10.4. Management and Capacity Planning/Utilization

Capacity planning is one side of the coin; capacity management is the other. Plans need to be executed, and this needs to be done effectively; moreover, well-developed demand management can provide conditions that are much more favorable to routine execution. For example, Toyota and several other Japanese auto manufacturers develop production plans with a stable rate of output (cars per day). Product mix variations are substantially less than those for other auto companies because they carefully manage the number and timing of option combinations. The result is execution systems that are simple, effective, and easy to operate with minimal inventories and fast throughput times. Capacity planning is straightforward and execution is more easily achieved, not only for the company itself but for its suppliers as well. That is, well-managed front-end planning can rationalize the entire supply chain.

10.4.1. Capacity Monitoring with Input/Output Control

One key capacity management issue concerns the match between planning and execution. This implies monitoring on a timely basis to see whether a workable capacity plan has been created and whether some form of corrective action is needed. The best-known approach to this issue is **input/output control,** where the work flowing through a work center is monitored: the planned work input and output are compared to the work actual input and output.

10.4.1.1. Input/Output Control

The capacity planning technique used delineates the planned input. Planned output results from managerial decision making to specify the capacity level; that is, planned output is based on staffing levels, hours of work, and so forth. In capacity-constrained work centers, planned output is based on the rate of capacity established by management. In non-capacity-constrained work centers, planned output is equal to planned input (allowing for some lead-time offset).

Capacity data in input/output control are usually expressed in hours. Input data are based on jobs' expected arrivals at a work center. For example, a CRP procedure would examine the status of all open shop orders (scheduled receipts), estimate how long they'll take (setup, run, wait, and move) at particular work centers, and thereby derive when they'll arrive at subsequent work centers. A finite loading system would do the same, albeit with better results. The approach would be repeated for all planned orders from the MRP database. The resultant set of expected arrivals of exact quantities would be multiplied by run time per unit from the routing file. This product would be added to setup time, also from the routing file. The sum is a planned input expressed in standard hours.

Actual input would use the same routing data, but for the *actual* arrivals of jobs in each time period as reported by the shop-floor control system. Actual output would again use the shop-floor control data for exact quantities completed in each time period, converted to standard hours with routing time data.

The only time-data not based on the routing file are those for planned output. In this case, management has to plan the labor-hours to be expended in the work center. For example, if two people work nine hours per day for five days, the result is 90 labor-hours per week. This value has to be reduced or inflated by an estimate of the relation of actual hours to standard hours. In our example, if people in this work center typically worked at 80 percent efficiency, then planned output is 72 hours.

A work center's actual output will deviate from planned output. Often deviations can be attributed to conditions at the work center itself, such as lower-than-expected productivity, breakdowns, absences, random variations, or poor product quality. But less-than-expected output can occur for reasons outside the work center's control, such as insufficient output from a preceding work center or improper releasing of planned orders. Either problem can lead to insufficient input or a "starved" work center. Another reason for a variation between actual input and planned input was shown by our capacity planning model comparisons —some models don't produce realistic plans!

Input/output analysis also monitors backlog. Backlog represents the cushion between input and output. Backlog decouples input from output, allowing work center operations to be less affected by variations in requirements. Arithmetically, it equals prior backlog plus or minus the difference between input and output. The planned backlog calculation is based on planned input and planned output. Actual backlog uses actual input and output. The difference between planned backlog and actual backlog represents one measure of the total, or net, input/output deviations. Monitoring input, output, and backlog typically involves keeping track of cumulative deviations and comparing them with preset limits.

The input/output report in Figure 10.14 is for work center 500 shown in weekly time buckets with input and output measured in standard labor-hours. The report was prepared at the end of period 5, so the actual values are current week-by-week variations in planned input. These could result from actual planned orders and scheduled receipts; that is, for example, if the input were planned by CRP, planned inputs would be based on timings for planned orders, the status of scheduled receipts, and routing data. The *actual* input that arrives at work center 500 can vary for any of the causes just discussed.

Figure 10.14 Sample Input/Output for Work Center 500 (as of the end of period 5)*

		Week				
		1	2	3	4	5
Planned input		15	15	0	10	10
Actual input		14	13	5	9	17
Cumulative deviation		−1	−3	+2	+1	+8
Planned output		11	11	11	11	11
Actual output		8	10	9	11	9
Cumulative deviation		−3	−4	−6	−6	−8
Actual backlog	20	26	29	25	23	31
Desired backlog: 10 hours						

**In standard labor-hours.*

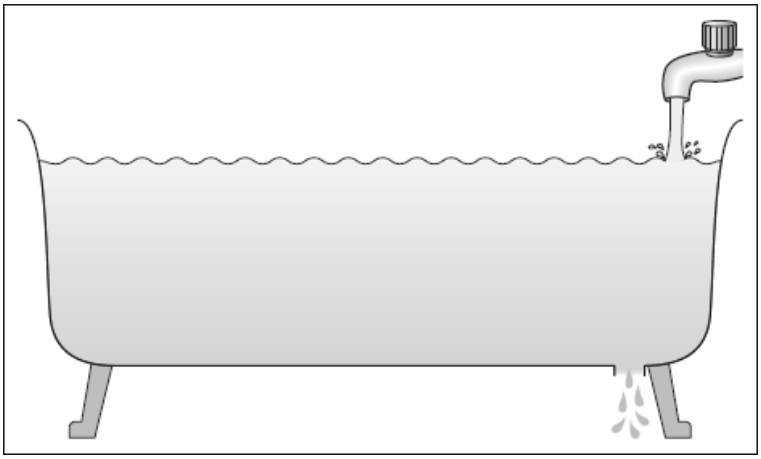
Work center 500's planned output has been smoothed; that is, management decided to staff this work center to achieve a constant output of 11 hours per week. The results should be to absorb input variations with changes in the backlog level. Cumulative planned output for the five weeks (55 hours) is 5 hours more than cumulative planned input. This reflects a management decision to reduce backlog from the original level of 20 hours. The process of increasing capacity to reduce backlog recognizes explicitly that flows must be controlled to change backlog; backlog can't be changed in or of itself.

Figure 10.14 summarized the results after five weeks of actual operation. At the end of week 5, the situation requires managerial attention. The cumulative input deviation (+8 hours), cumulative output deviation (−8 hours), current backlog (31 hours), or all three could have exceeded the desired limits of control. In this example, the increased backlog is a combination of more-than-expected input and less-than-expected output.

One other aspect of monitoring backlog is important. In general, there's little point in releasing orders to a work center that already has an excessive backlog, except when the order to be released is of higher priority than any in the backlog. The idea is to not release work that can't be done, but to wait and release what's really needed. Oliver Wight summed this up as one of the principles of input/output control: "Never put into a manufacturing facility or to a vendor's facility more than you believe can be produced. Hold backlogs in production and inventory control." With today's APS system, a similar dictate results: concentrate on executing the most immediate schedule—exactly. The APS system will take care of the future schedules.

Figure 10.15 depicts a work center "bathtub" showing capacity in hydraulic terms. The input pipe's diameter represents the maximum flow (of work) into the tub. The valve represents MPC systems like MPS, MRP, and JIT, which determine **planned input** (flow of work) into the tub. Actual input could vary because of problems (like a corroded valve or problem at the water department) and can be monitored with input/output analysis. We can determine **required capacity** to accomplish the planned input to the work center with any of the capacity planning techniques. The output drain pipe takes completed work from the work center. Its diameter represents the work center's planned or **rated capacity,** which limits planned output. As with actual input, actual output may vary from the plan as well. It too can be monitored with input/output analysis. Sometimes planned output can't be achieved over time even when it's less than maximum capacity and there's a backlog to work on. When that occurs, realized output is called **demonstrated capacity.** The "water" in the tub is the **backlog** or **load,** which can also be monitored with input/output analysis.

Figure 10.15 The Capacity Bathtub



10.4.2. Managing Bottleneck Capacity

Eliyahu Goldratt developed a key capacity management idea that he popularized more than 25 years ago in *The Goal*. Fundamentally, one needs to find the bottlenecks in any factory, and thereafter manage their capacities most effectively. Goldratt's maxim is that an hour of capacity lost in a bottleneck work center is an hour of capacity lost to the entire company—worth a fortune. But an hour of capacity gained in a nonbottleneck work center will only increase work-in-process inventory and confusion. Eli Goldratt has gone on to other things, but this fundamental concept remains at the base of his work. Today, he and his colleagues have generalized the ideas into what they refer to as "theory of constraints" (TOC). For the purposes of capacity planning and management, TOC teaches that the capacities of bottleneck work centers need to be planned and managed much more carefully than those of nonbottlenecks. In fact, Goldratt points out that for nonbottlenecks it may not be important to even have decent data. If sufficient capacity exists, execution of capacity plans is easy. Spend the time and energy on execution of what at first seems impossible.

Goldratt has many suggestions for how to execute the impossible. For example, why shouldn't bottleneck work centers run through lunch hours and coffee breaks—others can run these work centers while the primary personnel eat lunch and drink coffee. Alternative routing is another solution, and this is a good idea even when it "costs" much more. Usually the costs are calculated with unrealistic assumptions. Extra work done in an underutilized work center has no real cost, and if the bottleneck workload is thereby reduced, it is an excellent idea to do it.

The TOC approach to capacity planning is essentially to first determine the bottleneck work centers. This can be done with a rough-cut capacity planning model or with CRP. Where are the bottlenecks? Next, TOC would try to find the quick solutions for eliminating bottlenecks. Finally, scheduling will concentrate on best managing bottleneck capacity. Essentially, TOC will separate those jobs that pass through the bottlenecks from those that do not. Only the jobs or work orders requiring capacity in the bottleneck resource are finite scheduled, using horizontal loading and back scheduling for the most critical jobs.

If we return to the hole in the schedule of Figure 10.13, the TOC approach would definitely front-schedule component D for the reason described there: it is the schedule for component C that constrains the start of end product A. Do not let component D become a constraint to this overall product schedule. TOC treats this early schedule (front-loaded) as a buffer in order to reduce the possibility of missing the overall goal: ship the end product!

TOC uses APS systems, but concentrates their attention on what is truly critical. For nonbottleneck work centers it is more than unimportant to utilize their capacity—it is fundamentally *wrong*. Increasing utilization of nonbottlenecks will result in more work being in the factory than necessary, yielding higher inventories and confusion. Nonbottleneck work will be done easily because there is basically no constraint to it. Restricting the use of APS systems to focus on the bottlenecks allows smart users to examine the best ways to "skin the cat."

The most critical capacity requirements need to be identified and thereafter utilized to maximum effectiveness. Capacity planning techniques can help with the former, but effective management is needed for the latter. Moreover, managerial policies can also create environments that are easier to execute—environments where capacities are utilized in a predictable and stable fashion.

10.4.3. Capacity Planning in the MPC System

To illustrate the importance of the interrelationships in designing and using the capacity planning system, let's consider the impact of production planning and resource planning decisions on shorter-term capacity planning decisions. To the extent that production planning and resource planning are done well, problems faced in capacity planning can be reduced, since appropriate resources have been provided. If, for example, the production plan specifies a very stable rate of output, then changes in the master production schedule (MPS) requiring capacity changes are minimal. If the material planning module functions effectively, the MPS will be converted into detailed component production plans with relatively few unexpected execution problems.

A quite different but equally important linkage that can affect capacity planning system design is the linkage with shop-floor execution systems. A key relationship exists in scheduling effective use of capacity. With sufficient capacity and efficient use of that capacity ensured by good shop-floor systems, we'll see few unpleasant surprises requiring capacity analysis and change. Effective shop-floor procedures utilize available capacity to process orders according to MRP system priorities, provide insight into potential capacity problems in the short range (a few hours to days), and respond to changes in material plans. Thus, effective systems reduce the necessary degree of detail and intensity of use of the capacity planning system. The result is a better match between actual input/output and planned input/output. Again, we see attention to the material planning side of the MPC system, in this case the shop-floor module, having an effect on the capacity planning side.

10.4.4. Choosing the Measure of Capacity

The choice of capacity measures is an important management issue. Alternatives run from machine-hours or labor-hours to physical or monetary units. The choice depends on the constraining resource and the firm's needs. In any manufacturing company, the "bundle of goods and services" provided to customers increasingly includes software, other knowledge work, after-sales service, and other customer services. In every case, providing these goods and services requires resources —"capacities" that must be planned, managed, and developed. Appropriate measures of capacity must be established and changed as evolution in the bundle of goods and services occurs.

Several current trends in manufacturing have a significant bearing on the choice of capacity measures. Each can have a major impact on what's important to measure in capacity. One important trend is considerable change in the concept of direct labor. Direct labor has been shrinking as a portion of overall manufacturing employment. Distinctions between direct and indirect labor are becoming less important. The ability to change labor capacity by hiring and firing (or even using overtime) has been reduced; notions of "lifetime employment" have further constrained this form of capacity adjustment.

One objective in JIT systems is continual improvement, so the basis for labor capacity is constantly changing. This mandates control procedures for identifying and changing the planning factors as improvements take place.

Another important trend is decreased internal fabrication and increased emphasis on outside purchasing, i.e., outsourcing. This trend can alter the conception of what capacity requirements are important. Procurement analysis, incoming inspection, and engineering liaison may become the critical capacities to be managed, as well as planning and scheduling the capacities in vendor firms. In fact, one of the major benefits ascribed to major outsourcing companies is their ability to more flexibly respond to changing capacity needs.

For many firms engaged in fabrication, machine technology is changing rapidly. Flexible automation has greatly increased the range of parts that can be processed in a machine center. Future product mixes are likely to be much more variable than in the past, with a marked effect on the equipment capacity required. Moreover, as equipment becomes more expensive, it may be necessary to plan and control the capacity of key pieces of equipment at a detailed level.

To the extent that cellular technologies are adopted as part of JIT manufacturing, the unit of capacity may need to change. Usually the entire cell is coupled and has only as much capacity as its limiting resource. Often, the cell is labor limited, so the unit of capacity is labor-hours (continually adjusted for learning). Sometimes, however, the capacity measure needs to be solely associated with a single aspect of the cell. Also, when dissimilar items are added to the cell for manufacture, it's necessary to estimate each new item's capacity requirements in terms of individual processing steps.

The first task in choosing a capacity measure is to creatively identify resources that are critical and in short supply. Capacity control is too complicated to apply to all resources. The next step is to define the unit of measure. If the key resource is people, then labor-hours may be appropriate. In other instances, such measures as tons, gallons, number of molds, number of ovens, hours of machine time, square yards, linear feet, lines of code, customer calls, and cell hours have been used. In some cases, these are converted to some "equivalent" measure to accommodate a wider variety of products or resources.

After the resources and unit of measure have been determined, the next concern is to estimate available capacity. The primary issue here is theory versus practice. The engineer can provide theoretical capacity from the design specifications for a machine or from time studies of people. A subissue is whether to use "full" capacity or some fraction thereof (often 75 to 85 percent). A further issue is "plasticity" in capacity. For almost *any* resource, if it's *really* important, more output can be achieved. We've seen many performances that fall short of or exceed capacity calculations.

Choice of capacity measure follows directly from the objective of providing capacity to meet production plans. The appropriate measure of capacity that most directly affects meeting these plans. The measure, therefore, should be appropriate to the critical limited resources and be based on what's achievable, with allowances for maintenance and other necessary activities. It must be possible to convert the bundle of products and services into capacity measurement terms. The results must be understood by those responsible, and they should be monitored.

10.4.5. Choice of a Specific Technique

In this chapter's discussion, the capacity planning techniques for converting a material plan into capacity requirements include three different methods for rough-cut capacity planning (CPOF, capacity bills, and resource profiles). We also examined capacity requirements planning, CRP, which is particularly useful for medium range planning. For the detailed day-to-day capacity planning APS systems can be valuable under some circumstances. The choice of method depends heavily on characteristics of the manufacturing environment.

The three rough-cut methods are most general, being applicable even in companies using JIT methods for shop-floor control. Rough-cut approaches can be useful in JIT operations to estimate the impact of changes in requirements called for by revisions to the master production schedule. For example, under level scheduling conditions, a change from a production rate of 480 units per day (one unit per minute) to 528 units per day (1.1 units per minute) might be needed. A rough-cut procedure could be used to examine the impact on each work center or manufacturing cell through which this volume would pass (including those of suppliers). Any indicated problems or bottleneck conditions could be addressed *before* the crisis hits. Similarly, a planned reduction in MPS could be evaluated to determine resources that might be freed to work on other tasks.

Rough-cut approaches do vary in accuracy, aggregation level, and ease of preparation. There's a general relationship between the amount of data and computational time required, and the quality and detail of the capacity requirements estimated. The issue is whether additional costs of supporting more complex procedures are justified by improved decision making and subsequent plant operations.

The capacity bills procedure has an advantage over capacity planning using overall factors (CPOF) because it explicitly recognizes product mix changes. This can be important in JIT operations, particularly where the level schedule is based on assumptions of product mix and where different products have different capacity requirements. On the other hand, if changes in mix are easily accommodated, and there are minimal differences in capacity requirements for different products, then CPOF's simplicity can be exploited. Under JIT operations, however, there's often little need to incorporate the added sophistication of the resource profile procedure. There simply won't be any added advantage to making lead time offsets in the planning process. Work is completed at virtually the same time as it's started.

Capacity requirements planning is only applicable in companies using time-phased MRP records for detailed material planning and shop-order-based shop scheduling systems. CRP is unnecessary under JIT operations anyway because minimal work-in-process levels mean there's no need to estimate the impact in capacity requirements of partially processed work. All orders start from "raw material" with virtually no amount of "capacity" stored in component inventories. Also, under JIT, there's no formal PAC procedure. There are no work orders. Thus, there are no status data on work orders.

Input/output control isn't usually an issue under JIT operations because attention has been shifted from planning to execution. As a result, actual input should equal actual output. Actual input becomes actual output with an insignificant delay. The backlog is effectively a constant zero. However, planned input can indeed vary from actual input and so can planned output vary from actual output. These variations should be achievable without violating the equality between actual input and actual output–with backlog remaining at zero. To the extent that plan-to-actual variations are possible, the result reflects the flexibility, or bandwidth, of the JIT unit.

10.4.6. Using the Capacity Plan

All the techniques we've described provide data on which a manager can base a decision. The broad choices are clear—if there's a mismatch between available capacity and required capacity, either the capacity or the material plan should be changed. If capacity is to be changed, the choices include overtime/undertime authorization, hiring/layoff, and increasing/decreasing the number of machine tools or times in use. Capacity requirements can be changed by alternate routing, make-or-buy decisions, subcontracting, raw material substitutions, inventory changes, or revised customer promise dates.

Choice of capacity planning units can lead to more effective use of the system. Capacity units need not be work centers as defined for manufacturing, engineering, or routing purposes. They can be groupings of the key resources (human or capital) important in defining the factory's output levels. Many firms plan capacity solely for key machines (work centers) and gateway operations. These key areas can be managed in detail, while other areas fall under resource planning and the shop-floor control system.

Capacity planning choices dictate the diameter of the manufacturing pipeline. Only as much material can be produced as there's capacity for its production, *regardless of the material plan.* Not understanding the critical nature of managing capacity can lead a firm into production chaos and serious customer service problems. In the same vein, the relationship between flexibility and capacity must be discussed. You can't have perfectly balanced material and capacity plans *and* be able to easily produce emergency orders! We know one general manager who depicts his capacity as a pie. He has one slice for recurring business, one for spare parts production, one for downtime and maintenance, and a final specific slice for opportunity business. He manages to pay for this excess capacity by winning lucrative contracts that require rapid responses. He *does not add* that opportunity business to a capacity plan fully committed to the other aspects of his business.