1

OVERVIEW OF THE THEORY OF CONSTRAINTS

Every now and then, a completely new idea comes along that can be described as either refreshing, disturbing, or both. Within the accounting profession, the theory of constraints is that change. It originated in the 1980s through the writings of Eliyahu Goldratt. His training as a physicist, rather than as an accountant, appears to have given him a sufficiently different mind-set to derive several startling changes to the concepts of operational enhancement and cost accounting. The theory of constraints is based on the concept that a company must determine its overriding goal, and then create a system that clearly defines the main capacity constraint that will allow it to maximize that goal. This chapter describes the operational and financial aspects of the theory of constraints.

DEFINITIONS FOR THE OPERATIONAL ASPECTS OF THE THEORY OF CONSTRAINTS

Comprehending the operational aspects of the theory of constraints requires some understanding of a new set of terms that are not used in traditional company operations. The terms are as follows:

- *Drum.* This is the element in a company's operations that prevents the company from producing additional sales. This is the company's constrained capacity resource or bottleneck operation. It will most likely be a machine or person, though it also might

be a short supply of materials. Because total company results are constrained by this resource, it beats the cadence for the entire operation—in essence, it is the corporate drum.

- *Buffer.* The drum operation must operate at maximum efficiency in order to maximize company sales. However, it is subject to the vagaries of upstream problems that impact its rate of production. For example, if the drum is located in the production department, then if the stream of work-in-process generated by an upstream work center is stopped, the inflow of parts to the drum operation will cease, thereby halting sales. To avoid this problem, it is necessary to build a buffer of inventory in front of the drum operation to ensure that it will continue operating even if there are variations in the level of production created by feeder operations. The size of this buffer will be quite large if the variability of upstream production is large, and correspondingly smaller if the upstream production variability is reduced.
- *Rope.* This term refers to the timed release of raw materials into the production process to ensure that a job reaches the inventory buffer before the drum operation is scheduled to work on it. In essence, the rope is the synchronization mechanism driving the flow of materials to the drum operation. The length of the rope is the time required to keep the inventory buffer full, plus the processing time required by all operations upstream of the drum operation.

These three terms are frequently clustered together to describe the theory of constraints as the drum-buffer-rope (DBR) system. The following section discusses the mechanics of the DBR system.

THE OPERATIONAL ASPECTS OF THE THEORY OF CONSTRAINTS

Pareto analysis holds that 20 percent of events cause 80 percent of the results. For example, 20 percent of customers generate 80 percent of all profits, or 20 percent of all production issues cause 80 percent of the scrap. The theory of constraints, when reduced down to one guiding concept, states that one percent of all events cause 99 percent of the results. This conclusion is reached by viewing a company as one

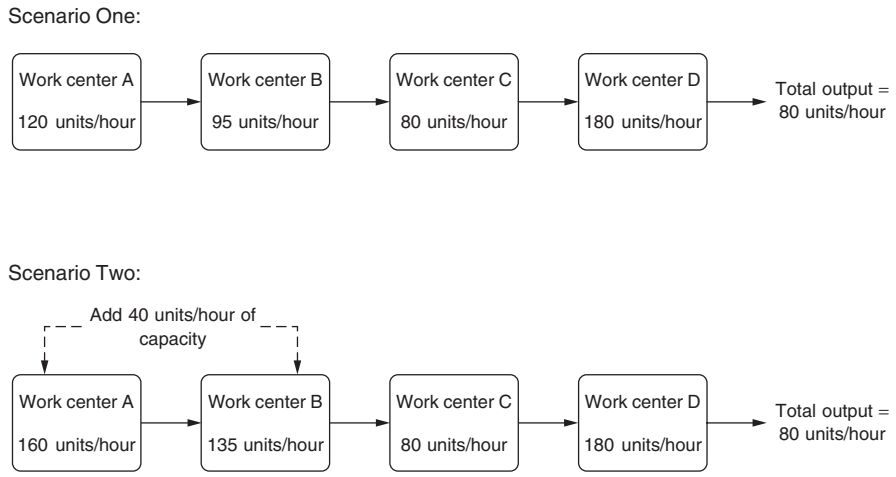

EXHIBIT 1.1 IMPACT OF THE DRUM OPERATION ON TOTAL OUTPUT

giant system designed to produce profits, with one bottleneck operation controlling the amount of those profits.

Under the theory of constraints, all management activities are centered on management of the bottleneck operation, or drum. By focusing on making the drum more efficient and ensuring that all other company resources are oriented toward supporting the drum, a company will maximize its profits. The concept is shown in Exhibit 1.1, where the total production capacity of four work centers is shown, both before and after a series of efficiency improvements are made. Of the four work centers, the capacity of center "C" is the lowest, at 80 units per hour. Despite subsequent efficiency improvements to work centers "A" and "B," the total output of the system remains at 80 units per hour, because of the restriction imposed by work center "C."

This approach is substantially different from the traditional management technique of local optimization, where *all* company operations are to be made as efficient as possible, with machines and employees maximizing their work efforts at all times.

The key difference between the two methodologies is the view of efficiency—should it be maximized everywhere, or just at the drum? The constraints-based approach holds that any local optimization of a non-drum resource will simply allow it to produce more than the drum operation can handle, which results in excess inventory. For

example, a furniture company discovers that its drum operation is its paint shop. The company cannot produce more than 300 tables per day, because that maximizes the capacity of the paint shop. If the company adds a lathe to produce more table legs, this will only result in the accumulation of an excessive quantity of table legs, rather than the production of a larger number of painted tables. Thus, the investment in efficiencies elsewhere than the drum operation will only increase costs without improving sales or profits.

The preceding example shows that not only should efficiency improvements *not* be made in areas other than the drum operation, but that it is quite acceptable to not even be efficient in these other areas. It is better to stop work in a non-drum operation and idle its staff than to have it churn out more inventory than can be used by the drum operation.

Given the importance of focusing management attention on maximization of drum efficiencies, the use of buffers becomes extremely important. An inventory buffer should be positioned in front of the drum operation, and is used to provide a sufficient amount of stock to the drum to keep it running at maximum efficiency, even when variations in upstream work centers create short-term reductions in the flow of incoming inventory. The need for a buffer brings up a major operational concept in the theory of constraints, which is that there will be inevitable production failures that will alter the flow of inventory through the facility. Buffers are used to absorb the shock of these production failures, though it is also possible to increase the level of sprint capacity to offset the need for large buffers.

Sprint capacity is excess capacity built into a production operation that allows the facility to create excess inventory in the short term, usually to make up for sudden shortfalls in inventory levels. Sprint capacity is extremely useful for maintaining a sufficient flow of inventory into the drum operation, since the system can quickly recover from a production shortfall. If there is a great deal of sprint capacity in a production system, then there is less need for a buffer in front of the drum operation, since new inventory stocks can be generated quickly.

The concept of sprint capacity brings up an important point in the theory of constraints—that it is not only useful, but necessary to

have excess capacity levels available in a system. This controverts the traditional management approach of eliminating excess capacity in order to reduce the costs associated with maintaining that capacity. Instead, management should be aware of those work centers with high levels of sprint capacity, which require much lower levels of inventory buffer, and primarily focus its attention on areas with low sprint capacity, which require larger buffer stocks.

Thus far, we have seen that the theory of constraints places a premium on maximum utilization of the drum operation, as well as the use of inventory buffers to support that utilization. One additional requirement is needed to ensure that the drum operates at maximum capacity at all times, which is the concept of the rope. The rope is the method used to release inventory into upstream production processes just in time to ensure that the drum operation and its buffer are fully supplied with the appropriate levels of work-in-process. If the rope releases inventory into the system too late, then the drum will be starved of input, and will produce less than its maximum amount. Conversely, the release of inventory too early will result in a large backlog of unfinished parts in front of the drum, which both represents an excessive investment in inventory and may result in confusion regarding which jobs to process next through the drum operation.

These factors comprise the drum-buffer-rope (DBR) elements of the theory of constraints, and will be explained more fully in Chapter 2, Constraint Management in the Factory. Having covered an overview of DBR, we will diverge briefly to address the nature of the constraint and then proceed to the financial aspects of the theory of constraints.

NATURE OF THE CONSTRAINT

The theory of constraints is based on the existence of a constraint, so it is useful to delve into the nature of this core concept. A constraint is a resource that limits a company's total output. For example, the constraint may be a machine that can only produce a specified amount of a key component in a given time period, thereby keeping overall sales from expanding beyond the maximum capacity of that machine. The key question to ask in locating this type of constraint is: "If we had more of it, could we generate more sales?" Physical constraints

of this type tend to be easy to locate within a company because there is usually a large amount of work-in-process piled up in front of it, waiting to be processed.

The most common system constraint cannot be seen or touched—it is the operational policy. A policy is a rule that dictates how a system is operated. Examples of policies are batch sizing rules and resource utilization guidelines. For instance, a policy may state that a work station completely fill a pallet with work-in-process before sending it on to the next work station, since this makes it more efficient for the materials handling staff to move inventory through the factory. The trouble is that the next work station may be the constrained resource, which has to halt operations while waiting for the pallet to be filled. In this case, the policy should have allowed a more continuous flow of inventory to the constrained resource, which means that much smaller batch sizes would have improved the utilization of the constrained resource.

Policy constraints are usually difficult to find and eliminate. *Finding* them is difficult because policies are not physical entities that can be readily observed; instead, they must be deduced from the operational flow of the production system. *Eliminating* them can be even more difficult, since they may be strongly supported by employees, who require considerable convincing before agreeing to change a policy that they may have used for years. Though there may be considerable resistance to a policy change, the actual fix can be extremely inexpensive. Once eliminated, a policy constraint can result in a larger degree of system improvement than the elimination of any physical constraint.

A concept impacting the presence of policy constraints is the paradigm constraint. This is a belief that causes employees to follow a policy constraint. A classic paradigm constraint is the belief that every work center must be run at full tilt in order to increase its efficiency, which is a teaching of traditional cost accounting theory. However, this paradigm can result in a policy constraint to create a bonus plan that rewards factory managers for running all equipment at as close to 100 percent capacity as possible. The result is an excessive investment in inventory, and the divergence of resources away from the constrained resource. Thus, a paradigm constraint can be a powerful roadblock to the elimination of a policy constraint.

Another constraint may be a raw material, if there is not enough to ensure that all orders can be filled. This less common problem tends to arise during bursts of peak industry-wide sales, when materials suppliers are caught with insufficient production capacity to meet all demand (which means that the constraint has now shifted to the supplier!). This type of constraint will be immediately evident to the materials management staff, which cannot schedule jobs for release to the production area until sufficient materials are available.

Another possible constraint is the sales staff, if there are not enough people to bring in all possible customer orders. This constraint is evident when a large potential market or a significant number of sales prospects exist at the top of the sales funnel, but very few actual sales are being generated.

A company may improve its operations so much that its current capacity is able to handle all orders currently placed by customers. If so, the constraint has now shifted into the marketplace. The company must now use its higher capacity to offer better pricing deals or service levels to customers in order to increase its share of the market.

A company can also intentionally position a constraint on a specific resource. This happens when the capacity of a particular resource would be extremely expensive to increase, so managers prefer to focus their attention on maximizing the efficiency of the work center without actually adding capacity to it. It is also useful to avoid positioning the constraint on a resource that requires complex level of management, such as one where employee training or turnover levels are extremely high. Thus, the positioning of the constrained resource should be a management decision, rather than an accident.

DEFINITIONS FOR THE FINANCIAL ASPECTS OF THE THEORY OF CONSTRAINTS

To explain the financial aspects of the theory of constraints requires the use of several new terms (or old terms with new definitions), so we will define them first, before delving into the mechanics of the system. They are as follows:

- *Throughput.* The contribution margin that is left after a product's price is reduced by the amount of its totally variable costs (which

is explained in the next bulleted point). There is no attempt to allocate overhead costs to a product, nor to assign to it any semi-variable costs. As a result, the amount of throughput for most products tends to be quite high.

- *Totally variable costs.* A cost that will only be incurred if a product is created. In many instances, this means that only direct materials are considered to be a totally variable cost, though subcontracting costs, commissions, customs duties, and transportation costs may also apply. Direct labor is not totally variable unless employees are only paid if a product is produced. The same rule applies to all other types of costs, so one will not find any type of overhead cost in the "totally variable cost" category.
- *Operating expenses.* The sum total of all company expenses, excluding totally variable expenses. Expenses usually categorized here are direct and indirect labor, depreciation, supplies, interest payments, and overhead. As a general rule, all expenses incurred as a result of the passage of time (rather than through the production process) are operating expenses. This group of expenses is considered to be the price a company pays to ensure that it maintains its current level of capacity. The theory of constraints does not care if a cost is semi-variable, fixed, or allocated—all costs that are not totally variable are lumped together into the Operating Expenses category.
- *Investment.* This definition is the same as one would find under standard accounting rules. However, there is a particular emphasis on a company's investment in working capital (especially inventory). The value of a company's investment in inventory does not include the value added by the system itself; so it does not include the value of direct labor or manufacturing overhead. The investment in inventory only includes amounts paid for components that are purchased from outside suppliers and used in the manufacture of inventory.
- *Net profit.* Throughput minus operating expenses.

These definitions are used to describe the financial aspects of the theory of constraints in the next two sections.

THE FINANCIAL ASPECTS OF THE THEORY OF CONSTRAINTS

The earlier discussion of the operational aspects of the theory of constraints might not appear to have a great deal of application to the work of the accountant, but its financial aspects reverse many long-standing principles of cost accounting. Since we are now covering an aspect of the theory of constraints that deals directly with the work of the accountant, we will refer to this area as throughput accounting.

A key concept of throughput accounting is the use of profitability analysis at the system level instead of gross margin analysis at the product level. In a traditional cost accounting system, costs from all parts of the production process are compiled and allocated by various means to specific products. When subtracted from product prices, this yields a gross margin that is used to determine whether a product is sufficiently profitable to be produced. Throughput accounting almost entirely ignores gross margin analysis at the product level. Instead, it considers the production process to be a single system whose overall profitability must be maximized.

The key reason for this difference in perspective is that most production costs do not vary directly with the incremental production of a single unit of a product. Instead, most production costs are required to maintain a system of production, irrespective of the number of product units created by it. For example, a traditional cost accounting system will assign the depreciation cost of a production machine to an overhead account, from which it is allocated by various means to each unit of a product manufactured. However, if one unit were not produced, would this result in a proportionate drop in the amount of overhead cost? Probably not. Instead, the same amount of overhead would now be assigned to the fewer remaining units produced, which raises their costs and lowers their gross profits.

To avoid this costing conundrum, throughput accounting uses an entirely different methodology, which is comprised of three elements: throughput, operating expenses, and investment. The key element of the three is throughput. To arrive at throughput, we subtract all *totally* variable costs from revenue. In reality, the only cost that varies totally with a product is the cost of its direct material. (Remember, even the

cost of direct labor does not usually vary with the number of units produced.) In how many companies does the staff immediately go home when the last product is completed, or do employees get paid solely based on the number of units of production they create? Instead, the staff is employed on various other projects during downtime periods, to ensure that the same experienced staff is available for work the next day. The result of the throughput calculation is a very high level of throughput—much higher than a product's gross margin, which includes both labor and overhead costs.

The result of using throughput instead of gross margin is that hardly any products will *not* be produced due to a negative margin. This will only occur in a throughput accounting environment if a product's revenue is matched or exceeded by its raw material cost, which is rarely the case. Instead, products with a low throughput will still be included in the product mix, since they contribute to some degree to the total throughput of a company's production system.

The next element of throughput accounting is the concept of operating expenses. This category includes all other expenses besides the totally variable ones used to calculate throughput. Operating expenses are essentially all costs required to operate the production system. In throughput accounting, there is no distinction between totally fixed or partially fixed costs—again, they are either totally variable costs or part of operating expenses. By avoiding the considerable level of analysis required to deduce the variable elements of most largely fixed costs, financial analysis is greatly simplified, as will be seen in the multitude of examples in Chapter 4, Throughput and Financial Analysis Scenarios.

Throughput accounting also places considerable emphasis upon investment, which is the amount of money added to a system to improve its capacity. When combined with throughput, totally variable costs, and operating expenses, throughput accounting uses the following formulas for a wide array of accounting decisions:

$$\text{Revenue} - \text{totally variable expenses} = \underline{\text{throughput}}$$

$$\text{Throughput} - \text{operating expenses} = \underline{\text{net profit}}$$

$$\text{Net profit/investment} = \underline{\text{return on investment}}$$

When making a decision involving changes to revenue, expenses, or investments, these three formulas can be used to arrive at the correct decision, which must yield a positive answer to one of the following three questions:

- Does it increase throughput?
- Does it reduce operating expenses?
- Does it improve the return on investment?

If a localized decision yields a positive answer to any one of these questions, then it will also improve the company-wide system, and so should be implemented.

When answering the three questions, it is best to favor decisions resulting in increased throughput, since there is potentially no upper limit to the amount of throughput that a company can generate. Decisions resulting in reduced operating expenses should be given the lowest action priority, since there is a limited amount of operating expense that can be reduced; also, a reduction of operating expenses may limit the production capacity of the system, which in turn may yield less throughput.

THE OPPORTUNITY COST OF OPERATIONS

A major concept of throughput accounting is to determine the true cost to a company of its capacity constraint. The capacity constraint is the drum operation, as described at the beginning of this chapter. If the use of the drum is not maximized, what is the opportunity cost to the company?

In a traditional cost accounting system, the cost would be the foregone gross margin on any products that could not be produced by the operation. For example, a work center experiences down time of one hour, because the machine operator is on a scheduled break. During that one hour, the work center could have created 20 products having a gross margin of $4.00 each. Traditional cost accounting tells us that this represents a loss of $80. Given this information, a manager might very well not back-fill the machine operator, and allow the machine to stay idle for the one-hour break period.

However, throughput accounting uses a different calculation of the cost of the capacity constraint. Since the performance of the constraint

drives the total throughput of the entire system, the opportunity cost of not running that operation is actually the total operating expense of running the entire facility, divided by the number of hours during which the capacity constraint is being operated. This is because it is not possible to speed up the constrained operation, resulting in the permanent loss of any units that are not produced. For example, if the monthly operating expenses of a facility are $1.2 million and the constrained resource is run for every hour of that month, or 720 hours (30 days multiplied by 24 hours/day), then the cost per hour of the operation is $1,667 ($1,200,000 divided by 720 hours). Given this much higher cost of not running the operation, a manager will be much more likely to find a replacement operator for break periods.

What about the cost of not running a nonconstrained resource operation? As long as its downtime does not impact the operation of the constrained resource, it has no opportunity cost at all. In fact, the situation is reversed, for it is actually better to only run nonconstraint resources at the pace of the drum operation, since any excess inventory produced will only increase the amount of inventory in the production system—and this represents an additional investment in the system for which there is no offsetting increase in throughput.

Thus, there are substantial differences in the opportunity cost of running various operations, which can be interpreted differently with different accounting systems. Throughput accounting focuses attention on the high cost of not running a constrained resource, while showing that there is a negative opportunity cost associated with running a nonconstrained resource more than it is needed.

SUMMARY

This chapter was designed to give a general overview of the operational and financial underpinnings of the theory of constraints and throughput accounting. Here are the key issues covered so far:

- A company's results are largely driven by its management of a single constrained resource.

- The drum-buffer-rope system can be used to manage the constrained resource.
- Throughput accounting focuses on the total throughput of the system, rather than the gross margins of individual products.

In Chapter 2, we will expand upon the constraint management concept as it applies to a factory environment, and then devote the remainder of the book to an examination of a multitude of throughput accounting issues, including overhead allocation in Chapter 3, financial analysis in Chapter 4, budgeting and capital budgeting in Chapter 5, generally accepted accounting principles in Chapter 6, control systems in Chapter 7, performance measurement and reporting systems in Chapter 8, and accounting management issues in Chapter 9.