CHAPTER 78

Advanced Planning and Scheduling for Manufacturing

Kenneth Musselman
Frontstep, Inc.

Reha Uzsoy
Purdue University

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy

1. INTRODUCTION

The problem of how to allocate a company's resources and material effectively among competing activities over time to optimize the company's market and financial positions is encountered in all industries producing goods or services. Unless a company has significant excess capacity and high inventories, decisions have to be made as to which orders it will accept, which it will turn away, and which products and customers will be given priority over others—in other words, how its available inventory and production capacity will be allocated among revenue-generating activities.

Several factors combine to make this a difficult task. The company must respond to often rapidly changing market conditions and technological developments. A number of different, often conflicting objectives, such as filling customer orders and maintaining low inventory levels and lead times, must be traded off against each other. Different amounts of variability in the production processes and customer demand must be managed. However, there is considerable evidence from various industries that effective execution of this task can provide a significant competitive advantage.

Today a number of trends are combining to render the area of production planning rather more active than it has been for several decades. Increasing competitive pressures have forced companies to forgo the expensive luxury of large amounts of excess capacity and high inventories, making effective allocation of manufacturing capacity and coordination of production activities throughout the supply chain a critical component of market success. The strong demand from industry for these services is demonstrated by the rapidly increasing number of software products and consulting companies specializing in this area that have emerged in the last five years. The explosive expansion of information technology fueled by Moore's law, manifested in the Internet, better databases, and faster, cheaper computers, has rendered feasible a whole range of solutions that could not even be imagined 10 years ago.

In this chapter we shall examine developments in the area of production planning, focusing on the case where a single factory is considered. However, much of the discussion and many of the modeling issues remain valid when production systems involving multiple plants are considered. We do not, however, make any attempt to consider the larger supply chain, which by its nature must consider such aspects as transportation, warehousing, and interactions with other companies. The issue of supply chain management is a fast-growing area of both research and practice and is discussed in more detail in Chapter 82, as well as in Tayur et al. (1998).

We begin by discussing the relationship between the production planning and scheduling functions and their importance to the manufacturing firm. In this context we discuss the effects of congestion on the shop floor and the relationship between workload and lead times, which is fundamental to the relationship between planning and scheduling. We introduce the topic of production planning by discussing at some length the well-known and widely used Material Requirements Planning (MRP) algorithm (Orlicky 1975) and its extensions. This approach is widely used in industry, and much of the current effort in developing advanced planning and scheduling (APS) systems is aimed at remedying its various deficiencies. After defining our view of APS, we present a range of production planning algorithms that have been proposed in industry and academia over the last several years, discussing their strengths and weaknesses. Finally, we identify a number of issues that in our experience must be addressed to implement an APS system successfully.

2. THE PLANNING AND SCHEDULING FUNCTIONS

Generally speaking, planning and scheduling jointly determine how, when, and in what quantity products will be manufactured or purchased. In essence, planning establishes *what* should be done and scheduling determines *how* to do it. The conventional approaches to both these functions are explained below, together with fundamental issues associated with both.

2.1. Planning

Planning determines when to manufacture and purchase parts and how many in order to satisfy future demand for end products. The process is externally focused since the demand comes from both actual and anticipated customer orders. It is controlled by higher-level attributes, such as end-item due dates and order types, and takes an aggregate view of the production process. This aggregation takes several forms: individual machines are aggregated into workcenters for the purpose of representing capacity, time is aggregated into discrete buckets, and the flow time for a number of operations required to produce a component or subassembly is often aggregated into a single lead time. The details associated with how work actually flows through the plant, such as the specific timing of individual operations and production sequences on individual machines, are left unresolved. The time frame, or horizon, over which the plan is made is normally on the order of weeks or months. The result of the process is a time-phased projection of inventory levels, production quantities, and workcenter requirements to satisfy independent demand.

In this incarnation, the production plan serves a number of functions. It represents a decision as to how the company's manufacturing capacity will be allocated among competing products and customers and hence the point where the company's strategy is turned into concrete actions visible by employees and customers. Secondly, it provides management with some visibility into the future status of production, allowing them to identify at least some problematic situations such as mismatches between demand and capacity in time for remedial action to be taken. It thus provides critical information for a number of activities, such as negotiating due dates with customers and deciding the timing and quantity of raw material purchases from suppliers.

It is important to note that the production plan is by nature somewhat tentative, being subject to significant uncertainty. In many companies, at least part of the demand considered in the plan is based on forecasts rather than firm orders and is hence subject to varying degrees of change over time as orders are modified, added, or canceled by customers. Even when demand uncertainty is minimal, the actual execution of the plan on the shop floor is subject to random disruptions such as machine failures, quality problems, and unexpected rush orders. In many companies, production plans are developed and used on a rolling horizon basis, with decisions in the early periods being considered binding but those further out being revised and altered as new information on the realizations of demand and production become available.

It is also important to note that the nature of the planning problem may differ quite substantially depending on the nature of the company's business. At one extreme is the make-to-stock environment, where demand is relatively high and stable. The lead time to produce an end item is sufficiently high that the company maintains substantial finished goods inventory to meet customer orders and produces to replenish this inventory based on demand forecasts. Another situation is a make-to-order environment, where customized items are produced upon receipt of the order. Here the planning system must allow the company to assess intelligently whether or not customer orders can be completed by the requested date and allow detailed coordination of material and other resources to achieve this. Assemble-to-order systems are intermediate in nature, where a number of basic subassemblies are produced to stock and then combined in different ways in response to customer orders. A company's planning and scheduling needs may differ quite substantially depending on the environment in which it is operating. Hendry and Kingsman (1989) discuss the needs of make-to-order companies and the relevance of many planning and scheduling approaches such as kanban (e.g., Monden 1983) and theory of constraints (Goldratt and Fox 1986) in this environment.

An important role in production planning belongs to the master production schedule (MPS), which specifies the quantity of each final demand item required in each time period and drives the requirements planning calculation of how many of each component and subassembly to produce to meet this demand over time and thus the scheduling system that moves the work through the individual operations to meet this plan. Originally MPS was treated as an exogeneous input to the system that developed the quantities and timing of releases to the shop floor. The goal of many APS systems, especially transaction-oriented systems, is to integrate planning and scheduling decisions to a much higher degree. Hence, much of the functionality of the MPS is fulfilled by the plan developed by the APS system.

The MPS, or the planning function that fulfills this role, is a crucial element of the production-planning process for several reasons. First, if the MPS is not realistic with respect to the various constraints such as production and supplier capacity and material availability faced by the manufacturing organization, it is unlikely that the best APS software or shop-floor scheduling package can save the situation. Secondly, the MPS is where a crucial set of decisions determining how the company's limited production capacity will be allocated among competing orders and customers is made. These decisions involve trade-offs whose effective resolution requires an understanding of the strategic and tactical goals of the company as well as negotiation among various functional groups within the company. For example, the manufacturing organization is likely to prefer an MPS that allows them to use equipment efficiently by having long production runs with few setup changes. On the other hand, sales and marketing are likely to push for an MPS that emphasizes delivery to key customers, as well as perhaps to customers who are getting ready to take their business elsewhere. Hence effective, thoughtful procedures for developing an MPS are critical to the company's long-term performance. However, in practice we find that in many cases companies develop their MPS using the intuition and knowledge of a few key employees and simple spreadsheet-based tools for data management. There is currently far more art than science to the development of an MPS, a situation that renders this area attractive for future research.

2.2. Scheduling

While the production plan lays out what mix and quantity of products the company is expected to produce over a certain time horizon, the schedule describes the detailed execution of this plan, giving a step-by-step work list in the form of a dispatch list or a specification of the times at which every operation should start and end. In contrast to the production plan, whose focus is on independent demand for end items, the scheduling process is internally focused, driven by the need to ensure that all the components, subassemblies, and assemblies needed to produce the end items are completed according to the plan as efficiently (in terms of resource usage) as possible. Inputs to the scheduling process are product definition information (e.g., routes), facility information (e.g., workcenter availability), and shop-floor status. Shop-floor status defines the current state of production, identifying what orders are still open, their current locations (e.g., what machines are working on them), and the yield or scrap rates for each of the operations associated with these orders. Depending on the capabilities of the particular scheduling algorithm used, the process can also address work order resequencing for more efficient workcenter processing and improved plant throughput. The process is controlled by lower-level attributes, such as manufacturing order due dates and work order selection logic at a workcenter. It also takes a more detailed view of the production process, working with individual resources (e.g., machines) vs. workcenters, operational level routes for parts vs. fixed lead times, shift times vs. planned hours, and often even continuous time vs. discrete time buckets. The time horizon is typically short, usually on the order of a shift, a day, or a week. The usefulness of the schedule generally decreases rapidly in the future since responses to shop floor disruptions and changes to the production plan will result in significant revisions.

In summary, planning focuses on allocating production resources and material to various end products or customers, while scheduling focuses on how to meet component level deliveries without sacrificing efficiency. In general, one of these functions tends to dominate, in the sense that its decisions are considered to be more important to satisfy and the other is forced to adhere to them. Many of today's APS systems have their roots in the attempt to remedy the deficiencies of traditional planning systems by integrating planning and scheduling more closely.

3. RELATIONSHIPS BETWEEN PLANNING AND SCHEDULING

It should by now be apparent that planning and scheduling are tightly intertwined. Planning starts with high-level demand and produces a "schedulable" plan. Scheduling then generates a time-sequenced allocation of individual resources to tasks over time that efficiently supports this plan. Closing the loop, planning then honors these allocations as it replans. In other words, planning affects scheduling and vice versa.

Effective planning leads to effective scheduling. No amount of clever scheduling can overcome the effects of a poor plan. Another way of saying this is that if a plan calls for the plant to build the wrong things at the wrong time, efficiently executing this plan may still leave the plant in considerable trouble. If the plan is so tightly determined that it does not allow the scheduling procedure to adapt work order (job) sequences to the detailed needs of the shop floor, such as sequence-dependent setup times, then the task of scheduling is obviated and processing efficiency is likely reduced. At the other extreme, a plan that fails to constrain the scheduling task properly risks a decline in on-time performance in the name of efficiency. Planning needs to provide the shop floor enough direction to maintain overall order performance without unduly limiting opportunities for efficient workcenter processing. Recall that planning is where we prioritize orders and customers; the scheduling function does not usually have access to the right information to make these decisions in the company's best interest. However, it is unfortunately quite common to see these decisions being made by scheduling personnel due to dysfunctionalities in the planning system such as inaccurate data on workcenter capabilities.

Just as proper planning leads to good scheduling, proper scheduling leads to good planning. Efficiently rearranging the work on the floor allows production gains to be made that translate into more available capacity in which to plan the work. This also offers the added benefit of closer promise dates. Conversely, poor scheduling can undo the work of a good plan when the inefficiencies incurred on the floor cause exceptions to the plan. These exceptions are often addressed by ad hoc remedies on the shop floor based on local objectives such as machine utilization, resulting in a reduction in global performance.

In spite of this close relationship between planning and scheduling, inherent differences exist between them. These manifest themselves in many ways, including their function, their treatment of capacity, and their representation of the manufacturing process. An interesting discussion of the relation between planning and scheduling can be found in Pritsker and Snyder (1997).

3.1. Function

In manufacturing there is a basic dualism at play between synchronization and sequencing. Synchronization is the process of prioritizing independent demand from customer orders and demand forecasts and deriving all dependent demand for the necessary components, materials, and subassemblies accordingly. Workcenter reservations, allocation of work-in-process to specific orders (pegging), and purchase requisitions are made in support of this coordinated plan. This is an order-driven process and is the primary function of planning. Sequencing, on the other hand, which is the primary function of scheduling, is workcenter driven in that it locally ranks demand on a time-phased basis and projects order completions accordingly.

Under APS, planning, with its synchronization emphasis, dominates and sequencing (or scheduling) is subservient to it. The operations to manufacture a part are initially placed in the plan to support a global criterion, say customer order due date, and then the sequencing of these operations is done to adhere to the plan as closely as possible.

3.2. Treatment of Capacity

Since the basic problem of production planning is that of allocating manufacturing capacity among various customers and products over time, it would seem natural to assume that such a basic notion as capacity should be well understood. However, once one tries to get specific, it turns out that capacity is a remarkably elusive concept. Elmaghraby (1989) gives an insightful discussion of this issue. In this section we first discuss some basic aspects of capacity and its relation to the outputs of the planning process, such as resource utilization, batch sizes, and lead times.

Despite the difficulty of achieving a rigorous definition of capacity, it is widely accepted that the ability to produce a given set of orders on a given set of equipment by specified due dates is affected by product mix, shop-floor scheduling decisions, and the stochastic nature of events on the shop floor. A basic driver of shop-floor dynamics is the phenomenon of congestion, in which the rate of response of the manufacturing system degrades as more work is introduced into the system, even when the demand rate for the system is well below its nominal capacity. This is due to variations in the arrival rates of jobs at workcenters causing short-term saturation, where the workcenter is unable to process all jobs arriving within a short period of time and queues form. These variations arise from the myriad random events that determine the outcome of most manufacturing processes, such as inherent variability in processing and setup times, machine failures, and quality problems.

A fundamental relationship that holds in a broad range of environments states that the average time to process an order at a workstation is a highly nonlinear function of the workload in the system and that as the workload approaches a nominal capacity both the mean and variance of the lead time increase exponentially, as illustrated in Figure 1. A second fundamental relationship is Little's law (Hopp and Spearman 1996), which states that the average work-in-process inventory (WIP) level and the average time to process a job through the system (i.e., lead time or cycle time) are directly proportional.

The most important aspect of congestion for planning purposes is that a plan that is feasible with respect to manufacturing capacity when aggregated over a specified time period may not be capacity feasible at all times during that interval. To see this, consider a situation where the planning time period is a day and we have a single machine that is available for eight hours. In this sense, placing four orders, each of which requires two hours of processing, on this machine on this day is perfectly feasible. However, it may well be that all four jobs arrive at this workcenter at 10 a.m. This causes significant short-term congestion, which renders it impossible to complete all orders in the time period assigned.

Another area in which this shop-floor dynamic is becoming better understood is that of the effect of batch sizes, which has been studied by Karmarkar (1987). This work shows that initially, increasing batch sizes lead to rapid reductions in flow times as excessive setups are eliminated, essentially increasing the capacity of the system. However, as the batch size continues to increase, the flow time begins to increase linearly with batch size due to the additional queueing and processing time incurred by the larger batch sizes.

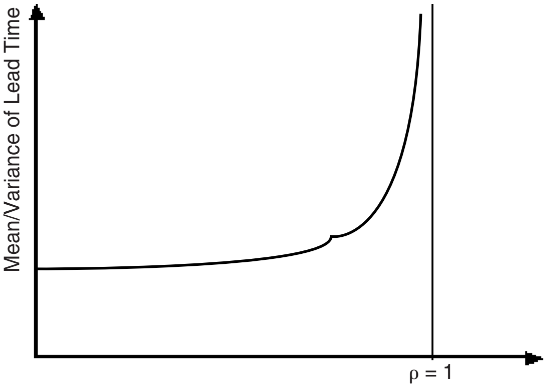

Figure 1 Relationship between Utilization (ρ) and Mean and Variance of Lead Time.

The important aspect of these relationships with regard to planning and scheduling is that both the workload of a given workcenter in a given time and the batch size used in a manufacturing facility are often outputs of the planning process. Hence, the planning process defines to a large extent the basic performance we can realistically expect from the shop floor. The specific scheduling algorithm we use will clearly affect this, at least in the sense that clever job sequencing will make the best possible use of available capacity. However, the high-level response of the system, in the sense of performance measures such as the average lead time, are defined to a large degree by the planning process and cannot be fundamentally altered by scheduling.

Another difference between planning and scheduling stems from their different views of how much capacity should be committed. In planning, it is often desirable to leave some capacity idle to allow time for the workcenters to handle contingencies that may arise, such as failures or rush orders. This is especially true for bottleneck or near-bottleneck workcenters that have demonstrated unreliable performance or whose cycle times vary widely. In planning for these workcenters, it is often advantageous to load the workcenter to a level below that of its nominal availability, yielding a more sparsely populated plan that can tolerate a certain degree of disruption before overall order performance begins to degrade. The extra capacity made available by this undercapacity approach to planning allows the scheduling algorithm, which tries to use the workcenter's full capacity, an opportunity to adjust to the disruptions encountered on the floor. However, while the advantages of this undercapacity planning approach may be significant, too much idle capacity in the plan may well lead to the company making inefficient use of its capacity, quoting overly conservative due dates and hence losing competitive advantage.

3.3. Representation

The advent of the computer and the powerful information technology tools it has brought with it have created many new possibilities for the planning and scheduling process. One area where this is evident is in the level of detail at which the manufacturing process is represented for planning and scheduling. Many more factors can be considered simultaneously with a computerized approach than with a manual one.

Yet even with this ability, planning and scheduling, in practice, still differ in how they represent the manufacturing process. Planning, which coarsely sets the boundaries within which to schedule, can take liberties in its representation. Average setup times, for example, are usually sufficient when generating a plan. The intricate logic that can accompany a setup time calculation is typically superfluous when the objective is to develop a synchronous, as opposed to an executable, plan. Appropriate decisions as to which resources must have their capacity explicitly modeled and which others are nonconstraining and can be treated as having infinite capacity can often improve execution speed without significantly compromising the quality of the plans generated.

This is not to say that detail is not appropriate at the planning level. When capacity is a major determinant of performance, incompatibilities can be introduced as a result of relaxing too many constraints. Consideration of tooling and overlapped operations, for example, may be necessary during planning to obtain a realistic picture of workcenter capacity and provide appropriate goals to the scheduling function. In some industries, such as integrated steel mills, the dynamics of the shop floor affect the capacity of the shop to such a degree that in order to be viable a plan must consider detailed shop-floor dynamics. Again, a well-constructed plan gives scheduling the opportunity to refine the sequence of operations to improve workcenter efficiencies without sacrificing, and hopefully enhancing, global performance to plan. If the plan is wildly inconsistent with the realities of the shop floor, then scheduling may be forced to make radical changes, which in turn cause major adjustments to the plan. This can be viewed as the scheduling function usurping some of the functionality of a dysfunctional planning system, which it may well not have sufficient information to perform adequately.

4. PLANNING ALGORITHMS

We can classify planning algorithms on two basic characteristics. The first of these is how manufacturing capacity is modeled. This yields two basic classes of algorithms: those that consider capacity within a given time period to be unconstraining, that is, essentially infinite, and those that recognize some limitation on the amount of available capacity. We will refer to these two classes of algorithms as infinite capacity and finite capacity, respectively.

The second classification concerns how the algorithm models the timing of production events. Some algorithms do not consider the issue of congestion at all, essentially assuming that the time for a task to be processed at a workcenter is independent of its workload, that is, is a property of the product being manufactured. Note that this does not necessarily imply that the underlying model assumes infinite capacity, only that it is incapable of modeling the congestion effects discussed above. A second set of algorithms considers congestion effects explicitly. We will refer to these two classes of algorithms as noncongested and congested, respectively. While noncongested algorithms work well when the production system is at low utilization, the ability of this model to predict job completion times accurately deteriorates rapidly as system utilization increases.

We will begin with a discussion of material requirements planning, which is the most widely used infinite capacity algorithm today. We will then introduce a variety of finite capacity algorithms, beginning with enhancements to the basic MRP algorithm and continuing through fundamentally different approaches such as optimization and artificial intelligence techniques.

4.1. Infinite Capacity Algorithms: Material Requirements Planning

Material requirements planning (MRP) was developed in the 1960s to apply the computational power of computers to production and inventory management problems. The basic paradigm adopted was that manufacturing was fundamentally a problem of coordinating the complex material flows involved in producing large, assembled products with deep bills of material. This basic logic, with a number of extensions and additions, is the common ancestor of most of today's enterprise resource planning (ERP) systems (Ptak and Schragenheim 1999). As such, MRP is prevalent in industry and continues to be the basic planning mechanism for many manufacturing companies as well as the driver for a multibillion-dollar software and consulting industry.

We will first briefly describe the basic MRP procedure and then discuss its inherent strengths and weaknesses. Details of this procedure and its many variants and enhancements can be found in texts such as Vollmann et al. (1988) and Nahmias (1993). Efforts to remedy these deficiencies will then lead us to a discussion of several alternative approaches to production planning that form the kernel of several of today's most successful APS systems.

The MRP planning procedure has three main inputs:

1. *The master production schedule (MPS)* or some other planning document that specifies how much of each end product is required in each time period t, over some specified planning horizon involving T periods. In most practical applications, the basic time period is a week, although longer periods of a month or so may be used for periods far in the future where there is more uncertainty in the demand process (e.g., the MPS is based more on forecasts than on firm customer orders). The relationship of the MPS to the production plan was discussed in Section 2.
2. *The bill of material (BOM)*, which specifies the structure of each product in terms of the components, subassemblies, and assemblies that constitute it. This structure is usually represented graphically as a tree whose root node represents the complete product, leaf nodes purchased components or raw materials, and intermediate nodes subassemblies and assemblies. For the sake of brevity we shall refer to these items collectively as modules. An example of such a BOM tree is shown in Figure 2. For each node (i.e., module) the BOM specifies which items are combined to make it and the quantity of each such item. This formalization allows MRP to draw a distinction between two types of demand. Independent demand originates with customers outside the company and is typically for end items. Dependent demand, on the other hand, refers to the demand for the BOM items used to produce the end item. Once the BOM and the independent demand are given, the dependent demand can easily be computed—if we know we need to build four cars, we know we will need four engines, four back seats, and 16 wheels, for instance. This distinction is of great importance since it allows us to focus on forecasting and managing the independent demand with confidence that the dependent demand can be generated easily if the BOM and estimates of independent demand are correct.

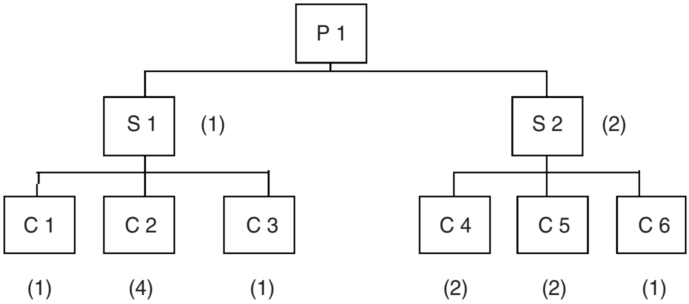

Figure 2 Example BOM Tree.

3. *Inventory status:* This is usually a database specifying the amount of each module on hand or on order. For parts that are on order, it will usually specify when the parts are expected to arrive.

The MRP algorithm combines these three inputs to generate planned order releases for all items in the BOMs of products that occur in the MPS. This will specify how many of each BOM item are needed in each time period covered by the MPS. We will assume that the nodes of the BOM have been indexed such that no node with a lower index occurs at a lower level of the tree than a node with a higher index. Well-defined algorithms to generate this type of indexing, known as level coding, exist. The MRP algorithm can be stated as follows:

For each independent demand item in each time period t, perform the following:

- *Step 1 (initialization):* Set $i = 0$, where node 0 denotes the root node of the BOM tree. Let d_{it} denote the number of item i required in period t.
- *Step 2 (time-phased order releases):* For each descendant j of node i in the BOM tree, perform the following:
  - Calculate the gross requirement $g_{jt} = a_{ij}d_{it}$, where a_{ij} is the number of units of module j required to form module i.
  - Calculate the net requirements for j as $n_{jt} = g_{jt} - I_{it}$, where I_{it} denotes the projected inventory of module j at the beginning of period t. I_{jt} is actually made up of inventory carried over from the last period and items on order expected to arrive in that period. For the sake of brevity, details of how the inventory status is updated can be found in any of several books on this subject (e.g., Nahmias 1993; Vollmann et al. 1988).
  - Generate the planned order release $p_{j,t-L_j} = n_{jt}$, where L_j denotes the number of periods after an order is placed that will elapse before the order is filled. We assume that the L_j are known constants that depend only on the item being ordered. Note that L_j may represent either the production lead time for a part manufactured in house or the supplier lead time for a part or material ordered from a supplier.
  - Once the total quantity of each module to be started in each period has been determined, we can combine the requirements for a number of periods into a release for a single period. This procedure, known as lot sizing, can serve to reduce the amount of setup time required for production changeovers, or may be dictated by process concerns.
- *Step 3 (stopping criterion):* Mark node i as expanded. Select the unexpanded node in the BOM tree with the lowest-level code and go to step 2. If no unexpanded nodes exist, stop.

The basic idea of the above procedure is almost absurdly simple: starting with the requirements for the end product for a given period of the MPS, subtract the estimated amount in inventory in that period to determine how many we actually need to make (the net requirements). Net requirements at one level of the tree become the gross requirements for the next level down. Once we have the net requirements for a given module in a given period, we then schedule the planned order release L_j periods earlier, such that the material will be where it is needed at exactly the right time. This is referred to as backward planning, where the order release date is obtained by working backward from the date the material or product is required and subtracting an estimate of the time needed to complete it.

To illustrate the operation of this algorithm, consider the product whose BOM is given in Figure 2, and let inventory availability and lead times be as listed in Table 1. In addition, we expect 400 units of P1 to become available in week 6, 100 units of S2 in week 4, and 400 units of C4 in week 5.

Based on this information, we can show the results of the MRP calculations in Table 2. Considering the end item P1, we will need 400 units in period 6, but we currently have none available and none expected to become available. Hence, our net requirement for period 6 is 400 units. Since the

TABLE 1 Inventory Status and Lead Times for Example

Item	Lead Time	On Hand
P1	1	–
S2	1	10
C4	2	50

TABLE 2 MRP Calculations for Example

Product P1	0	1	2	3	4	5	6
Gross requirements							400
On hand							
Scheduled receipts							
Net requirements							400
Order releases						400	
Assembly S2	0	1	2	3	4	5	6
Gross requirements						800	
On hand	10	10	10	10	10	110	
Scheduled receipts					100		
Net requirements						690	
Order releases					690		
Component C4	0	1	2	3	4	5	6
Gross requirements					1380		
On hand	50	50	50	50	50		
Scheduled receipts					400		
New requirements					930		
Order releases			930				

lead time is one period, this means that the order must be released at the beginning of period 5 for the end items to be ready when needed at the end of period 6. Since each unit of P1 requires two units of S2, this release causes a gross requirement of 800 units for S2 in period 5. Netting out the 110 units on hand at the beginning of the period, this yields a net requirement for S2 of 690 units, released as an order in period 4. Since, again, each unit of S2 requires 2 units of C4, we obtain a gross requirement of 1380 units of C4 in period 4, resulting in an order for 930 units placed in period 2.

It is interesting to examine the MRP algorithm from a number of different perspectives. First of all, its basic goal appears to be to achieve just-in-time material flow—if the lead times L_j are accurate, material will arrive exactly in the time period in which it will be used. From an optimization standpoint, it attempts to minimize deviation from the MPS (in terms of items completed after their request date) subject to the constraints of the lead times, the BOM structure, and the initial inventory status. Clearly, if the MPS is unrealistic, the backward scheduling procedure in step 2 may indicate a need to release an order in a time period that is already in the past. In this situation, the basic MRP logic offers no help—it is up to the user to revise the MPS to achieve a feasible situation. Many commercial systems provide the option to plan problematic orders forward in time, assuming that required modules at the lowest level of the BOM are started immediately and using the lead times to work forwards to an earliest achievable completion time. In this context, however, it should be noted that neither the backward planning algorithm used by MRP nor the forward algorithm used to remedy deficiencies are rigorously correct. Hence, there may exist a pattern of releases that renders the MPS feasible even though the MRP calculations show them to be otherwise. Moreover, we cannot be certain that if the MRP system claims a plan to be feasible it will indeed turn out to be so when executed on the shop floor.

The basic MRP logic, although widely used and still actively promoted in industry, has a number of fundamental flaws that seriously limit its usefulness as a planning tool. The most important is the treatment of manufacturing capacity. The only reflection of manufacturing capacity in the MRP algorithm is the lead times L_j. These are viewed as being known constants that are an attribute of the module j only. Specifically, they are assumed to be independent of the product mix in the shop and the loading or utilization of the shop at the time the order is released. As we discussed in Section 3.2, the lead time is a function of how the shop is loaded relative to its capacity at the point in time the order is released. Hence, while the fixed lead time assumption may be reasonably accurate for shops at low levels of utilization, as utilization increases, congestion will become more significant. Both the mean and the variance of the time in system at heavily loaded workcenters will increase, rendering the fixed lead times an increasingly inaccurate representation of the actual situation. Inaccurate lead times, in turn, will result in the release of orders to the shop that the factory will not be able to complete on time.

In many environments where the shop is heavily loaded and cannot achieve the lead times quoted in the MRP system, manufacturing will often try to make the case that the lead times should be extended to allow them more time to get work through the plant. However, when the lead times are extended, we are essentially releasing orders to the shop earlier than we used to, and hence we actually increase the number of orders on the floor. Congestion increases, the actual lead times again increase, and manufacturing goes back to planning to ask for another lead time extension.

Another disadvantage of the MRP algorithm is that when a capacity infeasible plan is recognized, it does not offer any help as to how to repair the problem. In most cases the user has two options—go with the infeasible plan and hope for the best (an option often adopted when the time available for plan revisions expires), or examine the MPS and try to move production requirements between periods so that the MRP algorithm can generate a feasible release plan. The latter is difficult for even an experienced planner to do, especially for a number of products with large, complex BOMs.

On the positive side, MRP correctly distinguishes between dependent and independent demand and provides a useful framework for requirements planning, the calculation of production requirements for dependent demand items given the BOM structure. MRP is also easy to understand, and its adoption forces companies to systematize a great deal of data about their operations, which is a beneficial exercise in and of itself, regardless of what planning algorithm the data are used in. For better or worse, MRP has become a de facto industry standard against which alternative approaches must be measured.

4.2. Finite Capacity Algorithms

The basic source of many of the problems identified with MRP lies in its fundamentally material-centric view of production planning. Other than the extremely indirect representation through the lead times, MRP does not try to represent manufacturing capacity at all. One would expect the infinite capacity approach to work well when capacity is plentiful and the main concern is to coordinate the flow of work through the factory. However, in environments where capacity is expensive and highly utilized, we would expect the infinite capacity model to be increasingly inaccurate in its predictions of job completion times. It did not take long for these problems to be noted in practice, and a number of extensions to the basic MRP procedure have been developed over the years to provide some form of capacity check on MRP calculations. A variety of *finite capacity* algorithms have evolved out of different attempts to address these deficiencies. These algorithms take various approaches to modeling capacity and to generating the actual production plan but are mostly noncongested. We shall present a broad overview of the basic approaches that have been developed in academia and industrial practice.

4.2.1. Extensions to MRP

The two best-known approaches to adding some capacity checks to the basic MRP calculations are rough-cut capacity planning and capacity requirements planning. Both of these approaches have essentially the same philosophy: to estimate the amount of capacity required at each workcenter in each time period and notify the user of any violations. It is up to the user to decide how to modify the MPS to obtain a capacity-feasible order release scheme. They differ in the amount of data required and the accuracy of the capacity profile generated.

Rough-cut capacity planning (RCCP) is intended to be performed on the MPS before the actual MRP run is made. In this approach, we associate a bill of resources with each item in the MPS. These data specify how much time on the various types of resources a given MPS item requires. The RCCP procedure then multiplies each entry in the bill of resources by the number of units required by the MPS in a given time period to estimate the total workload implied for each resource. Note that this calculation is performed before the MRP run, so the timing of planned order releases is unavailable. Hence, lead time information is not used in RCCP. By the same token, neither does it consider the amount of available inventory in estimating this workload. Hence, the accuracy of the workload predictions made by RCCP is often quite poor, although it may allow the user to identify gross capacity violations before the actual MRP run is made. Its chief virtue is that its data requirements are modest and the computations simple.

Capacity requirements planning (CRP) is performed after the MRP run and essentially converts the planned order releases into a capacity profile. This approach in its pristine form considers lead times for individual processing steps rather than for the production of the BOM item and thus requires a lot more data than RCCP. It is also substantially more time consuming to perform. However, being done after the MRP run, it considers both lead times and inventory status. Unfortunately, the arguments above against constant lead times also hold here, rendering the capacity profile generated increasingly inaccurate at high utilization levels.

Another approach to enhancing MRP is the capacitated MRP (MRP-C) approach proposed by Tardif and Spearman (1997). This approach has a number of advantages: it specifically pinpoints the reason for the infeasibility as being due to the current WIP distribution, which does not allow the work in process to be finished in time to meet demand, or lack of capacity at a specific point in time. The basic idea of this approach is that capacity is explicitly considered while doing the calculation itself. Hence, net requirements are calculated in a given time period and are checked against available capacity to determine whether they can actually be manufactured in that period. If they cannot be produced, then the current plan is infeasible and action must be taken to render it feasible, such as delaying demand or adding overtime. An interesting aspect of this approach is that it identifies infeasibilities as being due to either poor positioning of the WIP in the line or mismatches of demand and capacity in time. The authors show that if their algorithm generates a feasible plan, that plan will minimize the total inventory over the planning horizon. Similarly, they provide algorithms to delay demand or add overtime in a manner that minimizes inventory and lateness costs. This algorithm is of considerable interest in that it integrates the two steps of generating material requirements and checking capacity feasibility that are performed serially under conventional MRP approaches. The fact that it identifies the source of infeasibilities and gives guidance as to what to do about it is also important.

4.2.2. *Optimization Approaches*

There is a long history of optimization models for production planning, almost all of them based on linear programming. The capacity of each workcenter is recognized to be finite in each time period. The planning problem is then that of assigning orders to time periods, subject to a subset of the relevant resource and relational constraints. Those most commonly considered are workcenter capacity, that is, the number of hours available in each time period and the routing of the products through those workcenters. In most cases the lead times at the workcenters are assumed to be independent of workload. Hence these models are classified as finite capacity noncongested. The objective is usually to maximize some combination of revenue and costs.

A simple model of this form to illustrate the basic types of issues that can be modeled is given below. Let the decision variables x_{it} denote the amount of product i to be produced in period t. We will let I_{it} denote the amount of inventory on hand at the end of period t and S_{it} the amount of backlog at the end of period t. The product j produced in period t is assumed to become available τj periods later, that is, we assume a fixed lead time of τ_j for product j. The model will also determine the amounts of regular and overtime labor to be used, which will be denoted by LR_t and LO_t respectively. Costs are also incurred when we increase and decrease our labor force, and are proportional to the amount of the increase or decrease. Denoting the amount of labor force increase or decrease in period t by λ_t^+ and λ_t^-, respectively, we can then state a basic model as follows.

$$\min \sum_{t=1}^{T} \left[\sum_{i=1}^{n} (p_{it}x_{it} + h_{it}l_{it} + w_{it}S_{it}) + C_{Rt}LR_t + C_{ot}LO_t + c_{lt}\lambda_t^+ + c'_{lt}\lambda_t^- \right]$$

subject to

$$NI_{it} = NI_{i,t-1} + x_{i,t-\tau_1} - c_{it}^i \text{ for all } i, j, t$$

$$NI_{it} = I_{it} - S_{it} \text{ for all } i, t$$

$$LR_t = LR_{t-1} + \lambda_t^+ - \lambda_t^- \text{ for all } t$$

$$LO_t - LU_t = \sum_{i=1}^{n} m_i x_{it} - LR_t \text{ for all } t$$

All variables are assumed to be nonnegative. LU_t is a slack variable denoting the amount of excess labor available in period t. Note that LU_t and LO_t cannot both be positive in the same period. The first set of constraints ensures that the inventory levels are consistent across periods, where NI_{it} is the net inventory of product i in period t. The second set of constraints defines the net inventory to be either positive or negative, since both variables on the right-hand side cannot be positive in any optimal solution. The third set of constraints models the evolution of the labor level over time, while the fourth set indicates the relationship among overtime, undertime, and the amount of regular labor on hand.

While this particular model views the production facility as a single stage whose production capacity is limited by the available workforce, it is easy to extend this approach to situations with multiple products following different routings through multiple-stage production systems. This basic model has been extended in many directions, such as the inclusion of variables and constraints related to working capital, marketing, and promotion decisions (Shapiro et al. 1993).

This model considers capacity at an aggregate level, in the sense of recognizing that the number of hours available on a given resource in a given time period is limited, but the modeling of congestion is still inadequate. As in the MRP algorithm, we are assuming that lead times (the τ_j in the above model) are known a priori and independent of the loading of the shop in that time period. While this may well be an acceptable approximation in lightly utilized facilities, it rapidly degenerates as the level of resource utilization increases. A number of authors, notably Leachman and his coworkers (Hackman and Leachman 1989; Leachman 1993) have significantly extended the capability of linear programming models in this regard, allowing lead times to vary over time as the workload in time periods changes. Hung and Leachman (1996) use an iterative approach where the results of the linear program are fed into a simulation model of the production facility to obtain updated lead time estimates, which are then input into the linear program for another run.

The model as stated here is also more suitable for MPS generation in that it deals with end-item demand translated into demand for workcenter capacity. Optimization models that explicitly address the multilevel nature of the BOM have been proposed (e.g., Billington et al. 1983) but are generally mixed-integer programs that are significantly more complex than the model above.

A number of authors, such as Graves (1986) and Karmarkar (1989), have attempted to develop models of manufacturing capacity that reflect the effects of congestion using the idea of clearing functions. As the name implies, a clearing function represents the amount of inventory at a workcenter that can be moved out of it in a given period. While a number of different forms for such clearing functions have been suggested in the academic literature, they remain an unexplored possibility in terms of industrial implementation. A particular area of research is how to obtain such clearing functions empirically for specific industrial scenarios.

4.2.3. Artificial Intelligence Approaches

Another approach that appears to be used in some current planning systems is to model the capacity of at least a subset of near-bottleneck workstations in an aggregate manner, as a total number of available hours per planning period. A plan is constructed by calculating the amount of time each operation of the order will require on each workcenter and assigning each operation to a specific planning period. The lead time between nonbottleneck operations is modeled as infinite-capacity, the idea being that these stations have such low utilization that the infinite capacity model is a reasonable approximation. These procedures tend to use quite sophisticated heuristic search procedures to assign orders to periods. Examples of such algorithms are the ReDS system (Hadavi et al. 1989) and the system developed at Texas Instruments described by Fargher and Smith (1994). Smith (1993) presents an excellent review of artificial intelligence approaches to production planning and scheduling problems, and Zweben and Fox (1994) provide a number of case studies.

4.2.4. Congestion Models

All the approaches considered so far have been based on an aggregate model of manufacturing capacity, in the sense that capacity constraints are checked only at an aggregate level over a specified time period. This clearly does not prevent infeasibilities due to mismatches between the arrival and completion of tasks at the workcenters. A number of procedures have been developed to address this problem by making a detailed scheduling model an integral part of the planning algorithm. These models address the congestion effect directly by modeling the operation of the shop in considerable detail. We shall discuss two basic flavors of this approach, asking the reader to bear in mind that in many implementations the distinction between them may become blurred, with any particular procedure exhibiting characteristics of both groups.

4.2.5. Detailed Scheduling as Part of the Planning Process

As discussed above, a major difficulty in most finite-capacity production planning models is that of modeling the nonlinear effects of congestion at the shop floor, caused by variability in the arrival patterns of jobs to the workcenters. A number of planning systems in both the research community and industry have attempted to address this by making the generation of a detailed schedule of shop-floor operations a part of the planning process. The basic idea is to generate a plan using some planning procedure and then try to construct a detailed schedule that meets the deadlines given by the plan. If such a schedule can be generated successfully, it shows that the plan is capacity feasible. In general, one would not expect the schedule generated during the planning process actually to be executed, as both plan and schedule will be revised. Rather, the function of the schedule is to verify that the proposed plan is indeed capacity feasible in the sense that at least one schedule that satisfies its demands can be constructed.

This approach has been illustrated in a research environment by Dauzère-Pérès and Lasserre (1994), who use an optimization algorithm for production planning. Once a plan has been developed, they use a sophisticated optimization-based scheduling heuristic to build a schedule that meets the deadlines set by the plan. In the event that the deadlines cannot be met, they revise the plan and iterate until a satisfactory solution is reached.

While intuitively attractive, this approach has a number of problems. While the idea of using a scheduling algorithm to verify the capacity feasibility of a proposed plan is attractive, the correctness of the conclusion depends on the quality of the scheduling algorithm used. A simple dispatching heuristic may well be unable to find a schedule that satisfies the plan that a more sophisticated exact solution procedure could identify. Moreover, the essentially discrete nature of the scheduling problem and its well-known computational intractability render the task of generating high-quality schedules very time consuming. This is an important drawback in an environment where the plan must be revised and schedules generated repeatedly before a satisfactory solution is reached.

A common approach widely discussed in industry is to feed the output of an infinite capacity planning system into a detailed discrete-event simulation model of the plant and use the simulation to determine whether the proposed start times will allow all orders to be completed by their due date. If some orders are identified as being late, the plan is modified and the simulation rerun until an acceptable result is achieved. While conceptually attractive, this approach has a number of difficulties. First of all, the time and effort involved in developing and maintaining a detailed simulation model of the facility may be significant. Secondly, the time required to obtain results from the simulation, especially if several replications must be run to obtain statistically valid results, may be quite substantial. Thirdly, if the simulation identifies an infeasibility, the user is often reduced to manual intervention and experimentation to identify a feasible plan, which may be difficult to achieve in the time available for the decision to be made. Finally, this approach requires some fairly detailed assumptions as to how the shop does the scheduling—in effect, the simulation model must generate the schedule as executed on the shop floor to ensure the capacity feasibility of the plan driving the schedule. All in all, this approach may work well in relatively simple manufacturing environments, but is unlikely to be practical in complex multistage environments where the time to make multiple simulation runs is substantial and the effects of changes in the schedules are hard to predict. However, despite the drawbacks of this basic approach, the APS procedures described in the following section are often quite successful using variants of this approach that use scheduling ideas to maintain records of how much capacity is available at each critical resource in a given time period and use this information to drive planning procedures.

5. ADVANCED PLANNING AND SCHEDULING

APS is the process of simultaneously coordinating material and capacity constraints at the operational level to best meet market demand. APS offers the planning function a detailed representation of the production process that was formerly found only in more advanced scheduling systems. As for the scheduling function, it links work orders to customer orders, permitting direct tracking of their progress. Arguably its most important advantage over traditional planning approaches is that material and capacity are *simultaneously* considered as elements that may constrain production. This ensures that the material plan, as it is being generated, is in agreement with the capacity schedule down to the level of individual resources such as machines, and the capacity schedule is in agreement with the material plan throughout all BOMs. This stands in marked contrast to the conventional MRP approach of independently planning material and then subsequently checking this plan against capacity to identify violations.

The computational complexity of this task is quite daunting—to achieve this, one needs to consider detailed scheduling information as well as customer information, essentially creating a detailed schedule in parallel with a production plan to ensure that the plan is indeed achievable. The need to do this for all levels of the potentially complex BOM adds to the difficulty. The computational complexity of scheduling problems on their own is well recognized (Pinedo 1995). Hence, many APS systems do not try to develop plans that are optimal in a rigorous sense. Instead, they are focused on transaction processing, determining at the time the order is placed whether the order can be completed on time or what the earliest possible completion date is if the original request date cannot be met. A complete resynchronization of the plan is then done periodically to optimize resource usage and material plans.

APS systems also differ in how they approach the generation of plans and schedules. In some systems, planning is done periodically in batch mode, with integrated plans being developed for a set of products and workcenters. On the other hand, others are focused on transaction processing, where a plan is constructed and capacity allocated incrementally as orders arrive.

In its pristine form, APS integrates three key processes: advanced planning, advanced scheduling, and order promising.

5.1. Advanced Planning

The goal of advanced planning is the synchronization of constrained material and resources to independent demand. Its purpose is to create a plan that is feasible with respect to all resources required (machines, material, tooling etc.) with sufficient operational slack to permit resequencing of work orders to enhance production efficiency. The independent demand comes from several sources, including customer orders, demand forecasts, master production schedules, transfer orders (i.e., orders from other plants), and the company's policies on safety stock. The advanced planner also considers the work order schedule already released to the shop floor as presented by the advanced scheduler. For each end-item demand, a complete requirements explosion is done using that item's BOM and backward scheduling dependent demand based on component routes, available resource capacity (not simply workcenter capacity), and available and projected inventory. Some advanced planning engines are refined enough to consider resource and material requests at their point-of-use, which can be helpful in environments where long operation times create extended delays for material usage. Typically, the APS horizon is similar to that of conventional planning tools, ranging from several weeks to months.

5.2. Advanced Scheduling

The second APS component, advanced scheduling, involves the detailed sequencing of operations and material in support of the aforementioned plan. Its purpose is to provide properly sequenced work orders, under possibly more refined constraints than those considered in the plan (e.g., sequence-dependent setups, maintenance schedules, more detailed machining constraints, additional operator restrictions) while still attempting to hold to the plan dates. It serves to efficiently load the workcenters and present a more discriminating schedule to the advanced planner. While advanced planning produces a detailed allocation of resources and material to orders, some applications require further refinement, especially where work sequencing can significantly affect workcenter production rates. Advanced scheduling produces a schedule constrained by both material and capacity. This schedule serves as a projection of what the shop floor should be doing and is used in the advanced planner as a basis for component supply. Exceptions to the advanced plan are identified for resolution in the advanced scheduler or for adjustments to the advanced plan. The advanced scheduling horizon tends to be short, as with conventional scheduling tools, but may need to be extended to support better the needs of the advanced planning process.

The reader should keep in mind that one may not need to run an advanced schedule to produce what the shop floor needs for execution. The main difference between the advanced plan and the advanced schedule is not necessarily the detail used in the representation, but rather the amount of emphasis placed on sequencing. If it is enough to determine what needs to be produced over a given time frame (e.g., a shift) without regard to sequence, then the advanced plan may be sufficient. If, on the other hand, the sequence in which this work is done can significantly affect the workcenter's production rate, an advanced schedule with a more refined sequenced execution list is required. This type of situation often arises in manufacturing systems where the decisions at various stages of production are tightly coupled and setup times between products are significant.

5.3. Order Promising

The third component of APS is order promising, which lies at the center of the transaction-based aspect of APS systems. This component is designed to suggest realistic promise dates for customer orders. The process, sometimes referred to as capable-to-promise (CTP), involves testing the customer's request date for feasibility and, if the date cannot be met, calculating the earliest date that it can be met. This is done based on available and projected inventory and available resource capacity.

CTP functions at two levels: disruptive and nondisruptive. Nondisruptive promising uses available capacity and material to determine the order's projected completion date without altering the planned completion times of orders currently in the system. Disruptive promising, on the other hand, reallocates capacity and material to determine the feasibility of meeting a particular date (presumably earlier than is possible under nondisruptive promising) and identifies what orders are affected as a result. In either case, CTP differs significantly from conventional available to promise (ATP), which considers only the uncommitted inventory balance by period. With CTP, the process extends through the bill of materials and part routings to examine the potential for manufacturing the item if it is not available.

The transaction focus of APS systems renders their speed of execution critical to their effectiveness. Taking advantage of the many advances in hardware and software technology over the last decade, APS systems have execution speeds orders of magnitude faster than those of traditional MRP. Most APS engines have their data downloaded, either in batch or transaction mode, to a dedicated server that is architected to run memory-resident programs and databases. Under this scheme, they are able to deliver real-time order promising and make multiple runs to test various actions in an effort to further improve the plan.

6. APS IMPLEMENTATION ISSUES

Compared to MRP alone, APS has a broader functional impact, is less tolerant of inaccurate and incomplete information, requires more data and in greater detail, and affects more people more directly every day. Beyond this, it causes a cultural change. These factors make APS implementations more challenging than MRP. Nevertheless, remaining cognizant of the value received will enable the technical and organizational hurdles that will inevitably appear throughout the course of the APS implementation to be overcome. Areas in which to be particularly vigilant are described below.

6.1. ERP Integration

The comprehensive nature of today's APS algorithms drives the need for copious amounts of data—data that typically reside in an ERP system. This means that attaining the full benefits of APS is largely predicated on how well it is integrated with ERP. When done well, both systems benefit.

The importance of the ERP system is seen in two key roles it plays. First, ERP provides the necessary infrastructure, holding and managing information about orders, parts, resources, and status. It houses the modules that feed and are fed by the APS system, including forecasting, customer order management, product definition, inventory management, purchase order management, work order reporting, dispatching, and costing. When APS is planning and scheduling, it needs to know what orders to consider, what jobs are finished, what work is in process, what purchased materials are coming in and when, and what capacity is available. Secondly, ERP functions as an execution system, firming and releasing manufacturing orders, cutting purchase orders, and communicating schedules. The orders to be considered are the purview of APS; the actions themselves are the purview of ERP.

This need for ERP integration does, however, pose a number of challenges. First of all, the full benefits of APS can only be realized through a two-way exchange. Closed-loop functionality gives APS the data it needs to plan and schedule and gives ERP the information it needs to take appropriate action. The amount of data transferred and the numerous ERP modules affected require a complex web of communications, as illustrated in Figure 3. Another challenge is the need to disable specific functionality. Some traditional ERP modules, being superseded by APS, are no longer relevant and require circumvention. This must be done with ERP vendor expertise, for these modules often contain utilities that are still germane to the process. This can raise issues of functional ownership.

Finally, determining the type of APS system to be adopted, stand-alone or preintegrated, also deserves some thought. Stand-alone systems, using straightforward handshakes through specially written programming utilities, are less intrusive, require less vendor involvement whenever customized changes are required, and offer well-targeted value. Preintegrated systems, on the other hand, provide a broader range of functionality and are in proven agreement with the ERP system (and in some cases actually share the same database). They may also offer a better upgrade path as improvements in both systems are made. Preintegrated systems also lower the cost and time to implement, thus improving the return on investment.

6.2. Timing, Access, and Quality of Data

6.2.1. Timing

In general, communications between ERP and APS are asynchronous. This means that data are sent to the associated system and no immediate response is required. This is particularly true in systems where the APS engine is transaction oriented. In these systems, updated information is sent on a continual basis in anticipation of a new plan or schedule being generated. Then, at the time of execution, APS knows the current conditions without requiring a time-consuming download, as may be the case with batch-oriented systems. The timing of a full replan or reschedule is often dependent on when the shop floor has consistently reported across the facility (e.g., end of shift) or when a major disruption has occurred (e.g., machine failure). More frequent planning and scheduling is possible with today's technology but is seldom realized given the requirement for the reported information to be consistent.

In the case of order promising, the demands on this timing change. Here, when an order is entered, it triggers the APS engine for the expected completion date and an immediate response is expected. This is a synchronous process. Moreover, as new orders are entered, they need to be promised based on the latest information, including those orders that have just been accepted. Thus, the planning system, upon order acceptance, must immediately reserve materials and capacity so that all future promises can reflect the impact of even the most recently accepted orders.

6.2.2. Access

Planning and scheduling often differ in their data requirements. Capturing some of the subtleties of the production process may be pertinent to scheduling because it seeks to refine the sequence of events, whereas planning may not need to be as precise and thus may not require this same level of detail. In practice, the information required to portray these scheduling subtleties properly usually resides outside the ERP system. For example, traditional ERP systems typically define the manufacturing process at the workcenter level. The scheduling engine, on the other hand, may be forced to examine the detailed differences among the various machines within a workcenter. When more refined information such as this is required, it is imperative that the scheduling system have access to sources of information outside the traditional footprint of ERP or at least be able to circumvent, as necessary, the ERP data definition.

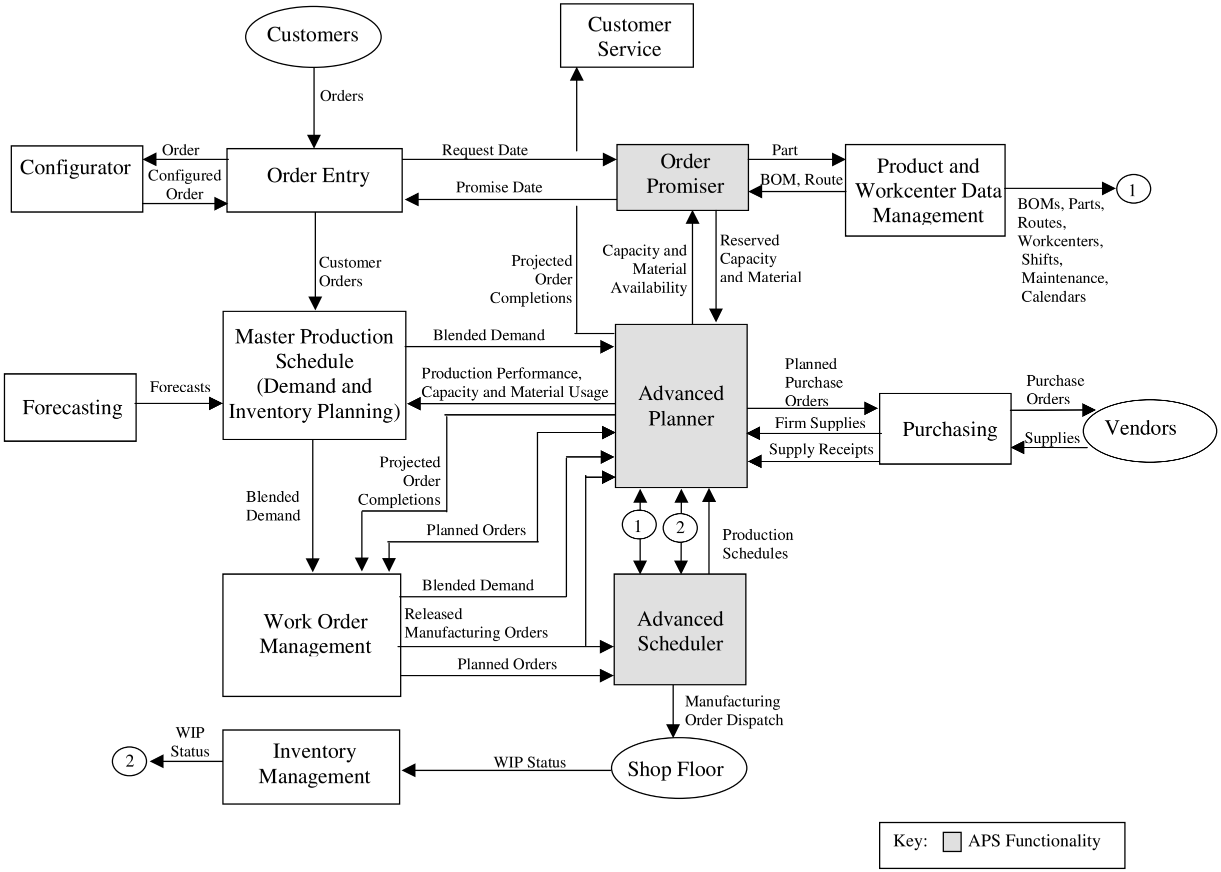

Figure 3 APS/ERP Integration.

6.2.3. *Quality*

Data quality strongly influences APS results. This is especially true of shop-floor data, which are not always available at the level, at the time, or in the form most appropriate for planning and scheduling.

Work-in-process (WIP) accuracy is particularly problematic. The data-collection system may be, for example, recording only an order's closed operations. Time remaining on partially completed operations may not be known, thus forcing the APS engine to replan and reschedule the operation as if nothing has been done. The significance of this obviously depends on the operation's duration. Similarly, backflushing, which records inventory changes based on assumptive issuing at the end of the process vs. discrete issuing at time of use, can also pose a problem, for it fails to record an order's intermediate progress. If an order's route requires key workcenters for extended lengths of time, this method of capturing status can lead to an erroneous representation of important near-term capacity. Fortunately, advances in data-collection systems have made it easier and more cost effective to track WIP.

Reporting frequency can also adversely affect results. Work accomplished between reports is not recognized until reported. The significance of this misrepresentation increases as the reporting interval widens. More frequent reporting is desirable.

However, frequent reporting does not guarantee valid and consistent status information. False reporting and an inability to hit an order's exact quantity levels at an operation can also mislead the APS engine. As a result, some filtering and interpretation of the shop-floor information may be required. Moreover, where consistent information is a problem (e.g., a downstream station reporting work on an order for which an upstream station has yet to acknowledge), the frequency of the APS resynchronization is forced to match those points in time when the information is most likely to be in agreement across the facility, such as the end of a shift.

Startup conditions (i.e., order and machine status) form yet another critical data source for the planning and scheduling process. As the APS engine is being asked to provide nearer-term dispatch lists (e.g., the next eight hours), the status of resources and WIP at the beginning of the planning and scheduling horizon takes on added importance. This is particularly true in applications where sequence-dependent setups are consequential, for failing to capture a machine's initial state can dramatically alter the resulting sequence. A false start can have a lasting impact.

6.3. Business Process Reengineering

Business process reengineering (Hammer and Champy 1994) is the single biggest factor affecting the success of APS and is what ultimately governs the company's true return on investment. Technology alone cannot guarantee APS success. While APS affords a company the opportunity to improve its order fulfillment process dramatically, it can only happen if the business processes change to accommodate and exploit it. There needs to be a conceptual match between these business processes and the APS system. When the current processes are based on antiquated planning and scheduling practices, changes are in order.

Unfortunately, these changes are not always easily understood and take time to implement. The new planning and scheduling paradigm of APS cuts across organization boundaries and threatens the company's traditional, and now obsolete, mode of operation. The real challenge becomes changing the way people currently think and operate. Confidence is required to move forward. It takes a strong sense of need to implement the procedural changes required. Enlistment of top-level management may be necessary to overcome the pushback that occurs when people are asked to change how they work.

Consequently, the implementation team should take time, in the early stages of the project, to understand better the business processes that will need to be changed in sales, customer service, purchasing, engineering, production planning and scheduling, and manufacturing. Sales will be challenged with sobering promise dates and with reconciling, at the time of order entry, a customer's request with the realities of the production plan. Customer service will wrestle with realistic (and variable) customer order projections. Purchasing will be immediately impacted by sales, and conversely, sales by purchasing. Engineering will be directed to keep BOMs more up to date and more in line with how items are actually built and could be asked to support a configurator to enable order-entry clerks to take direct advantage of the order promise capability. Planning and scheduling will be challenged with having a more comprehensive and integrated view of the production process and with needing improved communications with various departments to better induce the necessary and frequent changes. Manufacturing will be asked to be more disciplined (i.e., follow the schedule and report status) and more flexible (i.e., heedful of new orders and changes to existing orders). In addition, the dynamic reallocation of work will be commonplace.

6.4. A Well-Defined Manufacturing System

Importance must also be placed on properly representing the manufacturing process in the APS engine. This means the manufacturing data need to be brought up to a level of accuracy commensurate with the APS task. As the manufacturing data deviate from this ideal, the effectiveness of the APS engine (or any other planning and scheduling tools) diminishes accordingly. The data that are most relevant to this process include:

6.4.1. As-Manufactured, Indented Bills of Material

If the item to be produced is not correctly defined, the APS engine is not able to plan and schedule it properly. The issue is usually not whether it is defined (unless in a configure-to-order environment), for most manufacturers have BOMs. Rather, the issue is *how* it is defined. Does an item's BOM reflect the sequence in which that item is to be manufactured, or does it simply represent a listing of what components are required or how the item was engineered? Do the bill's components mirror only those items of importance for planning and scheduling, or do they include a complete listing of all the items in the bill, regardless of their relevance?

6.4.2. Accurate Routes

An item's route directly controls when the requisite resources and materials used in its production are engaged. Obviously, correct identification of these resources and materials is important. This does not mean, however, that every resource or material used in the item's production needs to be identified; only if it supports the representation for purposes of planning and scheduling. Readily available material and secondary as well as tertiary resources at an operation are often not necessary. Furthermore, the operation times themselves are important, for they not only hold the resources for the correct apportionment of time, but they also influence when the other associated operations in the route request their resources and materials.

6.4.3. Supportive Operational Buffer Times

The APS plan may need to allow for the resequencing of work at an operation. When this is the case, operational or resource buffer times are used. This gives a certain degree of latitude to the advanced scheduler for work order resequencing. It is important to set these resequence buffers large enough to allow for some scheduling adjustment, but not so large as to misrepresent the item's lead time.

6.4.4. Consistent Workcenter Definitions

Often, workcenter definitions in ERP reflect a costing orientation more than a planning and scheduling orientation. This can substantially limit the options available to the APS engine. For example, instead of recognizing the more complete set of machines that can actually do a particular operation, the route may only identify a specific (e.g., least-cost) machine, thus misrepresenting the true scheduling options available and potentially artificially extending the time it should take to produce the item. When this is the case, circumvention of the ERP workcenter definition may be required.

6.4.5. Representative Purchasing Lead Times

Purchasing lead times are often inflated to trigger early action by purchasing. In a materially constrained APS application, this can cause overly conservative projections. Flagging an item for early purchase is different than constraining on it. Under a nonconstraining paradigm like MRP, cautious lead times are acceptable, for the purpose is to expose potential problems. In a constraining paradigm like APS, more aggressive (i.e., shorter) lead times generally present a more realistic picture of what is possible.

6.4.6. Appropriate Workcenter and Material Constraints

Bottleneck resources and critically scarce materials need to be modeled in a way that reflects their limited availability, for they are largely responsible for setting the manufacturing flow rate. Failure to represent these limitations properly can render the resulting plans and schedules useless. Nonbottleneck resources and noncritical materials, on the other hand, need not be represented to this same level of detail. This can save on execution time and obviate the need for their precise representation in the model.

6.4.7. Representative Scheduling Rules

Seldom does a manufacturer operate according to one rule (e.g., due date). Rather, each machine or workcenter has its own individualized set of rules (e.g., highest priority, then dynamic slack, and lastly minimum setup). Taken collectively, these rules define the company's manufacturing strategy. Correctly defining these rules and capturing them in the scheduling engine not only ensures valid schedules but also helps APS gain acceptance by those responsible for executing these schedules.

6.5. Usual Suspects

Some APS systems are particularly challenged to represent or work with certain aspects of a manufacturing process adequately. This is not to say it cannot be done, just that it may require additional thought, necessitate extensions to the software, or run counter to their planning and scheduling paradigm. Examples of these challenges include batching, outsourcing, nonsequential operations, flexibility, and early order-release strategies.

Batching, in a discrete process, refers to the combining of like orders for the purposes of processing them together through a portion of their routes. This often surfaces in applications where there are painting or heat-treat operations. The difficulty lies not only with synchronizing these orders appropriately but also with maintaining their unique identities (for these orders may be routed separately after their batched sequence has completed), establishing the batch criteria (e.g., item attributes, timing constraints), associated batching rules (e.g., number in a batch, sum of the batched items' attributes), and gaining access to the requisite information to form the batch.

Outsourcing refers to the practice of having certain operations (or parts) done outside the plant. From a planning perspective, this can be effectively captured. The appropriate delay is simply factored into the item's route. The problem arises when a particular operation or set of operations is done outside the plant before the item reenters for its final set of operations. Precise scheduling of these remaining operations is predicated on having accurate status information from the outside source.

Routing operations that can be done in any order are defined as nonsequential. To bring some structure to the process, an ERP system assumes a specific operational sequence. Since the APS system receives its routing information from the ERP system, the same implied sequence is applied in APS. To function otherwise would require additional information outside the traditional bounds of the ERP system. Nonsequential operations also give rise to significant combinatorial issues.

Most often, the issue of flexibility arises with operators. As more latitude is given to dynamically reassigning people on a line, the demands on the planning and scheduling system increase accordingly. The rules on when to engage a change in assignment and what to change to can be quite complex. Moreover, depending on the number of people assigned, the station times on the line may change. Capturing these dependencies in the planning and scheduling model can be an arduous task and should only be considered when commensurate value results.

Releasing work orders as late as possible is in the best interest of the manufacturing facility and APS. Later releases commit manufacturing at the last possible moment and give planning more flexibility to adjust to ever-changing conditions. Keeping the planning engine in control for as long as possible increases the ability to satisfy "drop-in" demand. Unfortunately, many ERP systems require the order to be released before it can be changed. This forces orders to be released early and prevents the planning engine from making last-minute adjustments, since manufacturing is now committed to these orders.

6.6. Implementation Strategies

APS implementation can be a daunting task. The sophistication of the software, the effort to get the necessary data to effectively drive the system, and the need to make fundamental procedural changes across the organization present significant challenges. Moreover, the people involved, their understanding of the business needs, and their willingness to change vary. This makes every APS implementation a unique endeavor. Aside from general project-management guidelines, little is offered to address the challenges that are faced.

Two vastly different APS implementation approaches have been advanced: big bang and evolutionary. With the big bang approach, the strategy is to leap to a complete implementation as quickly as possible. Emphasis is placed on time to value more than fidelity. The approach initially uses the data as it is currently defined rather than going through an extensive modification process to better support the requisite APS functionality. This makes it most appropriate in those environments where the manufacturing data are well defined. With this approach, reliance on the APS system occurs with deliberate speed, immediately exposing opportunities for improvement and abruptly forcing people to work within the new system to effect change. Because of this, procedures need to be in place to quickly and regularly maintain the APS data. This approach usually results in a faster return on investment, but with considerably more trepidation.

With the evolutionary approach, the APS system is initially set up to honor just the BOM constraints, leaving resources and purchased parts unconstrained. If good routing data are not initially available, standard manufacturing lead times are invoked. Under this configuration, the APS system mimics an order-based MRP system. Then, as routes are introduced, key workcenters are constrained, giving an improved representation of the process. As understanding is gained, more constraints are added, such as other workcenters or key purchased parts, until a realistic representation is achieved. This approach, which tends to be more palatable in a proven MRP environment, provides for a more gradual and systematic implementation but extends the time to reach full-value APS functionality and runs the risk of stopping short of this goal.

7. CONCLUSIONS

We have attempted to present a review of the basic planning algorithms, their strengths and weaknesses, and the role of APS in integrating the planning, scheduling, and demand management functions more closely. The issues to be addressed in implementing APS systems have also been addressed. The field of production planning and control has been extremely active in the last several years, and it promises to remain a challenging and interesting area for both researchers and practitioners in the decades to come.

REFERENCES

Billington, P. J., McClain, J. O., and Thomas, L. J. (1983), "Mathematical Approaches to Capacity-Constrained MRP Systems: Review, Formulation and Problem Reduction," *Management Science*, Vol. 29, pp. 1126–1141.

Dauzère-Pérès, S., and Lasserre, J.-B. (1994), *An Integrated Approach in Production Planning and Scheduling*, Lecture Notes in Economics and Mathematical Systems, Springer, Berlin.

Elmaghraby, S. E. (1991), "Manufacturing Capacity and its Measurement: A Critical Evaluation," *Computers and Operations Research*, Vol. 18, pp. 615–627.

Fargher, H. E., and Smith, R. A. (1994), "Planning in a Flexible Semiconductor Manufacturing Environment," in *Intelligent Scheduling*, M. Zweben and M. Fox, Eds., Morgan Kaufmann, San Francisco.

Goldratt, E., and Fox, R. (1986), *The Race*, North River Press, Croton-on-the-Hudson, New York.

Graves, S. C. (1986), "A Tactical Planning Model for a Job Shop," *Operations Research,* Vol. 34, pp. 522–533.

Hackman, S. T., and Leachman, R. C. (1989), "A General Framework for Modeling Production," *Management Science*, Vol. 35, pp. 478–495.

Hadavi, K., Shahraray, M. S., and Voigt, K. (1989), "ReDS: A Dynamic Planning, Scheduling and Control System for Manufacturing," *Journal of Manufacturing Systems*, Vol. 9, pp. 332–344.

Hammer, M., and Champy, J. (1994), *Reengineering the Corporation: A Manifesto for a Business Revolution*, Harper Business, New York.

Hendry, L. C., and Kingsman, B. G. (1989), "Production Planning Systems and Their Applicability to Make-to-Order Companies," *European Journal of Operational Research*, Vol. 40, pp. 1–15.

Hopp, W., and Spearman, M. L. (1996), *Factory Physics*, Irwin, Chicago.

Hung, Y-F., and Leachman, R. C. (1996), "A Production Planning Methodology for Semiconductor Manufacturing Based on Iterative Simulation and Linear Programming Calculations," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 9, pp. 257–269.

Karmarkar, U. S. (1987), "Lot Sizes, Lead Times and In-Process Inventories," *Management Science*, Vol. 33, pp. 409–423.

Karmarkar, U. S. (1989), "Capacity Loading, Release Planning and Master Scheduling with WIP and Lead Times," *Journal of Manufacturing and Operations Management*, Vol. 2, pp. 105–132.

Leachman, R. C. (1993), "Modeling Techniques for Automated Production Planning in the Semiconductor Industry," in *Optimization in Industry*, T. A. Ciriani and R. C. Leachman, Eds., John Wiley & Sons, New York.

Monden, Y. (1983), *Toyota Production System*, Industrial Engineering and Management Press, Norcross, GA.

Nahmias, S. (1993), *Production and Operations Analysis*, 2nd Ed., Richard D. Irwin, Homewood, IL.

Orlicky, J. (1975), *Material Requirements Planning: The New Way of Life in Production and Inventory Management*, McGraw-Hill, New York.

Pinedo, M. (1995), *Scheduling: Theory, Algorithms and Systems*, Prentice Hall, Englewood Cliffs, NJ.

Pritsker, A. A. B., and Snyder, K. (1997), "Production Scheduling Using FACTOR," in *The Planning and Scheduling of Production Systems*, A. Artiba and S. E. Elmaghraby, Eds., Chapman & Hall, London.

Ptak, C. A., and Schragenheim, E. (1999), *ERP: Tools, Techniques and Applications for Integrating the Supply Chain*, St. Lucie Press, St. Lucie, FL.

Shapiro, J. F. (1993), "Mathematical Programming Models," in *Handbooks in Operations Research and Management Science,* Vol. 4, *Logistics of Production and Inventory*, S. C. Graves, A. H. G. Rinnooy Kan, and P. Zipkin, Eds., North-Holland, Amsterdam.

Smith, S. F. (1993), "Knowledge-Based Production Management: Approaches, Results and Prospects," *Production Planning and Control* Vol. 3, pp. 350–380.

Tardif, V., and Spearman, M. L. (1997), "Diagnostic Scheduling in a Finite-Capacity Environment," *Computers and Industrial Engineering*, Vol. 32, pp. 867–878.

Tayur, S., Magazine, M., and Ganesham, R., Eds. (1998), *Quantitative Models for Supply Chain Management*, Kluwer, Dordrecht.

Vollman, T. E., Berry, W. L., and Whybark, D. C. (1988), *Manufacturing Planning and Control Systems*, 2nd Ed., Richard D. Irwin, Homewood, IL.

Zweben, M., and Fox, M., Eds. (1994), *Intelligent Scheduling*, Morgan Kaufmann, San Francisco.

ADDITIONAL READING

Baker, K. R., "Requirements Planning," in *Handbooks in Operations Research and Management Science, Logistics of Production and Inventory*, S. C. Graves, A. H. G. Rinnooy Kan, and P. Zipkin, Eds., North-Holland, Amsterdam, 1993.

Bermudez, J., "Advanced Planning and Scheduling Systems: Just a Fad or a Breakthrough in Manufacturing and Supply Chain Management?" Report on Manufacturing, Advanced Manufacturing Research, Inc., Boston, December 1996.

Johnson, L. A., and Montgomery, D. C., *Operations Research in Production Planning, Scheduling, and Inventory Control*, John Wiley & Sons, New York, 1974.

Leachman, R. C., Benson, R. F., Liu, C., and Raar, D. J., "IMPReSS: An Automated Production Planning and Delivery Quotation System at Harris Corporation—Semiconductor Sector," *Interfaces*, Vol. 26, pp. 6–37, 1996.

Venkataraman, R., "Frequency of Replanning in a Rolling Horizon Master Production Schedule for a Process Industry Environment: A Case Study," *Production and Operations Management*, Vol. 5, pp. 255–265, 1996.