SCHOOL OF OPERATIONS RESEARCH
AND INDUSTRIAL ENGINEERING
COLLEGE OF ENGINEERING
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853-3801

TECHNICAL REPORT NO. 1017

July 1992

A HIERARCHICAL APPROACH FOR METAL PARTS FABRICATION

by

Eleftherios Iakovou[1] , Kavindra Malik[2]
and John A. Muckstadt[3]

[1] Department of Industrial Engineering, University of Miami, P.O. Box 248294, Coral Gables, FL 33124. All correspondence should be sent to this author.

[2] Johnson Graduate School of Management, Cornell University, Ithaca, NY 14853.

[3] This author's research was supported in part by the National Science Foundation grant DDM-8819542.

A Hierarchical Approach for Metal Parts Fabrication

Eleftherios Iakovou* Kavindra Malik† John A. Muckstadt‡

July 1992

Abstract

When planning the design and operation of a manufacturing system in which numerically controlled (NC) machine tools are used to produce a variety of parts, a number of important factors must be taken into consideration. This paper addresses certain of these design and operational planning problems and develops a supporting optimization framework . A methodology is proposed for machining batches of parts on groups of machines in a way that accounts for machine loads and tool assignments. The models we develop consider the effects of variability and correlation of the demands among the parts. This is done so that the manufacturing system could accommodate wide fluctuations in demand without the machines suffering from significant over- and under-capacity utilization during the system's dynamic operation. The proposed approach is an integrative one, which recognizes the decisions that have to be made, the sequence in which they must be made, the computational complexity of the problem, and various operational constraints.

*Department of Industrial Engineering, University of Miami, P.O. Box 248294, Coral Gables, FL 33124. All correspondence should be sent to this author.

†Johnson Graduate School of Management, Cornell University, Ithaca, NY 14853

‡School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY 14853.

1 Introduction

When planning the design and operation of a manufacturing system in which numerically controlled (NC) machine tools are used to produce a variety of parts, a number of important factors must be taken into consideration. The purpose of this paper is to address certain of these design and operational planning problems and to develop a supporting optimization framework.

We describe below the manufacturing environment that we encountered in several plants of a major manufacturing company that fabricates industrial parts used in hydraulic control systems found in aircraft, automobiles, farm, construction and mining equipment. Recently, this company has reorganized its manufacturing facilities into focused factories, each of which has several manufacturing cells. The parts were assigned to these cells based on group technology concepts. The environment we observed in this and similar systems is organized based on product rather than processs considerations. This work focuses on operational issues in the management of this type of system.

More specifically, we consider a portion of a manufacturing cell in which the bottleneck operations are performed on a set of parallel identical flexible manufacturing machines. Since we assume there is a single bottleneck operation in the cell, we can focus on this single stage of production. While parts may require other operations, such as plating or anodyzing, which take place in other portions of the cell, we assume the system is designed such that NC machines are the bottleneck resource and determine the parts' grouping and production scheduling.

Flexible manufacturing machines (FMMs) are versatile NC machines which are able to manufacture a wide variety of parts. Each machine performs a number of operations as long as the necessary tools are loaded in its limited capacity tool magazine. Typical NC machining centers have permanently attached tool magazines, with capacities often ranging from 30 to 120 cutting tools and are equipped with automatic tool changers. In fact, the efficient use of such sophisticated and capital intensive equipment is crucial to the success of a metal parts fabrication company in an environment where competitors have similar equipment at their disposal.

The grouping of parts into families comes in accordance with the basic principles conveyed by group technology. Group technology (GT) is a manufacturing philosophy that achieves economies throughout the manufacturing cycle by grouping similar parts into families. It is applied mainly to small- and medium-sized batch production systems. Through classification and coding systems, parts that have similar design and/or manufacturing characteristics are grouped into families. The formation of part families based on design similarity results in the reduction of component variety. On the other hand, when families are formed based on the parts' manufacturing similarities, there is an impact upon the production process itself. In some cases, families identified by design similarity will also have similar manufacturing requirements; for example, parts in a family may require the same material and have the same specifications in terms of surface finish. Cellular manufacturing (CM) is a specific application of GT, which involves processing collections of similar parts, called part families, on dedicated clusters of machines, called cells.

Focused factories make use of principles conveyed by GT and CM. The manufacturing arena for the industrial environments under study is typically characterized by rapid proliferation of products with short life cycles. Consequently, a facility has to produce at the same time low volume specialty parts, high volume parts with relatively stable demand patterns, along with parts that exhibit rapidly increasing or decreasing demand rates. For such production environments, Skinner (1974) discusses the notion of *focused factory* in which segments of the manufacturing system are dedicated to the fabrication of parts with similar production volumes and manufacturing characteristics.

Since we assume parts have been clustered together into groups because they are quite similar in terms of geometry, raw materials and required fixturing, it is the tooling requirements along with machining times that will drive the sequencing and scheduling decisions for this production environment. The parts have to be grouped into families so that parts belonging to the same family share a common major setup. A switch in production from one family to another requires a major setup.

The purpose of this paper is to develop a hollistic approach for planning production in an environment characterized by cellular manufacturing and production of parts in families in a flexible machining cell, which consists of M identical machines. The primary issues that have to be resolved are :

- How should we group parts into families ?
- How should we allocate parts for production to different machines ?
- How should we schedule the production of the parts?

A major goal in the real systems we have studied has been to produce each part in a family one or more times in what is called a manufacturing cycle. The scheduling of parts in this manufacturing cycle is an important paradigm in cyclic scheduling which is fully discussed in Hall (1988) and Wittrock (1985). The length of a cycle is determined by the number of parts in a family, the production requirements for the parts, and the setup times. Since setup times consume a relatively small portion of a cycle's length, due to the construction of the families and the nature of the NC machining centers, the allocation of parts to machines is based largely on processing times and, of course, on the geometric, material, and fixturing similarities among the parts. An objective in the real application is to keep the manufacturing cycle as short as possible so that inventories can be kept low for both cycle and safety stocks. Hence a reasonable objective for our optimization problem is to make part assignments to machines such that the overall cycle length for all parts is as small as possible, that is, we would want to minimize the makespan.

The makespan for a machine is its total workload, which consists of the total processing time that has been allocated to the machine plus the setup time. The setup time in our environment is a function of the production sequence of the parts assigned to the machine and their respective tooling requirements. The system makespan is the maximum of the machine makespans. This system makespan objective incorporates two concepts: *workload balance* and *setup minimization.* As we will see, these two quantities may be in conflict with each other; hence, our goal is to identify the best trade-off between them. Let us define precisely what is meant by these two concepts.

Workload balance : Parts are assigned so that the resulting workload is evenly distributed among the machines. By doing so, the total time to complete the processing of all the parts is minimized, that is, the makespan is minimized.

Setup minimization : Each part requires a subset of tools which must be placed in the machine's tool carousel before the part can be processed. Each machine has a tool magazine with limited capacity C, and, in general, the number of tools needed to produce all the parts exceeds this capacity. Therefore, it is sometimes necessary to change tools when a machine switches from one type of part to another. The manufacturing environment that we examined consists of machining centers that are equipped with automatic tool interchanging devices. These devices can switch a set of tools simultaneously between the tool magazine and the tool storage area. In this case, an effective performance criterion of setup minimization would be the minimum number of switching instants, where a *switching instant* is an instant at which at least one tool must be switched (Tang and Denardo, (1988, II)). These switching instants are equivalent to the occurrence of setups.

As we have described the makespan minimization problem, it is a static problem faced in managing this environment. However, reality is even more complex due to the uncertainty in demands of various parts. Typically in the systems we have examined, the parts exhibit both erratic and uncertain demand patterns. For these parts a number of periods with zero or very low demand are often followed by few periods of significant demand. The time varying and uncertain nature of the demand patterns of all the parts that are to be fabricated on the same machine significantly affect the distribution of the machine's workload . How the manufacturing system accommodates wide fluctuations in the demands without causing large over- and under-capacity loading of machines in its dynamic operation is of vital importance in scheduling this system. Therefore, in the design of our scheduling framework, we explicitly incorporate the objective of smoothing the workload in the system's dynamic operation. We do this by assigning parts to machines so that the effects of erratic and unknown demand patterns are minimized.

Let us give a brief outline of our approach. Having observed the demand patterns, we first partition the parts into two categories: high volume and low volume. The first account for 80% or even more of the total production volume. Obviously variability in demand of any of these parts has a much more significant impact on the variability of a machine's workload compared with the impact of demand variability resulting from low volume parts. We therefore propose to first assign the high volume parts to machines in a way that minimizes the variability of the workload over time for each machine. Subsequently, we assign the remaining low volume parts considering setup minimization and workload balance, along with economies of scale related to the tools that have to be maintained in the tool storage area.

Finally, following the assignment of parts to machines, the setup minimization objective leads to the clustering of parts into groups or *families* based on common tooling requirements. Tooling commonalities are exploited so that the total number of switching instants, and therefore the total setup time, is minimized.

The problems of workload smoothing, workload balancing, and setup minimization are complex problems even when considered individually. They have been tackled to varying degrees in the literature, which is briefly reviewed in Section 2. However, the totality of our problem involving a complex interaction of these issues has no precedent in the literature.

The remainder of the paper is organized as follows. In Section 3 we introduce and motivate a hierarchical approach to address the above mentioned problems. In Section 4 we formulate the assignment problem for the high volume parts, and in Section 5 we present a solution procedure for this problem. Section 6 presents the assignment problem for the low volume parts, and Section 7 describes its corresponding solution procedure. Section 8 describes the scheduling of the parts that have been assigned to the same machine. Section 9 provides concluding remarks.

2 Research Literature.

As mentioned, our approach is related to the makespan scheduling idea found in the literature. Minimizing makespan even for parallel identical machines with no setup times has been shown to be NP complete (Ullman, 1976). For this problem the MULTIFIT heuristic developed by Coffman et al. (1978), based on the " first fit decreasing " bin - packing heuristic, gives a schedule with makespan at most 22% greater than that of the optimal schedule.

The LPT (Longest Processing Time first) heuristic (Graham 1969) first assigns the M longest jobs to separate machines. The remaining jobs are assigned in order of decreasing processing times to the machine that would complete the job first, given the previously assigned workload. It is shown that the algorithm produces solutions with a makespan which is in the worst case 19/12 of the optimum makespan.

There is also a considerable amount of literature related to scheduling identical machines with setups. Geoffrion and Graves (1976) studied the problem in the context of sequence dependent changeover costs, and production costs. Their model arises in chemical processes environments. Parker et al. (1977) with an objective of minimizing total changeover cost, use a Vehicle Routing Heuristic since their model is a Generalized Assignment Problem. Tang (to appear) and Wittrock (1990) give heuristics for minimizing the makespan on parallel unrelated machines that require minor setups between part types of the same family and major setups between part types of different families. However, these models do not address the problems of composing the families and assigning parts to machines even though both of the problems are interrelated. For example, Tang (to appear) and Wittrock (1990) assume the composition of the families is already prespecified.

The problem of minimizing the time dedicated to setups on a single machine using different criteria has been addressed by Tang and Denardo (1988, I and II). In the first case, the performance criterion is the minimization of the total number of tool switches. Realizing the complexity of the problem the authors provide lower bounds along with heuristics for sequencing and grouping the parts through the machine. Unfortunately, these bounds can be quite poor. Furthermore, bounds produced using various lagrangean relaxations are not always tight, either (Bard (1988)). In the second case, the criterion that they use is minimization of the total number of instants at which tools are switched. A branch and bound procedure is presented.

To the best of our knowledge, the effect of the stochastic, dynamic nature of the demand process for different parts has not been taken into consideration when establishing the allocation of workload to machines.

Surveys on GT and CM are given by Burbridge (1979) and Hyer and Wemmerlov (1984). There is an extensive set of papers addressing the parts grouping problems. Surveys are provided by King and Nakornchai (1982), Kusiak (1985), and Wemmerlov and Hyer (1986). Most of the classification schemes produce a binary matrix that provides information about the machines required for the processing of each part. The entry (i, j) of the matrix is 1 if machine i is necessary for the production of part j, and 0 otherwise. This matrix is used for the formation of part groups and equivalently of machine cells.

The methods that are used can be classified into:

1. Schemes based on the similarity index (Carrie (1973), De Witte (1980), Rajagopalan and Batra (1982)).

2. Heuristics that involve rearrangement of columns and rows of the machine/ part matrix (McCormick et al. (1972), King (1980), King and Nakornchai (1982)).

3. Mathematical Programming Techniques (Barnes (1982), Kumar et al. (1986)).

Monden (1983) and Schonberger (1982) discuss the use of CM in Japan as a crucial step to achieve just-in-time manufacturing. Flexible manufacturing systems is a specific instance of CM. Jaikumar and Wassenhove (1989) classify FMSs into three categories based on the interdependence of machine operations for a given part and the space available for the storage of work in process.

3 An Integrative - Hierarchical Approach

We will now amplify some key ideas that were briefly discussed in the introduction. As we already mentioned, a typical pattern of monthly demand can be expressed by using the Product - Quantity Pareto graph, as shown in Figure 1. The graph shows that approximately 20% of the parts account for roughly 80% of the total production volume (that is, the P/Q ratio is 20 : 80). In our experience we have found P/Q ratios of 10:90 or 5:90 not to be unusual. This suggests that a small fraction of parts, the high volume parts (HVP), constitute a large fraction of the work content. Assume that a high volume part which accounts for 10% of the total production volume and a low volume part (LVP) which accounts for 1% both have the same high coefficient of variation. Suppose they have been assigned to the same machine. It is obvious that the variability of the high volume part would have a much more significant effect on the machine's workload than the variability of the low volume part. Therefore, fluctuations of the demands of the high volume parts assigned to a machine have a major impact on the variability of the workload assigned to the corresponding machine.

This leads us to suggest a decomposition of our parts assignment problem into the subproblem of assigning the high volume parts and the one of assigning the low volume parts. The high volume parts have a larger impact on the work content, and therefore, the objective of balancing and smoothing the machine workloads in their dynamic operation is largely determined by their assignment. The low volume parts are larger in number and lower in work content. Therefore, their assignment affects the setup minimization objective to a larger extent. In addition, this assignment can be used to improve the balancing of workloads among machines. Thus, we propose a hierarchy of decisions in which we first assign the high volume parts to machines and subsequently make the assignment of the low volume ones.

In order to achieve a relatively smooth workload for a machine, we have to take into consideration correlations among the demand histories of the high volume parts assigned to the same machine. Consider the case that two high volume parts with significant positive correlation have been assigned to the same machine. If we look at the total demand in machining time as an aggregate commodity, the new " aggregate " part would most probably exhibit higher variability than the individual parts, if their demands are indepedent, since periods of high (or respectively low) demand for both parts would coincide. A machine could experience several consecutive periods of very low demand, followed by periods of high demand that could go far beyond the available capacity. This phenomenon can promote erratic machine workload assignments. On the other hand, if the two high volume parts have a significant negative correlation, their demands in machining time would be synchronized in such a way that periods of high demand for the first part would be periods of low demand for the other part and vice versa. This allocation would provide a machine with a much less variable workload. The objective function for the assignment of the high volume parts reflects the desire to create a workload assignment which is relatively insensitive to demand variability. A model that reflects these ideas is presented in the next section.

Once the high volume parts are assigned, we then assign the remaining low volume parts to the machines, taking into consideration the allocations that already have been made. That is, the assignment of the low volume parts must take into account the total workload and the tooling requirements of the high volume parts assigned to each machine. Given the remaining available capacity, the goal is to allocate the remaining parts to machines so that the newly added parts require as few additional tools as possible. This helps to achieve two of our objectives. First, the total number of tools carried is reduced and corresponding economic objectives are met. Second, the scheduling of parts on machines subsequent to part assignment leads to a more manageable grouping problem. Since tool commonality is considered in part assignment, schedules with fewer switching instants are easier to find.

At the final step of the hierarchy all parts have been assigned among the different machines. For each machine we have to schedule the production of the assigned parts so that the total time dedicated to setups, which is a function of the number of switching instants, is minimized.

As we mentioned earlier, our goal in managing the manufacturing environment under study is to be able to respond quickly to customer demands, to keep inventory levels low, and therefore to reduce the length of the manufacturing cycle. This cycle is on the order of days, while the part allocations and the overall planning framework will be performed over a longer period, which we call planning horizon (PH). PH is on the order of weeks and often of months. For this horizon we have reliable estimates for part demands. In our hierarchical scheme the objective will be to complete the maximum number of cycles during the length of the planning horizon.

4 Assignment Problem for the High Volume Parts

Recognizing our earlier comments about the effects of positive and negative correlation among the high volume parts, our goal in assigning these parts is to smooth the workload for each machine throughout the planning horizon. As we mentioned earlier, correlations among demand histories for the low volume parts do not affect the distribution for the workload of a machine as much.

For each high volume part, we assume the following data are available : the average demand in machining time over the planning horizon, the set of tooling requirements, and a variance - covariance matrix of demands. The diagonal entry $(\cdot)_{ii}$ of the variance - covariance matrix gives the variance of demand for each part i and the $(\cdot)_{ij}$ entry, where $i \neq j$, measures the correlation for the demands between parts i and j.

Let W_{ij} be a reward when parts i and j are both assigned to the same machine. $W_{ij}(.)$ is a function of the correlation for the demands of parts i and j weighted by the average processing times for parts i and j. A candidate reward function would be:

$$W_{ij} = (Corr(i,j) - 1)^2 * (p_i + p_j)/P_{HV},$$

where p_i = average total processing time for part i (average demand over the planning horizon (PH) times the unit processing time), and P_{HV} is the total processing time for all the HVPs.

The function $W_{i,j}$ assigns the highest reward when the correlation is -1, and no reward for correlation +1. Alternate specifications of W_{ij} may also include information about the commonality of tooling requirements. In general, the weights W_{ij} would depend on the nature of the particular problem and the relative importance of exploiting tool commonality when assigning parts to machines versus the significance of smoothing the workload on each machine. If all the high volume parts have similar tooling requirements, then correlation is really the key factor in assigning parts to machines. Let

$$Y_{im} = \begin{cases} 1, & \text{if part } i \text{ is assigned to machine } m, \\ 0, & \text{otherwise .} \end{cases}$$

With the above-mentioned notation, we formulate the following part assignment problem, which we denote by HVPA :

(HVPA) :

$$max \sum_{m=1}^{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} W_{ij} Y_{im} Y_{jm} \qquad (1)$$

Subject to:

$$\sum_{m=1}^{M} Y_{im} = 1, \;\; i = 1, \ldots, N, \tag{2}$$

$$\sum_{i=1}^{N} p_i Y_{im} \leq CAP_m, \;\; m = 1 \ldots M, \tag{3}$$

$$Y_{im} \in \{0, 1\}, \; i = 1, \ldots, N, \;\; m = 1, \ldots, M. \tag{4}$$

We assume all parts are to be produced to meet demand for the planning horizon PH of time units (e.g., a week). This is the planning horizon for which the average processing times, p_i, of the parts are estimated. CAP_m denotes the estimated total available time on machine m. Each machine may experience non-negligible down-time due to both planned events (e.g. routine maintenance) and unexpected ones such as failures. Since the assignment of parts to machines is unknown at this point, we initially have no knowledge about the formation of part families for each machine. Clearly the setup time depends on the families that have been assigned to each machine. Since we do not know the composition of the families, we can only initially estimate the amount of setup time that will be required to implement a schedule. Once estimated, we subtract this amount of time from the total available time to obtain an estimate of the remaining run time capacity. Consequently, we assume that $CAP_m < PH$, but CAP_m is a significant portion of the length of the planning horizon.

The objective is the maximization of the total reward. Let's examine the problem constraints. Constraints (2) require each part to be assigned to exactly one machine. Constraints (3) are machine capacity constraints. This formulation is a quadratic assignment problem. We propose below a lagrangean optimization for solving that problem, which exploits the problem's structure. The approach is theoretically sound, intuitively appealing and computationally tractable.

5 A Solution Procedure for HVPA

To solve problem HVPA we construct a related problem in which the assignment constraints (2) and the machine capacity constraints (3) are relaxed. Let lagrangean multipliers λ_i and μ_m be the corresponding lagrangean multipliers, respectively. Then the relaxed problem can be written as :

$$L(\lambda,\mu) = max \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} W_{ij}Y_{im}Y_{jm} - \sum_{i=1}^{N}\sum_{m=1}^{M}(\mu_m p_i + \lambda_i)Y_{im} + \sum_{i=1}^{N}\lambda_i + \sum_{m=1}^{M}\mu_m CAP_m$$

s.t.

$$Y_{im} \in \{0,1\},\ i = 1,\ldots,N,\ \ m = 1,\ldots,M,$$

where $\mu_m \geq 0,\ \forall m$ while λ_i is unrestricted.

As is well known, the optimal value of the lagrangean dual $\min_{\mu_i\geq 0,\lambda} L(\lambda,\mu)$ is an upper bound on the optimum objective value for the original problem. $L(\lambda,\mu)$ can be solved using a method first given by Rhys (1970) and explained by Balinski (1970). The Y_{im} that constitute the solution can be efficiently computed by using a maximum flow algorithm on a bipartite graph. In the next section we give an adaptation of Rhys' method in solving $L(\lambda,\mu)$ and the HVPA problem.

5.1 Solving the Lagrangean Dual

Let $c_{im} = \mu_m p_i + \lambda_i$ $(i = 1,\ldots N;\ m = 1,\ldots,M)$. Observe that by definition $W_{ij} \geq 0$ and that if $c_{im} < 0$, it is optimal to set $Y_{im} = 1$, since c_{im} appears with a negative sign in the objective function of $L(\lambda,\mu)$. The remaining problem can be solved in polynomial time using a maximum network flow algorithm on the directed, bipartite network $G_{\lambda,\mu}$ associated with the problem $L(\lambda,\mu)$ as described below.

The vertex set consists of the source s, the sink t, and two additional sets of vertices $\Phi = \{Y_{im}Y_{jm},\ i,j = 1,\ldots,N\ ,m = 1,\ldots,M\}$ and
$\Psi = \{Y_{im},\ i = 1,\ \ldots,N\ ,m = 1,\ldots,M\}$.

The arc set consists of arcs $(s, Y_{im}Y_{jm})$, $(Y_{im}Y_{jm}, Y_{jm})$, and (Y_{im}, t). The capacities on the arcs of the network are as follows. The arcs emanating from the source, $(s, Y_{im}Y_{jm})$ have capacity W_{ij}, arcs incident to the sink (Y_{im}, t) have capacity c_{im}, and arcs $(Y_{im}Y_{jm}, Y_{jm})$ have an infinite capacity. Recall that a cut is a partition of the set of nodes into two sets, say M and $\bar{M}$, with the source $s \in M$ and the sink $t \in \bar{M}$, and that the value of a cut is the sum of the capacities on arcs (i,j) with $i \in M$ and $j \in \bar{M}$. In Figure 2, we show the network $G_{\lambda,\mu}$, with the arc capacities.

The problem can be solved in network terms by observing some key properties of the network $G_{\lambda,\mu}$. By the way that we constructed the network there is one-to-one correspondence between cuts containing no arc of type (Φ, Ψ) and feasible solutions to the problem. Moreover, no arc of type (Φ, Ψ) can belong to the minimum cut, since such an arc has infinite capacity. Finally, the minimum cut corresponds to the optimal solution. Therefore, an algorithm for finding an optimal solution $Y^\star$ is to use a labelling procedure (see, e.g. Ford (1962)) to maximize the flow in the network $G_{\lambda,\mu}$ which at the same time identifies a minimum cut.

The dual lagrangean problem $\min_{\mu_i \geq 0,\lambda} L(\lambda, \mu)$ is solved by using the subgradient algorithm (Held et al., (1978)). It is a standard procedure in lagrangean optimization that has been quite successful in solving many hard combinatorial problems. Fisher (1981) provides a broad and insightful review of the technique along with a number of its successful applications. We describe the algorithm for the problem, which uses the subgradient algorithm, in the following subsection.

5.2 Lagrangean Based Algorithm

Let us note some of the features that the lagrangean dual should exhibit:

1. As the lagrangean multipliers λ_i and μ_m are updated from iteration to iteration, the relaxed problem's objective function value provides an upper bound on the objective of the original problem. However, the optimal solution for the relaxed problem does not necessarily satisfy the two relaxed constraints (2) and (3). When λ, μ are near their optimum values, as a result of applying the subgradient optimization algorithm, the Y_{im} variable values that solve the coresponding problem $L(\lambda, \mu)$ will provide a feasible or close to feasible solution to HVPA. We begin finding the solution by employing a greedy heuristic to generate an initial feasible solution that also provides us with a lower bound.

   In the design of our heuristic we note that the relaxed capacity constraint (3) involves a rough cut estimate of machine's capacity and the low volume parts are yet to be assigned. As noted earlier, the violations of (3) in the lagrangean solution are small. Therefore, we do not necessarily find a solution which exactly matches the available capacity. Rather our procedure concentrates on finding a solution that removes the violations of assignment constraint (2).

   Greedy Procedure:

   Let $\lambda^\star$, $\mu^\star$ be the best (λ, μ) computed by the subgradient and $Y^\star_{im}$ the optimizers of $L((\lambda^\star, \mu^\star)$.

   - Step 1. Identify the sets J_1 and J_2 of parts that violate assignment constraints, i.e., $J_1 = \{i : \sum_{m=1}^{M} Y_{im} > 1\}$and $J_2 = \{i : \sum_{m=1}^{M} Y_{im} = 0\}$. Denote by Y^0_{im} the Y_{im}'s that satisfy ($Y^\star_{im} = 1$ and $\sum_{m=1}^{M} Y^\star_{im} = 1$).

- Step 2. Stop if $J_1 \bigcup J_2 = 0$; otherwise, pick the first $i \in J_1 \bigcup J_2$. For $i \in J_1$ do: for all m such that $Y_{im} = 1$ calculate the corresponding profit : $Profit_m = Reward_m - Cost_m = (\sum_{j:Y^0_{jm=1}} W_{ij}) - c_{im}$. Pick $\hat{m}$ with the largest profit and maintain $Y_{i\hat{m}} = 1$. For the specific part i, set the remaining Y_{im}'s equal to zero, then set $J_1 = J_1 \setminus i$, GOTO (2).

  For $i \in J_2$ do: for all m calculate the corresponding profit $Profit_m$ as before. If $\hat{m}$ is the machine with the largest profit set $Y_{i\hat{m}} = 1$, then set $J_2 = J_2 \setminus i$, GOTO (2).

2. After any iteration, we have the greatest lower bound (R^{LB}) and the least upper bound (R^{UB}) over all iterations carried out thus far. This allows us to compute a percentage error

$$(R^{UB} - R^{LB})/R_{LB} \times 100\%$$

In our tests we have chosen to terminate the heuristic the first time one of the following occurs:

- The percentage error is less than 5%
- The percentage error has not decreased for a prespecified number of iterations.
- The total number of iterations exceeds a prespecified number.

6 Assignment Problem for the Low Volume Parts

By assigning the high volume parts to machines, we also establish a set of tools assigned to each machine. We will call these tools the *seed* tools for every machine. Our goal is to allocate the low volume parts based on the composition of the seed tools and on the effective remaining capacity of each machine. Our objective will be to assign all remaining parts to machines by adding as few new tools as possible to each machine, and without exceeding the machine capacity constraints.

To formulate the problem we will use the following notation:

$$U_{tm} = \begin{cases} 1, & \text{if tool } t \text{ is added to machine } m, \\ 0, & \text{otherwise,} \end{cases}$$

$$Z_{im} = \begin{cases} 1, & \text{if part } i \text{ is added to machine } m, \\ 0, & \text{otherwise.} \end{cases}$$

For every machine, we will call S_m the set of tools that have been assigned to it, up to that stage. Initially this consists of the tools assigned from the solution to the HVPA problem. For every part we denote by T_i the set of its tooling requirements. Then the overall problem can be formulated as the following integer problem, which we denote by LVPA :

(**LVPA**) :

$$min \sum_{m} \sum_{t \notin S_m} U_{tm} \tag{5}$$

Subject to:

$$U_{tm} \geq Z_{im}, \; t \notin S_m, \; \forall t \in T_i, \; \forall i, \tag{6}$$

$$\sum_{m} Z_{im} = 1, \;\; \forall i, \tag{7}$$

$$\sum_{i} p_i Z_{im} \leq CAP_m, \;\; \forall m, \tag{8}$$

$$U_{tm}, \; Z_{tm} \in \{0,1\} \;\; \forall t, \;\; \forall m. \tag{9}$$

Constraints (7) force every low volume part to be assigned to exactly one machine, while constraints (8) are machine capacity constraints. Recall from the previous sections that certain high volume parts along with the tools that they require have been assigned to each machine. Constraints (6) require all tools to be on a machine if the part is assigned to the machine. The objective function counts the number of new tools that have to be added to the machine because of the low volume parts that are assigned to it. The objective is to minimize the total number of the new tools assigned to machines, thereby minimizing setup time.

In the next section we develop a solution scheme for this problem that appears in the second stage of our hierarchical decision-making process.

7 A Solution Procedure for LVPA

7.1 Introduction

When assigning the low volume parts to machines we have two objectives. Firstly, we strive to add as few new tools to machines as possible while exploiting tool commonality of the parts and secondly, to minimize the makespan.

Let us first look at the makespan objective. For this problem we are given a set of N jobs with integral processing times p_i to be scheduled on M identical machines. As we mentioned before, the minimum makespan problem is NP-complete; therefore, it is extremely unlikely that an efficient algorithm exists to find a schedule that achieves the optimal makespan . We will denote the optimal value of the makespan, for a given instance of processing times and number of machines, by OPT_{MS}. Because of the complexity of the problem, it is natural to consider algorithms that are guaranteed to produce solutions close to the optimum. Polynomial-time algorithms that produce solutions that are at most $(1+\epsilon)$ of the optimal value are called ϵ- *approximation algorithms.* Minimizing makespan is one of the problems that have been studied the most in the theory of approximation algorithms for NP-hard problems.

The first class of algorithms that have been proposed for the minimum makespan problem is the class of *list processing* algorithms. According to this class of algorithms, the jobs are ordered in a list, and the next job on the list is assigned to the next machine that will become idle. Graham (1966), showed that any such algorithm gives a schedule that has makespan at most $(2-1/m)OPT_{MS}$. Graham again (1969), showed that if the jobs are ordered with the Longest Processing Time rule (LPT), then the produced schedule has makespan at most $(4/3-1/(3m))OPT_{MS}$.

A closely related problem is the bin-packing problem. In this type of problem there are N pieces of size p_i, with $p_i \in [0,1]$. The objective is to pack the pieces into bins, under the constraint that the sum of the pieces packed to a specific bin would not exceed 1, so that the number of bins used is minimized.

Coffman et al. (1978) exploited the relationship between these two problems deriving their MULTIFIT algorithm for the minimum makespan problem. The MULTIFIT algorithm is an extension of the FIRST FIT DECREASING bin-packing problem. It is proved that it provides a schedule with makespan at most $1.22OPT_{MS}$. MULTIFIT-based algorithms have the best known bounds, among algorithms that are polynomial in the length of the input.

We will use a MULTIFIT-based algorithm with additional considerations to assign the low volume parts to machines.

7.2 MULTIFIT Algorithm

Our MULTIFIT algorithm uses a binary search on the makespan. Initially, upper and lower bounds on the makespan are computed. Then, at each iteration, the mean of the two bounds is used as a candidate makespan, MS. Then an allocation algorithm (ALLOCATE) tries to compute a feasible allocation for MS; that is, an allocation for which all jobs are completed before MS. If a feasible allocation is achieved, then the upper bound (MS_{UB})is set to MS. Otherwise, the lower bound (MS_{LB})is set to $MS + 1$. The search is terminated when the two bounds coincide.

The initial lower bound will be set to zero. We will use as an upper bound the maximum time capacity for each machine. This could be the length of the planning horizon (PH) (for example, this could be a week or a month). This is the length of the time for which the processing times p_i have been estimated. The makespan must satisfy $MS \leq PH$. Our goal is to assign the set of the low volume parts to the M machines. There is a bound k on the desired number of iterations. Then the MULTIFIT algorithm proceeds as follows:

1. Set $MS_{LB} \leftarrow 0$;

   $MS_{UB} \leftarrow PH$;

   $I \leftarrow 1$;

2. If $I > k$, halt.

   Otherwise, set $MS \leftarrow [MS_{UB}(I-1) + MS_{LB}(I-1)]/2$.

3. If ALLOCATE assigns all parts then, set $MS_{UB}(I) \leftarrow MS$;

   $MS_{LB}(I) \leftarrow MS_{LB}(I-1)$;

   $I \leftarrow I + 1$;

   and go to 2.

4. If ALLOCATE cannot assign all the parts, set

   $MS_{LB}(I) \leftarrow MS$;

   $MS_{UB}(I) \leftarrow MS_{UB}(I-1)$;

$I \leftarrow I + 1;$

and go to 2.

7.3 Algorithm ALLOCATE

We now construct the procedure ALLOCATE, which at every iteration of the MULTIFIT algorithm tries to allocate parts (among the low volume ones) to machines for a given candidate makespan MS. The objective is to assign as many parts as possible to each machine, adding as few new tools and satisfying the capacity determined by the current makespan estimate, MS.

Recall that high volume parts have already been assigned to machines. The workload corresponding to this assignment differs from machine to machine, and therefore for the current MS and the remaining capacity CAP_m is different for each machine. ALLOCATE considers the machines sequentially, in decreasing order of assigned workload, and allocates parts given the capacity CAP_m, for every machine m.

We are going to use the following notation:

$$X_i = \begin{cases} 1, & \text{if part } i \text{ is chosen to be assigned,} \\ 0, & \text{otherwise,} \end{cases}$$

$$R_t = \begin{cases} 1, & \text{if tool } t \text{ is chosen to be assigned to the machine,} \\ 0, & \text{otherwise.} \end{cases}$$

$$\xi_t = \begin{cases} 0, & \text{if tool } t \text{ belongs to the set of the seed tools,} \\ 1, & \text{otherwise.} \end{cases}$$

T_i is the set of tooling requirements for part i.

At every step of the MULTIFIT algorithm, ALLOCATE is called. ALLOCATE goes through all the machines sequentially, and allocates parts given the remaining capacities CAP_m's that result from the current MS. After every assignment of a part to a machine, a list that includes all unassigned parts is updated, along with the remaining capacity for the machine.

For a *specific machine* the problem can be formulated as the following integer program :

$$max \sum_i X_i \tag{10}$$

subject to :

$$X_i \leq R_t, \; if \; t \in T_i, \tag{11}$$

$$\sum_t \xi_t R_t \leq M, \tag{12}$$

$$\sum_i p_i X_i \leq CAP_m, \tag{13}$$

$$X_i \in \{0,1\}, \;\; , \; R_t \in \{0,1\}. \tag{14}$$

Constraints (11) ensure that if part i is picked then all tools that it requires have to be picked, too. Constraint (13) is the machine capacity constraint. Constraint (12) imposes a constraint on the total number of new tools that are assigned to the machine, where M is just a parameter, measuring perhaps, the remaining tool magazine capacity. A tool will be characterized as new, if it is not one of the tools (seed tools) that have been allocated to the machine already as a result of the assignment of the high volume parts, done in the first stage of our optimization procedure. Through the parameter ξ_t only the allocation of new tools is considered.

The above optimization problem has a special structure, which we will try to exploit. We relax constraints (12) and (13) using nonnegative lagrangean multipliers α and β respectively.

Then the lagrangean relaxed problem is :

$$L(\alpha,\beta) = max \sum_i (1-\beta p_i)X_i - \alpha \sum_t \xi_t R_t + \beta CAP_m + \alpha M$$

s.t.

$$X_i \leq R_t,\ if\ t \in T_i.$$

Notice that $L_{\alpha,\beta}$ has the same form as the optimal selection problem, already used in the allocation of the high volume parts to machines. Observe that the lagrangean function consists again of a positive and a negative part, as the objective function of the optimal selection problem does. Therefore, we can maximize this function in polynomial time using a maximum network flow algorithm. Let's describe now the network $\Theta(m)_{\alpha,\beta}$ that is associated with the problem $L_{\alpha,\beta}$ for machine m.

The vertex set consists of the source s, the sink r, and two additional sets of vertices $\Phi = \{i : i \text{ is an unassigned low volume part }\}$ and $\Psi = \{t : a_{ti} = 1,\ i\ \in \Phi\}$, that is, the set of tools that are required by the unassigned low volume parts. The arc set consists of arcs (s,i), (i,t), and (t,r). The capacities on the arcs of the network are as follows. Arcs emanating from the source, (s,i), have capacity $b_i = (1-\beta p_i)$, arcs incident to the sink, (t,r), have capacity α, and arcs going from set Φ to Ψ, (i,t), for $t\ \in T_i$, have infinite capacity. In Figure 3, we formulate the network. On every arc we assign its capacity. We again use subgradient optimization to solve the lagrangean problem $L_{\alpha,\beta}$.

Recall that at each iteration of the MULTIFIT algorithm we assign parts to each machine and then proceed to the next machine sequentially. For every machine m, given the lagrange mulitipliers α and β, we solve a maximum flow algorithm on the network $\Theta(m)_{\alpha,\beta}$. This algorithm will also identify the minimum cut. As mentioned in Section 5, the minimum cut identifies the parts along with their tools that are to be chosen to be assigned to the machine;

this selection is the optimal one, that is, it maximizes $L_{\alpha,\beta}$.

The parts assigned to machine m are deleted from the set Φ of the low volume parts that are still to be assigned. At the same time the set Ψ is updated,too. These updated sets are then used for the formation of $\Theta(m+1)_{\alpha,\beta}$ at the next step of the algorithm (the assignment of parts to the next machine).

To reduce the computational effort involved in assigning the LVPs we propose the following procedure. As it can been seen in Figure 1, the majority of the LVPs (approximately 50% of the total number of parts) have a marginal effect on the cumulative demand. Therefore we partition the set of LVPs into two subsets. The first consists of the LVPs with the relatively highest contribution to the total cumulative demand (approximately 30% of the parts), and the second with the LVPs that contribute the least to the total demand. For the first subset of parts we use the MULTIFIT algorithm along with the ALLOCATE procedure, as is described above. For the remaining ones a simpler procedure, rather than ALLOCATE, can be used at each iteration of the MULTIFIT algorithm. This procedure would be a FIRST FIT DECREASING heuristic with the extra requirement that the tooling constraint (12) should be satisfied every time a part is allocated to a machine. If this is not the case, then the heuristic attempts to allocate the part to the next machine on the list, that is the one with the immediately higher assigned workload.

8 Clustering of Parts into Families

At the first stage of the hierarchy we assigned the high volume parts to machines based on the correlation among the parts and on similarities on tooling. With the high volume parts and their corresponding tools assigned, we subsequently allocated the low volume parts to machines with the dual objectives of minimizing makespan along with the total number of new tools added to each machine. Finally all parts have been assigned for fabrication to the different machines. We now have to schedule the production of the parts through each machine so that we reduce the time dedicated to setups by exploiting tool commonalities.

For a given assignment of parts to machines we will use the following natural way for grouping the parts into families. Parts would be assigned to the same family if they can be processed with the same tools in the magazine of the machine. Therefore, there is a one-to-one correspondence between these part families and specific tool configurations of machine magazines. These *families* are groups of parts that were formed based on the parts' tooling requirements. For a given allocation of parts to a single machine, Tang and Denardo (1988, II) address the issue of grouping parts into families with the objective of minimizing the total number of instants at which tools are switched. They show that the particular scheduling problem generalizes the classical bin packing problem. A branch-and-bound procedure that terminates with an optimal solution is developed, and quite satisfactory computational results are presented.

9 Conclusions

In this paper we studied a problem faced by manufacturers of industrial products, which are subject to varying demands over time. In this environment parts are grouped according to the principles of group technology and cellular manufacturing. Based on these concepts, we have developed an approach for machining families of parts on a set of machines considering the effects of machine loads and tool assignments. In these environments, both workload smoothing and setup time reduction are crucial for increasing the effective capacity of the system and therefore minimizing the cycle's length. The issues we consider are at the operational level for a multi-machine manufacturing cell. More specifically, we showed how to assign parts to machines and group parts to families based on their tooling requirements. The procedure we developed is a hierarchical scheme that first assigns the high volume parts to machines, and then the low volume ones.

One of the major contributions of this paper is the inclusion of variability and correlation for the demands among the parts in the models that determine the assignment of parts to machines. In addition, in our models we formulate a quadratic assignment problem, which has many applications relevant to the grouping of parts in other settings. An important contribution of this paper is to show the equivalence of the assignment problem to a maximum network flow problem. This equivalence shows that a simple and efficient solution method can be used to obtain a good allocation.

References

[1] Balinski, M.L., 1970, On a Selection Problem. *Management Science* 17, 230-231.

[2] Bard J.F., 1988, A Heuristic for Minimizing the Number of Tool Switches on a Flexible Machine. *IIE Transactions* 20, 382-391.

[3] Barnes, E.R., 1982, An Algorithm for Partitioning the Nodes of a Graph. *SIAM J. Alg. Disc. Math*, 3, 541-550.

[4] Burbridge, J.L., 1979 *Group Technology in the Engineering Industry.* (London: Mechanical Engineering Publications).

[5] Carrie, A.S., 1973, Numerical Taxonomy Applied to Group Technology and Plant Layout. *International Journal of Production Research*, 11, 399-416.

[6] Carrie, A.S. and D.T.S. Perera, 1986, Work Scheduling in FMS Under Tool Availability Constraints. *International Journal of Production Research*, 24 (6) 1299-1308.

[7] Coffman E.G., M. Carey, and D. Johnson, 1978, An Application of Bin Packing to Multi-processor Scheduling. *SIAM Journal of Computing*, Vol. 7, 1-16.

[8] De Witte, J., 1980, The Use of Similarity Coefficients in Production Flow Analysis. *International Journal of Production Research*, 18, 503-514.

[9] Fisher, M.L., 1981, The Lagrangean Relaxation Method for Solving Integer Programming Problems. *Management Science*, 27, 1-18.

[10] Ford, L.R., Jr. and D.R. Fulkerson, 1962, *Flows in Networks.* Princeton University Press.

[11] Geoffrion A. and G. Graves, 1976, Scheduling Parallel Production Lines with Changeover Costs: Practical Application of a Quadratic Assignment/ LP Approach. *Operations Research*, Vol. 24, 595-610.

[12] Golden, B.L. and W.R. Stewart. Empirical Analysis of Heuristics. *The Traveling Salesman Problem*,(E.L. Lawler et al. (eds). John Wiley & Sons, Chichester, 207-249).

[13] Graham, R.L., 1966, Bounds for Certain Multiprocessing Anomalies. *Bell Syst. Tech. J.*, 45, 1563-1581.

[14] Graham, R.L., 1969, Bounds for Multiprocessing Timing Anomalies. *SIAM J. Appl. Math.*, 17, 263-269.

[15] Hall, R.W.,Cyclic Scheduling for Improvement, 1988, *International Journal of Production Research*, 26, no. 3, 457-472.

[16] Held, M., P. Wole, and H.P. Crowder, 1974, Validation of Subgradient Optimization. *Mathematical Programming*, 6, 62-88.

[17] Hyer, N.L. and U. Wemmerlov, 1984, Group Technology and Productivity. *Harvard Business Review*, 140-149.

[18] Jaikumar R., and Wassenhove L.N.V. , 1989, A Production Planning Framework for Flexible Manufacturing Systems. *Journal of Mfg. and Operations Mgt.* 2, 52-79.

[19] King, J.R., 1980, Machine-Component Grouping in Production Flow Analysis: An Approach Using a Rank Clustering Algorithm. *International Journal of Production Research*, 18, 213-232.

[20] King, J.R., and V. Nakornchai, 1982, Machine-Component Group Formation in Group Technology: Review and Extension. *International Journal of Production Research*, 20, 117-133.

[21] Kumar, K.R., A. Kusiak and A. Vannelli, 1986, Grouping of Parts and Components in Flexible Manufacturing Systems. *European Journal of Operations Research*, 24, 387-397.

[22] Kusiak, A., 1985, The Part Families Problem in Flexible Manufacturing Systems. *Annals of Operations Research* 3, 279-300.

[23] McCormick, W.T., P.J. Schweitzer, and T.W. White, 1972, Problem Decomposition and Data Reorganization by a Clustering Technique. *Operations Research*, 20, 993-1009.

[24] Monden Y., 1983, *Toyota Production System: Practical Approach to Production Management* (Industrial Engineering and Management Press).

[25] Parker R.G., R.H.Deane, R.A. Holmes, 1977, On the use of A Vechicle Routing Algorithm for the Parallel Processor Problem with Sequence Dependent Changeover Costs. *AIIE Transactions*, Vol. 9, No. 2, 155-160.

[26] Rajagopalan, R., and J.L. Batra, 1982, Design of Cellular Production Systems : A Graph-Theoritic Approach. *International Journal of Production Research,* 13, 567-579.

[27] Rhys, J.M.W., 1970, A Selection Problem of Shared Fixed Costs and Network Flows. *Management Science*, 17, 200-207.

[28] Schonberger, R.J., 1982, *Japanese Manufacturing Techniques.* (New York : The Free Press).

[29] Skinner, W., 1974, The Focused Factory. *Harvard Business Review*, 111-122.

[30] Tang C., Scheduling Batches on Flexible Manufacturing Machines. To appear in *European Journal of Operations Research.*

[31] Tang C.S, E.V. Denardo, 1988, Models Arising from a Flexible Manufacturing Machine, Part I: Minimization of the Number of the Tool Switches. *Operations Research*, 36, 767-777.

[32] C.S. Tang, E.V. Denardo, 1988, Models Arising from a Flexible Manufacturing Machine, Part II: Minimizing the Number of Switching Instants in an FMS. *Operations Research*, 36, 778-784.

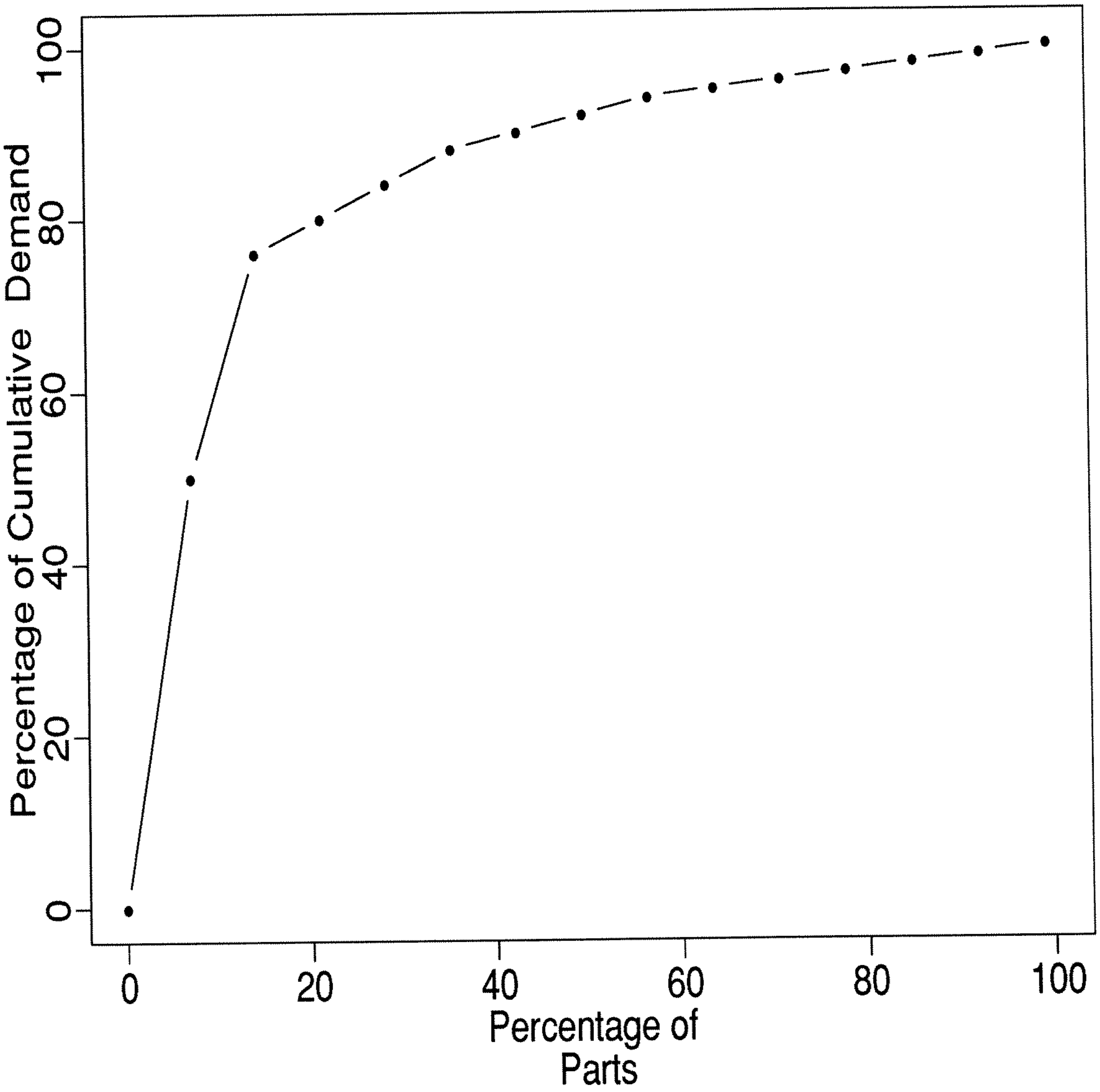
[33] Ullman J.D., 1976, Complexity of Scheduling Problems. *Computer and Job/Shop Scheduling Theory.* (E.G. Coffman (Ed.) , John Wiley, pp. 139-164).

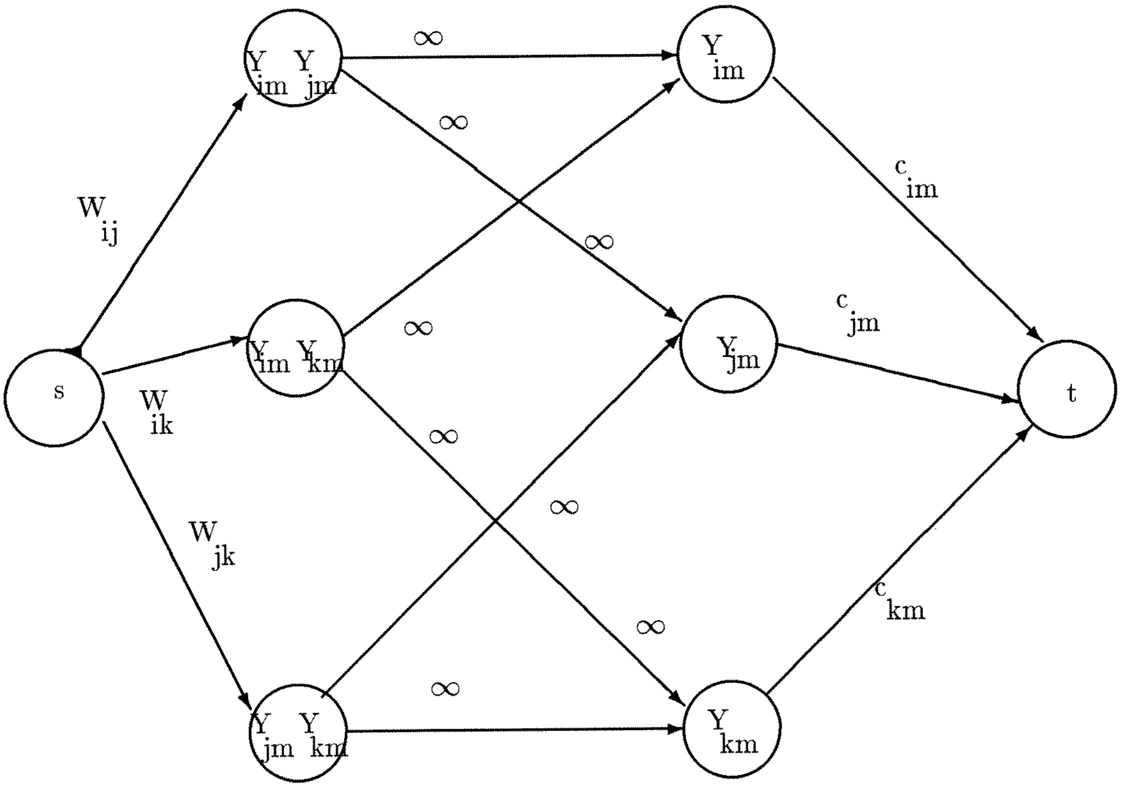
[34] Wemmerlov, U. and N.L. Hyer, 1986, Procedures for the Part Family - Machine Group Identification Problem in Cellular Manufacturing. *Journal of Operations Management* 6, 125-147.

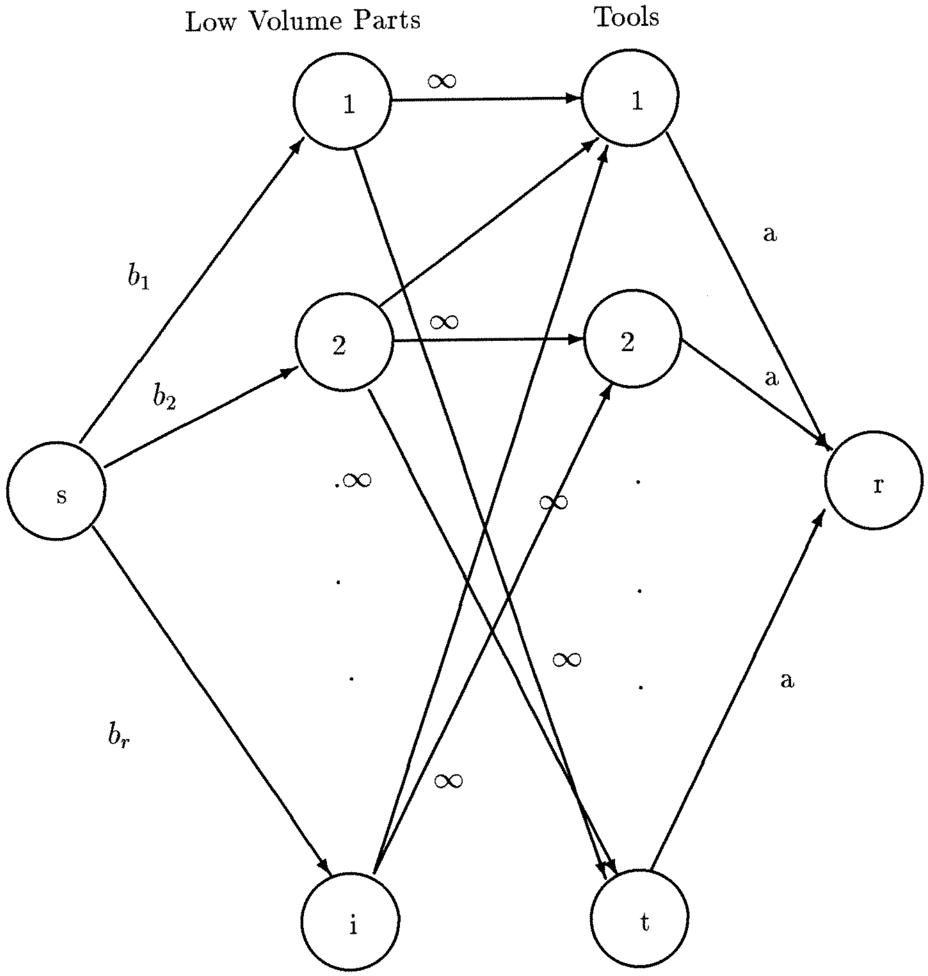
[35] Wittrock, R. J., 1985, Scheduling algorithms for flexible flow lines, *IBM J. Res. Develop.*, 29, 4, 401-411.

[36] Wittrock, R. J., 1990, Scheduling Parallel Machines with Major and Minor Setup Times. *The International Journal of Flexible Manufacturing Systems*, 2, 329-341.

Captions of the Figures:


Percentage of Cumulative Demand
0
20
40
60
80
100
0
20
40
60
80
100
Percentage of
Parts


Y_{im} Y_{jm}
∞
Y_{im}
∞
W_{ij}
c_{im}
∞
c_{jm}
∞
Y_{im} Y_{km}
Y_{jm}
s
t
W_{ik}
∞
∞
W_{jk}
c_{km}
∞
∞
Y_{jm} Y_{km}
Y_{km}


Low Volume Parts
Tools
1
2
i
1
2
t
s
r
b_1
b_2
b_r
∞
∞
.∞
∞
∞
∞
a
a
a