January 7, 1988 LIDS-P-1734

HIERARCHIES IN PRODUCTION MANAGEMENT AND CONTROL: A SURVEY

by

Camille M. Libosvar

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, Massachusetts

CONTENTS

INTRODUCTION

The present paper 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 [81] to the hierarchical production planning of HAX and MEAL [49] or the decentralization by pricing of BAUMOL and FABIAN [9]. Similarly, in MESAROVIC et al. [71] -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 that present paper 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 PAPER

The first part of this paper 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. [71], the concepts of multilayer and multilevel control systems are surveyed in Sections 1.2 and 1.3. Section 1.1 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 2.2 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 2.3, 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 2.4 focuses on the difficult problem of integrating detailed scheduling in a hierarchical system: the coordination scheme described is iterative. Finally, section 2.5 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. [71] 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 behaviour 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 1.2 and 1.3 respectively, whereas Section 1.1 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. [80]. Moreover, [94] (especially chapters 1,4,5 and 8) gathers a representative selection of the work concerning multilevel systems and aggregation, whereas SINGH [84] presents an extensive synthesis of the work in dynamic multilevel systems and FINDEISEN et al. [28] review 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.

1.1 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 1.2 will provide enough evidence that there is no systematic procedure to design models that would satisfy the requirement stated hereabove.

AOKI [4],[5],[94] 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 [5] 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 ([7]).

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 closely approximate those of the original one.

The particular case of a quadratic objective and a feedback control law $u = Kx^*$ 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 $u = Kx^*$ 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 behaviour 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 [4].

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. [80], 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 2.3 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 "diretion" along which aggregation is performed: in [4] and [5], 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.

1.2 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 [19] or SANDELL et al. [80]).

In the case of a singular perturbation (i.e. "a perturbation to the left-hand side of a differential equation" [80]) 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. [17] 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 ε. 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. [71] first introduce the concept of multilayer decisional hierarchies and exemplify it by the early work of ECKMANN and LEFKOWITZ [25]. 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.[28].

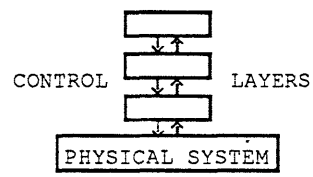
The other multilayer concept (termed multi-horizon in FINDEISEN et al.[28]) is also introduced in MESAROVIC et al.[71] and related to the hierarchical management systems in which the controller is decomposed "into algorithms operating at different time intervals"[28],[29].

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 2.3).

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. [80] point out that the literature in this field is rather scanty and qualitative.

FUNCTIONAL MULTILAYER HIERARCHIES



In [25], 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 behaviour in the vicinity of the operating point" [25].

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 [64]. 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 other variables. This idea of combining multilayer and multilevel concepts was already present in MESAROVIC et al.[71] 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 [29], 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.

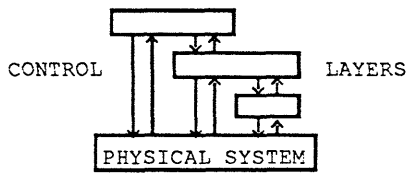
The multilayer concept is introduced in that a boundary set B included in the state space S is defined so that whenever a boundary state is reached, a new strategy can be defined. Thereby the controller is divided into a regulator and a supervisor, where the regulator applies the strategy imposed by 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 contro 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 [64] of finding a trade-off between loss in performance and computation cost to determine the sampling rate is further developed in DONOGHUE and LEFKOWITZ [23], 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. [28] describe a multi-layer control structure that is very similar to those presented in Section 2.3 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 higher layer and still meet the "external" requirements.

This example illustrates the difference between ordinary multi-variable systems and multi-variable systems with significantly

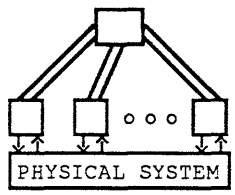
different response times to changes in the control variables. For this latter class of systems (and the water supply system belongs to it since changing the level of one of the main reservoirs will take much more time than regulating the flows), the multi-horizon structure is particularly well suited 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 2.3 in particular suggests that manufacturing sysytems are particularly amenable to a multihorizon control structure. Along this vein, the approach suggested by DONOGHUE and LEFKOWITZ [23] 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 that most influence 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.[28], 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.

1.3 DECOMPOSITION - COORDINATION: MULTILEVEL HIERARCHICAL CONTROL



Unlike the multilayer systems literature, the literature introducing multilevel systems is vast; most of MESAROVIC et al. [71] already addresses the central issue in multilevel systems, namely 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 respects 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 [87] 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 in 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 in 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 π_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:

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

When the optimum is reached, the values assumed by the interaction variables π_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:

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

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

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. [71].

1. The prediction principle: the supremal unit predicts the values of the interactions $\pi(t)$ and supplies the infimals with both $\beta(t)$ and $\pi(t) = \underline{\pi}(\beta)$. The infimals then satisfy their local problems and determine the optimal $x_i^*(t)$ and $u_i^*(t)$. The global solution is reached when $\underline{\pi}(\beta) = g(x^*,u^*)$, 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 [94], 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 $\underline{\pi}(\beta,u^*)$ 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 $\underline{\pi}(\beta,u^*) = g(x^*,u^*)$.

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

It is proved that if ß* is the value of the coordination vector that solves the overall problem and, for a given value ß, each subproblem has a unique solution $u_i(ß)$, $x_i(ß)$, $\pi_i(ß)$ then $J(ß) < J(ß^*)$ or $ß = ß^*$. 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 [60] and PEARSON [94], 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, $J_d(ß)$ 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 [85]. 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 [94]. The same approach is adopted to decompose a time-delay system and an application in traffic control is presented in [85].

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 [84] 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. [27] 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 [84] 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 imprcatical 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 [84] 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 [86] 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 [84] 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 [20] 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 [18] or LOOZE [65] claim a unified approach for decomposition-coordination techniques.

Furthermore, the evaluation of multilevel techniques in SANDELL et al. [80] 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. [85]) cannot be adadpted to management problems. The only exception known to the author is LASDON [94], that is, the application of decomposition to a very specific problem, namely determine the number of machines to setup 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 [20], [23], or [71] 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. [39]). It seems however that management and control have not reached the same degree of maturity.

For instance, already in 1960, LEFKOWITZ [64] 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.

For instance, in the wafer fabrication industry, where random yields, failures and reentrance complicates 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 [16]- 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.

How the managerial decision process is decomposed

HAX [46] reviews several frameworks to classify the logistic decisions, with a particular interest in the hierarchical taxonomy described by ANTHONY. In [3], 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 [95] 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".

Coordination / integration of decisions

These problems encountered in production management cannot be solved indepedently. The need for an integration of the tools developed to solve each of them was thus felt very early. HOLSTEIN [50] argues that inefficiencies that appear at the short term control level can be due to bad longer term decisions and he describes the informations 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 a model 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.

The computational aspect

One of the guidelines suggested in [64] 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 [14]).

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 [61]). Unfortunately, Section 2.1 gives extensive evidence of the gap there can be between mathematical decomposition and managerial decentralization.

2.1 MONOLITHIC MODELS FOR MULTI-LEVEL DECISIONS

The first work entering this category is probably GELDERS and KLEINDÖRFER [31],[32]. 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 2.3. 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 2.3.

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 [47]. However, the basic assumption in HAX and MEAL [49] 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 [68], DZIELINSKI and GOMORY [24], KLEINDÖRFER and NEWSON [56], LASDON and TERJUNG [61] and NEWSON [73],[74].

All these works are based on MANNE's result on "dominant" schedules (independently found by WAGNER and WHITIN [92]), 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 [22] 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 [22], 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 [61] 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 [94] proposes a multilevel decomposition of the problem formulated in DZIELINSKI and GOMORY [24]. 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 [31] [32], 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 [68] 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.

2.2 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. [71]) 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 [9] 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 [79]: 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 in 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 [15] 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 [59] 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 [57] 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 [71]. 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.

2.3 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 [33] and DEMPSTER, FISHER et al. [21]. Both groups of authors consider the work initiated by HAX and MEAL [49] 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.

HAX AND MEAL'S AND DERIVED MODELS

Although designed for a particular implementation, the model described in HAX and MEAL [49] -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 [45], 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 <u>families</u> group items sharing the same major setup.
  - . product <u>types</u> 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 <u>types</u> 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 <u>family</u> in the period considered; since <u>types</u> group several <u>families</u>, 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 [28]) 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 [53] 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 [30] 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 [40] 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 the average.

BITRAN and HAX [11] 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 [10]. 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 [93] investigates 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, is showed experimentally to give good results and require little computation.

HAAS, HAX and WELSCH [43] then 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 [12] prove that 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 consists in 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 showed 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 [26] 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).

Then, the system proposed in BITRAN, HAAS and HAX [12] is considered. In the "one step look-ahead" procedure, 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 [12], 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.

EXTENSIONS OF THE MODEL

Introduction of Feedback

In GRAVES [41], 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 [40]), 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 iterative solving of the two models and updating of the multipliers.

Multi-stage Systems

Other enhancements of HAX and MEAL's model were devised to adapt hierarchical planning to multistage systems. CANDEA [14] reviews the theoretical results existing in production planning of multistage systems and identifies a need for further research. Thus, 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 [13] 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

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

Advantages claimed in [46] are:

- reduction of the computational size, compared to a monolithic optimization with detailed data over the longest 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 suit a reasonably wide range of production systems.

Major shortcomings of the approach are:

- the product aggregation proposed in HAX and MEAL [49] 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 [70], 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.

IMPLEMENTATIONS

A good illustration of this statement is given in MACKULAK, MOODIE and WILLIAMS [66] 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 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). Another tends to minimize the volume of unassigned product. The last limits the amplitude of 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. [69], 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 [8] 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 [2] consider an implementation in an assembly industry. They choose to adapt an MRP system in order that it satisfies the 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. [69] 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 [62] 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 [78] applies HAX's design technique -described in [46] - to the case of a production and distribution firm, and OLSON [76] 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 [33] 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.

SOFTWARE SYSTEMS

HAX and GOLOVIN [48] 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 [49]. However, the user is not limited to the algorithms described in [49], [11] and [12]; 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 [72] and LAWRENCE and MORTON [63] 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 suboptimizating algorithms with simulation capabilities and rules of thumb. Two issues considered of significant importance (e.g. in [69],[8]) 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.

SUBOPTIMALITY OF THE HIERARCHICAL APPROACH

This issue has been addressed in DEMPSTER, FISHER et al. [21]. 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 [6]. 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.

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 [58] 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 [95] 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 [31],[32]. 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 [26], one can find the issue of double aggregation/disaggregation over parts and machines addressed in AXSÄTER [7]. 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 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.

2.4 INTERACTION BETWEEN AGGREGATE- AND DETAILED-SCHEDULING MODELS

GREEN [42] 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 [83] 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 in 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 [6] 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 [51] proposed replacing 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 in the area, it appears that the conclusions drawn at the end of Section 2.3 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.

2.5 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 [52]). On the other hand, new management structures are being progressively developed in order to match these new requirements.

O'GRADY and MENON [75] 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. [89] 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 fits the models described in FINDEISEN et al. [28].

According to the terminology used in management science, the two problems to be solved in the hierarchical frame 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 [91], whereas the routing problem is addressed in VILLA, CANUTO and ROSSETTO [88]. 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 [88], 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 [90] 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 [88] 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 [82] 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 [77]).

In [90], 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 some sort of pattern-matching.

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 [54] 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 [54], 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, 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 is 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 [55] , which summarizes the most innovative results of [54], 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 [64], since the decision tables have to be up-dated if the values of the machines failures parameters come to change.

Both in [54] and [55] 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 [67].

In GERSHWIN, AKELLA and CHOONG [38], 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 behaviour of the system, which considerably simplifies its computation. Additionally, whereas in KIMEMIA [54] and KIMEMIA and GERSHWIN [55] 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 [1] to the performance achieved by using other policies. GERSHWIN [34] 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 [35] with the addition of a control level to determine the setup frequency.

GERSHWIN [36] synthesizes and extends these ideas in a novel hierarchical framework for production scheduling. Production is represented as the occurrence of different events (some controllable, others random) affecting the resources of the system and indicating the beginning or the end of activities. 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 τ_{ij} and frequency u_{ij} on each resource i.

Two assumptions are made :

→ flow conservation: 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, $\sum_j u_j \tau_{ij} < 1$

→ frequency separation: 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 $f_{k-1} << u_j << f_k$; k is then called the level 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:

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

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 [54] and subsequent work. The idea is to define a surplus for each activity and to keep it at a hedging point as much as possible, in order to avoid that uncontrolled activities take the resources and prevent the objectives from being reached.
- → the staircase strategy (basically, the loading policy of [1] and [38]) 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 [37] 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 computationsif 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 [36] can be successfully applied.

CONCLUSION

The objective of this paper 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 [23]), 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 th 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 [85]).

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. [55] 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 paper: one is a "static" answer ([45],[49]), 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 ([89],[90],[36],[37]) 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 ([24],[31],[40]): 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 ([11],[26],[49]). 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 [55], 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.

Further work should therefore be aimed at answering these questions, especially the last two.

REFERENCES

[1] 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.

[2] ANDERSSON, H., S. AXSÄTER and H. JÖNSSON, "Hierarchical Material Requirement Planning.", *Intern. Journ. Prod. Res.*, Vol. 19, No. 1, 1981.

[3] ANTHONY, R.N., "Planning and Control Systems : A Framework for Analysis.", *Harvard University, Graduate School of Business Administration*, Boston, MA., 1965.

[4] AOKI, M., "Control of Large Scale Dynamic Systems by Aggregation.", *IEEE Transactions on Automatic Control*, Vol. AC-13, No. 3, 1968.

[5] AOKI, M., "Some Approximation Methods for Estimation and Control of Large Scale Systems.", *IEEE Transactions on Automatic Control*, Vol. AC-23, No. 2, 1978.

[6] 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.

[7] AXSÄTER, S., "Aggregation of Product Data for Hierarchical Production Planning.", *Operations Research*, Vol. 29, No. 4, 1981.

[8] 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.

[9] BAUMOL, W.J. and T. FABIAN, "Decomposition, Pricing for Decentralization and External Economies.", *Management Science*, Vol. 11, No. 1, 1964.

[10] 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.

[11] BITRAN, G.R. and A.C. HAX, "On the Design of Hierarchical Planning Systems.", *Decision Sciences*, Vol. 8, 1977.

[12] BITRAN, G.R., E.A. HAAS and A.C. HAX, "Hierarchical Production Planning : A Single Stage System.", *Operations Research*, Vol. 29, No. 4, 1981.

[13] 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.

[14] 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.

[15] CHARNES, A., R.W. CLOWER and K.O. KORTANEK, "Effective Control through Coherent Decentralization with Preemptive Goals.", *Econometrica*, Vol. 35, No. 2, 1967.

[16] 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.

[17] 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.

[18] COHEN, G., "Optimization by Decomposition and Coordination : A Unified Approach.", *IEEE Transactions on Automatic Control*, Vol. AC-23, No. 2, 1978.

[19] 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.

[20] DELEBECQUE, F. and J.P. QUADRAT, "Optimal Control of Markov Chains Admitting Strong and Weak Interactions.", *Automatica*, Vol.17, No. 2, 1981.

[21] 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.

[22] DIRICKX, Y.M.I. and L.P. JENNERGREN, "Systems Analysis by Multilevel Methods.", *International Series on Applied Systems Analysis*, Wiley, 1979.

[23] 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.

[24] DZIELINSKI, B. and R. GOMORY, "Optimal Programming of Lot Sizes, Inventories and Labor Allocation.", *Management Science*, Vol. 7, No. 9, 1965.

[25] ECKMAN, D.P. and I. LEFKOWITZ, "Principles of Model Techniques in Optimizing Control.", *Proceedings of the first IFAC Congress*, Butterworths, 1960.

[26] ERSCHLER, J., G. FONTAN and C. MERCE, "Consistency of the Disaggregation Process in Hierarchical Planning.", *Operations Research*, Vol. 34, No. 3, 1986.

[27] 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.

[28] FINDEISEN, W., F.N. BAILEY, M. BRDYS, K. MALINOWSKI, P. TATIEWSKI and A. WOZNIAK, *Control and Coordination of Hierarchical Systems*, Wiley, 1979.

[29] FORESTIER, J.P. and P. VARAYIA, "Multilayer Control of Large Markov Chains", *IEEE Transactions on Automatic Control*, Vol. AC-23, No. 2, 1978.

[30] GABBAY, H., "A Hierarchical Approach to Production Planning.", Technical report No. 120, *Operations Research Center*, M.I.T., Cambridge, MA., 1975.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] GOLOVIN, J.J., "Hierarchical Integration of Planning and Control.", Technical report No. 116, *Operations Research Center*, M.I.T., Cambridge, MA., 1975.

[41] GRAVES, S.V., "Using Lagragian Techniques to Solve Hierarchical Production Planning Problems.", *Management Science*, Vol.28, No.3, 1982.

[42] GREEN, R.S., "Heuristic Coupling of Aggregate and Detailed Models in Factory Scheduling.", unpublished P.H.D. thesis, M.I.T., Cambridge, MA., 1971.

[43] 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.

[44] HACKMAN, S.T. and R.C. LEACHMAN, "An Aggregate Model of Project Oriented Production.", *Operations Research Center*, U.C. Berkeley, February 1987.

[45] 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.

[46] 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.

[47] HAX, A.C., "Aggregate Production Planning.", in *Handbook of Operations Research*, J. Moder and S. Elmaghraby eds., Van Nostrand Reinhold, New York, 1978.

[48] HAX, A.C. and J.J. GOLOVIN, "A Computer Based Operations Management System (COMS*).", *Studies in Operations Management*

[49] 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.

[50] HOLSTEIN, W.K., "Production Planning and Control Integrated.", *Harvard Business Review*, Vol. 46, No. 3, 1968.

[51] IMBERT, S., "Interaction entre deux Niveaux de Décision en Planification de la Production.", thèse de 3ème cycle, Université Paul Sabatier, Toulouse, 1986.

[52] JAIKUMAR, R., "Postindustrial Manufacturing.", *Harvard Business Review*, Nov.-Dec. 1986.

[53] KARMARKAR, U.S., "Equalization of Run-out Times", *Operations Research*, Vol. 29, No. 4, 1981.

[54] 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.

[55] 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.

[56] KLEINDÖRFER, P. and E.F.P. NEWSON, "A Lower Bounding Structure for Lot Size Scheduling Problems.", *Operations Research*, Vol. 23, No. 2, 1975.

[57] KORNAI, J, and T. LIPTAK, "Two Level Planning.", *Econometrica*, Vol. 33, No. 1, 1965.

[58] KRAJEWSKI, L.J. and L.P. RITZMAN, "Disaggregation in Manufacturing and Service Organizations : Survey of Problems and Research.", *Decision Sciences*, Vol. 8, 1977.

[59] KYDLAND, F., "Hierarchical Decomposition of Linear Economic Models.", *Management Science*, Vol. 21, No. 9, 1975.

[60] LASDON, L.S., "Duality and Decomposition in Mathematical Programming", *IEEE Transactions on Systems Science and Cybernetics*, Vol. SSC-4, No. 2, 1968.

[61] LASDON, L.S. and R.C. TERJUNG, "An Efficient Algorithm for Multi-item Scheduling.", *Operations Research*, Vol. 19, 1971.

[62] 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.

[63] LAWRENCE, S.R. and T.E. MORTON, "Patriarch: Hierarchical Production Scheduling.", *N.B.S. Special Publication 724*, R. Jackson and A. Jones eds., 1986.

[64] LEFKOWITZ, I., "Multilevel Approach Applied to Control System Design.", *Trans. ASME*, Vol. 88, 1966.

[65] LOOZE, D.P., "Hierarchical Control and Decomposition of Decentralized Linear Stochastic Systems.", unpublished P.H.D. thesis, M.I.T., Cambridge, MA., 1978.

[66] 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.

[67] 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.

[68] MANNE, A.S., "Programming of Economic Lot Sizes.", *Management Science*, Vol. 4, No. 2, 1958.

[69] 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.

[70] MEAL, H.C., "Putting Decisions where they Belong.", *Harvard Business Review*, Mar.-Apr. 1984.

[71] MESAROVIC, M.D., D. MACKO and Y. TAKAHARA, *Theory of Multilevel Hierarchical Systems*, New York : Academic, 1970.

[72] 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.

[73] NEWSON, E.F.P., "Multi-Item Lot Size Scheduling by Heuristic Part I : with Fixed Resources.", *Management Science*, Vol. 21, No. 10, 1975.

[74] NEWSON, E.F.P., "Multi-Item Lot Size Scheduling by Heuristic Part II : with Variable Resources.", *Management Science*, Vol. 21, No. 10, 1975.

[75] 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.

[76] OLSON, C.D., "A Prototype System for Hierarchical Production Planning.", unpublished M.S. thesis, M.I.T., Cambridge, MA., 1983.

[77] PAPAS, P.N., "ISIS Project in Review.", *N.B.S. Special Publication 724*, R. Jackson and A. Jones eds., 1986.

[78] PENDROCK, M.E., "A Hierarchical Approach to Integrated Production and Distribution Planning.", unpublished M.S. thesis, M.I.T., Cambridge, MA., 1978.

[79] RUEFLI, T.W., "A Generalized Goal Decomposition Model.", *Management Science*, Vol. 17, No. 8, 1971.

[80] 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.

[81] SAATY, T.L., *The Analytic Hierarchical Process : Planning, Priority Setting, Manpower Allocation*, Mc Graw-Hill 1980.

[82] 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.

[83] 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.

[84] SINGH, M.G. *Dynamical Hierarchical Control*, Elsevier, rev. 1982.

[85] 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.

[86] SINGH, M.G. and A. TITLI, "Closed Loop Hierarchical Control for Non-linear Systems Using Quasilinearisation.", *Automatica*, Vol. 11, 1975.

[87] SMITH, N. and A.P. SAGE, "An Introduction to Hierarchical Systems Theory.", *Computers and Electrical Engeneering*, Vol. 1, 1973.

[88] 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.

[89] 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.

[90] 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.

[91] 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.

[92] WAGNER, H.M. and T.M. WHITIN, "Dynamic Version of the Economic Lot Size Problem.", *Management Science*, Vol. 5, 1958.

[93] WINTERS, P.R., "Constrained Inventory Rules for Production Smoothing.", *Management Science*, Vol. 8, No. 4, 1962.

[94] WISMER, D.A. (ed.) *Optimization Methods for Large Scale Systems… with applications*, Mc Graw Hill, 1971.

[95] ZOLLER, K., "Optimal Disaggregation of Aggregate Production Plans.", *Management Science*, Vol. 17, 1971.