December 2008

Uncertainty and Production Planning

Stephen C. Graves
MIT
77 Massachusetts Avenue, E53-347
Cambridge MA 02139-4307
sgraves@mit.edu

The intent of this chapter is to review and discuss how uncertainty is handled in production planning. We describe and critique current practices and then prescribe possible improvements to these practices. In particular, we argue that there are a set of tactical decisions that are critical to the proper handling of uncertainty in production planning. We observe that current planning systems do not provide adequate decision support for these tactical decisions; we regard this shortcoming as an opportunity for new research and development, which could significantly improve the practice of production planning.

The chapter is based primarily on personal observations of production planning practices in a variety of industrial contexts. As such it is written more as an essay than as a scientific research article. We make no effort to review or survey the research literature on production planning under uncertainty; we do provide a commentary on the research literature, and identify a few points of entry for the interested reader. We also cite a few illustrative references, albeit primarily from our research. Similarly, our observations on practice are not the outcome of a carefully designed field study, but rather are derived from a potpourri of projects over many years. Again, our intent is to provide a framework and set of observations on current practice, and to provoke some new thinking on how we might do better.

We organize the chapter into five sections. In the first section, we briefly describe and discuss major sources and types of uncertainty, how these uncertainties get realized and how these uncertainties affect the production plan.

In the second section, we provide a commentary on the research in production planning as it relates to the theme of this chapter. We note that much of the research literature is based on deterministic models, and does not explicitly account for uncertainty. With regard to the research that does include uncertainties, we discuss its applicability and the challenges to transfer to practice.

In the third section, we introduce a stylized framework for describing current production planning systems; the intent here is to provide a basis for the subsequent discussion and critique of production planning under uncertainty.

In the fourth section, we make a series of observations on the generic treatment of uncertainty in production planning in practice, and note that most systems for production planning do not recognize or account for uncertainty. Yet these systems are implemented in uncertain contexts; thus the planning organizations need to develop coping strategies. We describe and critique the most common coping strategies.

In the final section, we identify a set of tactical decisions that we view as being critical for handling uncertainty in production planning. We describe how these tactics can be incorporated into production planning systems as proactive countermeasures to address various forms of uncertainty. We provide perspective on the key trade-offs in making these decisions, and identify both examples from the research literature of relevant work, as well as opportunities for new research on developing effective decision support for these tactics.

Types of Uncertainty

In this section we identify and discuss the three major types of uncertainty that arise in manufacturing contexts and that can affect a production plan. We contend that the production plan needs to account in some ways for these uncertainties. In subsequent sections we discuss and critique the research literature and current practice, and then will propose possible tactics for handling these uncertainties.

Uncertainty in demand forecast

This is usually the largest source of uncertainty. All production plans rely on a demand forecast or a demand plan as an input. A demand forecast extends over a multi-period planning horizon and represents the firm's best guess at the future demand. The forecast is based on a combination of inputs, which depend on the context. These inputs include a projection of historical demand data, as might be done by a statistical forecasting package; advanced orders in contexts where at least some of the production is make-to-order; a corporate demand plan for firms that operate with a sales and operations planning (S&OP) process; any customer forecasts that the customer is willing to share with the firm; and market intelligence, often in the form of expert judgments.

As a firm gets more and better information about future demand, it updates the forecast. Indeed, in most planning systems, there is regular cycle in which the forecasts are moved forward and revised; for instance, at the start of each week, a new forecast for demand over the next thirteen weeks is released. The new forecast might reflect information inferred from the observed demand since the last forecast update, any changes to customer orders or forecasts, as well as any changes on the market outlook.

Forecasts are never perfect, and the actual demand realization will differ from the forecast, resulting in a forecast error. To address the uncertainty due to forecasts, we need to characterize the forecast errors. Typically we will view the forecast errors as random variables for which we will want to know (at least) the first two moments. It is important to recognize here that the forecast for a particular product is usually a vector of forecasts, which cover the planning horizon. That is, at any time t, for each product we have a forecast for future time periods t+i, for i = 1, … H, where H is the planning horizon. Thus, we have H forecasts; for instance if we have weekly time periods, then we have a one-week forecast, a two-week forecast and so on. We then need to characterize the errors for each type of forecast, as each forecast has a different impact on the production plan.

Uncertainty in external supply process

A second type of uncertainty is associated with the external supply process. A production plan results in orders placed on outside suppliers. Furthermore, the production plan has expectations on the fulfillment of these orders. That is, a plan might initiate an order for ten steel plates of a certain dimension and grade, and then expect that these plates will arrive and be available for processing according to a stated lead time of, say, eight weeks. Nevertheless, there can be uncertainty in the delivery date due to uncertainty and capacity constraints in the supplier's manufacturing and distribution processes; for instance, the order might not arrive until ten weeks due to a work stoppage or delays attributable to the weather. Furthermore, in many contexts there can be uncertainty in the amount of the delivery. For instance, a supplier might be permitted by

contract to deliver plus or minus 10% of the amount ordered; in other contexts, the buyer might reject some portion of the delivery due to quality considerations. To model this uncertainty, one needs to characterize the uncertainty in the replenishment lead times and in the replenishment quantities.

Uncertainty in internal supply process

A production plan needs also to account for uncertainty in the internal supply process, which is similar to the uncertainty in the external supply process. A production plan results in work or job orders placed on the internal manufacturing, transportation and supply processes. Furthermore, the production plan has expectations on how these processes will perform. That is, a plan might set the number of wafer starts into a semi-conductor fabrication (fab) facility with expectations on both the yield from these wafers and the flow time or process duration for these wafers within the fab. Again, there is uncertainty on both accounts. The actual flow or completion time will deviate from the expectation depending upon the work-in-process in the shop, the equipment availability and the dispatch rules; the wafer yield is inherently random and depends on numerous process factors and conditions. Again, one needs to characterize the uncertainty in the flow or process lead times and in the yield quantities for each process step.

Observations on Production Planning Research

In this section I provide general comments on the operates-research-based research in production planning; I do not attempt a survey of the literature, but try to provide relevant references for a few entry points into this vast literature.

Deterministic models

The dominant thrust of the research literature has been the formulation of deterministic models for production planning, and the development of solution procedures, both optimal and heuristic, for these models. The primary intent of these models has been to specify the requirements for a feasible production plan and to capture the key cost tradeoffs that depend on the production plan. Typically a feasible production plan is one that satisfies the given demand over the planning horizon with no backorders or lost sales, that abides by specified production recipes for each final product, and that does not violate any capacity constraints. The models attempt to optimize total costs, somehow defined. Costs will often include: sourcing and production costs, including setup-related costs; holding costs for pipeline inventory and cycle stock; costs for adjusting production capacity, such as hiring and overtime costs; and logistics costs for transportation and warehousing. For reviews of this research, we suggest Thomas and McLain (1993), Shapiro (1993), Graves (2002).

This literature is largely oblivious to uncertainty. Much like research on the economic-order-quantity (EOQ) model, the contention is that the value of these models is in optimizing critical cost tradeoffs, often in the context of tight constraints. The research perspective is that dealing with uncertainty is of secondary importance to getting the

tradeoffs right; furthermore, there is the assumption that the uncertainties can be handled by other measures, which are independent of the determination of the production plan. Nevertheless, there is also the recognition that the deterministic assumptions are a shortcoming of this research, but were necessary in order to keep the models tractable.

Hierarchical production planning

Hax and Meal (1975) introduced hierarchical production planning (HPP) as a framework for production planning and scheduling, motivated by the desire to create an applicable structure for developing effective planning systems. A hierarchical approach partitions the production planning problem into a hierarchy of subproblems, often corresponding to the organizational hierarchy of the planning organization. In any planning period the subproblems are solved sequentially, with solutions from upper-hierarchy subproblem(s) imposing constraints on the lower hierarchy subproblem(s). The planning system implements only the solutions for the immediate period, and re-solves the subproblems each period in a rolling horizon fashion. See Bitran and Tirupati (1993) for a review of research literature on HPP, and Fleischmann and Meyr (2003) for a review of HPP and advanced planning systems.

The literature identifies three advantages for HPP relative to the alternative of solving a monolithic problem: it is computationally simpler; depending on the formulation, it can have less onerous data requirements; and it has implementation advantages to the extent that the subproblems are aligned to the hierarchy of decision makers.

The HPP approach has primarily been applied to deterministic models for planning and scheduling problems. As such, it is subject to the same criticisms raised in the prior sub-section. Yet, the approach would seem to have an advantage in the consideration of uncertainties, in that it might be possible to tailor the lower-hierarchy subproblems to account for uncertainty, e.g., short-term demand uncertainty. Indeed, there is some research along this premise: Bitran, Haas and Matsuo (1986), Lasserre and Mercé (1990) and Gfrerer and Zapfel (1995). In each case the research attempts to develop within a HPP framework a deterministic aggregate plan that is robust to item-level demand uncertainty; that is, as the item-level demand uncertainty is realized, there is some assurance that the lower hierarchy subproblem can disaggregate the aggregate plan into a good detailed schedule. This is an interesting line of research, but so far has been limited to fairly specific, single-stage production contexts.

Production planning with quadratic costs

One of the earliest production-planning modeling efforts was that of Holt, Modigliani, Muth and Simon (1960), who developed a production-planning model for the Pittsburgh Paint Company. They assume a single aggregate product, and have three sets of decision variables for production, inventory and work force level in each period. More notable are their assumptions on the cost function, entailing four components. The regular payroll cost is a linear function of the work force level. The second component is the hiring and layoff costs, which are assumed to be a quadratic function in the change in work force from one period to the next. The production cost is also modeled by a quadratic function. HMMS assume for a given work-force level that there is an ideal production target and

that the incremental cost of deviating from this target (representing either overtime or idle time) is symmetric and quadratic around this target. Finally they model inventory and backorder costs in a similar way. The inventory and backorder cost is a quadratic function of the deviation between the inventory and an inventory target that depends on the demand level.

With these cost assumptions, HMMS minimize the expected costs over a fixed horizon, where the expectation is over the demand random variables. The analysis of this optimization yields two key and noteworthy results. First, the optimal solution can be characterized as a *linear decision rule*, whereby the aggregate production rate in each period is a linear function of the future demand forecasts, as well as the work force and inventory level in the prior period. Second, the optimal decision rule is derived for the case of uncertain demand, but only depends on the mean of the demand random variables. That is, we only need to know (or assume) that the demand forecasts are unbiased in order to apply the linear decision rule.

This research stands out from other production planning research in that it explicitly allows for uncertain demand, and it develops an easy-to-implement plan, namely the linear production rule. However this line of research has fallen out of favor for a couple of reasons. One is discomfort with the assumptions on the cost functions. Two is that the simplicity of the result depends on the restriction to one aggregate product with a single capacity; the form of the production plan gets more complex with more products or resource types.

Stochastic programming

Over the past ten to fifteen years there has been an increasing interest in adding uncertainty to production planning models. Mula et al. (2006) provide an extensive review of this research. A good portion of this research examines the applicability of stochastic optimization, particularly stochastic programming methods, to production planning models. Stochastic programming is notoriously computational–intensive for many problem contexts; production planning is no exception. Yet, with the ever-increasing computational power and the improvements in our optimization algorithms, there has been more exploration of the feasibility of using stochastic programming for production planning. Escudero et al. (1993) show how to formulate a multi-stage stochastic program for production planning and explore its computational implications. Graves et al. (1996) report on the application of two-stage stochastic programming for the optimization of production and supply chain planning for the Monsanto Crop Protection business.

Whereas stochastic programming methodology has promise as a methodology for capturing uncertainty, it also has significant limitations. In many contexts, it remains computationally prohibitive when there are many periods in the planning horizon and frequent re-planning. Also, the resulting scenario-based production plans can be difficult to communicate and hence difficult to implement.

A Generic Framework for Production Planning

In this section I present a highly-simplified generic framework to describe current production planning practices. My intent in introducing this framework is to create a "straw man" on which to comment and present some observations about planning practices. In the following section, I will use this framework to characterize and critique how many planning systems address their uncertainty.

Almost all text books that discuss production planning provide a framework that lays out the various elements of a planning system (e.g., Hopp and Spearman, 2007; Nahmias, 2008; Silver et al., 1998). In Figure 1 I provide a stylized version, given my intent as discussed above.

A planning system starts with a forecast of future demand over some forecast horizon of length H periods. The long-term portion of this forecast is an input into a capacity and/or aggregate planning module that assesses whether or not there is sufficient capacity to satisfy the demand forecast. This is the aggregate capacity plan as it is usually done using aggregate products and large time buckets, and must account for the key capacity considerations within the manufacturing system and/or supply chain. To the extent that there is a mismatch between the available capacity and the long-term demand forecast, the module needs to examine and decide how to rectify this gap.

In general there are four common ways that this might be done. First, in contexts with seasonal demand, one might develop an aggregate production plan that builds inventory during the off season in anticipation of a seasonal demand peak. Second, the mismatch might be addressed by adding to or augmenting current capacity, for instance through overtime or subcontracting. Third, when it is not possible to build anticipation

inventory or add capacity, one would delay meeting demand. This is usually done in terms of extending the backlog by quoting longer and longer delivery lead times to customers. Finally, there may be some downward revision of the forecast so as to eliminate the gap between the firm's supply capacity and the anticipated demand; this might be an outcome from a sales and operations planning (S&OP) process that equates the forecast to a sales plan and then aligns the sales plan to the production plan. In practice, a firm would rely on a mixture of these tactics in developing its capacity plan.

The next step in Figure 1 is to convert the forecast over the shorter term into a detailed master production schedule, subject to the guidelines or constraints from the aggregate capacity plan. The aggregate capacity plan determines, at a gross level, how and when customer demand is met. Given the aggregate capacity plan and the current finished goods (FG) inventory, the master production schedule determines the necessary production output to meet the demand forecast over the short term. Relative to the aggregate capacity plan, the master production schedule is at a much more detailed level, both in terms of products and time periods. In some contexts, production lot or batch sizing is done as part of the master scheduling, so as to account for economies of scale in the production process. Also, there is sometimes an additional iteration to check the feasibility of the master schedule relative to the available capacity.

To determine the inputs into the production system, we need to convert the master production schedule into a plan for the raw materials (RM) and intermediate products. This is traditionally done with the logic of materials requirement planning (MRP), based on a bill of materials and process recipes for each final product. The key assumption is that we have planned lead times for each required activity to produce the product; that is,

for the procurement of each raw material, as well as for each process and transportation step, we assume that the activity takes a known deterministic amount of time, termed the planned lead time. With this assumption, we can readily translate the master production schedule into time-phased requirements for each of the raw materials, intermediate products and subsystems required to produce each final product. For the raw materials, these requirements trigger replenishment requests from outside suppliers. For the intermediate products and subsystems, these requirements are input to the shop floor control system, which determines the work releases into production and the job priorities throughout the production system. Again, these decisions are based on and guided by the planned lead times for the process steps in the production operation.

The MRP step might also account for lot sizing considerations, whereby a lot sizing heuristic or algorithm is applied for each process step or component in the product bill of material. In some planning systems, this step of the planning process might consider some capacity constraints on production, usually by means of heuristics such as a forward loading scheme. However, even in these cases, the system must rely to some extent on planned lead times in order to coordinate the replenishments for multiple components and sub-systems whose requirements derive from the demand for an assembly product or multi-product order.

Observations on Common Approaches for Handling Uncertainty

In this section I provide a series of observations on how uncertainty is addressed in practice by many planning systems, where we will use the framework from the prior section. Admittedly, these observations are largely anecdotal but are based on a large set of industry-based projects and other interactions.

Most systems for production planning do not explicitly account for uncertainty. The planning system operates as if its world were deterministic. That is, plans are created based on assumptions that the forecast is perfect, the internal production processes are perfect and the outside suppliers are perfect. The plans presume that the supply and production processes perform exactly as prescribed by their planning parameters, and that customer demand occurs as predicted by the demand forecast.

Most implementations for production planning systems make no effort to recognize the uncertainty in their environment. Indeed, in many planning systems that I have observed, there is limited, if any, attempt to track and measure the uncertainty in the demand forecast or the replenishment processes. For instance, we find that most planning systems do not retain and measure forecast errors. Once a forecast has been revised, we observe that the old forecast is written over by the new forecast, and the old forecast is lost. As a consequence, there is no record from which to measure the forecast errors. Similarly, we still find implementations of planning systems that do not track the actual replenishment times from suppliers, although this is slowly changing as lead-time performance becomes a more common metric in supply contracts. The execution component of most planning systems now has the capability to track the actual flow times within the production operations; however, again we seldom see this data used to

understand the uncertainty in these processes. On the other hand, yield data does get recorded more routinely, and at least the statistics on average yield seem to be more routinely used in the planning systems.

Even though these planning systems do not explicitly recognize or account for uncertainty, they do operate in an uncertain world. The planning organization thus develops various strategies and tactics for coping with this reality. We describe our observations on the major tactics as follows.

Safety stocks: Most planning systems do allow the users to set safety stocks for finished goods and for raw materials, as depicted in Figure 1. Safety stocks for raw materials can protect against supplier uncertainty, both in lead times and in yield. In a make-to-stock environment, it is possible to install a finished goods safety stock that can buffer against uncertainty from the production processes as well as uncertainty due to forecast errors.

Re-planning: A second common practice is re-planning. That is, at some regular frequency, the planning system is rerun to create a new plan. Often the re-planning frequency corresponds to the frequency with which the demand forecast gets updated; in other cases, re-planning is done even more frequently in order to capture the dynamics from the internal and external supply processes.

This re-planning might be done each week or month, even though the plan might extend for three to twelve months into the future. The revised plan would account for not just the changes in the demand forecast, but also the realizations from both the external supply and internal production processes. In this way, the system reacts to the uncertainty as it occurs, by revising its plans and schedules based on the updated information on demand and the supply processes.

Time fences and frozen schedules: One consequence of this re-planning is that it induces additional uncertainty in the form of schedule churn. That is, with each forecast revision, we generate a new master production schedule (MPS), which then can result in new detailed schedules for all raw materials, components and sub-systems. The due dates for some replenishment orders and production jobs are accelerated, while others are delayed. But a change in priority or due date in one revision might often be reversed by the next revision in the next time period. This churn or schedule nervousness leads to additional costs as changing priorities inevitably lead to inefficiencies in any production operation, as lower priority work gets put aside in order to expedite the higher priority work.

The schedule churn also leads to dissatisfaction and distrust of the planning system. Indeed, we find that suppliers (both internal and external) will often develop their own forecast of requirements, rather than accept and follow the requirements schedule that gets passed to them; in effect, they "second guess" the requirements schedule that is given to them. The suppliers know the requirements schedule from the planning system will keep changing, and they think they can do a better job with their own forecast. Nevertheless, it is not at all clear whether this second guessing helps or hurts.

One tactic to protect against the induced uncertainty from re-planning is the freezing of the master production schedule. A firm might decide that no changes are permitted for part of the master schedule, say for the next four weeks. The frozen schedule provides some stability in the short term, as any short-term changes in the demand forecast get accumulated and then deferred until beyond the frozen period. A

related tactic is to use time fences to establish varying limits on the amount of change permitted to the master production schedule. For instance, a firm might set a time fence at week 4, 8 and 13, and then specify that the MPS is frozen within the first time fence of 4 weeks, but permitted to change by at most (say) +/- 10 % for the weeks 5 to 8 and (say) +/- 25% for the weeks 9 to 13. Beyond the last time fence there might be no restrictions on how much the MPS could change. Again, this type of policy can help to mollify and dampen the dynamics introduced by production re-planning. However, these time fences and frozen schedules act as a constraint on the re-planning; thus, they necessarily limit the effectiveness of re-planning as a tactic for handling and responding to uncertainties in the demand forecast and in the supply and production replenishment processes.

Flexible capacity: In some contexts, firms maintain a capacity buffer to respond to uncertainty. In effect there is a reserve capability that permits the manufacturing system to recover from supply disruptions and/or to respond to unanticipated changes in demand volume or mix. A common example is overtime work, which is employed on an as needed basis to handle contingencies. However, even though many firms will rely on flexible capacity as a way to cope with uncertainties, their planning systems do not formally recognize or plan for any capacity buffer; that is, there is no means within the planning system to record or track utilization of the capacity buffer, let alone provide guidance on its size and deployment.

Backlog management: Another tactic for dealing with uncertainty is how a firm manages its order backlog. In particular, a firm might vary the size of the backlog or correspondingly, vary the delivery or service times quoted to customers. Thus, the backlog grows (falls) if demand is greater (less) than forecast and/or if the supply process

is slower (faster) than planned. This is only possible in contexts in which the manufacturer has sufficient market power to do this.

Inflated planned lead times: The last tactic that I have observed is the use of planned lead times to indirectly create additional safety stock throughout the production system. As noted above, most planning systems rely extensively on planned lead times. There are several good reasons for this to be the case.

- Planned lead times greatly simplify any complex planning problem by means of decomposition. The assignment of a planned lead time to each process step permits the scheduling of a multi-step serial production activity to be separated into a series of single-step activities, whereby each process step has a specific time window within which to accomplish its task. In effect, the planned lead times convert the final due date for a multi-step activity into intermediate due dates, one for each process step.
- Planned lead times facilitate coordination whenever multiple components or subsystems need to be joined or assembled together into a final product or assembly. As above, the planned lead times permit a decomposition by which we can establish intermediate due dates for each component or subsystem and then can manage each replenishment process independently.
- Planned lead times often serve as a proxy for dealing with capacity constraints (see Graves, 1986 for a discussion and analysis). Many planning systems do not explicitly account for capacity constraints; instead, they rely on the planned lead times to compensate for this oversight. The planned lead times are set to reflect the impact of limited capacity. A constrained work center that is heavily loaded

will have a longer planned lead time than one with a lighter load. A highly-utilized work center requires more smoothing of its work arrivals in order to level its load to match its capacity. A longer planned lead time results in a larger queue at the work center, which permits more smoothing.

Yet, beyond these reasons, we also observe that firms use their planned lead times as a way to create another buffer to protect against uncertainty. As noted earlier, the planning systems use the planned lead times to determine work and order releases; as a consequence, a planned lead time translates directly to a level of work-in-process (WIP). If a work center has a planned lead time of three days, then we might expect that it will have three days of WIP on average. (This follows from Little's Law, if we assume that the actual realized lead time is on average equal to the planned lead time.) As noted above, for capacity-constrained work centers, this level of WIP might be dictated by the need to smooth the work load through the work center. However, we often observe that the WIP (and planned lead times) exceeds that which is needed for work smoothing. In these cases, we find that the WIP is actually acting as a safety stock for the production system; its purpose is to provide an additional buffer to protect against uncertainties in demand and/or in the supply processes.

This WIP safety stock differs from the raw material (RM) and finished good (FG) safety stocks in a couple of important ways. First, the RM and FG safety stocks are established by setting their planning parameters in the planning system. In contrast, there are no planning parameters to directly set a WIP safety stock; rather, the WIP safety stock is a byproduct from the planned lead times, and as such is not the result of any deliberate planning decisions. Second, the WIP safety stock typically does not reside in a

warehouse, but sits on the shop floor in the form of work-in-process. One consequence of these two observations is that this buffer is usually under the radar of the "inventory police" and is not recognized as a planned safety stock. Indeed, we have seen operations eliminate their safety stocks and close their warehouses as part of an inventory reduction initiative, only to have that inventory recur on the shop floor in WIP, as the inherent uncertainties of demand and supply remain unchanged and thus a (unrecognized) safety stock is still required.

The inflated lead times also result in a control phenomenon known as *launch and expedite*. Based on a demand forecast, a firm releases work orders to initiate production to meet the demand forecast; that is, the firm pushes or *launches* the work into the shop based on the shop lead time. This creates a large WIP in various stages of completion, depending on how inflated are the planned lead times. The actual demand deviates from the forecast. As actual orders come in, the firm matches the orders with the available WIP and pulls or *expedites* the work out of the shop to meet the true demand. (see Sahney, 2005 for a case study)

This type of operation is often subject to an unhealthy dynamic. A longer planned lead time results in more work getting pushed into the shop, creating a larger WIP safety stock. This is attractive to the planner as there is more work from which to select to expedite to meet demand, once it has been realized. However, a longer planned lead time results in more uncertainty in the demand forecast; with more uncertainty, the shop needs even more safety stock, which results in pressure to increase further the planned lead times. This can lead to a so-called vicious cycle.

In summary, we have argued in this section that most systems for production planning do not recognize or account for uncertainty. Plans are typically created based on assumptions that the forecast is perfect, and the production and supply processes are perfect. Yet these systems are implemented in uncertain contexts; as a result, the planning organization needs to develop coping strategies. These coping strategies take the form of rapid and regular re-planning (subject to time fences and frozen short-term schedules), with the resulting churn; explicit FG and RM safety stocks as well as hidden safety stock in the form of WIP; and a mixture of fluctuating service or delivery times along with flexible capacity. We find substantial inefficiency in the deployment of these tactics. This is not surprising, as these responses are generally reactive ad-hoc measures. In the next section, we review the key tactics for dealing with uncertainty and discuss how these might be explicitly incorporated into the planning system.

A Proactive Approach to Uncertainty in Production Planning

In this section I discuss possible counter measures and practices for handling uncertainty. We take as given that the current practices and systems for production planning are not likely to change radically in the immediate future. There is a huge installed base of planning systems and the accompanying IT support systems; as described above, these planning systems are largely oblivious to uncertainties. The question here is what might help. What can be done to help these systems be more proactive with respect to the uncertainties in their environments?

The first step is certainly to do *more routine tracking and measurement of the uncertainty,* in whatever form it occurs. This includes the measurement of forecast errors, lead-time variability and yield variability. This by itself is not overly challenging, as it primarily entails keeping track of deviations from the norm or a target. But to use this data in planning requires some characterization of each type of variability so as to be able to model its occurrence. For instance, should the random deviations from a target yield be modeled as an additive or multiplicative process? Are these yield deviations correlated over time? Alternatively, the yield process might be better modeled by a Bernoulli process, for which there is some probability of having a yield bust, namely a zero yield. Systems are needed to not just capture the data but also to help in building models to characterize the uncertainty.

Similarly, modeling the lead-time variability for the purposes of inventory planning can have some subtlety. Consider an example based on a project to size a finished goods inventory for a semi-conductor wafer fabrication facility. We measured the lead times for production lots in the facility and found the coefficient of variation

(ratio of standard deviation to mean) to be on the order of 0.20. Based on this, we made inventory recommendations, which were rejected by the factory as being excessive. Upon closer inspection, we saw why. There was substantial amount of "order-crossing" within the wafer fab, as the production lots for the same product were not consistently processed in a first-in, first-out sequence for a variety of reasons. As a consequence, the completion order of the lots differed significantly from the order in which the lots were released into the shop. But from the standpoint of the finished goods inventory, the order of the output did not matter; what did matter was the cumulative output process, indicating how much had been completed by any point in time. We then re-defined the lead times to reflect this[1], and recalculated the lead time statistics to find that the coefficient of variation was less than half of the original number. This resulted in a more reasonable inventory recommendation. (see Johnson, 2005 for this case study)

We need also to understand the forecast evolution process so as to decide how best to characterize and model it. That is, given a forecast at one point in time, how should we think about the update process in the next time period? We know that with new market information and advanced orders the forecast will change, but does the forecast improve? How does a forecast change or evolve over a number of update cycles? And ultimately, does the forecast improve and by how much, as it is updated from period to period? Graves, Meal et al. (1986), Heath and Jackson (1994), and Gallego and Ozer (2001) provide some examples and approaches for the characterization of the forecast evolution process.

[1] We order the start times and completion times for all of the production lots, from earliest to latest. Then we define the n^{th} lead time as the difference between the n^{th} completion time and the n^{th} start time.

Under the assumption that we can characterize the uncertainties, the second step is to develop *a more explicit consideration of the tactical decisions that provide counter measures to the uncertainty*. We identify and discus five categories of tactical decisions. We contend that most planning systems address these decisions in an ad hoc way, and that there is a great opportunity to do better. Indeed, we think the key to improving our existing planning systems is to devise more systematic, proactive ways to find the right mix of tactics that match the uncertainty.

We will describe how these tactical decisions interface with the planning system depicted in Figure 1; for the most part, the connection is by means of setting planning parameters. Also, as will be clear, the tactical decisions are highly interdependent and their deployment will depend very much on the context.

Customer Service Times: One tactic is to decide the service or delivery times to quote to customers, and how to adapt this in light of the uncertainty in demand volume and mix. This tactical decision would be incorporated into the aggregate capacity planning module in Figure 1, as it is key to deciding how to match demand and capacity.

Allowing the customer backlog to vary with demand permits more efficient utilization of the production and supply system and/or less inventory buffers. However, varying the customer service times can result in market-related costs, such as lost sales; indeed, in some contexts this option is not economically feasible. To set these service times requires an examination of these trade-offs. Whereas in theory this is not difficult, this is not the case in practice. The trade-off requires an understanding of the customer sensitivity to the service times, as well as the cost inefficiencies from varying the production and supply processes.

MPS Smoothing: A second tactical decision is how to convert the short-term forecast into a detailed master production schedule in a way that is consistent with the aggregate capacity plan and cognizant of the demand forecast uncertainty. There are two sets of questions to consider. First, should the MPS smooth the demand forecast and by how much? Second, when the forecast changes, how should the MPS be updated to accommodate these changes?

The MPS acts as a gatekeeper between the demand forecast process and the upstream production system and supply chain. The MPS determines how much of the uncertainty in the forecast and in the forecast updates gets seen by the production and supply system. Hence, a key tactical decision is to decide how wide this gate should be. The more smoothing of the demand forecast by the MPS, the less uncertainty gets sent upstream. The same is true with the response by the MPS to forecast updates; the revision to the MPS can dampen the forecast updates and reduce the schedule churn, and thus provides more stable signals to the production and supply system.

There is a cost when the MPS buffers the uncertainty in this way. The replenishment schedules for production and supply are, by design, less responsive to actual changes in demand, and thus a larger finished good inventory is required to assure some level of customer service. Again, we have a trade-off between the cost impact from passing the forecast uncertainty to the production and supply system and the cost of buffering the uncertainty with the MPS. We noted earlier that time fences and freezing of the MPS are current practices for smoothing the MPS. However, these approaches tend to be deployed in an ad hoc fashion, with limited consideration of the trade-offs and the alternatives. I think there is an opportunity to develop decision support tools for master

scheduling, which would account for the uncertainty in the demand forecast process and provide a more complete treatment of the trade-offs; one example of such an approach is Graves et al. (1998).

Inventory Buffers: An important tactical decision is where to position inventory buffers within the production and supply system. The stylized model in Figure 1 assumes an inventory buffer of raw material and of finished goods, but with no other buffers between them. This seems to be non-optimal in many settings with any level of complexity. We noted earlier that some planning organizations circumvent this shortcoming in their planning systems by using inflated lead times to create a WIP safety stock. In other cases, we suspect that excessive finished goods inventory and/or underutilized capacity is required for the systems to function.

We contend that a better approach is to designate several inventory buffers, strategically located across a production and supply system. These buffers would act as de-coupling buffers; that is, each buffer is sized to protect the upstream supply processes from the noise or uncertainty in the downstream demand and to protect the downstream supply processes from any uncertainty in the upstream replenishment times or quantities. In effect, these buffers create a safety stock that allows the downstream to operate independently from any hiccups in the upstream process and vice versa.

Thus, the placement of these buffers can define relatively independent operating units within the production and supply system. Depending on the context, this can be an important consideration in deciding how many buffers and where to locate. Beyond this, one would of course need to account for the inventory holding cost; there is a cost for

each buffer that depends on the size of the buffer and the value of the contents. The buffer size depends on the magnitude of both the downstream demand uncertainty as well as the variability in the upstream replenishment process over its lead time. The value of the inventory depends on where it is in the process. Graves and Willems (2000) provide one framework for determining the location and size of these buffers. Schoenmeyr and Graves (2008) show how to extend the framework and analyses to account for a forecast evolution process. This remains a fruitful area for developing decision support tools for guiding these tactical decisions.

Capacity Buffers: In some contexts, it may be more economical to employ a capacity buffer, rather than build an inventory buffer. For instance, in a make-to order assembly operation, it may not be feasible to have a finished goods inventory buffer, due to all of the possible combinations that can be built. Instead, the daily variability in demand might be handled by varying the production capacity. To do this requires that there be some reserve capacity to respond to upswings in demand; this reserve is often the capability to expand or lengthen the work day, by working a longer shift.

A capacity buffer can be more flexible than an inventory buffer, as it can be used to create multiple types of inventory. A capacity buffer can take several forms. One, as noted above, is the ability of a production unit to expand or flex its capacity by working longer hours. In a one or two-shift operation, this might occur by extending the length of each shift, say, from eight hours to ten hours. Alternatively, in a five-day operation, there might be an option to work a sixth or seventh day. A second way of creating a capacity buffer is by explicitly scheduling "idle" time on a work center. That is, we reserve time in a schedule where the actual use of the time will be decided later. In this way we

postpone the decision of what product will be produced until we have a better resolution of the demand or process uncertainties. A capacity buffer might also take the form of an option. We might contract with a supplier or contract manufacturer to reserve a certain amount of capacity, which can be exercised at a later date at some exercise price.

One common context for a capacity buffer is when there is yield uncertainty in a manufacturing process. Miller (1997) provides one example based on a project examining the production planning practices for film manufacturing at Kodak. The bottleneck operation is the film sensitizing operation, which at the time was subject to significant process variability. A single capital-intensive machine performs the film sensitizing operation for multiple film products. The machine is highly utilized and changeovers are expensive. The machine is operated with a cyclic schedule that sequences through each type of film; the size of the batch run for each film is set to meet its short term requirements.

One element of the process variability was the occurrence of incident failures, whereby there is a major discrepancy between the actual output of good film and the planned output. Over the course of the year, there were about one incident failure per week, with 95% of the weeks having zero, one or two incident failures. As all film types are vulnerable to these incident failures, using inventory buffers to protect against this uncertainty was deemed unreasonable. Instead, Miller (1997) implemented a capacity buffer that could be used for any film type; in each production cycle, a certain amount of capacity was reserved at the end of the cycle. If there were one or two incident failures during the cycle, then the reserve capacity would be used to run a second batch of each of the affected films. If there were no incident failures, then the reserve capacity would not

be used and either the machine would be idled or the next production cycle would be moved forward in time.

This example is indicative of how one might deploy a capacity buffer. However, I am not aware of any systematic approach to thinking about this tactic, especially for a complex multi-stage production system. In particular, one would want to identify which process steps are good candidates for a capacity buffer, and how best to create and size these buffers. There is also the question of how to use capacity buffers in conjunction with inventory buffers: where and how should they be positioned and for what types of uncertainties would each buffer be deployed. This seems like a good opportunity for research and the development of decision support tools.

Planned Lead Times: Planned lead times are critical planning parameters in current planning systems. As discussed earlier, a planned lead time establishes the level of WIP in a manufacturing process step or the pipeline stock in a supply step. This inventory serves to dampen or absorb variability; in particular, it permits the smoothing of time-varying requirements. The longer is the planned lead time, the more smoothing is possible.

Typically the planned lead times are set in correspondence to the actual lead times. Sometimes they are set to equal the average or median observed lead time. In other instances, we have seen the planned lead times set "conservatively" so as to cover the actual lead time with high probability; for instance, the planned lead time might be set to match the 80^{th} percentile for the actual lead times.

There does not seem to be a standard practice for setting these parameters. Furthermore, I question the validity of setting a planned lead time based on observations

of the actual lead time, as there should be a strong interdependence between the planned and actual lead times. If the planned lead time sets the WIP at a work center, then Little's Law would say that the actual lead time should equal the planned lead time on average. That is, if the planned lead time at a work center were three days, then the planning system pushes work to the work center to create three-days of WIP. If the work center is staffed according to its work load, then it will process roughly one day of work each day. One then expects the actual lead time to match, at least on average, the planned lead time. Indeed, we find that the planned lead times can often become self-fulfilling prophecies.

We regard the determination of the planned lead times to be a critical tactical decision as these parameters dictate the local tactics for dealing with uncertainty within a series of process steps. At each process step, a longer planned lead time translates into using more WIP for damping the effects from demand and process uncertainty; in contrast, a shorter planned lead time requires more flexible capacity as the means to handle the uncertainty. We think there is a great opportunity for developing decision support to help planners in understanding the trade-offs and in setting these parameters in a more scientific way. Hollywood (2000) and Teo (2006) provide model developments of one line of approach to finding the planned lead times, based on the framework from Graves (1986). However, this work requires assumptions that might not apply to every setting. We expect there would be great value to practice from a more concerted effort on this problem domain.

An alternative approach is to replace the planned lead times with load-dependent lead times. This would entail a significant modification to current planning systems; yet this could result in a more accurate formulation of the planning problem. One challenge

here is to model the relationship between the work load at a production step and its lead time, capturing the congestion effects and supply uncertainties. Another challenge is then to incorporate this relationship into a planning model. Asmundsson et al. (2006) and Ravindran et al. (2008) provide viable approaches to these challenges and establish this as a promising avenue for future research and development.

In summary, we have identified five tactical decisions for handling uncertainty in the context of production planning. We have discussed each of these in terms of the trade-offs and considerations, and pointed out opportunities for developing more explicit approaches for making these decisions. We contend that getting these decisions right presents a huge opportunity for improvement to the current practice of production planning.

Acknowledgement: The preparation of this chapter has been supported in part by the MIT Leaders for Manufacturing Program, a partnership between MIT and global manufacturing firms; and by the Singapore-MIT Alliance, an engineering education and research collaboration among the National University of Singapore, Nanyang Technological University, and MIT. The author acknowledges and thanks the anonymous reviewer for helpful comments that have improved the chapter.

References

Asmundsson, J. M., R. L. Rardin and R. Uzsoy, 2006. "Tractable Nonlinear Production Planning Models for Semiconductor Wafer Fabrication Facilities." *IEEE Transactions on Semiconductor Manufacturing* 19: 95-111.

Bitran, G. R., E. A. Haas, and H. Matsuo. 1986. "Production planning of style goods with high set-up costs and forecast revisions," *Operations Research*, Vol. 34, 226 – 236.

Bitran, G. R. and D. Tirupati, 1993, "Hierarchical Production Planning," In *Handbooks in Operations Research and Management Science, Volume 4, Logistics of Production and Inventory*, edited by S. C. Graves, A. H. G. Rinnooy Kan and P. H. Zipkin, Amsterdam, Elsevier Science Publishers B. V., pp. 523-568.

Escudero, L. F., P. V. Kamesam, A. J. King, and R. Wets, 1993, "Production planning via scenario modeling," *Annals of Operations Research*, Vol. 43 Issue 1-4, p311-335.

Fleischmann, B. and H. Meyr, 2003, "Planning hierarchy, modeling and advanced planning systems," In *Handbooks in Operations Research and Management Science, Volume 11, Supply Chain Management: Design, Coordination and Operation*, edited by A. G. de Kok and S. C. Graves, Amsterdam, Elsevier Science Publishers B. V., pp. 457-523.

Gallego G. and Ö.Özer, 2001, "Integrating replenishment decisions with advance demand information," *Management Science*, Vol. 47, No. 10, 1344-1360.

Gfrerer, H. and G. Zapfel, 1995, "Hierarchical model for production planning in the case of uncertain demand," *European Journal of Operational Research*, Vol. 86, pp. 142-161.

Graves, S. C., 1986, "A Tactical Planning Model for a Job Shop," *Operations Research*, July-August, Vol. 34, 522-533.

Graves, S. C., 2002, "Manufacturing Planning and Control," in Handbook of Applied Optimization, edited by P. Pardalos and M. Resende, Oxford University Press, New York, pp. 728 - 746.

Graves, S. C., C. Gutierrez, M. Pulwer, H. Sidhu and G. Weihs, 1996, "Optimizing Monsanto's Supply Chain under Uncertain Demand," *Annual Conference Proceedings - Council of Logistics Management*, Orlando FL, October 1996, pp. 501-516.

Graves, S. C., D. B. Kletter, and W. B. Hetzel, 1998, "A dynamic model for requirements planning with application to supply chain optimization," *Operations Research*, 46, S35–S49.

Graves, S. C., H.C. Meal, S. Dasu, and Y. Qiu, 1986, "Two-Stage Production Planning in a Dynamic Environment," in Lecture Notes in Economics and Mathematical Systems, *Multi-Stage Production Planning and Inventory Control*, edited by S. Axsater, Ch. Schneeweiss, and E. Silver, Springer-Verlag, Berlin, 1986, Vol. 266, 9-43.

Hax, A. C. and H. C. Meal, 1975, "Hierarchical Integration of Production Planning and Scheduling," In *Studies in Management Sciences, Vol. 1: Logistics*, edited by M. A. Geisler, New York, Elsevier, pp. 53-69.

Heath, D.C., and P.L. Jackson 1994, "Modeling the evolution of demand forecasts with application to safety stock analysis in production/distribution systems," *IIE Transactions*, Vol. 26, No. 3, 17-30.

Hollywood, J. S., 2000, "Performance Evaluation and Optimization Models for Processing Networks with Queue-Dependent Production Quantities," Ph.D. Thesis, MIT Operations Research Center, Cambridge MA.

Holt, C. C., F. Modigliani, J. F. Muth and H. A. Simon, 1960, *Planning Production, Inventories and Work Force*, Englewood Cliffs NJ, Prentice-Hall.

Hopp, W. J. and M.L. Spearman, 2007, Factory Physics: Foundations of Manufacturing Management, Burr Ridge, IL: Irwin/McGraw-Hill, third edition 2007.

Johnson, Jeffrey D., 2005, "Managing Variability in the Semiconductor Supply Chain," M.S. thesis, MIT Engineering Systems Division, Cambridge MA.

Lasserre, J.B., and Merce, C. (1990), "Robust hierarchical production planning under uncertainty," *Annals of Operations Research*, Vol. 26, 73-87.

Miller, Michael P., 1997, "Business System Improvement Through Recognition of Process Variability," M.S. thesis, MIT Leaders for Manufacturing Program, Cambridge MA.

Mula. J., R. Poler, J.P. García-Sabater, and F.C. Lario, 2006, "Models for production planning under uncertainty: A review," *International Journal of Production Economics*, Vol. 103, 271–285.

Nahmias, S., 2008, Production and Operations Analysis, Irwin/McGraw-Hill, sixth edition.

Ravindran, A., K. G. Kempf, and R. Uzsoy, 2008, "Production Planning with Load-Dependent-Lead Times and Safety Stocks", working paper, February 2008 (32 manuscript pages).

Sahney, Mira K., 2005, "Building Operational Excellence in a Multi-Node Supply Chain," M.S. thesis, MIT Leaders for Manufacturing Program, Cambridge MA.

Schoenmeyr, T. and S. C. Graves, 2008, “Strategic safety stocks in supply chains with evolving forecasts, “ to appear in *Manufacturing & Service Operations Management*.

Shapiro, J. F., 1993, “Mathematical Programming Models and Methods for Production Planning and Scheduling,” In *Handbooks in Operations Research and Management Science, Volume 4, Logistics of Production and Inventory*, edited by S. C. Graves, A. H. G. Rinnooy Kan and P. H. Zipkin, Amsterdam, Elsevier Science Publishers B. V., pp. 371-443.

Silver, E. A., D. F. Pyke, and R. Peterson, 1998, Inventory Management and Production Planning and Scheduling, Wiley, 3rd Edition.

Teo Chee Chong, 2006, “A Tactical Planning Model for Make-To-Order Environment under Demand Uncertainty,” Ph.D. thesis, NTU, May.

Thomas. L. J. and J. O. McClain, 1993, “An Overview of Production Planning,” In *Handbooks in Operations Research and Management Science, Volume 4, Logistics of Production and Inventory*, edited by S. C. Graves, A. H. G. Rinnooy Kan and P. H. Zipkin, Amsterdam, Elsevier Science Publishers B. V., pp. 333-370.

Figure 1: Framework for classical production planning