Hierarchical production planning for complex manufacturing systems†

A. Mehra†, I. Minis
Department of Mechanical Engineering and Institute for Systems Research, University of Maryland, College Park, MD 20742, USA

&

J. M. Proth
INRIA-Lorraine, 4 Rue Marconi, Metz 2000, 57070 Metz, France

(Received for publication 22 November 1995)

A hierarchical approach to production planning for complex manufacturing systems is presented. A single facility comprising a number of work-centers that produce multiple part types is considered. The planning horizon includes a sequence of time periods, and the demand for all part types is assumed known. The production planning problem consists of minimizing the holding costs for all part types, as well as the work-in-process and the backlogging costs for the end items. We present a two-level hierarchy that is based on aggregating parts to part families, work-centers to manufacturing cells and time periods to aggregate time periods. The solution at the aggregate level is imposed as a constraint to the detailed level problems which are formulated for each manufacturing cell separately. This architecture uses a rolling horizon strategy to perform the production management function. We have employed perturbation analysis techniques to adjust certain parameters of the optimization problems at the detailed level to reach a near-optimal detailed production plan. Numerical results for several realistic example problems are presented and the solutions obtained from the hierarchical and monolithic approaches are compared. The results indicate that the hierarchical approach offers major advantages in computational efficiency, while the loss of optimality is acceptable. Copyright © 1996 Elsevier Science Limited

Key words: manufacturing systems, production planning, aggregation.

1 INTRODUCTION

Production planning consists of determining the quantities of products to manufacture in a sequence of time periods in order to optimize a certain criterion while satisfying constraints, such as those imposed by the capacity of resources. This problem is complicated by endogenous (e.g. resource failures) as well as exogenous (e.g. delayed receipts of raw material and unexpected changes in demand) random events. The resulting optimization problem is extremely large and complex. Maxwell et al.¹ stress the need to develop and implement a generic production planning system by stating: 'Billions of dollars are wasted in US each year by manufacturers of discrete parts because of inadequate procedures for controlling inventory and production.'

Two distinct approaches to production planning have been adopted in the literature. The first is a monolithic approach, wherein the entire problem is formulated as a large mixed-integer linear programming-type problem. The second is a hierarchical approach which partitions the global problem into a series of sub-problems that correspond to different hierarchical levels of the manufacturing system and are solved sequentially, such that the solution at each level imposes constraints on the solution of the subsequent lower level. The major advantages of the hierarchical approach are: (1) reduction of complexity and (2) gradual absorption of random events. For other advantages see Dempster et al.²

†This paper was first presented at the ASME Computers in Engineering Conference, 1994.
‡Current address: Gensym Corporation, 125 Cambridge Park Drive, Cambridge, MA 02140, USA.
Hax and Meal\textsuperscript{3} conducted pioneering work in the area of hierarchical production planning (HPP). They considered a multi-plant firm with four decision levels. The highest level distributes the products to be manufactured to individual plants, and is solved once. The next three decision levels address the management of a single plant. The concept of a rolling horizon is employed to perform repetitive open-loop optimizations. Bitran and Hax\textsuperscript{4,5} further refined Hax and Meal's methodology for a single-stage production system, and Bitran et al.\textsuperscript{6} extended this methodology to a two-stage process. These models considered three levels of aggregation for products: (1) items, (2) families and (3) types. The problem is formulated as a convex knapsack problem which minimizes the total set-up costs. Graves\textsuperscript{7} introduced feedback between decision levels. The monolithic problem is formulated as a mixed integer problem, which is decomposed by Lagrangian relaxation. The Hax and Meal model, and its extensions, are based on a special cost structure and consider different aggregation schemes, such as similarity of set-up costs and of demands. Furthermore, no spatial and time decompositions of the production system are proposed.

Axsater,\textsuperscript{8} Hillion et al.,\textsuperscript{9} Meier,\textsuperscript{10} Nagi,\textsuperscript{11} Thompson and Davis\textsuperscript{12} and Thompson et al.\textsuperscript{12} address a double aggregation scheme over products and machines, where at the top layer the entities of relevance are machine cells or departments and part families. This important class of literature does not address temporal aggregation issues in their formulation. We believe that temporal aggregation (i.e. aggregation of planning time periods) is an important aspect of hierarchical production planning, since it facilitates the use of aggregate forecasted demand information, thus reducing forecasting errors stemming from the unavailability of detailed demand information. Furthermore, temporal aggregation allows both aggregate forecast information and 'firm' customer orders to be planned using different time scales in the planning problem.

Designing an HPP system which provides good solutions for a given manufacturing system remains an open problem. To address it we propose a two-level hierarchy which is based on aggregating parts to part families, work-centers to manufacturing cells and time periods to aggregate time periods. The optimization problems at the aggregate and detailed level of the proposed hierarchy have been formulated. The solution at the aggregate level is imposed as a constraint to the detailed level problem, which employs a decomposition based on manufacturing cells. This architecture uses a rolling horizon strategy to perform the production planning. We adjust some parameters at the detailed level to obtain the best possible performance of the HPP system. These parameters are typically related to local decisions at each level, which cannot be derived from the solution of the previous upper level. We propose an iterative approach that is closely linked to perturbation analysis (PA) to determine near optimal values of these parameters. We determine cost gradient estimates with respect to the parameters of interest and use them to obtain near-optimal values of the parameters.

We have presented several numerical examples to demonstrate the effectiveness of the proposed hierarchical approach. To the best of our knowledge, it is the first time that simultaneous aggregation of parts, work-centers and time periods is considered, and PA is employed in HPP. The numerical results indicate that the hierarchical approach offers major advantages in computational efficiency, while the loss of optimality is acceptable.

### 2 Hierarchical and Monolithic Production Planning Models

The two levels in the proposed hierarchical structure are (1) the aggregate (or high) level that plans production for part families on cells within sub-periods and (2) the detailed (or low) level that disaggregates the high-level solution to determine the production plan for part types on work-centers during elementary periods. At the high level the criterion is to minimize inter-cell work-in-process (WIP) costs as well as inventory and backlogging costs for finished part families. At the low level the criterion is to minimize WIP costs as well as inventory and backlogging costs of finished part types. Thus, we introduce a hierarchy among inventories: priority is given to inter-cell and end-product inventories. The disaggregation is performed over a short-term horizon. Production planning is usually performed on a rolling horizon basis. That is, although the high-level solution is computed over $H$, only a part of it (over the short-term horizon) is implemented. After the implementation of the plan in this short-term horizon, the plan is recomputed based on the actual state of the system. This is done in order to incorporate the future demands/forecasts progressively.

#### 2.1 Monolithic production planning problem

We consider a single manufacturing facility consisting of a number of manufacturing resources or work-centers. A number of part types are manufactured in the facility. Each part type is produced following a certain sequence of operations (production routing), and each operation is performed on a particular work-center. The processing time of each operation is given; set-up times are ignored in this treatment.

We limit ourselves to a two-level hierarchy in order to present the proposed approach and to evaluate its effectiveness. In this paper we assume that the temporal
aggregation has been defined a priori. Furthermore, the aggregation of parts to families and machines to cells satisfy the following conditions: (1) parts belonging to the same family follow a common sequence of cells during their manufacture. Note that it is not necessary for parts to follow the same machine sequence; (2) Parts belonging to the same family have similar processing times within each cell; this requirement will result in a good estimation of aggregate processing times. Assumptions (1) and (2) do not over-restrict the model and are in agreement with industrial practice.

We consider a planning horizon $H$ consisting of $Z$ sub-periods (aggregate time periods). Each sub-period is divided into $z$ elementary periods (detailed time periods) of duration $T$ each. Parts are due and new orders arrive at the beginning of these elementary periods. $T$ is supposed to be much larger than the maximal time needed to perform one operation. Furthermore, $zT$, which is the duration of a sub-period, is much larger than the time needed to complete the set of operations performed on any type of part in any cell. We also assume that: (1) at most, one operation related to a certain part is performed during each elementary period; (2) a part visits at most one cell during each sub-period.

The production planning problem consists of determining the number of operations of each part type to be performed on the system’s work-centers during each elementary period of the planning horizon in order to minimize the total holding and backlogging costs. The notation employed in this problem is as follows: $P = \{P_1, P_2, \ldots, P_N\}$ is the set of $N$ part types; $M = \{m_1, m_2, \ldots, m_M\}$ is the set of $M$ work-centers; $R_j$ is the manufacturing routing, i.e. the sequence of operations that part type $P_j$ must undergo. The total number of operations required for part type $P_j$ is denoted by $n_j$, $I_{j,w}$ is the holding cost of one unit of part type $P_j$ after its $w$th operation is completed for one elementary period; $B_j$ is the backlogging cost of one unit of finished part type $P_j$ for one elementary period; $t_{j,w}$ is the processing time of the $w$th operation of part type $P_j$; $d_{j,w}^k$ is the indicator function. It assumes a value of 1 if machine $m_k$ is required for the $w$th operation of part type $P_k$, and 0 otherwise; $d^k$ is the demand of part type $P_k$ at the beginning of the $k$th elementary period, which must be satisfied by the end of that period; $s^k_{j,0}$ is the raw-material inventory for part type $P_j$ at the end of the $k$th elementary period; $s^k_{j,0}$ is the initial inventory of part type $P_j$ at the end of the $k$th elementary operation; $s^k_{j,w}$ is the inventory of part type $P_j$ at the end of the $k$th elementary period and at the end of the $w$th operation (at the beginning of the first elementary period); $s^k_{j,w}$ is the inventory of part type $P_j$ at the end of the $k$th elementary period and at the end of the $w$th operation, negative values indicate backlog; $u^k_{j,w}$ is the production volume of part type $P_j$ related to the $w$th operation during the $k$th elementary period.

The monolithic production planning problem can be formally modeled by the following linear program:

**Problem P₁:**

minimize \[
\sum_{k=1}^{Z} \sum_{j=1}^{N} \left\{ \sum_{w=1}^{n_j-1} I_{j,w} s^k_{j,w} \right\} + I_{j,n_j} [s^k_{j,n_j}]^+ + B_j [-s^k_{j,n_j}]^+
\]

subject to:

\[
s^k_{j,w} = s^k_{j,w-1} + I_{j,w} - u^k_{j,w+1} \quad (2)
\]

\[
s^k_{j,n_j} = s^k_{j,n_j-1} + I_{j,n_j} - d^k_j \quad (3)
\]

\[
\sum_{j=1}^{N} \sum_{w=1}^{n_j} d_{j,w}^k u^k_{j,w} t_{j,w} \leq T \quad (4)
\]

\[
u^k_{j,w} \leq s^k_{j,w-1} \quad (5)
\]

\[
u^k_{j,w} \geq 0 \quad (6)
\]

for $j = 1, 2, \ldots, N, k = 1, 2, \ldots, Z$ in constraints (2)–(6);

for $w = 1, 2, \ldots, n_j$ in constraints (5) and (6);

for $w = 1, 2, \ldots, n_j - 1$ in constraint (2); for $i = 1, 2, \ldots, M$ in constraint (4).

The symbol $[\cdot]^+$ implies $\max\{0, \cdot\}$. The objective function (1) consists of both (1) the inventory holding costs of WIP (for $w = 1, 2, \ldots, n_j - 1$) and of finished parts (for $w = n_j$) and (2) the backlogging costs of finished parts over the entire planning horizon. Constraints (2) and (3) represent the state equations of the inventory levels. Note that the inventory level at any operation of a part is only updated at the end of the corresponding elementary period. Also, no more than one operation can take place on a specific part in a single elementary period. The capacity constraints on the work-centers for all elementary periods of the horizon are represented by constraint (4). Constraint (5) denotes that the total number of instances of an operation on a certain part within an elementary period cannot exceed the number of parts contained in the upstream buffer at the end of the previous elementary period. Finally, constraint (6) represents the non-negativity of production.

Note that although the objective function in eqn (1) is nonlinear, it can be transformed into a linear one by adding constraints as shown below.

minimize \[
\sum_{k=1}^{Z} \sum_{j=1}^{N} \left\{ \sum_{w=1}^{n_j-1} I_{j,w} s^k_{j,w} \right\} + P^k_j
\]

subject to:

\[
P^k_j \geq I_{j,n_j} s^k_{j,n_j} \quad (7)
\]

\[
P^k_j \geq -B_j s^k_{j,n_j} \quad (8)
\]

\[
P^k_j \geq 0 \quad (9)
\]

for $j = 1, 2, \ldots, N$, and $k = 1, 2, \ldots, Z$ in constraints (7)–(9); where $P^k_j$ is the inventory/backlogging cost related to the $k$th elementary period.
Although this LP problem is simple to formulate, there are several reasons why it cannot be solved easily or implemented in practice: (i) the LP is of a very large dimension for a typical manufacturing system and planning horizon; (ii) detailed information about the demand of part types is not known for the entire horizon; (iii) the demand is subject to change due to order cancellations and acceptance of new orders, and thus the monolithic formulation requires frequent recomputation; (iv) the monolithic formulation does not allow random events to be absorbed with a computational effort that is proportional to the impact of those events; (v) the monolithic formulation does not allow different criteria to be used at different levels of the system managerial hierarchy.

These issues suggest a hierarchical approach to the production planning problem. In addition, the hierarchical structure is parallel to a corporate management hierarchy and thus provides significant assistance to the overall management function. The following paragraphs outline the hierarchical approach.

2.2 Hierarchical production planning: aggregate level problem

The problem at the aggregate level considers the production planning of part families on manufacturing cells during sub-periods of the planning horizon \( (H) \). Notation related to the aggregate level is as follows: \( \mathcal{F} = \{ p_1, p_2, \ldots, p_N \} \) is the set of \( N \) part families, \( N \leq N; \mathcal{C} = \{ c_1, c_2, \ldots, c_M \} \) is the set of \( M \) manufacturing cells, \( M \leq M; T_z \) is the length of a sub-period; \( \mathcal{A}_f \) is the macro-manufacturing process, i.e. the sequence of macro-operations (performed in cells) a part family \( p_f \) must undergo, \( \bar{n}_f \) represents the number of macro-operations for part family \( p_f \); \( s_{f,q}^w \) is the inventory cost of one part unit in family \( p_f \) after the \( q \)th macro-operation is completed; \( \bar{B}_f \) is the backlogging cost of one part unit in family \( p_f \) at the final production stage; \( \bar{u}^w_{f,q} \) is the production volume of parts in family \( p_f \) related to the \( q \)th macro-operation during the \( k \)th sub-period; \( \tau_{f,q}^w \) is the processing time related to the \( q \)th macro-operation of one unit of a part type in family \( p_f \) during the \( k \)th sub-period; \( D^*_{f,k} \) is the demand of parts in family \( p_f \) at the beginning of the \( k \)th sub-period and is given from

\[
D^*_{f,k} = \sum_{k=(n-1)z+1}^{\infty} \sum_{p_f \in \mathcal{P}_f} d^k_{p_f}
\]

The macro-manufacturing processes (macro-routings) are developed by aggregating work-centers into cells according to the aggregation rules presented in Section 2.1. In order to obtain this aggregation, we consider the routing \( \mathcal{A}_f \) of each part \( p_f \in \mathcal{P}_f \), and define a set of \( \bar{n}_f \) subsets (sub-routings) \( \{ \mathcal{A}_f^q; q = 1, \ldots, \bar{n}_f \} \), where \( q \) is the rank of the macro-operation in the macro-manufacturing process of family \( p_f \). Note that several values of \( q \) may correspond to the same cell. Each sub-routing \( \mathcal{A}_f^q \) corresponds to one of the cells visited by part \( p_f \), and, alternatively, each sub-routing corresponds to one macro-operation of its family \( p_f \). That is:

\[
\mathcal{A}_f^q = \bigcup_{q=1}^{\bar{n}_f} \mathcal{A}_f^q \text{ and } \mathcal{A}_f^q \bigcap \mathcal{A}_f^b = \emptyset \text{, for } a \neq b
\]

Note that \( \mathcal{A}_f^q \) corresponds to the \( q \)th macro-operation of family \( p_f \), if \( p_f \in \mathcal{P}_f \). Let \( \mathcal{F}(j,q) \) be the last operation of the sub-routing \( \mathcal{A}_f^q \) and \( \mathcal{F}(j,q) \) be the set of operations of \( \mathcal{A}_f^q \), except the last one.

The computation of inventory/backlogging costs and processing times for families can be obtained from eqns (11)–(13). In fact, these parameters depend on the production volume of part type \( u^w_{f,q} \). However, since the production volumes are known only after solving the detailed level problem, we use the weighted average of the costs and processing times with respect to the parts’ demand, i.e.

\[
\tau_{f,q}^w = \left( D^*_{f,k} \| \mathcal{A}_f^q \| \right)^{-1} \sum_{k=(n-1)z+1}^{\infty} \sum_{p_f \in \mathcal{P}_f} \sum_{w \in \mathcal{F}(j,q)} \left( d^k_{p_f} \sum_{i \in \mathcal{A}_f^q} I_{f,i(w)} \right)
\]

(11)

\[
\bar{B}_f = \left( \sum_{k=1}^{\infty} \sum_{p_f \in \mathcal{P}_f} d^k_{p_f} \right)^{-1} \sum_{k=1}^{\infty} \sum_{p_f \in \mathcal{P}_f} (d^k_{p_f} B_f)
\]

(12)

\[
\bar{I}_{f,q} = \left( \sum_{k=1}^{\infty} \sum_{p_f \in \mathcal{P}_f} d^k_{p_f} \right)^{-1} \sum_{k=1}^{\infty} \sum_{p_f \in \mathcal{P}_f} (d^k_{p_f} I_{f,i(w)})
\]

(13)

for \( w \in \mathcal{F}(j,q) \) in eqn (13).

The symbol \( \| \cdot \| \) represents cardinality. The aggregate level problem can be formally stated as a linear program similar to Problem \( P_1 \). The problem consists of determining the production \( u^w_{f,q} \) over the entire planning horizon for all families at each macro-operation. The production parameters at this level (inventory/backlogging cost, processing time) correspond to macro-operations of families over sub-periods. The criteria at the aggregate level is minimization of: (1) the inventory holding costs of WIP between cells and finished families, (2) the backlogging costs of finished families over the entire planning horizon. The inventory is estimated at the end of each sub-period (instead of elementary period in \( P_1 \)) for all part families \( f \in \mathcal{F} \) at the end of each macro-operation. The capacity constraints of the aggregate level problem are applied to cells for all sub-periods of the planning horizon.

The aggregate level problem is presented below.

**Problem \( P_2 \)**

\[
\text{minimize } \sum_{w} \sum_{f} \left\{ \sum_{q=1}^{\bar{n}_f} (\bar{I}_{f,q} \bar{s}_{f,q}^w) + \bar{B}_f [\bar{s}_{f,q}^w - s_{f,q}^w] \right\}^{+}
\]

(14)
Hierarchical production planning for complex manufacturing systems

subject to:

\[ \hat{s}^*_{f,q} = \hat{s}^*_{f,q-1} + \bar{u}^*_{f,q} - \bar{u}^*_{f,q+1}; \quad q = 1, 2, \ldots, n_f - 1 \]  
(15)

\[ \hat{s}^*_{f,q} = \hat{s}^*_{f,q-1} + \bar{u}^*_{f,q} - D_f \]  
(16)

\[ \sum_{j \in \mathcal{F}} \sum_{q \in \mathcal{Q}} \bar{u}^*_{f,j,q} T_{j,q} \leq T \quad \forall j \in \mathcal{J} \]  
(17)

\[ \bar{u}^*_{f,q} \leq \hat{s}^*_{f,q-1}, \quad q = 1, 2, \ldots, n_f \]  
(18)

\[ \bar{u}^*_{f,q} \geq 0, \quad q = 1, 2, \ldots, n_f \]  
(19)

for \( f = 1, 2, \ldots, N \) in constraints (15), (16), (18), (19), for \( k = 1, 2, \ldots, \kappa \) in constraints (15)–(19).

The aggregate objective function (14) minimizes the aggregate inventory and backlogging costs. The aggregate inventory state equations are represented by eqns (15) and (16). Constraint (17) is the resource capacity constraint for each manufacturing cell. Constraint (18) represents that the aggregate production cannot be more than the inventory of the input buffers, and constraint (19) ensures the non-negativity of aggregate production.

2.3 Hierarchical production planning: detailed level problem

The detailed level problem consists of determining the production plan for part types on work-centers during elementary periods of the first sub-period. Recall that in our rolling horizon approach we solve only the first sub-period to determine the production volume \( u^*_{f,w} \) (for the \( w \)-th operation of part type \( p_j \), during the \( k \)-th elementary period; \( k = 1, 2, \ldots, \kappa \)), respecting the aggregate high level solution \( \hat{u}^*_{f,q} \). The criterion is to minimize the holding cost of all part types (end items and WIP), and backlogging costs for end items over this first sub-period. In addition, the low-level problem adopts a decomposition-based approach in which the overall planning problem is solved by considering the manufacturing cells independently. This is also referred to as spatial decomposition and is consistent with CIM industrial practice. In order to make the problem definition concise we have defined some sets presented below.

\( \mathcal{J}(j,v) \) represents the set of operations for part \( p_j \) that are performed in cell \( c_v \). Let \( J^*(j,v) \) be an indicator function which takes the value of 1 if the last operation in the routing of part \( p_j \) is performed in cell \( c_v \) and 0 otherwise.

Since the demand for each part type is known only for the cell that is in charge of the final stage of the part’s manufacturing process, we define the parameters \( \alpha^*_f(j,q) \) as the ratios of the expected production \( u^*_{f,w} \) of part type \( p_j \) to the expected production of parts in family \( \beta_f \) during the \( \kappa \)-th sub-period, where

\[ k = (\kappa - 1)z + 1, \ldots, \kappa z \]

\[ \alpha^*_f(j,q) = \frac{\bar{u}^*_{f,w}}{u^*_{f,q}}, \quad w \in \mathcal{J}(j,q) \]  
(20)

Note that:

\[ \sum_{k=(\kappa-1)z+1}^{\kappa z} \sum_{p_j \in \beta_f} \alpha^*_f(j,q) = 1 \]  
(21)

For each cell \( c_v \), where \( c_v \in \mathcal{C} \), the detailed level problem can be formally stated by the following linear program:

**Problem \( P_3(v) \):**

minimize \( \sum_{k=1}^{\kappa} \sum_{j=1}^{N} \left( \left\{ \sum_{w \in \mathcal{J}(j,v)} I_{j,w} \left[ a^k_{j,w} \right]^+ \right\} + B_j J^*(j,v) \left[ -s^k_{j,n_j} \right]^+ \right) \)  
(22)

subject to:

\[ s^k_{j,w} = s^k_{j,w-1} + u^k_{j,w} - d^k_{j,w} \]  
(23)

for \( w \in \mathcal{J}(j,v) \) such that the \((w+1)\)-th operation is not performed in \( c_v \),

\[ s^k_{j,w} = s^k_{j,w-1} + u^k_{j,w} - d^k_{j,w} \]  
(24)

for \( \text{the last operation of } \mathcal{J}_j \) is performed in \( c_v \),

\[ s^k_{j,w} = s^k_{j,w-1} + u^k_{j,w} - \alpha^*_f(j,q) \bar{u}^*_{f,q+1} \]  
(25)

for \( w \neq n_j \) and \( w \in \mathcal{J}(j,q) \); \( q \) being the index of the macro-operation for part \( p_j \) performed in \( c_v \),

\[ \sum_{j=1}^{N} \sum_{w \in \mathcal{J}(j,v)} \left\{ s^k_{j,w} u^k_{j,w} T_{j,w} \right\} \leq T, \quad m_j \in c_v \]  
(26)

\[ u^k_{j,w} \leq s^k_{j,w-1}, \quad w \in \mathcal{J}(j,v) \]  
(27)

\[ \sum_{k=1}^{\kappa} \sum_{p_j \in \beta_f} u^k_{j,w} = \bar{u}^*_{f,q}, \quad \beta_f \in \mathcal{F} \]  
(28)

if the last macro-operation of \( \mathcal{J}_f \) is performed in \( c_v \),

\[ \sum_{k=1}^{\kappa} u^k_{j,w} = u^k_{j,q} \sum_{k=1}^{\kappa} \alpha^*_f(j,q) \]  
(29)

for all macro-operations of \( \mathcal{J}_f \) performed in \( c_v \), \( w \in \mathcal{J}(j,q) \),

\[ u^k_{j,w} \geq 0, \quad w \in \mathcal{J}(j,v) \]  
(30)

for \( j = 1, 2, \ldots, N; \quad k = 1, 2, \ldots, \kappa \) and \( p_j \in \beta_f \) in constraints (23)–(25), (27), (29), (30), for \( k = 1, 2, \ldots, \kappa \) in constraints (26), (28).

The objective function (22) consists of: (1) the inventory holding costs of WIP and (2) the holding and backlogging costs of finished parts, over the first sub-period. Constraints (23)–(25) represent the state
equations of the inventory levels. The capacity constraints on the work centers $m_i$ belonging to cell $c_j$ for all elementary periods of the first sub-period are represented by constraint (26). Constraint (27) denotes that the total instances of a certain operation in an elementary period cannot exceed the number of parts present in the upstream buffer at the end of the previous elementary period. Constraints (28) and (29) are the production volume constraints imposed by the aggregate level. Constraint (28) represents the cumulative production of family $f$ related to the $n_j$th operation for the entire first sub-period. The value $m_{ijq}$ in this constraint is imposed by the solution of the aggregate level problem. Constraint (28) is applied only if macro-operation $m_i$ is performed in cell $c_j$. Constraint (29) represents the production of part $p_j$ related to the last operation in sub-routing $s_j$ for the entire first sub-period. This value is determined from the solution of the aggregate level problem and the parameter $\alpha_{ijq}$ for that part type summed over all elementary periods of the first sub-period. Constraint (30) represents the non-negativity of production.

In this formulation, we observe that the demand is known only for the cell that is in charge of the final stage of the manufacturing process. Thus, we use the parameters $\alpha_{ijq}$ (see constraint 25) to determine the demand requirement of parts during the intermediate macro-operations. A perturbation analysis-based algorithm is used to determine the near-optimal value of these parameters. This algorithm is presented below.

3 DEMAND DISAGGREGATION USING PERTURBATION ANALYSIS

As discussed in the previous section, the optimization Problems $P_3(w)$ at the detailed level include the control parameters $\alpha_{ijq}$, the value of which are not known a priori. We seek to obtain the near-optimal value of these parameters by adjusting them iteratively. At each iteration, we update $\alpha_{ijq}$ by moving along the gradient of the cost function.

The gradient is estimated by using PA. Consider the cost function (22), say $Y$, which is a function of the vector of parameters $\alpha$, i.e. $Y = Y(\alpha)$, where $\alpha = \{\alpha_{ijq}\}$, $q = 1, 2, \ldots, n_j$, $p_j \in s_j \in S$, $k = 1, 2, \ldots, z_2$. While using PA, the optimization is done by recursive reassignment of the vector $\alpha$ of control parameters to the series of vectors $\{\alpha_{m}\}_{m=0}^{\infty}$ as follows:

$$\alpha_{m+1} = \alpha_m + A_m \nabla_\alpha Y(\alpha_{m})$$

Here $\nabla_\alpha Y(\alpha_{m})$ is the estimate of the gradient and $A_m = \text{diag}(a_{m}^1, a_{m}^2, \ldots, a_{m}^k)$. Blum has shown that if we define the sequences $\{a_{m}^i\}_{m=0}^{\infty}$, $i = 1, \ldots, k$, such that $\lim_{m \to \infty} a_{m}^i = 0$, $\forall i$; $\sum_{m=1}^{\infty} a_{m}^i = \infty$, $\forall i$; $\sum_{m=1}^{\infty} (a_{m}^i)^2 < \infty$, $\forall i$ and $\nabla_\alpha Y(\alpha)$ is an unbiased estimate of $\nabla_\alpha Y(\alpha)$, then the sequence $\{a_{m}^i\}_{m=0}^{\infty}$ converges asymptotically to at least a local optimum of $Y$. For our simulation we chose $a_{m}^i = \frac{a_{m}^i}{m}$, $\forall i$ which satisfies all the conditions discussed above.

The gradient estimation proposed in this paper is carried out in two steps. We first perturb the parameters $\beta_{ijq}$

$$\beta_{ijq}^n = \sum_{k=(n-1)z+1}^{k=zn} a_{ijq}^k$$

of each sub-period and estimate the value of the gradient $\nabla_\beta Y(\beta^m)$, such that the inventories of parts at the end of a macro-operation remain invariant at the end of all successive sub-periods. Secondly, we perturb parameters $\alpha$ and estimate the value of the gradient $\nabla_\alpha Y(\alpha^m)$, such that the part inventories at the end of a macro-operation remain invariant at the end of that sub-period, and all successive sub-periods. The small perturbations are represented by $\Delta \beta_{ijq}^n$ and $\Delta \alpha_{ijq}$ for the first and second step, respectively. Below, we present the procedure for the gradient estimation of the second step, since the first step is performed in a similar manner.

3.1 Gradient estimation with respect to parameters $\alpha$

Consider the perturbations $\Delta \alpha_{ijq+1}^k$ that satisfy eqn (31):

$$\sum_{k=(n-1)z+1}^{k=zn} \Delta \alpha_{ijq+1}^k = 0, \quad \forall j, k, q$$

Note that $\Delta \alpha_{ijq+1}^k$ may have both positive and negative values. Let $\Delta s_{ijw}^k = [w \in s_j(\alpha_{ijq+1}^k)]$ be the change in inventory $s_{ijw}^k$ due to the perturbations $\Delta \alpha_{ijq+1}^k$. It will be seen below that due to eqn (31), the effect of $\Delta \alpha_{ijq+1}^k$ on the inventories of part types related to operations $w \in s_j^q$ which are performed during the $k$th sub-period. The change in inventory $s_{ijw}^k$ due to the perturbations $\Delta \alpha_{ijq+1}^k$ can be obtained from eqn (25) as follows:

$$\Delta s_{ijw}^k = -\Delta \alpha_{ijq+1}^k \bar{u}_{ijw}^k \quad \Delta s_{ijw}^{k+1} = -\Delta \alpha_{ijq+1}^k \bar{u}_{ijw}^{k+1}$$

Furthermore, from eqn (31)

$$\Delta s_{ijw}^{k+1} = -\Delta \alpha_{ijq+1}^k \bar{u}_{ijw}^{k+1} + \Delta \alpha_{ijq+1}^{k+1} \bar{u}_{ijw}^{k+1} + \Delta \alpha_{ijq+1}^{k+2} \bar{u}_{ijw}^{k+2}$$

Substituting the new value of inventories in criterion
(22) and taking derivatives, we obtain
\[
\begin{align*}
\frac{\partial Y}{\partial \alpha_j^{(n-1)+1}} & \approx -((\Delta_j^{(n-1)+1} + \Delta_j^{(n-1)+1} + \Delta_j^{(n-1)+1}) \\
& + [s_j^{(n-1)+2} + \Delta_j^{(n-1)+2}] \\
& + \cdots + [s_j^{(n-1)+1} + \Delta_j^{(n-1)+1}]^*) \tilde{a}_{j,q+1} I_{j,w}  \\
\frac{\partial Y}{\partial \alpha_j^{(n-1)+2}} & \approx -((s_j^{(n-1)+2} + \Delta_j^{(n-1)+2}) \\
& + \cdots + [s_j^{(n-1)+1} + \Delta_j^{(n-1)+1}]^*) \tilde{a}_{j,q+1} I_{j,w} \\
& \ldots \\
\frac{\partial Y}{\partial \alpha_j^{(n-1)+2}} & \approx 0
\end{align*}
\]
where
\[
[s_j^k + \Delta s_j^k]^* = 1 \text{ if } s_j^k + \Delta s_j^k \geq 0 \\
\text{and } \Delta s_j^k \neq 0 \\
0 \text{ if } \Delta s_j^k = 0 \\
-\infty \text{ if } s_j^k + \Delta s_j^k \leq 0
\]
Note that the inventory at the end of the \(k\)th elementary period does not change by \(\Delta \alpha_{j_{-i-1}}^{(n-1)+1}, \ldots, \Delta \alpha_{j_{+i+1}}^{(n-1)+1}\); in fact \(s_j^{(n-1)+1} = 0\). Hence, the impact of \(\Delta \alpha_{j_{+i+1}}^{(n-1)+1}\) which satisfies eqn (31) is local and it affects only the production volume of the part-types corresponding to the macro-operation and sub-period under consideration.

To illustrate the procedure of estimating the gradient with respect to the variable \(\alpha_j^k\), we consider the following example. Let for part type \(p_j \in \bar{p}_j\), \(s_j^1 = 80\), \(s_j^2 = 50\), \(s_j^3 = 20\), \(z = 3\), \(I_j^1 = 1\), \(a_j^1 = 100\). If \(\Delta \alpha_{j_{-i-1}}^{k} = 0.2, \Delta \alpha_{j_{+i+1}}^{k} = -0.15\) and \(\Delta \alpha_{j_{+i+1}}^{k} = -0.05\), then the changes in inventories are: \(s_j^2 = (0.2) \times 100 = -20; \Delta s_j^2 = (-0.2 - 0.15) \times 100 = -5; \Delta s_j^3 = 0\), and the value of the derivatives are: \(\partial Y/\partial \alpha_j^k \approx -100; \partial Y/\partial \alpha_j^{k+1} \approx 0\).

3.2 Algorithm to update \(\alpha\) iteratively

We present an algorithm to update the values of the parameters \(\alpha\). We first evaluate perturbations \(\Delta \beta_j^k\) and then perturbations \(\Delta \alpha_j^k\).

**Algorithm APA**

I. Update parameters \(\beta\)

II. Update parameters \(\alpha\)

**Step 1.** Estimate the gradient of the cost function \(Y\) with respect to the parameters \(\alpha_j^k\) for all \(p_j \in \mathcal{F}, q = 1, 2, \ldots, \bar{p}_j - 1\) and \(k = 1, 2, \ldots, Z\)

- choose a value for the step size \(ss\)
- estimate \(\Delta \beta_j^{(n-1)+1}, \ldots, \Delta \beta_j^{(n-1)+2}\) using the equations presented in Section (3.1)
- estimate \(\partial Y/\partial \alpha_j^{(n-1)+1}, \ldots, \partial Y/\partial \alpha_j^{(n-1)+2}\) using the equations presented in Section (3.1).

**Step 2.** Sort the positive and negative gradients. Let
\[
\bar{W}_q^k(j) \left\{ k: p_j \in \bar{p}_j, \partial Y/\partial \alpha_j^k \geq 0, \alpha_j^k \neq 0; \right\}
\]
and \(W_q^k(j) \left\{ k: p_j \in \bar{p}_j, \partial Y/\partial \alpha_j^k < 0; \right\}
\]
Note that the inventory at the end of the \(k\)th elementary period does not change by \(\Delta \alpha_{j_{-i-1}}^{(n-1)+1}, \ldots, \Delta \alpha_{j_{+i+1}}^{(n-1)+1}\); in fact \(\Delta \alpha_{j_{+i+1}}^{(n-1)+1} = 0\). Hence, the impact of \(\Delta \alpha_{j_{+i+1}}^{(n-1)+1}\) which satisfies eqn (31) is local and it affects only the production volume of the part-types corresponding to the macro-operation and sub-period under consideration.

**Step 3.** Estimate \(\Delta \alpha_j^k\)

\[
\Delta \alpha_j^k \left\{ \begin{array}{ll}
\frac{\partial Y}{\partial \alpha_j^k} \min (ss, W_q^k(j)) & \text{if } k \in \bar{W}_q^k(j) \\
\frac{\partial Y}{\partial \alpha_j^k} \min (ss, W_q^k(j)) & \text{if } k \in W_q^k(j) \\
0 & \text{if } \partial Y/\partial \alpha_j^k \geq 0 \\
\text{and } \alpha_j^k = 0
\end{array} \right.
\]

**Step 4.** Check

- if \(\alpha_j^k > 1\) then
- \(ss \leftarrow ss/2\)
- \(W_q^k(j) \leftarrow W_q^k(j)/2\)
- go back to Step 2
- else
- return the vector of parameters \(\alpha\)
- end

4 ALGORITHM FOR HIERARCHICAL PRODUCTION PLANNING

In this section we present the algorithm used to solve the production planning problem. The aggregate level optimization problem presented in Section 2.2 is solved first. The detailed level problems presented in Section 2.3 are then solved for all manufacturing cells at the beginning of each sub-period. The estimates of the parameters \(\alpha_j^k\) are determined by the algorithm APA.

A rolling horizon technique is employed to adjust the production plan according to the most recent forecast of the demand.

**Step 1.** Estimate the attributes of the part families using eqns (10)-(13)

- \(I_j, \bar{I}_j, \sigma_j^k, \sigma_j, D_j^k\) for all \(p_j \in \mathcal{F}, q \in \mathcal{F}_j\) and \(k = 1, 2, \ldots, Z\)
Step 2. Initialize
- parameters \( \alpha_{j,q}^k \)
- \( K \leftarrow 1 \) (first sub-period)
- \( \text{iter} \leftarrow 1 \)
- \( \text{sum(iter)} \leftarrow 0 \)
- \( \text{stepsize} \leftarrow \text{initialstepsize} \)

Step 3. Solve aggregate level problem
- for the production values \( u_{j,w}^k \) of part families in sub-periods \( k=K,K+1,\ldots,K+Z-1 \)

Step 4. Solve detailed level problem for the \( K \)th sub-period
- for each cell \( c \in C \) individually
- for production values \( u_{j,w}^k \) of part-types in elementary periods \( k=(K-1)z+1, (K-1)z+2,\ldots,Kz \)
- \( \text{sum(iter)} \leftarrow \text{sum(iter)} + \text{objective function of detailed level Problem } P_3(v) \)

Step 5. Roll the planning horizon by one sub-period
- If \( K = Z \) Then go to Step 6
- Else \( K \leftarrow K + 1 \), estimate \( s_{j,w}^k \) for all \( p_j \in P, w \in W_j \)
- and \( k=(K-1)z+1, (K-1)z+2,\ldots,Kz \), go to Step 3

Step 6. Stopping criterion
- If \( \| (\text{sum(iter)} - \text{sum(iter-1)})/\text{sum(iter-1)} \| \leq \epsilon \)
  where \( \epsilon \leq 0.01 \) Then stop. The production plan is given by the detailed solution
- Else go to Step 7

Step 7. Iterate
- \( \text{iter} \leftarrow \text{iter} + 1 \)
- \( a \leftarrow 1 \)
- \( \text{sum(iter)} \leftarrow 0 \)
- \( \text{stepsize} \leftarrow \text{initialstepsize}/\text{iter} \)
- estimate new value for parameters \( \alpha_{j,q}^k \) from algorithm APA
- go to Step 4

5 NUMERICAL RESULTS

We first demonstrate the model of our hierarchical production planning approach using a numerical example. Subsequently, we examine the performance of the algorithm by solving sets of realistic production planning problems.

5.1 A numerical example

The sample production planning problem under consideration comprises six part types and six work-centers which are to be managed over a horizon that consists of 13 sub-periods. Each sub-period consists of four elementary periods. The routings of part types 1–6 are:

- \( \{m_1,m_2,m_3\}, \{m_3,m_4,m_5\}, \{m_1,m_3,m_5\}, \{m_1,m_2,m_3\}, \{m_1,m_3,m_5\} \), respectively. Part-types \( p_1 \) and \( p_2 \), \( p_3 \) and \( p_4 \), and \( p_5 \) and \( p_6 \) are grouped in pairs to form part families \( f_1, f_2 \) and \( f_3 \) respectively. Similarly, work-centers \( m_1, m_2, m_3 \) and \( m_4, m_5, m_6 \) form manufacturing cells \( c_1, c_2 \) and \( c_3 \) respectively. For this grouping the macro-routings of part families 1–3 are \( \{c_2,c_3\}, \{c_1,c_2,c_3\} \) and \( \{c_1,c_2,c_3,c_2\} \), respectively.

The parameters \( c_{j,w}, f_{j,w} \) and \( d_{j}^k \) of the planning problem were obtained using a random number generator; \( s_{j,w}^k = 9999 \) for \( j=1,\ldots,6 \) and \( k=1,\ldots,13 \); \( s_{j,w}^0 = 100 \) for \( j=1,\ldots,6 \) and \( w=1,2,\ldots,n_j-1 \). Note that a large quantity of raw-material is available for all part types. The cost coefficients and processing times were also obtained from the random number generator. The capacity of all work-centers \( (T) \) was selected to be 750 units. The cost coefficients, the processing times \( (T_{f_{j,w}}, f_{j,w}) \) and the initial inventory states of the aggregate level are estimated from the low level parameters \( c_{j,w}, f_{j,w}, d_{j}^k \) and \( s_{j,w}^0 \) by using eqns (11)–(13). For the first iteration, the value of \( \alpha_{j,q}^k \) is assumed to be \( 0.125 \).

The numerical results for this example were obtained by executing the two-level hierarchical production planning algorithm of Section 4 on a SUN/SPARC station. The programs were coded in C and the LPs were solved by XMP. The HPP problem is solved on a rolling horizon basis, i.e. the aggregate level LP is first solved for the entire planning horizon for all aggregate time-periods. Its solution is passed through the corresponding parameters to the detailed level LPs. The latter are solved for each cell separately. The planning horizon at the detailed level is one sub-period.

The monolithic LP (Problem \( P_1 \)) has 3120 structural constraints and 2080 decision variables. The number of structural constraints and decision variables of the HPP problems are: (1) 39 and 52 for cell \( c_1 \), (2) 104 and 80 for cell \( c_2 \), (3) 59 and 48 for cell \( c_3 \) and (4) 351 and 234 for the aggregate LP. The total number of structural constraints and decision variables for the HPP problem is 553 and 414, respectively.

The value of the criterion and the computational times corresponding to the solution of the monolithic and hierarchical approaches are presented in Table 1. It is emphasized that the optimality of the solution depends on the selection of the initial step size, which is critical for the PA portion of the algorithm. The step size should go to zero at a rate that is slow enough to avoid premature convergence of the criterion to the wrong value and fast enough to lead to convergence. Results show that the converged solution was 3.17% away from the optimum criterion value obtained by solving the monolithic problem. The computational time required by the hierarchical approach was 6.8% of the time required by the monolithic approach.
Hierarchical production planning for complex manufacturing systems

Table 1. Comparison of HPP approach vs monolithic approach

<table>
<thead>
<tr>
<th>Approach</th>
<th>Optimal value of criterion ($\times 10^3$)</th>
<th>Computational time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monolithic approach</td>
<td>193.85</td>
<td>416.0</td>
</tr>
<tr>
<td>Hierarchical approach</td>
<td>777.07</td>
<td>2.4</td>
</tr>
<tr>
<td>(after first iteration)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hierarchical approach</td>
<td>200.05</td>
<td>28.2</td>
</tr>
<tr>
<td>(converged)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.2 Performance evaluation

An extensive numerical study was performed to evaluate the performance of the HPP algorithm presented in Section 4. The monolithic and the hierarchical approaches were compared by evaluating the CPU times required to obtain the production plans and the resulting values of the objective function.

Specifically, two problem sets were considered: the first set comprised small-size shop examples, while the second set comprised medium-size shop examples. The parameters selected for these problem sets are presented in Table 2. For all problems: (1) the maximum number of operations (macro-operations) on any part type (family) is assumed to be the total number of machines (cells) for that problem instance; (2) the number of elementary periods in a sub-period is three; (3) the number of sub-periods in the planning horizon is three; (4) the number of machines in a cell is five; (5) the number of part types in a part family is four; (6) the routing, the holding and backlogging costs, and the processing times of each part type belonging to a part family are similar.

The small-size shop in the sample problems consists of 10 machines that manufacture between 8 and 28 part types. The medium-size shop consists of 30 machines that manufacture between 100 and 160 part types. The parameters of all sample planning problems were obtained using a random number generator. Overall 50 example problems were solved: 30 for the small-size shop case and 20 for the medium-size shop case.

Let $C_h$ and $C_m$ be the average objective function values obtained from the hierarchical and monolithic approaches, respectively, for all problems with the same value of problem size parameters $x$ and $y$. Furthermore, let the average CPU times of the hierarchical and monolithic approaches be $T_h$ and $T_m$, respectively. The comparison of these metrics is shown in Figs 1 and 2. The results show that the hierarchical approach is computationally faster than the monolithic approach even for small-size problems. For larger problems, the hierarchical approach requires very small computational

Table 2. Parameters used in sample problems

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of part types</td>
<td>$x$</td>
</tr>
<tr>
<td>Number of machines</td>
<td>$y$</td>
</tr>
<tr>
<td>Number of families</td>
<td>$x/4$</td>
</tr>
<tr>
<td>Number of cells</td>
<td>$y/5$</td>
</tr>
<tr>
<td>Number of sub-periods in $H$</td>
<td>4</td>
</tr>
<tr>
<td>Number of elementary periods in $H$</td>
<td>12</td>
</tr>
</tbody>
</table>

$x$, $y =$ problem size parameters.
effort compared to the monolithic approach. The sub-optimality of the planning problem resulting from the hierarchical approach increases as the problem size increases, but remains within acceptable limits.

6 CONCLUSIONS

This paper presents a hierarchical approach for solving the production planning problem. Aggregation of part types, machines and time-periods is considered. A two-level hierarchy is presented. At the aggregate level the production planning of part families on manufacturing cells over aggregate elementary periods is performed. At the detailed level, the production planning of parts on work-centers over the first aggregate time period is performed. The rolling horizon strategy is utilized. Perturbation analysis is employed to adjust some control parameters at the detailed level. An algorithm is presented that provides a 'near-optimal' solution for several real-life example problems with small computational effort compared to the monolithic approach.

The proposed hierarchical scheme permits the computation of aggregate as well as detailed production plans when detailed demand/forecast information is not known (or considered necessary) over long planning horizons. It also allows absorption of random events without frequent recomputations. For faster computation, parallel processing can be employed to solve the set of independent detailed-level problems concurrently.

ACKNOWLEDGEMENTS

This work is supported by the Institute for Systems Research of the University of Maryland under Grant No. NSFD CD 8803012.

REFERENCES