entered the low-end of the market, capitalizing either on governmental help -in certain countries-, or on the fact that, in the production of aluminum, most of the economies of scale are realized at the plant level.

The consequence is that *large integrated companies need to improve the operations of their upstream activities* (electrolysis, casting, rolling) in order to keep them competitive in spite of increasingly cumbersome logistics, and in order to ensure the best possible service to the downstream activities, which will realize most of their profit.

This observation has motivated to a large extent the work presented in this thesis. In fact, the problem stated has been addressed at two levels: on one hand, the system to be considered typically consists of several plants producing hundreds of products and falls in the class of Large Scale Systems; therefore the control of such a system must be hierarchical. On the other hand, the objective pursued is to achieve economies of scale beyond the plant level, and a necessary condition for this objective to be achievable is that the operations of the different plants be coordinated.

The organization of this thesis followed from that analysis: the first chapter was aimed at substanciating the assertion that the management of a multiple-plant firm must be hierarchical. The second chapter is a survey of previous attempts to design hierarchical control frameworks, which showed that in most of the work considered, the production system is modelled as a single facility instead of a set of subsystems whose activity must be coordinated.

Since this was the major issue to be addressed in the context for which this research was devised (i.e. the upstream stages of an integrated aluminum company), Chapter three introduces a model that addresses it and was meant as one layer of a hierarchical controller. The physical system is represented as a network of subsystems of limited capacity through which product families "flow" to meet external demand requirements. These products undergo the multiple phases of their process in the different subsystems and can be stored between subsystems, the objective being to minimize their total flow-time through the system.

It was shown in Chapter four that the solution to this problem could be characterized sufficiently -in the most general case- to be determined by means of a numerical method, dynamic programming. Furthermore, if the inventory holding costs of the products increase with their added value, this solution can be determined analytically. More precisely, it was shown that this optimal flow-control problem would be globally solved if each subsystem determined locally the flow rates that minimize its inventory holding cost subject both to its capacity constraint and to the demand requirements imposed by the downstream subsystems. (These local problems were shown to have an analytical solution.)

Finally, Chapter five illustrates on two numerical examples the difference in complexity between the general numerical solution and the analytical solution under the restrictive cost assumptions.

## Conjectures and Future Work

A first and straightforward extension of these results would be to consider the case when the objective to minimize is a combination of linear production and inventory holding costs. It is easy enough to prove that the production plan that minimizes the total inventory holding cost also minimizes the total volume of production, and thus, the production cost if this one is linear. However, this result would not hold if the production cost was assumed concave (e.g. to account for setup costs), and the results to be found in [BP] and [BR] for concave-cost problems are slightly different from those presented here.

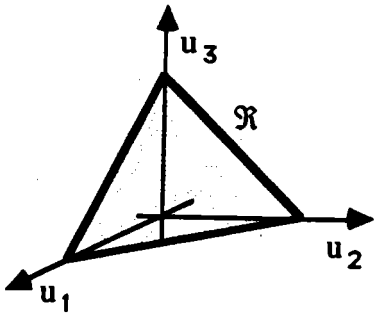
Another straightforward extension would be to consider the case of concave inventory holding costs and no production cost: the results would then probably be the same as in Chapter 4. However, the interest of the extension is questionable, because in most traditional manufacturing environments, the use of concave inventory-holding costs is not justified and, if it were, it is still very unlikely that the parameters of the cost function would be known for any application.

It was pointed out in Chap. 3 that the expression of the capacity constraint adopted for a system producing several part-types is not as general as it could be. In fact, it consists of a single inequality bounding the instantaneous production rates of these part-types:

$$\forall\, t \in I, \quad \sum_{i=1}^{p} a_i \cdot u_i(t) \leq C(t) \qquad \text{(C)}$$

(The coefficients $a_i$ weight the "capacity requirements" of the different part-types.)

This inequality defines as "feasible" -for the capacity constraint- a region of $\mathbb{R}^p$ limited by a hyperplane that intersects all the axes. Since the production rates are also bound to be positive, the feasible region $\mathfrak{R}$ considered in the work presented in Chapter 4 was a particular type of polyhedron:

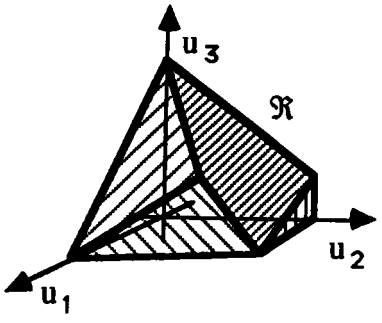



In [KI] and subsequent work, the capacity constraint for a multi machine system producing several part-types is expressed by means of a system of inequalities, one for each machine:

$$\forall k \in \{1,..,m\}, \forall t \in I, \quad \sum_{i=1}^{p} \tau_{ik} \cdot u_i(t) \leq \alpha_k(t) \qquad (S)$$

where $\tau_{ik}$ is the processing time of product i on machine k and $\alpha_k(t)$ is either the average time "up" of machine k per unit time -in a deterministic model- or a random variable equal to 1 when machine k is up and to 0 when it is down.

Each of these inequalities defines a feasible half-space bound by a hyperplane that may intersect 1 to p axes. As a result, the feasible region can be any convex polyhedron included in $\mathbb{R}_+^p$:




The physical justification of this expression of the capacity constraint requires the assumption of flow conservation. Some time should therefore be devoted to finding an aggregation procedure that would actually generate subsystems for which this assumption is valid.

Regardless of the outcome of this research, it is worth investigating the extension of the results of Chapter 3 to the case where the capacity constraint is assumed similar to (S). The optimal production plan is likely to be of the bang-bang type as it is in Chapter 3, but the important result of theorem 3.III.1 (which allowed to determine the optimal control analytically) will not hold.

In fact, this theorem, together with the results that followed, means -in geometrical terms- that the optimal production plan for a single-stage multi-product system is either equal to the demand or to a projection of the demand vector on the hyper surface defined by the capacity constraint. And its proof was based on the fact that, as regards the capacity constraint, products are interchangeable in a given ratio, regardless of the position of the production vector with respect to the capacity set. This property is no longer true in general with the type of capacity constraint defined by (S).

A more desirable extension of the work presented in this thesis would be achieved by introducing randomness in the flow-control models. It was argued in Chapter 3 that, in some cases, deterministic models were justified even in a stochastic environment. There are clearly cases, however, in which stochastic models for flow-control will prove superior.

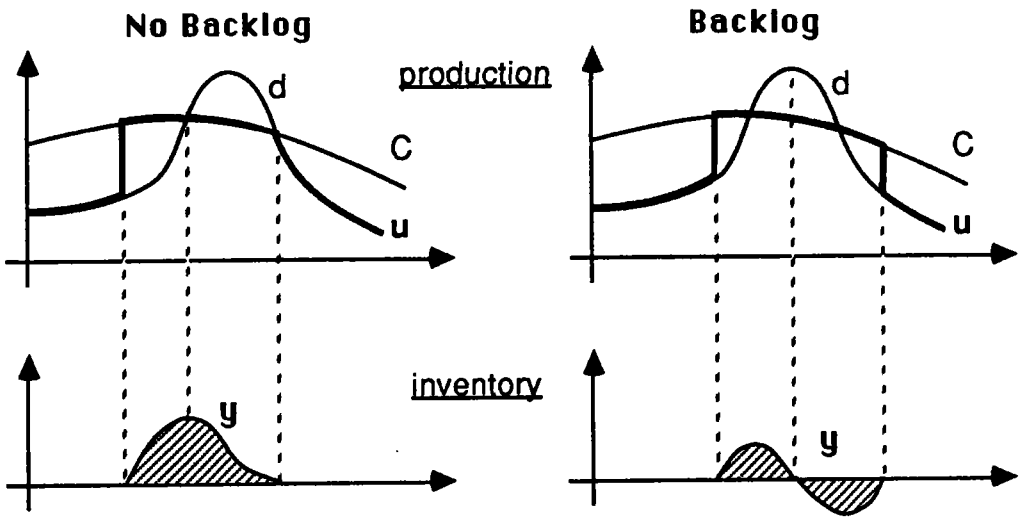
Randomness in production can be identified with "disruptions", that is, changes in the inputs required for decision-making, that actually cause a modification of previous decisions or prompt new decisions. Cast in that format, it is undestandable that randomness appear at all levels of a hierarchical management, but that the instances will be different depending on the level. On the shop-floor, for example, disruptions can be caused by machine failures, yield losses, changes in the production requirements, engineering changes, human operators' errors... At the strategic level, on the other hand, the disruptions are more likely to be related to "events" on the stock exchange or the financial markets...

The randomness typically modelled in the manufacturing literature concerns the volumes of demand, the instantaneous capacity of the system (as results from failures and repairs), and, to a lesser extent, the yield of a process or the lead-time of a product through the system. This last example illustrates a characteristic of the modelling of randomness: the variability of the lead-time of a product can be due to any (or several) of the disturbances listed hereabove. However, the effects of these disturbances can be aggregated in a model of the lead-time as a random variable.

In the industry of primary aluminum, the main source of randomness seems to be the variability of the demand mix: the total volume of demand is usually fixed, the company selling most of its production to hard-core clients on a contractual basis, and selling the remainder on the spot market to keep some flexibility. However, because the clients cannot specify long in advance the detailed product types they will need, the product-mix of the demand is known with little anticipation and can constantly be altered by the reception of a new order.

The introduction of randomness in the model of the demand would imply giving up the constraint of positive inventories, at least in all cases in which the demand rate can exceed the capacity. Backlog models have not been studied in the body of this thesis, for reasons developed in Chapter 3. However, it seems that some of the results of Chapter 4 could be extended to deterministic, models allowing backlog but penalizing deviations from a zero inventory. Consider for example the case of a single-stage, single-product system. If no backlog is allowed, the optimal production plan is obtained by the "backward smoothing" procedure of Section 4.1.

If backlog was allowed but penalized as inventory is (e.g. by a quadratic function), the optimal production plan would probably be obtained by smoothing the demand peaks both backwards and forwards, by resorting both to storage before and backlogging after the peak.




However, this research direction was not pursued, for reasons explicited in Chapter 3.

To the best of the author's knowledge, there has been no work devoted specifically to the study of the randomness of a demand mix. However, CHOONG has recently proposed a jump process model that could be appropriate to represent this particular type of randomness in demand. For a single stage, multi product system bound to satisfy this type of demand, the optimal control was shown to have the same characteristics as the optimal control in the case investigated by KIMEMIA (constant demand, and capacity modelled as a jump process).

An interesting question to be answered concerning this optimal control (in fact, a refined version of a bang-bang control), is whether or not it is a stochastic jump process and, if it is, what its characteristics are, given those of the demand. If it were a jump process as the final demand is, one can easily imagine a multi stage system in which each stage would be controlled individually to satisfy the stochastic demand imposed by the next downstream stage. It would then be particularly interesting also to determine whether the entire control is globally optimal for the multi stage system under cost conditions analogous to those introduced in Chapter 4...

[AC] AKELLA, R., Y.F. CHOONG and S.B. GERSHWIN, "Performance of Hierarchical Production Scheduling Policy.", *IEEE Trans. on Components, Hybrids and Manufacturing Technology*, Vol. CHMT-7, No. 3, 1984.

[AA] ANDERSSON, H., S. AXSÄTER and H. JÖNSSON, "Hierarchical Material Requirement Planning.", *Intern. Journ. Prod. Res.*, Vol. 19, No. 1, 1981.

[AN] ANTHONY, R.N., "Planning and Control Systems : A Framework for Analysis.", *Harvard University, Graduate School of Business Administration*, Boston, MA., 1965.

[AO1] AOKI, M., "Control of Large Scale Dynamic Systems by Aggregation.", *IEEE Transactions on Automatic Control*, Vol. AC-13, No. 3, 1968.

[AO2] AOKI, M., "Some Approximation Methods for Estimation and Control of Large Scale Systems.", *IEEE Transactions on Automatic Control*, Vol. AC-23, No. 2, 1978.

[AH] ARMSTRONG, R.J. and A.C. HAX, "A Hierarchical Approach for a Naval Tender Job Shop Design.", Technical report No. 101, *Operations Research Center*, M.I.T., Cambridge, MA., 1974.

[AX] AXSÄTER, S., "Aggregation of Product Data for Hierarchical Production Planning.", *Operations Research*, Vol. 29, No. 4, 1981.

[BA] BAKER, K.R., "An Experimental Study of the Effectiveness of Rolling Schedules in Production Planning.", *Decision Sciences*, Vol. 8, 1977.

[BC] BAKER, T.E. and D.E. COLLINS, "The Integration of Planning, Scheduling, and Control for Automated Manufacturing.", *N.B.S. Special Publication 724*, R. Jackson and A. Jones eds., 1986.

[BF] BAUMOL, W.J. and T. FABIAN, "Decomposition, Pricing for Decentralization and External Economies.", *Management Science*, Vol. 11, No. 1, 1964.

[BO] BECHLER, E., J.M. PROTH and K. VOYATZIS, "Aritificial Memory in Production Management', Research report No. 336, I.N.R.I.A., Centre du Rocquencourt, France, 1984.

[BR] BENSOUSSAN, A., M. CROUHY and J.M. PROTH, "Mathematical Theory of Production Planning.", Elsevier Science Pub. Co., 1983.

[BP] BENSOUSSAN, A. and J.M. PROTH, "Inventory Planning in a Deterministic Environment: Continuous Time Model with Concave Costs.", *European Journal of Operational Research*, Vol. 15, 1984.

[BX1] BITRAN, G.R. and A.C. HAX, "On the Solution of Convex Knapsack Problems with Bounded Variables", Technical report No. 121, *Operations Research Center*, M.I.T., Cambridge, MA., 1976.

[BX2] BITRAN, G.R. and A.C. HAX, "On the Design of Hierarchical Planning Systems.", *Decision Sciences*, Vol. 8, 1977.

[BS1] BITRAN, G.R., E.A. HAAS and A.C. HAX, "Hierarchical Production Planning : A Single Stage System.", *Operations Research*, Vol. 29, No. 4, 1981.

[BS2] BITRAN, G.R., E.A. HAAS and A.C. HAX, "Hierarchical Production Planning : A Two Stage System.", Technical report No. 179, *Operations Research Center*, M.I.T., Cambridge, MA., 1980.

[CA] CANDEA, D.I., "Issues of Hierarchical Planning in Multi-stage Production Systems.", Technical report No. 134, *Operations Research Center*, M.I.T., Cambridge, MA., 1977.

[CC] CHARNES, A., R.W. CLOWER and K.O. KORTANEK, "Effective Control through Coherent Decentralization with Preemptive Goals.", *Econometrica*, Vol. 35, No. 2, 1967.

[CH] CHEN, H., A. HARRISON, A. MANDELBAUM, A. van ACKERE and L.M. WEIN, "Queuing Network Models of Semiconductor Wafer Fabrication.", *Stanford University, Center for Integrated Systems*, 1986.

[CK] CHOW, J.H. and P.V. KOKOTOVIC, "Time Scale Modeling of Sparse Dynamic Networks.", *IEEE Trans. on Automatic Control*, Vol. AC-30, No. 8, 1985.

[CW] CODERCH, M., A.S. WILLSKY, S.S. SASTRY and D.A. CASTAÑON, "Hierarchical Aggregation of Linear Systems with Multiple Time-Scales.", *IEEE Transactions on Automatic Control*, Vol. AC-28, No. 11, 1983.

[CO] COHEN, G., "Optimization by Decomposition and Coordination : A Unified Approach.", *IEEE Transactions on Automatic Control*, Vol. AC-23, No. 2, 1978.

[DQ] DELEBECQUE, F. and J.P. QUADRAT, "Optimal Control of Markov Chains Admitting Strong and Weak Interactions.", *Automatica*, Vol. 17, No. 2, 1981.

[DF] DEMPSTER, M.A.H., M.L. FISHER, B. LAGEWEG, L. JANSEN, J.K. LENSTRA and A.H.G. RINNOY KAN, "Analytic Evaluation of Hierarchical Planning Systems.", *Operations Research*, Vol. 29, No. 4, 1981.

[DJ] DIRICKX, Y.M.I. and L.P. JENNERGREN, "Systems Analysis by Multilevel Methods.", *International Series on Applied Systems Analysis*, Wiley, 1979.

[DL] DONOGHUE, J.F. and I. LEFKOWITZ, "Economic Tradeoffs Associated with a Multilayer Control Strategy for a Class of Static Systems.", *IEEE Transactions on Automatic Control*, Vol. AC-17, No. 1, 1972.

[DG] DZIELINSKI, B. and R. GOMORY, "Optimal Programming of Lot Sizes, Inventories and Labor Allocation.", *Management Science*, Vol. 7, No. 9, 1965.

[EL] ECKMAN, D.P. and I. LEFKOWITZ, "Principles of Model Techniques in Optimizing Control.", *Proceedings of the first IFAC Congress*, Butterworths, 1960.

[EF] ERSCHLER, J., G. FONTAN and C. MERCE, "Consistency of the Disaggregation Process in Hierarchical Planning.", *Operations Research*, Vol. 34, No. 3, 1986.

[FB1] FINDEISEN, W., F.N. BAILEY, M. BRDYS, K. MALINOWSKI, P. TATIEWSKI and A. WOZNIAK, "On-line Hiearchical Control for Steady-state Systems.", *IEEE Transactions on Automatic Control*, Vol. AC-23, No. 2, 1978.

[FB2] FINDEISEN, W., F.N. BAILEY, M. BRDYS, K. MALINOWSKI, P. TATIEWSKI and A. WOZNIAK, *Control and Coordination of Hierarchical Systems*, Wiley, 1979.

[FV] FORESTIER, J.P. and P. VARAYIA, "Multilayer Control of Large Markov Chains", *IEEE Transactions on Automatic Control*, Vol. AC-23, No. 2, 1978.

[GA] GABBAY, H., "A Hierarchical Approach to Production Planning.", Technical report No. 120, *Operations Research Center*, M.I.T., Cambridge, MA., 1975.

[GK1] GELDERS, L. and P. KLEINDÖRFER, "Coordinating Aggregate Planning and Detailed Scheduling in the One-machine Job Shop : Theory.", *Operations Research*, Vol. 22, No. 1, 1974.

[GK2] GELDERS, L. and P. KLEINDÖRFER, "Coordinating Aggregate Planning and Detailed Scheduling in the One-machine Job Shop : Computation and Structure.", *Operations Research*, Vol. 23, No. 2, 1975.

[GW] GELDERS, L.F. and L.N. van WASSENHOVE, "Hierarchical Integration in Production Planning : Theory and Practice.", *Journal of Operations Management*, Vol. 3, No. 1, 1982.

[GE1] GERSHWIN, S.B., "A Hierarchical Scheduling Policy Applied to Printed Circuit Board Assembly.", *Report No. LIDS-R-1395*, Laboratory for Information and Decision Systems, M.I.T., Cambridge, MA., 1984.

[GE2] GERSHWIN, S.B., "Stochastic Scheduling and Set-ups in Flexible Manufacturing Systems.", *Proceedings of the 2nd ORSA/TIMS Conference on Flexible Manufacturing Systems : O.R. Models and Applications*, K. Stecke and R. Suri eds., 1986.

[GE3] GERSHWIN, S.B., "A Hierarchical Framework for Discrete Event Scheduling in Manufacturing Systems.", presented at the *IIASA Workshop on Discrete Event Systems: Models and Applications*, Sopron, Hungary, August 1987, to appear in the *I.E.E.E. Proceedings, Special Issue on Dynamics of Discrete Event Systems*, February 1988.

[GE4] GERSHWIN, S.B., "A Hierarchical Framework for Manufacturing Systems Scheduling.", *Proceedings of the 26th I.E.E.E. Conference on Decision and Control*, Los Angeles, California, December 1987.

[GC] GERSHWIN, S.B., R. AKELLA and Y.F. CHOONG, "Short Term Production of an Automated Manufacturing Facility.", *I.B.M. Journal of Research and Development*, Vol. 29, No. 4, 1985.

[GH] GERSHWIN, S.B., R.R. HILDEBRANDT, R. SURI and S.K. MITTER, "A Control Perpective on Recent Trends in Manufacturing Systems.", *IEEE Control Systems Magazine*, Vol. 6, No. 2, 1986.

[GD] GOLDRATT, E.M. and J. COX, *The Goal*, North River Press, Croton on Hudson, N.Y. 10520, U.S.A., 1986.

[GO] GOLOVIN, J.J., "Hierarchical Integration of Planning and Control.", Technical report No. 116, *Operations Research Center*, M.I.T., Cambridge, MA., 1975.

[GR] GRAVES, S.V., "Using Lagragian Techniques to Solve Hierarchical Production Planning Problems.", *Management Science*, Vol.28, No.3, 1982.

[GN] GREEN, R.S., "Heuristic Coupling of Aggregate and Detailed Models in Factory Scheduling.", unpublished P.H.D. thesis, M.I.T., Cambridge, MA., 1971.

[HH] HAAS, E.A., A.C. HAX and R.E. WELSCH, "A Comparison of Heuristic Methods Used in Hierarchical Production Planning.", Technical report No. 160, *Operations Research Center*, M.I.T., Cambridge, MA., 1979.

[HL] HACKMAN, S.T. and R.C. LEACHMAN, "An Aggregate Model of Project Oriented Production.", *Operations Research Center*, U.C. Berkeley, February 1987.

[HA1] HAX, A.C., "Integration of Strategic and Tactical Planning in the Aluminum Industry.", Technical report No. 86, *Operations Research Center*, M.I.T., Cambridge, MA., 1973.

[HA2] HAX, A.C., "The Design of Large Scale Logistics Systems : A Survey and an Approach.", W. Marlow ed., in *Modern Trends in Logistics Research*, Cambridge, MA. : M.I.T. Press, 1976.

[HA3] HAX, A.C., "Aggregate Production Planning.", in *Handbook of Operations Research*, J. Moder and S. Elmaghraby eds., Van Nostrand Reinhold, New York, 1978.

[HG] HAX, A.C. and J.J. GOLOVIN, "A Computer Based Operations Management System (COMS*).", *Studies in Operations Management*

[HM] HAX, A.C. and H.C. MEAL, "Hierarchical Integration of Production Planning and Scheduling.", in *Studies in the Management Sciences*, M.A.Geisler, ed., Vol.1, *Logistics*, North Holland - American Elsevier, 1975.

[HE] HILLION, H., K. MEIER and J.M. PROTH, "Production Subsystems and Part-Families: the Top-level Model in Hierarchical Production Planning Systems.", presented at the *11th Triennial Conference on Operations Research*, august 1987, Buenos Aires.

[HP] HILLION, H. and J.M. PROTH, "Performance Evaluation of Job-shop Systems Using Timed Event-Graphs.", submitted to the *IEEE Transactions on Automatic Control*.

[HO] HOLSTEIN, W.K., "Production Planning and Control Integrated.", *Harvard Business Review*, Vol. 46, No. 3, 1968.

[IE] I.E.E.E., "Data Driven Automation", *IEEE SPECTRUM*, May 1983.

[IM] IMBERT, S., "Interaction entre deux Niveaux de Décision en Planification de la Production.", thèse de 3ème cycle, Université Paul Sabatier, Toulouse, 1986.

[JA] JAIKUMAR, R., "Postindustrial Manufacturing.", *Harvard Business Review*, Nov.-Dec. 1986.

[KA] KARMARKAR, U.S., "Equalization of Run-out Times", *Operations Research*, Vol. 29, No. 4, 1981.

[KI] KIMEMIA, J.G., "Hierarchical Control of Production in Flexible Manufacturing Systems.", *Report No. LIDS-TH-1215*, Laboratory for Information and Decision Systems, M.I.T., Cambridge, MA., 1982.

[KG] KIMEMIA, J.G. and S.B. GERSHWIN, "An Algorithm for the Computer Control of Production in Flexible Manufacturing Systems.", *IIE Trans.*, Vol. 15, No. 4, 1983.

[KN] KLEINDÖRFER, P. and E.F.P. NEWSON, "A Lower Bounding Structure for Lot Size Scheduling Problems.", *Operations Research*, Vol. 23, No. 2, 1975.

[KL] KORNAI, J, and T. LIPTAK, "Two Level Planning.", *Econometrica*, Vol. 33, No. 1, 1965.

[KR] KRAJEWSKI, L.J. and L.P. RITZMAN, "Disaggregation in Manufacturing and Service Organizations : Survey of Problems and Research.", *Decision Sciences*, Vol. 8, 1977.

[KY] KYDLAND, F., "Hierarchical Decomposition of Linear Economic Models.", *Management Science*, Vol. 21, No. 9, 1975.

[LA] LASDON, L.S., "Duality and Decomposition in Mathematical Programming", *IEEE Transactions on Systems Science and Cybernetics*, Vol. SSC-4, No. 2, 1968.

[LT] LASDON, L.S. and R.C. TERJUNG, "An Efficient Algorithm for Multi-item Scheduling.", *Operations Research*, Vol. 19, 1971.

[LM] LASSERRE, J.B., J.P. MARTIN and F. ROUBELLAT, "Aggregate Model and Decomposition for Mid-term Production Planning.", *Intern. Journ. Prod. Res.*, Vol. 21, No. 6, 1983.

[LO] LAWRENCE, S.R. and T.E. MORTON, "Patriarch: Hierarchical Production Scheduling.", *N.B.S. Special Publication 724*, R. Jackson and A. Jones eds., 1986.

[LE] LEFKOWITZ, I., "Multilevel Approach Applied to Control System Design.", *Trans. ASME*, Vol. 88, 1966.

[LZ] LOOZE, D.P., "Hierarchical Control and Decomposition of Decentralized Linear Stochastic Systems.", unpublished P.H.D. thesis, M.I.T., Cambridge, MA., 1978.

[MM] MACKULAK, G.T., C.L. MOODIE and T.J. WILLIAMS, "Computerized Hierarchical Production Control in Steel Manufacture.", *Intern. Journ. Prod. Res.*, Vol. 18, No. 4, 1980.

[MG] MAIMON, O.Z. and S.B. GERSHWIN, "Dynamic Scheduling and Routing for Flexible Manufacturing Systems that have Unreliable Machines.", *Report No. LIDS-TH-1610*, Laboratory for Information and Decision Systems, M.I.T., Cambridge, MA., 1986.

[MA] MANNE, A.S., "Programming of Economic Lot Sizes.", *Management Science*, Vol. 4, No. 2, 1958.

[MU] MAXWELL, W., J.A. MUCKSTADT, J. THOMAS and J. VANDEREECKEN, "A Modeling Framework for Planning and Control of Production in Discrete Parts Manufacturing Systems and Assembly Systems.", *Interfaces*, Vol. 13, 1983.

[ME] MEAL, H.C., "Putting Decisions where they Belong.", *Harvard Business Review*, Mar.-Apr. 1984.

[MP] MEIER, K. and J.M. PROTH, "Scheduling in Large Scale Production Systems: a Medium Term Production Management Model", *Proc. of the 2nd International Conference on Production Systems*, Paris, 1987.

[MC] MESAROVIC, M.D., D. MACKO and Y. TAKAHARA, *Theory of Multilevel Hierarchical Systems*, New York : Academic, 1970.

[MS] MORTON, T.E. and T.L. SMUNT, "A Planning and Scheduling System for Flexible Manufacturing.", *Flexible Maufacturing Systems: Methods and Studies*, KUSIAK ed., North-Holland, 1986.

[NE1] NEWSON, E.F.P., "Multi-Item Lot Size Scheduling by Heuristic Part I : with Fixed Resources.", *Management Science*, Vol. 21, No. 10, 1975.

[NE2] NEWSON, E.F.P., "Multi-Item Lot Size Scheduling by Heuristic Part II : with Variable Resources.", *Management Science*, Vol. 21, No. 10, 1975.

[OM] O'GRADY, P.J. and U. MENON, "A Hierarchy of Intelligent Scheduling and Control for Automated Manufacturing Systems.", *N.B.S. Special Publication 724*, R. Jackson and A. Jones eds., 1986.

[OL] OLSON, C.D., "A Prototype System for Hierarchical Production Planning.", unpublished M.S. thesis, M.I.T., Cambridge, MA., 1983.

[PA] PAPAS, P.N., "ISIS Project in Review.", *N.B.S. Special Publication 724*, R. Jackson and A. Jones eds., 1986.

[PN] PARNABY, J., "Concept of a Manufacturing System", *Intern. Journ. Prod. Res.*, Vol. 17, No. 2, 1979.

[PE] PENDROCK, M.E., "A Hierarchical Approach to Integrated Production and Distribution Planning.", unpublished M.S. thesis, M.I.T., Cambridge, MA., 1978.

[PR] PROTH, J.M., "Gestion de Stocks avec Coûts Concaves, Notion d'Horizon de Planification.", *Revue Sciences de Gestion*, No. 2, 1981.

[RU] RUEFLI, T.W., "A Generalized Goal Decomposition Model.", *Management Science*, Vol. 17, No. 8, 1971.

[SA] SAATY, T.L., *The Analytic Hierarchical Process : Planning, Priority Setting, Manpower Allocation*, Mc Graw-Hill 1980.

[SV] SANDELL, N.R. Jr., P. VARAYIA, M.A. ATHANS and M. SAFONOV, "A Survey of Decentralized Control Methods for Large Scale Systems.", *IEEE Transactions on Automatic Control*, Vol. AC-23, No. 2, 1978.

[SH] SHAW, M., "A Two-Level Planning and Scheduling Approach for Computer Integrated Manufacturing.", *N.B.S. Special Publication 724*, R. Jackson and A. Jones eds., 1986.

[SO] SHOENBERGER, R.J., "Some Observations on the Advantages and Implementation Issues of Just-in-Time Production Systems.", *Journal of Operations Management*, Vol. 3, No. 1, 1982.

[SW] SHWIMER, J., "Interaction between Aggregate and Detailed Scheduling in a Job Shop.", Technical report No. 71, *Operations Research Center*, M.I.T., Cambridge, MA., 1972.

[SI] SINGH, M.G. *Dynamical Hierarchical Control*, Elsevier, rev. 1982.

[SD] SINGH, M.G., S.A.W. DREW and J.F. COALES, "Comparison of Practical Hierarchical Control Methods for Interconnected Dynamical Systems.", *Automatica*, Vol. 11, 1975.

[ST] SINGH, M.G. and A. TITLI, "Closed Loop Hierarchical Control for Non-linear Systems Using Quasilinearisation.", *Automatica*, Vol. 11, 1975.

[SS] SMITH, N. and A.P. SAGE, "An Introduction to Hierarchical Systems Theory.", *Computers and Electrical Engeneering*, Vol. 1, 1973.

[VC] VILLA, A., E. CANUTO and S. ROSSETTO, "A Hierarchical Part Routing Control Scheme for Flexible Manufacturing Systems.", *Proc. of the 3rd Bilateral Meeting G.D.R.-Italy on Advances in Informational Aspects of Industrial Automation*, Berlin, 1985.

[VO] VILLA, A., A. CONTI , F. LOMBARDI and S. ROSSETTO, "A Hierarchical Approach Model and Control Manufacturing Systems.", *Material Flow, Special Issue on Material Handling in Flexible Manufacturing Systems*, A. KUSIAK ed., 1984.

[VM] VILLA, A., R. MOSCA and G. MURARI, "Expert Control Theory : a Key for Solving Production Planning and Control Problems in Flexible Manufacturing.", *Proc. of the 1986 IEEE Conf. on Robotic and Automation*, San Francisco, CA.

[VR] VILLA, A. and S. ROSSETTO, "Towards a Hierarchical Structure for Production Planning and Control in Flexible Manufacturing Systems.", *Modeling and Design of Flexible Manufacturing Systems*, Kusiak ed., Elsevier, 1986.

[WW] WAGNER, H.M. and T.M. WHITIN, "Dynamic Version of the Economic Lot Size Problem.", *Management Science*, Vol. 5, 1958.

[WL] WASHINGTON, L.A. and A.H. LEVIS, "Effectiveness Analysis of Flexible Manufacturing Systems.", *Proc. of the 1986 IEEE Conf. on Robotic and Automation*, San Francisco, CA.

[WI] WINTERS, P.R., "Constrained Inventory Rules for Production Smoothing.", *Management Science*, Vol. 8, No. 4, 1962.

[WS] WISMER, D.A. (ed.) *Optimization Methods for Large Scale Systems… with applications*, Mc Graw Hill, 1971.

[ZA] ZANGWILL, W.I., "A Backlogging Model and a Multi-Echelon Model of a Dynamic Economic Lot Size Production System - A Network Approach.", *Management Science*, Vol. 15, No. 9, 1969.

[ZO] ZOLLER, K., "Optimal Disaggregation of Aggregate Production Plans.", *Management Science*, Vol. 17, 1971.

# APPENDIX :

# SYNTHESE EN FRANÇAIS

L'objectif d'une entreprise est de réaliser un profit tout en s'assurant qu'elle pourra continuer à le faire ; toute l'activité d'un tel système et toutes les décisions qui la gouvernent devraient donc être déterminées de façon à maximiser l'espérance du profit à réaliser sur l'horizon stratégique. Il est clair, cependant, que toutes les tâches qui seront effectuées durant la vie d'une entreprise ne sont pas planifiées le jour de sa création. En effet, la taille du problème qu'il faudrait résoudre est beaucoup trop importante, et, si elle ne l'était pas, les informations nécessaires à la prise de décisions détaillées n'étant connues avec une fiabilité suffisante qu'à très court terme, il est impossible de prendre ces décisions à l'avance.

L'approche adoptée par tout être humain face à un tel type de problème est essentiellement hiérarchique ; en effet elle consiste avant tout à aggréger les tâches en actions plus globales -autant de fois que la complexité du problème le requiert-, puis à prendre des décisions concernant ces actions et enfin seulement à décider et accomplir, en temps voulu, les tâches correspondant à ces actions. Cette approche présente un triple avantage : d'une part, les problèmes à traiter sont de taille plus modeste que le problème initial, d'autre part, le type d'information requis pour les résoudre est plus fiable, et enfin, cette approche permet de décomposer la résolution du problème, ce qui est particulièrement important dans la mesure où la spécialisation augmente l'efficacité.

On peut ainsi, par exemple, classer les décisions à prendre dans une entreprise en fonction de leur horizon (c'est la désormais célèbre classification en décisions opérationnelles, tactiques et stratégiques), mais on peut aussi les classer en fonction du type de connaissance et d'information qu'elles requièrent : connaissance du marché, connaissances techniques concernant le produit et le procédé de fabrication, et connaissance du système de production. La gestion de production concerne les décisions correspondant à cette dernière catégorie.

Dans la pratique, la gestion de production est donc hiérarchisée au même titre que tout processus décisionnel humain concernant un problème de taille importante et ayant une composante non déterministe. Ce ne serait pas là une justification pour l'adoption d'une approche hiérarchisée pour la gestion informatisée si les ordinateurs avaient une puissance de calcul infinie et si l'on pouvait modéliser tous les aléas auxquel est soumis un système de production. Comme il n'en est rien, l'approche hiérarchisée se trouve justifiée. C'est pourquoi la première partie de cette thèse est consacrée à un état de l'art en la matière. De cette étude il ressort que la plupart des travaux de recherche en gestion de production ont porté sur des sous-problèmes clairement identifiés par la pratique à différents niveaux (dimensionnement d'outil, ordonnancement…). En fait, il semble que trois "écoles" seulement se soient distinguées par leurs travaux au sujet des hiérarchies :

L'école MESAROVIC a introduit un vocabulaire pour la théorie de la commande hiérarchisée et s'est intéressée aux algorithmes à structure hiérarchique, c'est-à-dire permettant la résolution d'un problème par résolution coordonnée de plusieurs sous-problèmes.

L'école HAX a été la première à proposer un système de gestion hiérarchisée. Ce système avait pour objet de déterminer les volumes de production hebdomadaires d'une usine ayant à satisfaire une demande à variations saisonnières ; l'approche adoptée pour résoudre ce problème était hiérarchique dans la mesure où elle consistait à résoudre successivement trois problèmes à horizons décroissants, la solution de chacun des deux premiers définissant les contraintes du suivant. De plus, les entités manipulées à chaque niveau étaient fonction de l'horizon : ainsi, les prévisions de vente concernant des aggrégats de produit étant plus fiables que celles concernant les produit détaillés, c'est en produits aggrégés qu'on raisonnait pour prendre en compte les variations saisonnières. Egalement, les critères étaient choisis en tenant compte du type d'entité manipulé et du fait que, par construction, les décisions prises à chaque niveau contraignaient les décisions des niveaux plus bas et de ce fait avaient plus d'impact sur la qualité de la solution.

Le principal défaut de cette approche est sa trop grande spécifité.

Récemment, GERSHWIN a proposé une méthode de conception de systèmes de gestion hiérarchisée : l'hypothèse de base est qu'on doit pouvoir classifier par fréquence les évènements se produisant dans un atelier, qu'ils soient ou non contrôlables ; ainsi, chaque niveau de la hiérarchie fonctionne à une fréquence, c'est-à-dire décide des lancements d'activités ayant une fréquence donnée. La coordination entre les niveaux de la hiérarchie est assurée par un mécanisme de transmission d'objectifs exprimés comme des "débits d'activités". Par exemple, les lancements étant moins fréquents que les productions de pièces, le niveau "lancement" décide à quels instants les machines seront configurées pour la production des différents types de pièces et impose des débits de ces pièces au niveau "production" pour chaque intervalle de temps qui lui est alloué sur les machines. En effet, l'objectif de production reçu par le niveau "lancement" correspond à un certain débit des différentes pièces, mais comme les machines ne sont occupées à produire chaque type de pièce qu'une fraction de leur temps, il faut qu'elles produisent pendant ces intervalles à un débit plus élevé.

Cette approche est apparemment la première à présenter un certain caractère de généralité ; elle est encore très incomplète dans la mesure où elle ne considère que l'aspect temporel de la hiérarchie, et elle repose sur une hypothèse très restrictive, à savoir la "conservation des débits". GERSHWIN fait l'hypothèse que le débit d'une activité est le même sur toutes les resources requises par cette activité. En d'autres termes, cette hypothèse signifie que le système bascule sans transition entre différents régimes permanents où par exemple toutes les opérations de la gamme de chaque pièce produite sont effectuées au même débit.

Cette hypothèse est visiblement très restrictive si l'on a des gammes longues et, de façon plus générale, sa validité diminue avec le nombre d'entités (produits, machines) manipulées.

Une réponse possible à cette déficience serait la décomposition "spatiale" de la gestion ; cette solution consisterait simplement à décomposer le système en sous-systèmes et à assurer la coordination de la gestion de ces sous-systèmes.

L'objet de la seconde partie de cette thèse est l'étude d'une certaine classe de modèles pouvant correspondre au niveau "haut" d'un système de gestion hiérarchisée. En effet, le système de production est représenté comme un réseau de sous-systèmes écoulant des familles de produits (ce type de structure peut être obtenu par certaines méthodes de décomposition croisée), et l'objectif est de le faire fonctionner à flux tendus pour satisfaire une demande externe, sachant que les sous-systèmes ont des capacités finies. On fait donc l'hypothèse qu'un "sens de flux" peut être défini dans ce système, c'est-à-dire qu'il est du type "flow-shop". Le problème revient alors à minimiser le coût de stockage cumulé sur l'horizon de planification, sous des contraintes de capacité et de positivité des stocks. On fait l'hypothèse que les coûts de stockage sont linéaires.

Le modèle choisi est en temps continu avec demandes et capacités variables mais déterministes. On caractérise tout d'abord le flux optimal pour un système à un niveau et produisant un seul produit : à l'optimum, le débit de production est égal à la demande sur certains intervalles de l'horizon, et égal à la capacité sur le reste de l'horizon. De plus, le flux optimal est obtenu en produisant aussi tard que possible.

On considère ensuite un système composé de sous-systèmes en série, toujours avec un seul produit ; on démontre que si les coûts de stockage augmentent avec le stade d'achèvement du produit, alors le flux optimal est obtenu en résolvant des problèmes du type précédent, en commençant par le sous-système le plus en aval et en "remontant la demande". On montre aussi qu'en revanche, si l'on ne fait pas cette hypothèse sur les coûts, il n'y a pas de méthode de résolution simple, c'est-à-dire qui puisse être caractérisée indépendamment des données, comme elle l'est sous cette hypothèse.

On considère ensuite un système unique produisant plusieurs types de pièces et on démontre que le flux optimal s'obtient en résolvant plusieurs problème "mono-produit" : on détermine d'abord le flux du produit le plus cher au stockage, on diminue d'autant la capacité disponible pour les autres produits et on itère cette procédure. Les débits de production à chaque instant sont donc égaux soit à la demande à cet instant, soit au solde de capacité laissé par la production de produits plus chers au stockage, soit à zéro si la capacité a été épuisée.

Pour pouvoir étendre ces résultats relativement puissants au système le plus général, c'est-à-dire composé de plusieurs sous systèmes en série et produisant plusieurs produits, il est nécessaire de faire des hypothèses de coût assez restrictives. Il faut en effet que les coûts de stockage des produits soient dans le même ordre à tous les sous-systèmes, qu'ils augmentent avec la valeur ajoutée du produit, et surtout que les incréments de coût entre niveaux soient dans le même ordre que les coûts eux-mêmes. L'exemple le plus simple d'une structure de coût ayant ces propriétés est celui où chaque coût est le produit de deux facteurs, l'un propre au produit considéré, et l'autre propre au sous-système où on le considère.

Sans ces hypothèses, on peut quand même caractériser les flux optimaux suffisamment pour envisager de résoudre le problème par la programmation dynamique. Cependant, une évaluation de la complexité de l'algorithme montre que l'on se heurte à des problèmes de dimensionalité même pour des systèmes de taille réduite. Une application de chacune des deux méthodes (programmation dynamique et algorithme "séquentiel") illustre ce propos.

**Imprimé en France**
**par**
**l'Institut National de Recherche en Informatique et en Automatique**

## ABSTRACT

Production Management is concerned with a class of decisions to be made in a manufacturing firm in order to gear it towards its objective. Since this decision making problem is very large, it must be approached hierarchically. Hierarchical production management systems are characterized by several decision levels operating in a coordinated fashion. Designing such systems means defining the models to be used at each level (entities, objective, horizon), and a coordination procedure. The models studied in this work are devised for the higher levels of a hierarchy; the production system is represented as a network of subsystems with limited capacity and the objective sought is to minimize the flow time of product families. It is proved that under certain assumptions concerning the inventory holding costs, a very simple algorithm exists to solve this deterministic optimization problem. It is then shown that it is possible to relax this assumption by using dynamic programming but the amount of computations required increases dramatically.



key words: hierarchical control, flow control, finite capacity, inventory, deterministic optimizing.



## RESUME

La gestion de production s'intéresse à une classe de décisions à prendre dans une entreprise de production de façon à lui faire atteindre son objectif. Comme le problème à résoudre est très vaste, il faut l'aborder au moyen d'une approche hiérarchisée. Les sytèmes de gestion hiérarchiques se caractérisent par plusieurs niveaux de décision coordonnés. Concevoir de tels systèmes suppose de définir les modèles à utiliser à chaque niveau (entités, objectif, horizon), et une procédure de coordination. Les modèles étudiés dans ce mémoire sont destinés au niveau haut d'un système hiérarchique; l'outil de production est représenté comme un réseau de sous-systèmes à capacités finies et l'objectif à atteindre est la production à flux tendus de familles de produits. On démontre que pour certaines structures de coûts de stockage, il existe un algorithme très simple pour résoudre ce problème d'optimisation déterministe. On montre également qu'il est possible de relaxer cette contrainte et d'utiliser la programmation dynamique, mais le volume de calcul requis s'en trouve considérablement augmenté.



mots-clés: gestion hiérarchisée, controle de flux, capacité finie, stocks, optimisation déterministe.

ISBN 2 - 7261 - 0532 -7