Scheduling Products with Bills of Materials Using an Improved Lagrangian Relaxation Technique

Christopher S. Czerwinski, and Peter B. Luh, *Senior Member, IEEE*

***Abstract*—A bill of materials specifies the sequence in which parts are to be processed and assembled in order to manufacture a deliverable product. In practice, a bill of materials may be quite complex, involving hundreds of parts to be processed on a number of limited resources, making scheduling difficult. This has forced many practitioners to turn to Material Requirements Planning and heuristic rules to perform scheduling. These methods are seldom integrated, however, resulting in unreliable completion times for products and, hence, low customer satisfaction. This paper addresses the issue of integrally scheduling parts that are related through a bill of materials for the purpose of improving the on-time performance of products as well as reducing work-in-process (WIP) inventory. The technique presented here is based on an existing Lagrangian relaxation (LR) approach for the scheduling of independent parts in a job shop. The current problem, however, is more complicated than the job shop problem because of the constraints between parts, imposed by the bill of materials. In order to make Lagrangian relaxation a viable approach to this problem, an auxiliary problem formulation with a modified subgradient method are adopted to improve the computation time of the existing LR approach. This improved LR approach allows the bill of material constraints to be considered directly in the problem formulation. Results to date show that the above integration improves product tardiness and WIP levels, compared to techniques that do not integrate the bill of material constraints into the product scheduling problem. The improved ability of a manufacturer to meet promised delivery dates for products by the above integration will ultimately enhance its credibility and competitiveness in the marketplace.**

I. Introduction

A. The Product Scheduling Problem

A BILL of materials is often used to detail the raw material requirements of a product and to specify the order in which *parts* within the bill of materials are to be processed or assembled. Figure 1 illustrates the bill of materials for a Pump Assembly which requires the mating of a support assembly with a housing. The housing, in turn, requires the assembly of a left housing mold and a center housing mold. This hierarchical arrangement of parts is common to a bill of materials and is often more complex than the example provided here. The manufacturing of a part is further comprised of a sequence of *operations*, each requiring processing by a particular machine type for a certain processing time. For example, the operation sequence for the Ring and Support Part of the Pump Assembly requires weighing, cleaning, and boring, followed by a lathe operation, as depicted in Fig. 2. An operation may begin only when all its preceding operations have been completed. Furthermore, the capacity of each machine type is finite, and may be time varying. Hence, the problem of scheduling products with bills of materials entails satisfying customer requests, i.e., satisfying due dates for products, and reducing work-in-process inventory (WIP), subject to the precedence constraints among parts and operations, machine capacity constraints, and processing time requirements. The primary purpose for studying this problem is to improve the ability of a manufacturer to meet promised delivery dates for products and to reduce work-in-process inventory, thereby enhancing its credibility and competitiveness in the marketplace.

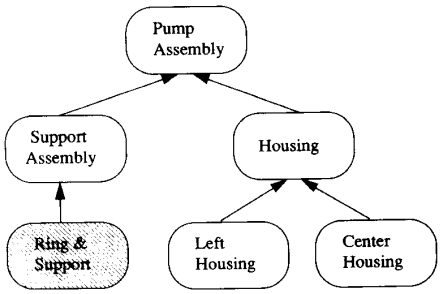

Fig. 1. A sample bill of materials for a pump assembly.

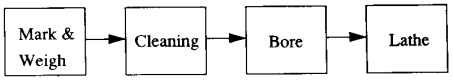

Fig. 2. The ring and support part in Fig. 1.

It is unlikely that a *practical* approach to this scheduling problem can yield an optimal solution. Most existing scheduling methodologies that give optimal solutions apply only to problems of very small size; i.e., a few parts to be scheduled on a few machines. This is true since a wide variety of machine scheduling problems are NP-hard, that is, the computational requirement grows exponentially as a function of the problem size (Lenstra, *et al.*, [8]). The job shop scheduling problem is one such example. It is reasonable to assume that the problem of scheduling products is NP-hard as well. Since a problem of practical size may involve hundreds of machines and hundreds of products, each comprised of thousands of parts to be scheduled over a period varying from several months to a few years, any method generating an optimal

Manuscript received November 3, 1992; revised May 30, 1993. This work was supported in part by the National Science Foundation under Grant DDM-9119074 and the Department of Higher Education, State of Connecticut and Pratt & Whitney under the Cooperative High Technology Research and Development Grant Program.

The authors are with the Department of Electrical and Systems Engineering, the University of Connecticut, Storrs, CT 06269-3157.

IEEE Log Number 9214694.

schedule will likely require excessive computation time and is not practical for use on a daily basis.

In order to cope with such computational complexity, practitioners often employ Materials Requirements Planning (MRP) in combination with heuristic rules to schedule individual work centers [3]. Typically, an MRP system is used to schedule parts so that precedence constraints among parts are satisfied. MRP ignores machine capacity, however, and may cause the schedule to violate capacity constraints [4]. In order to satisfy capacity constraints, simple heuristics (e.g., dispatching rules) are often used to determine sequencing of parts at individual work centers [11], [16]. Unfortunately, the heuristic rules usually are not well integrated with the MRP system, i.e., parts are sequenced according to rules that usually do not account for the relations among parts within each product as established by the MRP system. As a consequence, *parts are delayed during production*, and the product completion time becomes uncertain as the delay of only one of its parts may result in the delay of the entire product.

This Paper addresses the issue of integrally scheduling parts that are related through bills of materials by using the Lagrangian relaxation technique. Since the problem size may be quite large and the complexity grows exponentially with the problem size, the goal of scheduling is *not* to obtain the optimal solution. Rather, it is sufficient to obtain *near-optimal* schedules with *quantifiable* performance and within *reasonable* computation time. The method developed here is an extension of previous work on the scheduling of job shops [6]. The current problem, however, is much more complicated since the constraints imposed by the bill of materials increases the problem size dramatically. Accordingly, the Lagrangian relaxation technique used in Hoitomt, *et al.* [6] is modified to account for increased problem complexity.

B. The Lagrangian Relaxation Technique

Lagrangian relaxation is a mathematical programming technique for solving constrained optimization problems, and has recently emerged as an important method for solving complex scheduling problems. The technique has been used to obtain near-optimal solutions for the scheduling of single-operation, independent parts on parallel, identical machines ([10]; see also Hoitomt, *et al.* [5], for the multi-operation case with *simple* precedence constraints). There, the objective is to minimize weighted part tardiness. The "coupling" machine capacity constraints are relaxed using Lagrange multipliers, yielding a number of smaller, decomposed subproblems which are easier to solve [13]. The Lagrangian dual function is also formed, which is to be maximized with respect to the multipliers. Since the dual function is nondifferentiable, a subgradient method is employed to iteratively update the multipliers. In this case, the multipliers can be interpreted as prices that regulate the use of a limited resource over time. As the dual function is maximized, the decision variables (operation beginning times) obtained from solving the subproblems tend to an optimal feasible solution, while the dual function itself provides a lower bound on the optimal cost. Since a solution is needed within a limited amount of time, however, the subgradient method is often terminated before a feasible solution is obtained. Consequently, a list scheduling heuristic is used to generate a feasible solution based on the infeasible solution obtained from the dual problem. This technique resulted in an algorithm which generates near-optimal, feasible schedules with quantifiable performance in reasonable computation time.

The Lagrangian relaxation technique has been extended to solve the *job shop scheduling problem* in which *multi-operation, independent jobs (or parts)* with *general precedence constraints among operations* are to be scheduled on nonidentical machines [5]. Using the same objective of minimizing weighted part tardiness, Lagrange multipliers are used to relax *both* resource capacity constraints *and* operation precedence constraints. The multipliers relaxing the latter set of constraints act as prices to discourage two precedence constrained operations from overlapping. Unfortunately, the resulting decomposed subproblems exhibit an undesirable property. If the cost of overlap with succeeding operations is greater than the cost of overlap with preceding operations, an operation tends to be scheduled very early. Conversely, if the cost of overlap with preceding operations is greater than the cost of overlap with succeeding operations, the operation tends to be scheduled very late. This causes subproblem solutions to oscillate from iteration to iteration and, in many cases, prevents convergence of the algorithm. This *solution oscillation* can alternately be explained by the geometry of the dual function, in which multiplier trajectories generated by the subgradient method zigzag about a set of nondifferentiable points from iteration to iteration [15] (this is explained in more detail in Section II). This solution oscillation has been alleviated to some extent in recent work by using an *augmented Lagrangian approach* in combination with a *Gauss-Seidel iterative technique* to enforce separability of subproblems [5]. A consequence of employing Gauss-Seidel iterations, however, is to increase the computation time of the algorithm. Furthermore, a lower bound on the optimal cost cannot be obtained with the augmented approach. In order to generate a lower bound, Hoitomt adopts a second problem formulation (termed the *evaluation phase*), further increasing the computation time of the algorithm. Since the bill of materials introduces an additional layer of precedence constraints among parts in the scheduling problem, a direct application of the augmented Lagrangian approach to the product scheduling problem will require excessive computational effort.

C. Overview of the Paper

This paper first presents an improvement on the previous work concerning the scheduling of the *job shop* [5], where independent parts have multiple operations to be processed on non-identical machines. The intended objective is to minimize weighted part tardiness and earliness penalties, since on-time delivery and reduced work-in-process inventory are of upmost importance. In order to reduce the aforementioned oscillation, however, an *auxiliary problem formulation* is developed in Section II in which quadratic tardiness and earliness penalties *for operations* are included in the objective function. These

quadratic terms are intended to define a period of time in the time horizon in which operations are scheduled with zero penalty. When solving the subproblems, this zero-cost period is balanced against the cost of the multipliers, and tends to reduce the magnitude of oscillation. In the dual space, these quadratic terms modify the geometry of the dual function and this, in turn, reduces oscillation between solutions. The primary advantage of this approach is a substantial savings in computation time over the augmented Lagrangian approach, as demonstrated in Section II using actual factory data. This is partially due to the modified geometry of the dual function and subsequent reduction in oscillation. Time savings are also obtained from the fact that this problem formulation allows the relaxed problem to be directly decomposed into subproblems. This is in contrast to the augmented Lagrangian approach, which requires time-consuming Gauss-Seidel iterations to enforce separability of subproblems. Additional time savings are obtained by using a *modified subgradient method* to solve the dual problem, in contrast to the method of Hoitomt [6], which uses a standard subgradient method. The additional time savings is the result of a smaller angle between the search directions and the directions toward the maximum. Results in Section II also suggest that although the auxiliary objective function includes the quadratic penalty terms for operations, actual weighted part tardiness is also reduced, often surpassing the performance of the augmented Lagrangian approach. As an added benefit, a lower bound with respect to the auxiliary objective is obtained directly, thus providing an indirect measure of schedule quality.

Having demonstrated that the auxiliary problem approach requires less computational effort than the augmented Lagrangian approach, Section III considers the problem of scheduling products with bills of materials using the auxiliary problem approach. The issue of integrally scheduling parts that are related through bills of materials is addressed by including *part precedence constraints* in the problem formulation. Similar to the job shop problem, the main objectives are to minimize weighted product tardiness and earliness. In addition, an auxiliary objective function containing quadratic tardiness and earliness penalties for parts and operations is used to modify the geometry of the dual function and subsequently reduce solution oscillation. Results to date show that the penalties in the auxiliary function reduce oscillation and can obtain effective schedules. It is also demonstrated that integration of the bills of materials into the product scheduling problem can reduce product tardiness, compared to algorithms in which the part precedence constraints are disregarded during the optimization process. The improved ability to meet product due dates by the above integration will enhance a manufacturer's credibility and competitiveness in the marketplace.

II. The Improved Lagrangian Relaxation Technique

This section presents the improved approach to the *job shop scheduling problem*. The problem formulation is given in Section II-A, and the Lagrangian relaxation technique is presented in Section II-B. Section II-B is divided into four subsections as follows: The decomposition of the relaxed problem into smaller subproblems is given in Subsection II-B1. Subsection II-B2 describes the effect of the quadratic penalty terms for operations on the subproblems and why solution oscillation is reduced. This is followed in Subsection II-B3 by a detailed discussion on the geometry of the dual function, how the quadratic terms modify the geometry, and why the new geometry reduces solution oscillation. Subsection II-B4 shows how the modified subgradient method further capitalizes on dual function geometry, resulting in improved computation time performance. Finally, the quality of the schedules, as well as the computation time of the algorithm, are compared to the augmented Lagrangian approach in Section II-C.

A. Problem Formulation

Assume that there are N parts to be processed and part i, $1 \leq i \leq N$, requires a sequence of N_i operations where operation j of part i is denoted by (i,j). Each operation can be processed by a machine from the set of machine types H_{ij} that are capable of performing operation (i,j). The machine type that is selected to process operation (i,j) is denoted by $m_{ij} \in H_{ij}$. The objectives of scheduling are to ensure on-time part completion and low work-in-process inventory, modeled by weighted part tardiness and earliness penalties, respectively. In addition, tardiness and earliness penalties for operations are included in an auxiliary objective function to reduce oscillation. These penalties are to be minimized subject to operation precedence, machine capacity, and processing time constraints. The decision variables are beginning times and machine types selected for all operations. Each of the penalties in the auxiliary objective function is described in more detail below.

1) The Auxiliary Objective Function:

a) Part tardiness tenalty: The tardiness for part i, T_i, is defined as the amount the part completion time C_i passes the part due date D_i, i.e., $\max[0, C_i - D_i]$. The part tardiness penalty is then the weight W_i times the square of the part tardiness T_i. This penalty accounts for the value of the part, the importance of meeting the due date, and the fact that the penalty becomes more severe with each time unit past the due date.

b) Part earliness penalty: Low work-in-process inventory, as emphasized in the Just-in-Time (JIT) manufacturing concept, is desirable since it reduces holding costs. Low work-in-process is captured by penalizing the starting or releasing of part i earlier than necessary. Given the part due date D_i, an early start date s_i can be roughly estimated based on the *critical path* of the part,[1] i.e.,

$$s_i \equiv D_i - \gamma \sum_{j=1}^{N_i} t_{ij}, \quad 1 \leq i \leq N; \ \gamma > 1, \tag{1}$$

where t_{ij} is the processing time of operation (i,j) along the critical path, and γ is a coefficient greater than 1. Let the part earliness E_i be defined as the amount the beginning time B_i leads the part start date s_i, i.e., $\max[0, s_i - B_i]$. The part

[1] The critical path of a part is defined as the sequence of operations in the part having the greatest cumulative elapsed time.

earliness penalty can then be defined as the weight β_i times the square of the part earliness E_i.

For future reference, let the part tardiness and earliness penalties defined above be referred to as the *original objective function* as used in [5], i.e.,

$$J \equiv \sum_{i=1}^{N} (W_i T_i^2 + \beta_i E_i^2). \tag{2}$$

c) Tardiness and earliness penalties for operations: As mentioned, quadratic penalties for operations are included in the auxiliary objective function to reduce solution oscillation. These penalties are modeled as tardiness and earliness penalties for operations. The tardiness for operation j, T_{ij}, is defined as the amount the completion time c_{ij} passes the operation due date d_{ij}, i.e., max[0, $c_{ij} - d_{ij}$], and the operation earliness is defined as max[0, $s_{ij} - b_{ij}$] where $s_{ij} \equiv d_{ij} - \gamma t_{ij}, \gamma > 1$. These tardiness and earliness penalties together define a period of time in the time horizon in which an operation can be scheduled with zero penalty. The operation due dates d_{ij} can be selected, for example, by performing backward scheduling from the part due date D_i to reflect the operation sequence. This results in non-overlapping zero-cost periods for each operation of a part, which tend to enforce the operation precedence constraints.

The auxiliary objective function to be minimized is comprised of the sum of all the penalty terms described above, i.e.,

$$J_{AUX} \equiv \sum_{ij} (\overline{W}_{ij} T_{ij}^2 + \overline{\beta}_{ij} E_{ij}^2),$$

$$1 \le i \le N;\ 1 \le j \le N_i,\ with$$

$$\overline{W}_{ij} \equiv w_{ij} + W_i \Delta_{iN_i},\ \overline{\beta}_{ij} \equiv \beta_{ij} + \beta_i \Delta_{i1}. \tag{3}$$

In (3) above, the fact that $b_{ij} = B_i$ for $j = 1$ and $c_{ij} = C_i$ for $j = N_i$ have been used. As such, Δ_{i1} is defined as an integer variable equal to one if operation (i,j) is the first operation in part i and zero otherwise, and Δ_{iN_i} is similarly defined for the last operation in part i.

The coefficients in (3) are selected based on testing experiences, with the following guidelines. The tardiness weights for operations and parts are chosen to satisfy $w_{ij} \ll W_i$ since minimizing part tardiness is the foremost criterion (i.e., objective function (2) is of significant importance). The earliness coefficient β_i is inversely proportional to W_i since a part that has a high tardiness priority should be allowed to start early to meet its due date. The operation earliness coefficients satisfy $\beta_{ij} \ll \beta_i$, similar to the relationships among the tardiness weights. Lastly, the parameter $\gamma(\ge 1)$ is roughly proportional to the desired WIP level in the schedule and usually chosen to be relatively small.

2) Constraints The minimization of (3) is subject to three constraints: operation precedence constraints, machine capacity constraints, and processing time requirements. Each constraint is described below.

a) Operation precedence constraints: Let I_{ij} denote the set of operations in part i immediately following operation (i,j) in the operation processing sequence. The operation precedence constraints require the beginning times of the set of operations in I_{ij} to be greater than or equal to the completion time of operation (i,j) plus any required timeout S_{ijl} between operations (i,j) and (i,l), $l \in I_{ij}$, i.e.,

$$c_{ij} + S_{ijl} + 1 \le b_{il},$$

$$1 \le i \le N;\ 1 \le j \le N_i - 1;\ l \in I_{ij}. \tag{4}$$

Here, a beginning time refers to the start of a discrete time slot. In contrast, a completion time refers to the *end* of a discrete time slot, requiring the constraint in (4) to account for one time unit.

b) Machine capacity constraints: The machine capacity constraint requires the total number of operations active on machine type h at time k to be less than or equal to the number of type h machines available at time k, i.e.,

$$\sum_{ij} \delta_{ijkh} \le M_{kh},\ 1 \le i \le N;\ 1 \le j \le N_i;$$

$$1 \le k \le K;\ 1 \le h \le H, \tag{5}$$

where δ_{ijkh} is an integer variable equal to one if operation (i,j) is active on machine h at time k and zero otherwise, M_{kh} is the number of type h machines available at time k, K is the time horizon, and H is the number of machine types.

c) Processing time requirements: The processing time requirement for operation (i,j) states that the elapsed difference between the beginning time b_{ij} and the completion time c_{ij} should be $t_{ijm_{ij}}$, the required processing time for (i,j) on machine type m_{ij} selected for the operation, i.e.,

$$c_{ij} = b_{ij} + t_{ijm_{ij}} - 1,$$

$$1 \le i \le N;\ 1 \le j \le N_i. \tag{6}$$

Among the above variables, the precedence structure, part due dates, processing times, timeouts, all the penalty coefficients, and the number of machines per type available as a function of time are given. The decision variables are the beginning times of all operations, $\{b_{ij}\}$, and the machine type selected to process each operation, $\{m_{ij} \in H_{ij}\}$. Once the decision variables are selected, completion times, number of active operations at a given time, and part and operation tardiness and earliness can be easily derived. Note that the auxiliary objective function in (3) is operation-wise additive and the constraints are linear, but the machine capacity and precedence constraints couple across operations to make the problem intractable.

B. Solution Methodology

The complexity of the scheduling problem motivates a decomposition approach. An augmented Lagrangian relaxation approach has been used in Hoitomt [5] to achieve a decomposition of the job shop scheduling problem. In this section, Lagrangian relaxation is applied to the auxiliary problem formulation.

1) The Lagrangian Relaxation Approach: The operation precedence and capacity constraints in (4) and (5) are relaxed by using the Lagrange multipliers η_{ijl} and π_{kh}, respectively. The relaxed problem R is:

$$R: L \equiv \min_{\{b_{ij}\},\{m_{ij}\in H_{ij}\}} [\sum_{ij}(\overline{W}_{ij}T_{ij}^2 + \overline{\beta}_{ij}E_{ij}^2) + \sum_{kh}\pi_{kh}(\sum_{ij}\delta_{ijkh} - M_{kh}) + \sum_{ij,l\in I_{ij}}\eta_{ijl}(b_{ij} + t_{ijm_{ij}} + S_{ijl} - b_{il})], \quad (7)$$

subject to the processing time requirements (6). Note that this directly results in a minimization subproblem for each operation (i,j) with π and η given:

$$R_{ij}: L_{ij} \equiv \min_{b_{ij},m_{ij}\in H_{ij}} [\overline{W}_{ij}T_{ij}^2 + \overline{\beta}_{ij}E_{ij}^2 + \sum_{k=b_{ij}}^{c_{ij}}\pi_{km_{ij}} + \sum_{l\in I_{ij}}\eta_{ijl}(b_{ij} + t_{ijm_{ij}}) - \sum_{l:j\in I_{il}}\eta_{ilj}b_{ij}]. \quad (8)$$

In the above subproblem, the fact that $\delta_{ijkh} = 0$ for all $h \neq m_{ij}$ has been used. The dual problem is formed by maximizing R with respect to the multipliers:

$$D: \max_{\pi,\eta\geq 0} L(\pi,\eta), \ with \ L(\pi,\eta) = [-\sum_{kh}\pi_{kh}M_{kh} + \sum_{ij,l\in I_{ij}}\eta_{ijl}S_{ijl} + \sum_{ij}L_{ij}]. \quad (9)$$

2) Reducing Oscillation in the Subproblems: As mentioned, tardiness and earliness penalties for operations are included in the auxiliary objective function to reduce solution oscillation. To see why, consider an operation of part i constrained by preceding and succeeding operations, i.e., $1 < j < N_i$. For simplicity, assume $|H_{ij}| = 1$ and there are infinitenumber of such machines resulting in $\pi_{kh} = 0, \forall k, \forall h$, where $|H_{ij}|$ is the cardinality of set H_{ij}. The subproblem R_{ij} then becomes:

$$R_{ij}: L_{ij} \equiv \min_{b_{ij}} [w_{ij}T_{ij}^2 + \beta_{ij}E_{ij}^2 + (\sum_{l\in I_{ij}}\eta_{ijl} - \sum_{l:j\in I_{il}}\eta_{ilj})b_{ij} + \sum_{l\forall X_{ij}}\eta_{ijl}t_{ij}]. \quad (10)$$

Now suppose that the tardiness and earliness penalties are not included in the objective function, i.e., $w_{ij} = 0$ and $\beta_{ij} = 0$. A consequence of the removal of these terms is that the selection of b_{ij} depends heavily on the sign of

$$\sum_{l\in I_{ij}}\eta_{ijl} - \sum_{l:j\in I_{il}}\eta_{ilj}.$$

When this term is negative, the resulting b_{ij} would be as large as possible ($b_{ij} = K - t_{ij} + 1$). Similarly, when this term is positive, b_{ij} would be as small as possible ($b_{ij} = 1$). Neither result is acceptable since a small beginning time overlaps with preceding operations, and a large beginning time overlaps with succeeding operations. When solving the dual problem, (described in Section II-B4), the multipliers $\{\eta_{ijl}\}$ are updated to try to enforce the constraints. As such, if b_{ij} is large, $\sum_{l\in I_{ij}}\eta_{ijl}$ will increase from iteration to iteration until, finally,

$$\sum_{l\in I_{ij}}\eta_{ijl} - \sum_{l:j\in I_{il}}\eta_{ilj}$$

becomes positive and results in b_{ij} being chosen small. Similarly, if b_{ij} is small, $\sum_{l:j\in I_{il}}\eta_{ilj}$ will increase from iteration to iteration until,

$$\sum_{l\in I_{ij}}\eta_{ijl} - \sum_{l:j\in I_{il}}\eta_{ilj}$$

becomes negative, resulting in b_{ij} being chosen large. This solution oscillation between small and large beginning times from iteration to iteration therefore leads to oscillations in multiplier values, making convergence to any meaningful multipliers difficult. In contrast, when quadratic terms are included in the objective function, i.e., $w_{ij} > 0$ and $\beta_{ij} > 0$, intermediate beginning time solutions are possible, i.e., $(1 \leq b_{ij} \leq K - t_{ij} + 1)$. As a consequence, the oscillation between small and large beginning times is reduced.

Hoitomt [5] attempts to alleviate the oscillation problem by adopting an augmented Lagrangian in which the quadratic penalty terms

$$\sum_{ij,l\in I_{ij}}\frac{\rho_{ijl}}{2}(b_{ij} + t_{ijm_{ij}} + s_{ijl} - b_{il})^2, s_{ijl} \geq S_{ijl},$$

are appended to the Lagrangian in equation (7). The disadvantage of this approach, however, is that these quadratic penalties couple beginning times together, and prevent the problem from being directly decomposed into operation-level subproblems. Therefore, Hoitomt adopts a Gauss-Seidel iterative technique to enforce separability, where an intermediate loop is added between the high level multiplier updates and the low level iterations to find the optimal operation beginning times. One Gauss-Seidel iteration entails solving all the operation subproblems, in order from the first to the last operation, for each part. In solving a particular subproblem relating to operation (i,j), the latest available $\{b_{il}\}$ are used, and are treated as constant. Because the subproblems are solved in the above order, the beginning times of all preceding $(l < j)$ operations have been solved in the nth Gauss-Seidel iteration, while the beginning times of all succeeding $(l > j)$ operations are taken from the $n-1$ iteration. If the solutions to the nth Gauss-Seidel iteration are the same as the $n-1$ iteration, the Gauss-Seidel is said to have converged. Otherwise, the operation beginning times obtained are used to start the $n+1$ Gauss-Seidel iteration. If convergence is not obtained after a fixed number of iterations, say, N, the solution generating the smallest cost is used.

A consequence of having Gauss-Seidel iterations is to increase the computation time of the algorithm. Specifically, the complexity of solving R_{ij} in equation (8) is of order $K * |H_{ij}|$, since solving subproblem R_{ij} entails enumerating all possible beginning times for each possible machine type $m_{ij} \in H_{ij}$, and comparing the values of L_{ij} derived from each possibility. In contrast, the Gauss-Seidel approach may require solving R_{ij} N times in the worst case, therefore increasing the complexity to $N * K * |H_{ij}|$, where $N \geq 1$, N integer.

3) Geometry of the Dual Function: The dual function $L(\pi, \eta)$ in equation (9) is polyhedral concave and thus nondifferentiable. As such, the subgradient method is used to maximize the dual function (described in Section II-B4). The subgradient method, however, may converge very slowly. To illustrate, consider a simple example where one part with three consecutive operations, each requiring a processing time $t_{1,1} = t_{1,2} = t_{1,3} = 5$, is to be scheduled on three identical machines. In this case, $\pi_{kh} = 0, \forall k, \forall h$, and $|H_{ij}| = 1$. The due dates for the operations are $d_{1,1} = 5$, $d_{1,2} = 10$, and $d_{1,3} = 15$. Also, $W_1 = 5$ and $\beta_1 = 5$, but the quadratic penalties for operations are omitted, i.e., $w_{1j} = 0$, $\beta_{1j} = 0$. The dual function to be maximized for this case,

$$L(\eta_{1,1,2}, \eta_{1,2,3}) = \min_{b_{1,1}, b_{1,2}, b_{1,3}} [W_1 T_{1,3}^2 + \beta_1 E_{1,1}^2 + \eta_{1,1,2}(b_{1,1} + t_{1,1} - b_{1,2}) + \eta_{1,2,3}(b_{1,2} + t_{1,2} - b_{1,3})], \quad (11)$$

is three-dimensional and is plotted in Fig. 3(a). Notice that this function is described by two *facets* (or hyperplanes), each corresponding to a particular set of solutions to the three subproblems. These two facets intersect in a line, which is an *edge* of the function, or a set of nondifferentiable points. The existence of only two facets is caused by the fact that the optimal beginning time of the second subproblem, ($R_{1,2}$), can take only two values, i.e., minimum ($b_{1,2} = 1$) or maximum ($b_{1,2} = K - t_{1,2} + 1$). Since the dual function is to be maximized, the optimal solution for this problem is located at $\eta_{1,1,2} = \eta_{1,2,3} = 0$. Now, when the subgradient method is used to maximize the dual function from an arbitrary starting point, the multipliers are iteratively updated and form a trajectory. In Fig. 3(b), a trajectory is overlayed onto a contour plot of the dual function, where the starting point is chosen to be $\eta_{1,1,2} = 4.5$, $\eta_{1,2,3} = 3.5$. Notice that the trajectory zigzags across the edge, and that the multipliers alternate between the first facet and the second. This results in slow convergence to the optimum[2] Furthermore, the zigzagging from one facet to the other corresponds directly to the oscillation between small and large beginning times described earlier in Section II-B2.

In the auxiliary problem formulation, tardiness and earliness quadratics for operations are included in the objective function, i.e., $w_{1j} > 0$, $\beta_{1j} > 0$. The effect of these terms is to modify the geometry of the dual function such that there are numerously more facets and edges. As can be seen from equation (10), the optimal beginning times of the second

[2]For a more mathematically rigorous discussion, refer to Tomastik, *et al.*, 1993.

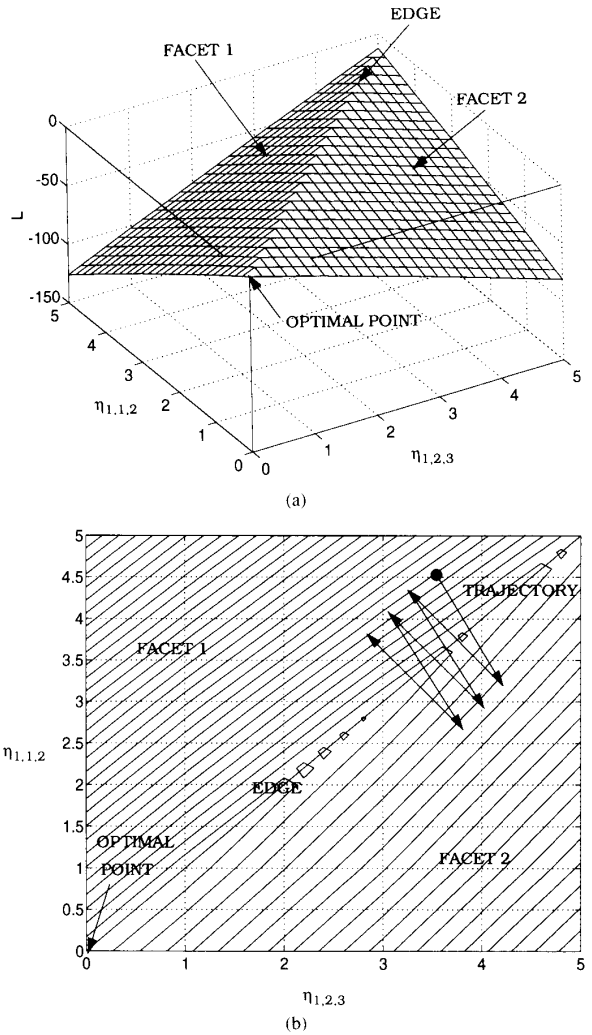

Fig. 3. (a) Dual function. (b) Example of oscillating trajectories in the dual space.

subproblem ($R_{1,2}$) can now take many discrete values as the magnitudes of $\eta_{1,1,2}$ and $\eta_{1,2,3}$ change. Each value corresponds to a particular facet in the dual space. Therefore there are numerously more facets and edges. The modified geometry for the example is depicted in Fig. 4(a), and a multiplier trajectory with the same initial point as before is overlayed onto a contour plot of the dual function in Fig. 4(b). Notice that this trajectory approaches the optimum more directly and does not zigzag between facets. This modified geometry more closely resembles the geometry of a differentiable function, and subgradient type algorithms could perform reasonably well. The solution oscillation difficulty is therefore reduced.

4) Solving the Dual Problem: Although the oscillation is reduced by using an auxiliary objective function, it may not be completely eliminated by this technique. Accordingly, a *modified subgradient method* is used to update the multipliers. The method, presented by Camerini, *et al.* [1], consists of computing the current search direction as a linear combination

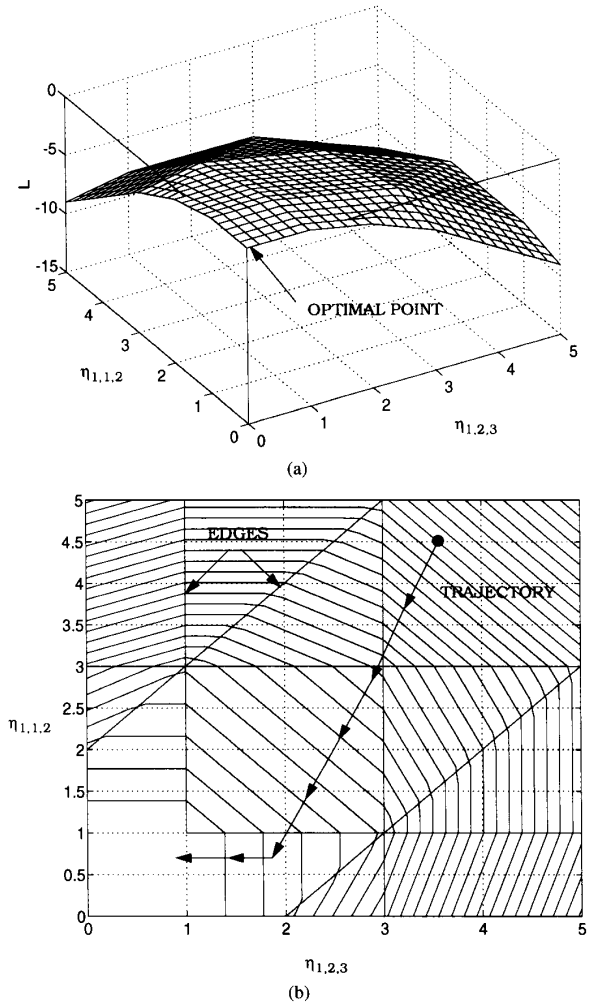

Fig. 4. (a) Modified geometry of the dual function. (b) New trajectory.

of the current subgradient and the direction used at the previous step, similar to the conjugate gradient method [9]. When an edge is encountered, this search direction forms a smaller angle with the direction towards the maximum than does the direction of the subgradient, thus enhancing the speed of convergence [14], [7]. The Lagrange multiplier π is updated by

$$\pi^{n+1} = \pi^n + \alpha^n s^n, \tag{12}$$

where α^n is the current stepsize and s^n is the current search direction. The stepsize α^n is given by

$$\alpha^n = \beta \frac{L^U - L^n}{(s^n)^T (s^n)}, \quad 0 < \beta \leq 1, \tag{13}$$

where L^U is an upper estimate of the optimal solution of (9) and L^n is the value of L at the nth iteration. The current search direction s^n is given by

$$\begin{aligned} s^0 &= g(\pi^0), \\ s^n &= g(\pi^n) + \gamma^n s^{n-1}, \quad n = 1, 2, \ldots, \end{aligned} \tag{14}$$

where $g(\pi^n)$ is the current subgradient of L, γ^n is given by:

$$\gamma^n = \max[0, -\epsilon^n \frac{(s^{n-1})^T (g(\pi^n))}{(s^{n-1})^T (s^{n-1})}],$$

$$0 \leq \epsilon \leq 2, \tag{15}$$

and s^{n-1} is the direction applied at the previous iteration. The subgradient component of the hth resource at the kth time is equal to

$$\sum_{ij} \delta_{ijkh} - M_{hk}.$$

Note that if $\gamma^n = 0$ in (14), (12)–(14) describe the subgradient method used in Hoitomt [6]. Equations (14) and (15) describe a modified subgradient approach only when $\gamma^n > 0$. This method converges at the rate of geometric progression (linear convergence rate). The other Lagrange multiplier η is updated in a similar fashion.

The cost obtained by solving the dual problem (9) is a lower bound on the optimal cost of (3), thus providing a measure of suboptimality of the schedule. In contrast, the augmented Lagrangian approach of Hoitomt [6] cannot provide a lower bound, since the Gauss-Seidel iterations may not converge to an optimal solution. In order to generate a lower bound, Hoitomt adopts a second problem formulation (termed the *evaluation phase*). The disadvantage of this approach, however, is to increase the computation time of the algorithm.

Since the solution to the dual problem at the termination of the subgradient algorithm is generally associated with an infeasible schedule in which a few constraints are violated, a feasible schedule is constructed using the same heuristic approach of Hoitomt [6]. The next section compares the quality of the schedules generated by the two approaches, as well as the computation time required by the algorithms.

C. Test Results

In the following test cases, the coefficients in the auxiliary objective function are assigned according to the comments in Section II-A1 and testing experiences. As such, the operation tardiness coefficients are initialized so that $10w_{ij} \equiv W_i$. The part earliness coefficients are chosen so that $10\beta_i \equiv \frac{1}{W_i}$, while the operation earliness coefficients satisfy $10\beta_{ij} \equiv \beta_i$. In each of the test cases, all multipliers are initialized to zero.

Example 1: The first example is taken from Hoitomt [5] and consists of three different machines and four equally weighted parts ($W_i = 10$). The planning horizon is 30 days (i.e., $K = 30$, and a time unit is one day). All machines are available on day one and throughout the planning horizon. All operations are available for processing on day one, but are due on day 0 or the current day. Each part is comprised of three operations, each to be processed on one of three machines without timeouts or mandatory slack time between operations. Data are shown in Table I.

A fixed number of iterations is used as a stopping criteria. For this example, 250 iterations is chosen as an acceptable

TABLE I
DATA FOR EXAMPLE 1.

Part i	Operation j	Mach. H_{ij}*	t_{ij}	I_{ij}
1	1	1	4	2
	2	2	3	3
	3	3	2	-
2	1	2	1	2
	2	1	4	3
	3	3	4	-
3	1	3	3	2
	2	2	2	3
	3	1	3	-
4	1	2	3	2
	2	3	3	3
	3	1	1	-

*The notation "i,j" refers to operation j of part i.

TABLE II
GANTT CHART OF THE SCHEDULE FOR EXAMPLE 1

Time	1	2	3	4	5	6	7
Mach. 1	1,1	1,1	1,1	1,1	2,2	2,2	2,2
Mach. 2	2,1	4,1	4,1	4,1	1,2	1,2	1,2
Mach. 3	3,1	3,1	3,1		4,2	4,2	4,2
Time	8	9	10	11	12	13	14
Mach. 1	2,2	4,3	3,3	3,3	3,3		
Mach. 2	3,2	3,2					
Mach. 3	1,3	1,3	2,3	2,3	2,3	2,3	

tradeoff between computation time and duality gap. The feasible schedule has a cost $J_{AUX} = 5,345$, where J_{AUX} is the auxiliary objective function in (3), while the lower bound on J_{AUX} is obtained as 5261, a relative duality gap of 1.6%. Note that the weighted part tardiness and earliness cost (equation (2)) of the feasible schedule is $J = 4750$. The resulting schedule is shown in the form of a Gantt chart in Table II. The time to solve the problem is 3.2 CPU seconds on a SPARCstation2 computer. This is summarized in column 1 of Table III.

The problem is now solved using the augmented Lagrangian approach of Hoitomt [6]. For a fair comparison between the two algorithms, 250 iterations are again used, 245 iterations during the optimization phase and 5 iterations during the evaluation phase to obtain a lower bound (see Section II-B4). The feasible schedule has cost $J_{GS} = 4750$, where J_{GS} is weighted part tardiness and earliness cost as in (2). The lower bound on J_{GS} is obtained as 4609, a relative duality gap of 3.1%. According to Hoitomt [5], this schedule is really optimal with respect to the objective J_{GS}. The time to solve the problem is 18.2 CPU seconds on the same computer. See column 2 of Table III.

Note that the weighted part tardiness and earliness cost of the feasible schedule, $J = 4750$, is identical to J_{GS}. This suggests that although J_{AUX} in equation (3) is the objective being minimized, improvements in J are obtained. This observation, coupled with the fact that the computation time of the first algorithm is substantially less, makes the auxiliary problem approach a competitive alternative to the augmented Lagrangian approach.

TABLE III
RESULTS FOR EXAMPLE 1

	Auxiliary objective	Gauss-Seidel
J_{AUX}	5345	-
J, J_{GS}	4750	4750
Lower bound	5261	4609
Duality gap	1.6%	3.1%
CPU seconds	3.2	18.2
Iterations	250	245 & 5

TABLE IV
RESULTS FOR EXAMPLE 2

	Auxiliary objective	Gauss-Seidel
J_{AUX}	1 648 600	-
J, J_{GS}	1 476 500	1 487 600
Lower bound	1 621 000	1 440 100
Duality gap	1.7%	3.4%
CPU seconds	39.8	155.5
Iterations	100	99 & 1

Example 2 This example draws data from a testbed at Pratt & Whitney's Development Operations shop. The data comes from about twenty work centers containing numerical control (NC) machines. There are a total of 29 different machines and 140 independent parts, each consisting of 1 to 7 operations, for a total of 187 operations. The parts are characterized by different due dates and may have one of five different weights. The planning horizon is 214 days. In this example, the number of iterations is reduced from 250 to 100 since the problem size is larger and requires a greater computation time. The feasible schedule has a cost $J_{AUX} = 1\,648\,600$, and the lower bound obtained is 1 621 000 with a relative duality gap of 1.7%. The CPU time on the same computer as in Example 1 is 39.8 CPU seconds. This is summarized in column 1 of Table IV.

The problem is now solved using the augmented Lagrangian approach of Hoitomt [5]. For a fair comparison between the two algorithms, 100 iterations are again used, 99 iterations during the optimization phase and 1 iteration during the evaluation phase to obtain a lower bound (see Section II-B4). The feasible schedule has a cost $J_{GS} = 1\,487\,600$, while the lower bound on J_{GS} is obtained as 1 440 100, a relative duality gap of 3.4%. The time to solve the problem is 155.5 CPU seconds on the same computer. See column 2 of Table IV.

Note that the cost of the schedule obtained with the improved approach, when evaluated using (2), is $J = 1\,476\,500$, which is *less than* J_{GS}. Here, the improved approach has found a superior schedule with respect to J in less than one fifth the CPU time. This again suggests that minimizing J_{AUX} in (3) can obtain significant improvements in J. Furthermore, the computation time of the improved algorithm is substantially less than the augmented Lagrangian approach.

To examine the effect of solution oscillation on algorithm performance, the operation tardiness and earliness penalties are removed from objective function (3). This is equivalent to setting $w_{ij} = 0$ and $\beta_{ij} = 0$. Using the same number of iterations as above, the feasible cost without the penalties is $J_{AUX} = J = 1\,489\,100$. This cost is higher than the cost obtained when the penalties are present ($J = 1\,476\,500$). This is the general trend in all test cases and indicates that the operation penalties reduce solution oscillation, enabling a lower cost J to be found.

Having demonstrated that the auxiliary problem approach requires less computational effort than the augmented Lagrangian approach while maintaining schedule quality, it is applied in the next section to the problem of scheduling products with bills of materials.

III. Scheduling Products with Bills of Materials

A. Introduction

As mentioned in Section I, a bill of materials is used specify the order in which parts are to be processed or assembled. Part precedence constraints imposed by the bill of materials are difficult to account for in a scheduling methodology since they increase the problem size dramatically. In order to cope with such complexity, practitioners often use MRP systems as described in Section I, or attempt to divide the task into independent subproblems for each part which are easier to solve. In the latter approach, parts are assigned due dates to reflect the structure of the bill of materials and are then scheduled in an *independent manner* so that capacity constraints are satisfied [12]. However, decomposing the product scheduling problem into *independent* part problems through the assignment of due dates obscures the relationships among parts. For example, a product may be delayed because of the unforeseen delay of a single component part, whereas other component parts will be rushed to meet their due dates, irrespective of the status of the product as a whole. Hence, independent subproblems for each part inhibits the responsiveness of schedules to unforeseen changes.

This section addresses the issue of integrally scheduling parts that are related through bills of materials by including *part precedence constraints* directly in the problem formulation. Similar to the job shop problem of Section II, an auxiliary objective function containing quadratic penalties is used to reduce solution oscillation. By comparing this method to another algorithm in which parts are scheduled independently, it is shown that integration of the bills of materials into the part scheduling problem can improve scheduling performance by reducing product tardiness.

B. Problem Formulation

Assume that there are P products to be processed and product p, $1 \leq p \leq P$, contains N_p parts, where part i of product p is denoted by (p,i). Without loss of generality, it is assumed that the manufacturing process of the product begins and terminates with a single part, (p,1) and (p,N_p), respectively. As in Section II-A, each part requires a sequence of N_{pi} operations where operation j of part (p,i), denoted (p,i,j), can be processed by a machine from the set of machine types H_{pij}.

Similar to the job shop problem, the primary objectives in the product scheduling problem are to ensure on-time *product* completion and low work-in-process inventory, modeled by weighted *product* tardiness and earliness penalties, respectively. In addition, tardiness and earliness penalties for operations *as well as parts* are included in the auxiliary objective function to reduce solution oscillation. These penalties are to be minimized subject to the constraints in Section II-A2, in addition to the *part precedence constraints* imposed by the bill of materials. The decision variables are beginning times and machine types selected for all operations, as in the job shop problem. Each of the penalties in the auxiliary objective function is briefly described below.

1) The Auxiliary Objective Function:

a) Product tardiness penalty: Similar to the part tardiness penalty in Section II-A1, the tardiness for product p, T_p, is defined as the amount the completion time C_p passes the product due date D_p, i.e., max[0, $C_p - D_p$]. The product tardiness penalty is then the weight W_p times the square of the product tardiness T_p.

b) Product earliness penalty: Also similar to Section II-A1, low work-in-process is captured by penalizing starting product p too early. Given the product due date D_p, an early start date s_p can be roughly estimated based on the critical path of the product; i.e.,

$$s_p \equiv D_p - \gamma \sum_{ij} t_{pij},\ 1 \leq p \leq P;\ \gamma > 1, \tag{16}$$

where t_{pij} is the processing time of operation (p,i,j) along the critical path, and γ is a coefficient greater than 1. Let the product earliness E_p be defined as the amount the beginning time B_p leads the product start date s_p, i.e., max[0, $s_p - B_p$]. The product earliness penalty can then be defined as the weight β_p times the square of the product earliness E_p.

For future reference, the product tardiness and earliness penalties defined above are referred to as the *original objective function*, i.e.,

$$J \equiv \sum_{p=1}^{N} (W_p T_p^2 + \beta_p E_p^2). \tag{17}$$

c) Tardiness and earliness penalties for parts and operations: As mentioned, tardiness and earliness penalties for *both parts and operations* are included in an auxiliary objective function to reduce solution oscillation. The tardiness for part (p,i), T_{pi}, is defined as the amount the completion time c_{pi} passes the part due date d_{pi}, i.e., max[0, $c_{pi} - d_{pi}$], and the part earliness is defined as max[0, $s_{pi} - b_{pi}$] where $s_{pi} \equiv d_{pi} - \gamma \sum_j t_{pij}, \gamma > 1$, similar to (1). The operation penalties are defined in a similar manner. These penalties are expressed as:

$$\sum_{pi} (w_{pi} T_{pi}^2 + \beta_{pi} E_{pi}^2) + \sum_{pij} (w_{pij} T_{pij}^2 + \beta_{pij} E_{pij}^2),$$

$$1 \leq p \leq P;\ 1 \leq i \leq N_p;\ 1 \leq j \leq N_{pi}. \tag{18}$$

The part and operation due dates required in (18) can be selected, for example, by performing backward scheduling from the product due date D_p to reflect the structure of the bill of materials.

The auxiliary objective function to be minimized is comprised of the sum of all the penalty terms described above, i.e.,

$$J_{AUX} = \sum_{pij} (\overline{W}_{pij} T_{pij}^2 + \overline{\beta}_{pij} E_{pij}^2),$$

$$1 \le p \le P;\ 1 \le i \le N_p;\ 1 \le j \le N_{pi},\ with$$

$$\overline{W}_{pij} \equiv w_{pij} + (w_{pi} + W_p \Delta_{pN_p}) \Delta_{piN_{pi}},\ and$$

$$\overline{\beta}_{pij} \equiv \beta_{pij} + (\beta_{pi} + \beta_p \Delta_{p1}) \Delta_{pi1}. \tag{19}$$

In (19) above, the fact that $b_{pij} = b_{pi}$ for $j = 1$, $c_{pij} = c_{pi}$ for $j = N_{pi}$, $b_{pi} = B_p$ for $i = 1$, and $c_{pi} = C_p$ for $i = N_p$ have been used. As such, Δ_{p1} is an integer variable equal to one if part (p,i) is the first part in product p and zero otherwise, and Δ_{pN_p} is similarly defined for the last part in product p. Additionally, Δ_{pi1} and $\Delta_{piN_{pi}}$ are similarly defined for the first and last operations of part (p,i), respectively. The coefficients in (19) are selected similar to the selection of coefficients for (3): $w_{pij} \ll w_{pi} \ll W_p$ and $\beta_{pij} \ll \beta_{pi} \ll \beta_p$.

2) Constraints: The minimization of (19) is now subject to the part precedence constraints imposed by the bill of materials as well as the constraints of Section II-A2. The reformulation of the constraints in Section II-A2 for the product scheduling problem simply entails duplicating the set of constraints for each product p. This is straightforward and is, therefore, omitted. The part precedence constraints are additional constraints in the product scheduling problem and are described below.

a) Part precedence constraints: One part in the bill of materials may be required for assembly with another part prior to a specific operation of the latter part. For example, the Ring & Support part in Fig. 1 is required to be assembled with the Support Assembly prior to the Lathe operation of the Support Assembly, i.e., part (p, i) must be finished prior to operation (p, q, r) of part q. Hence, the part precedence constraints require the beginning time of operation r of part q to be greater than or equal to the completion time of part (p, i) plus any required timeout S_{piq} between part (p, i) and operation r of part q, i.e.,

$$c_{pi} + S_{piq} + 1 \le b_{pqr},\ 1 \le p \le P;$$

$$1 \le i \le N_p - 1;\ (p, q, r) \in I_{pi}. \tag{20}$$

In the above, I_{pi} denotes the set of operations of other parts (not part i) immediately following part (p, i) in the processing sequence of the bill of materials.

Among the above variables, the part precedence structure, part due dates and associated penalty coefficients, and the variables listed in Section II-A are given. In the formulation presented here, the objective function in (19) is operation-wise additive. The part precedence constraints are also linear, but produce further coupling across parts, making the problem intractable.

C. Solution Methodology

The complexity of the product scheduling problem also motivates a decomposition approach. As in Section II, operation precedence and machine capacity constraints are relaxed using Lagrange multipliers. In the problem presented here, the additional part precedence constraints are also relaxed using Lagrange multipliers.

1) The Lagrangian Relaxation Approach: The operation precedence, part precedence, and capacity constraints are relaxed by using the Lagrange multipliers η_{pijl}, η_{piq}, and π_{kh}, respectively. The relaxed problem R is:

$$R : L \equiv \min_{\{b_{pij}\},\{m_{pij} \in H_{pij}\}} [\sum_{pij} (\overline{W}_{pij} T_{pij}^2 + \overline{\beta}_{pij} E_{pij}^2) + \sum_{kh} \pi_{kh} (\sum_{pij} \delta_{pijkh} - M_{kh}) + \sum_{pi,(p,q,r) \in I_{pi}} \eta_{piq} (c_{pi} + S_{piq} + 1 - b_{pqr}) + \sum_{pij, l \in I_{pij}} \eta_{pijl} (b_{pij} + t_{pijm_{pij}} + S_{pijl} - b_{pil})], \tag{21}$$

subject to the processing time requirements. This results in a minimization subproblem for each operation (p, i, j) with π and η given:

$$R_{pij} : L_{pij} \equiv \min_{b_{pij}, m_{pij} \in H_{pij}} [\overline{W}_{pij} T_{pij}^2 + \overline{\beta}_{pij} E_{pij}^2 + \sum_{(p,q,r) \in I_{pi}} \eta_{piq} (b_{pij} + t_{pijm_{pij}}) \Delta_{piN_{pi}} - \sum_{q:(p,i,j) \in I_{pq}} \eta_{pqi} b_{pij} + \sum_{l \in I_{pij}} \eta_{pijl} (b_{pij} + t_{pijm_{pij}}) - \sum_{l:j \in I_{pil}} \eta_{pilj} b_{pij} + \sum_{k=b_{pij}}^{c_{pij}} \pi_{km_{pij}}]. \tag{22}$$

In the above subproblem, the fact that $c_{pi} = b_{pij} + t_{pijm_{pij}} - 1$ for $j = N_{pi}$ and $\delta_{pijkh} = 0$ for all $h \ne m_{pij}$ have been used. The dual problem is formed by maximizing R with respect to the multipliers:

$$D : \max_{\pi, \eta \ge 0} L(\pi, \eta),\ with\ L(\pi, \eta) \equiv [-\sum_{kh} \pi_{kh} M_{kh} + \sum_{pi,(p,q,r) \in I_{pi}} \eta_{piq} S_{piq} + \sum_{pij, l \in I_{pij}} \eta_{pijl} S_{pijl} + \sum_{pij} L_{pij}]. \tag{23}$$

2) Scheduling Individual Operations: Similar to Section II-B2, solving subproblem R_{pij} in equation (22) entails enumerating all possible beginning times for each possible machine type $m_{pij} \in H_{pij}$, and comparing the values of L_{pij} derived from each possibility. The complexity of subproblem R_{pij} is therefore of order $K * |H_{pij}|$.

Once the operation subproblems have been solved, the dual problem is solved using the modified subgradient method described in Section II-B4. The cost obtained by solving the dual problem (23) is a lower bound on the optimal cost of (19), thus providing a measure of suboptimality of the schedule. Since the solution to the dual problem at the termination of the subgradient algorithm is generally associated with an infeasible schedule in which a few constraints are violated, a feasible schedule is constructed using an *extension* of the heuristic approach of Hoitomt [5]. The extension is straightforward and is omitted here.

D. Test Results

Three test cases representative of results to date are presented below. The effect of the part and operation tardiness and earliness penalties on solution oscillation is obtained by removing the penalties and examining the resulting weighted product tardiness and earliness. The effect of the earliness penalties is also obtained by examining the average cycle time of scheduled products, where the cycle time of a product is the time from the beginning of the first operation to the completion of the last operation of a product. In addition, the effect of the part precedence constraints in the formulation is examined by comparison with a method in which these constraints are not considered, i.e., parts are scheduled independently. Since the resulting schedule is not feasible with respect to the part precedence constraints, the heuristic described in Section III-C2 is used to provide a feasible schedule, facilitating a direct comparison between the two methods.

In all the test cases, the coefficients in the auxiliary objective function are assigned according to the comments in Section III-C1 and testing experiences. As such, the part and operation tardiness coefficients are initialized so that $100w_{pij} \equiv 10w_{pi} \equiv W_p$. The product earliness coefficients are chosen so that $100\beta_p \equiv \frac{1}{W_p}$, while the remaining earliness coefficients satisfy $100\beta_{pij} \equiv 10\beta_{pi} \equiv \beta_p$. In each of the test cases, all multipliers are initialized to zero.

Example 1: The first example is intended to demonstrate the ability of the algorithm to obtain a schedule using a simple bill of materials. There are four equally weighted products ($W_p = 10$), each comprised of one to three parts. In turn, each part is comprised of one to three operations, each to be processed on one of three machines without timeouts or mandatory slack time between operations. The planning horizon is 30 days (i.e., $K = 30$, and a time unit is one day). All machines are available on day one and throughout the planning horizon. All operations are available for processing on day one, but are due on day 0 or the current day. Data are shown in Table V.

A fixed number of iterations is used as a stopping criteria. For this example, 400 iterations are chosen as an acceptable tradeoff between computation time and duality gap. The feasible schedule has a cost $J_{AUX} = 5697.8$ and the lower bound is obtained as 5577.7 with a relative duality gap of 2.1%. According to the result of Hoitomt [5], this schedule is really optimal. The resulting schedule is shown in the form of a Gantt chart in Table VI. The time to solve the problem is 11.0 CPU seconds on a SPARCstation2 computer.

TABLE V
DATA FOR EXAMPLE 1

Prod. p	Part i	Op. j	H_{pij} [3]	t_{pij}	I_{pij}	I_{pi}
1	1	1	1	4	2	-
		2	2	3	3	-
		3	3	2	-	-
2	1	1	2	1	-	2
	2	1	1	4	2	-
		2	3	4	-	-
3	1	1	3	3	2	-
		2	2	2	-	2
	2	1	1	3	-	-
4	1	1	2	3	-	2
	2	1	3	3	-	3
	3	1	1	1	-	-

TABLE VI
GANTT CHART OF THE SCHEDULE FOR EXAMPLE 1

Time	1	2	3	4	5	6
Mach. 1	1,1,1	1,1,1	1,1,1	1,1,1	2,2,1	2,2,1
Mach. 2	2,1,1	4,1,1	4,1,1	4,1,1	1,1,2	1,1,2
Mach. 3	3,1,1	3,1,1	3,1,1		4,2,1	4,2,1
Time	7	8	9	10	11	12
Mach. 1	2,2,1	2,2,1	4,3,1	3,2,1	3,2,1	3,2,1
Mach. 2	1,1,2	3,1,2	3,1,2			
Mach. 3	4,2,1	1,1,3	1,1,3	2,2,2	2,2,2	2,2,2
Time	13	14	15	16	17	18
Mach. 1						
Mach. 2						
Mach. 3	2,2,2					

Example 2 [3]As in Section II, this example draws data from the same testbed at Pratt & Whitney's Development Operations shop, only now the parts are not independent, but are related through a bill of materials. There are a total of 15 different machines and 46 products, each consisting of 1 to 5 parts, for a total of 57 parts. Each part, in turn, consists of 1 to 7 operations, for a total of 93 operations. Products are characterized by different due dates and may have one of five different weights. The planning horizon is 193 days. In this example, the number of iterations is reduced from 400 to 100 since the problem size is larger and requires a greater computation time. The feasible schedule has a cost $J_{AUX} = 7\,696\,476$, while the lower bound obtained is 7 525 632, a relative duality gap of 2.3%. The CPU time on

[3]The notation "p,i,j" refers to operation j of part i of product p.

TABLE VII
RESULTS FOR EXAMPLE 2

	Constraints in formulation	Parts scheduled independently[4]
J_{AUX}	7 696 476	7 712 445
Lower bound	7 525 632	7 529 082
Duality gap	2.3%	2.4%
CPU seconds	43	43

TABLE VIII
RESULTS FOR EXAMPLE 2

	Without penalties	With penalties
J_{AUX}	6 838 655	7 696 476
J	6 838 655	6 812 465
Avg. cycle time	20.435	20.109

the same computer as in Example 1 is 43 CPU seconds. This is summarized in column 1 of Table VII.

To examine the effect of the bill of material constraints in the formulation, the schedule obtained with the methodology presented here is compared with the method of Section II in which parts are scheduled independently. Since the latter method does not consider bills of materials, it does not provide a schedule that is feasible with respect to the part precedence constraints. In order to facilitate direct comparison with the current method, the dual solution obtained with the method of Section II is modified with the heuristic described in Section III-C2, thus ensuring that part precedence constraints are satisfied. This technique for comparison is equivalent to setting the Lagrange multipliers that relax the part precedence constraints to zero, i.e. $\eta_{pil} = 0$, and keeping them zero during the subgradient update. Thus, scheduling parts independently by setting $\eta_{pil} = 0$ in the current method obtains a feasible schedule with cost $J_{AUX} = 7\,712\,445$ and a lower bound of 7 529 082 (using the same number of iterations as above) with a relative duality gap of 2.4%. This is summarized in column 2 of Table VII. The cost of the feasible schedule obtained using the current method is lower than the cost of a feasible schedule obtained by first scheduling parts independently via the method of Section II followed by the heuristic in Section III-C2. This illustrates that a feasible schedule with lower cost can be obtained by considering the part precedence constraints directly in the problem formulation.

To examine the effect of solution oscillation on algorithm performance, the part and operation tardiness and earliness penalties are removed from the auxiliary objective function. This is equivalent to setting $w_{pi} = 0$, $w_{pij} = 0$, $\beta_{pi} = 0$, and $\beta_{pij} = 0$. Using the same number of iterations as above, the feasible cost without the penalties is $J = J_{AUX} = 6\,838\,655$, while the feasible cost when the penalties are present is $J_{AUX} = 7\,696\,476$ ($J = 6\,812\,465$), as shown in Table VIII. This shows that weighted product tardiness and earliness cost J, the foremost criterion, is higher as a result of the oscillations that occur when the part and operation penalties are not present.

[4] $\eta_{pil} = 0$ during the subgradient updates

TABLE IX
RESULTS FOR EXAMPLE 3

	Constraints in formulation	Parts scheduled independently*
J_{AUX}	1 658 226	1 658 620
Lower bound	1 603 586	1 601 046
Duality gap	3.4%	3.6%
CPU seconds	145	145

*$\eta_{pil} = 0$ during the subgradient updates.

TABLE X
RESULTS FOR EXAMPLE 3

	Without penalties	With penalties
J_{AUX}	1 487 780	1 658 226
J	1 487 780	1 485 250
Avg. cycle time	9.207	9.143

The effect of the earliness penalties is further illustrated by considering the cycle time of a product, i.e., $C_p - B_p$. The average cycle time per product in the absence of the penalties is 20.435 days, but decreases to 20.109 days in the presence of the penalties (see row 3 of Table VIII).

Example 3 As a third example, data is used from the same testbed NC machine area, except that it is now a few months later. The center now has 140 products with 187 parts to be scheduled on 15 different machines. The planning horizon is 275 days. Again using 100 iterations, the current method obtains a feasible schedule with cost $J_{AUX} = 1\,658\,226$ with a lower bound of 1 603 586, a relative duality gap of 3.4%. This is summarized in column 1 of Table IX. If parts are scheduled independently, the method obtains a feasible schedule with cost $J_{AUX} = 1\,658\,620$ with a lower bound of 1 601 046, a relative duality gap of 3.6%. This is summarized in column 2 of Table IX. Again, these results show that inclusion of the part precedence constraints directly in the problem formulation improves scheduling performance.

The feasible cost without the part and operation penalties in this problem is $J = J_{AUX} = 1\,487\,780$, while the feasible cost when the penalties are present is $J_{AUX} = 1\,658\,226$ ($J = 1\,485\,250$), as shown in Table X. Again, the higher cost J obtained in the former case is a result of the oscillations that occur in the absence of the part and operation penalties. The average cycle time per product in the absence of the penalties is 9.207 days, but decreases to 9.143 days in the presence of the penalties (see row 3 of Table X).

IV. CONCLUSION

This paper has presented an improved Lagrangian relaxation technique with an application to the problem of scheduling products with bills of materials. In Section II, an auxiliary problem formulation was developed for the job shop scheduling problem. There, quadratic penalties are included in the objective function for the purpose of modifying the geometry of the dual function. This is necessary since the subgradient method can cause solutions to oscillate about a set of non-differentiable points, slowing convergence of the algorithm. Although solution oscillation is not completely eliminated, the auxiliary objective function reduces the magnitude of oscilla-

tion, making an augmented Lagrangian approach unnecessary. In addition, the relaxed problem is directly decomposed into operation-level subproblems. This is in contrast to the augmented Lagrangian approach which requires a Gauss-Seidel iterative technique to enforce separability. The advantage is a significant computation time savings, as demonstrated in the results. Furthermore, a lower bound with respect to the auxiliary objective is obtained, thus providing an indirect measure of schedule quality. The results show that the new algorithm can generate effective, feasible schedules using less computation time than the augmented Lagrangian approach.

In Section III, the improved Lagrangian relaxation technique was applied to the problem of scheduling products with bills of materials. The bills of materials, commonly used to describe the assembly of complex products, are integrated into the product scheduling problem by including part precedence constraints in the formulation. In addition, an auxiliary objective function containing quadratic penalties is used to reduce oscillation in the solution process. Results to date that compare this method to an algorithm in which parts are scheduled independently indicate that integration of the bills of materials into the scheduling problem improves scheduling performance by reducing product tardiness. This is significant since the improved ability of a manufacturer to meet promised delivery dates for products by the above integration will ultimately enhance a manufacturer's credibility and competitiveness in the marketplace. The subject of our current investigation is to develop methods that eliminate solution oscillation completely, as additional computation time savings are expected [15].

ACKNOWLEDGMENT

The authors would like to thank Dr. Debra Hoitomt, Thomas Owens, Robert Tomastik, and Scott Bailey for their invaluable suggestions and support.

REFERENCES

[1] P. Camerini, L. Fratta, and F. Maffioli, "On improving relaxation methods by modified gradient techniques," *Mathematical Programming Study 3*, Amsterdam, pp. 26–34, 1975.
[2] C. S. Czerwinski and P. B. Luh, "An improved Lagrangian relaxation technique for job shop scheduling," *Proceedings of the 1992 IEEE Conference on Decision and Control*, 1992.
[3] S. C. Graves, "A review of production scheduling," *Operations Research*, vol. 18, pp. 841–852, 1981.
[4] J. Harhen, "MRP/MRP II," in *Computer-Aided Production Management*, A. Rolstadas, Ed. Springer-Verlag, pp. 23–35, 1988.
[5] D. Hoitomt, P. Luh, E. Max, and K. Pattipati, "Scheduling jobs with simple precedence constraints on parallel machines," *IEEE Control Systems Magazine*, vol. 10, no. 1, pp. 34–40, 1990.
[6] D. J. Hoitomt, P. B. Luh, and K. R. Pattipati, "Job shop scheduling," *Proceedings of the First International Conference on Automation Technology*, Taipei, Taiwan, 1990, pp. 565-574; A revised version will appear in *1992 IEEE Transactions on Robotics and Automation*.
[7] S. Kim, and H. Ahn, "Convergence properties of the modified subgradient method of Camerini *et al.*," *Naval Research Logistics*, vol. 37, pp. 961–966, 1990.
[8] J. K. Lenstra, A. H. G. Rinnooy Kan, and P. Bruckner, "Complexity of machine scheduling problems," *Annals of Discrete Mathematics*, vol. 7, pp. 343–362, 1977.
[9] D. G. Luenberger, *Linear and Nonlinear Programming*. Reading, MA: Addison-Wesley, 1984, pp. 243–246.
[10] P. B. Luh, D. J. Hoitomt, E. Max, and K. R. Pattipati, "Schedule generation and reconfiguration for parallel machines," *IEEE Transactions on Robotics and Automation*, vol. 6, no. 6, pp. 687–696, 1990.
[11] K. N. McKay, F. R. Safayeni, and J. A. Buzacott, "Job-shop scheduling theory: What is relevant?" *Interfaces*, vol. 18, pp. 84–90, 1988.
[12] S. Miyazaki, "Combined scheduling system for reducing job tardiness in a job shop," *International Journal of Production Research*, vol. 19, no. 2, pp. 201–211.
[13] G. Nemhauser and L. Wolsey, *Integer and Combinatorial Optimization*. New York: John Wiley & Sons, Inc., 1988, pp. 323–337.
[14] N. Z. Shor, *Minimization Methods for Non-Differentiable Functions*. Berlin: Springer-Verlag, 1985, pp. 40–42.
[15] R. Tomastik and P. B. Luh, "The facet ascending algorithm for integer programming problems" in *Proceedings of the 32nd IEEE Conference on Decision and Control*, San Antonio, Texas, December 1993, pp. 2880–2884.
[16] K. P. White, Jr., "Advances in the theory and practice of scheduling," in *Advances in Industrial Systems, Control and Dynamic Systems*, vol. 37, C. T. Leondes, Ed. San Diego, CA: Academic Press, Inc., 1990, pp. 115–158.

Christopher S. Czerwinski received the B.S. and M.S. degrees in Electrical and Systems Engineering from the University of Connecticut, Storrs, in 1990 and 1992, respectively. His interests include manufacturing technology, schedule generation and reconfiguration for manufacturing systems, and optimization techniques. He is currently employed at the Otis Elevator Engineering Center in Farmington, Connecticut and is involved in new product development and cost reduction efforts.

Peter B. Luh (S'76–M'80–SM'91) received the B.S. degree in Electrical Engineering from National Taiwan University, Taipei, Taiwan, Republic of China, in 1973, the M.S. degree in Aeronautics and Astronautics Engineering from M.I.T., Cambridge, MA in 1977, and the Ph.D. degree in Applied Mathematics from Harvard University, Cambridge, MA, in 1980. From 1980 he has been with the University of Connecticut, and is currently a Professor in the Department of Electrical and Systems Engineering, and Director of Micropreci-sion Manufacturing within the Advanced Technology Center for Precision Manufacturing of the State of Connecticut. His major research interests include schedule generation and reconfiguration for manufacturing systems, scheduling of power systems, and distributed decisionmaking. He has been a principal investigator and consultant to many industry and government-funded projects in the above areas, and has published more than 140 papers. He has made significant contributions in manufacturing by developing a near-optimal and efficient schedule generation and reconfiguration methodology to improve on-time delivery of products and reduce work-in-progress inventory. The method has been adopted as the backbone of a new scheduling system developed by the Development Operations Shop of Pratt & Whitney, is currently in use for the scheduling of selected work centers of the shop on a daily basis, and has re-shaped the scheduling community. He has also made significant contributions in power systems by developing a near-optimal and efficient unit commitment and hydro-thermal coordination methodology that is currently in use by Northeast Utilities, a major electric utility company in New England, on a daily basis.

Dr. Luh is a Technical Editor for the *IEEE Transactions on Robotics and Automation* (1990–1994), Program Vice-Chairman for Invited Sessions for the 1993 IEEE International Conference on Robotics and Automation, was an Associate Editor the *IEEE Transactions on Automatic Control* (1989–1991); and has served in Program Committees and Operating Committees of many national, international, and inter-society conferences. He is a senior member of the IEEE, member of Sigma Xi, and listed in *Who's Who in Engineering, Who's Who in the East*, and *Who's Who in American Education*.