A COMPLEMENTARY INVENTORY-CAPACITY TAXONOMY

FOR THEORY OF CONSTRAINTS

Carl E. Betterton, School of Business, The Citadel, Charleston, SC 29409

(843) 571-1580, carl.betterton@citadel.edu

ABSTRACT

Theory of Constraints (TOC) literature typically addresses three categories of inventory: (1) productive inventory, protective inventory, and excess inventory. TOC also recognizes three corresponding categories of resource capacity: productive capacity, protective capacity, and excess capacity. These inventory and capacity terms are intended to correspond with three levels of resource action: (1) production by the constraint, or production by non-constraint resources matching the constraint resource's needs, (2) subordinate production by non-constraint resources to protect the constraint from interruptions, and (3) no production. Defects are common in manufacturing but such yield loss is rarely addressed in TOC literature, and not at all in the inventory and capacity terms used. This article extends inventory and capacity terms to address yield loss, and in so doing, demonstrates the complementary relationship of inventory and capacity.

INTRODUCTION

Betterton and Cox [5] called for systematic efforts to move the research community toward consensus on conceptual and operational definitions of the TOC framework, including adherence to appropriate operational and conceptual definitions of terms. Distinctions between terms used in the TOC literature are not always clear, and that literature neglects entirely definitions of capacity and inventory related to waste caused by yield loss.

In TOC literature the use of inventory and capacity terminology has centered on the need to support production at the system constraint resource and to protect that critical resource from any form of interruption. The usual characterization of the sources of potential interruption is reference to "statistical fluctuations" and "dependent resources" seen most often as variation in process times, failure occurrences, and repairs that may cause starvation, blockage, or downtime of the constraint [7]. If a machine breaks down it is abundantly clear that it is not producing. If that machine is upstream of the constraint, some amount of buffer stock is needed to avoid starving the constraint during the downtime period. This implies that the upstream machine needs some additional or reserve capacity if it is to build or replenish buffer inventory. If the machine is downstream of the constraint, no buffer inventory is needed but it may need a similar additional capacity to process jobs quickly to avoid blocking the constraint after any breakdown.

But if a machine produces a defective item, it has still "produced," albeit what may be retrospectively called *wasted inventory* by using *wasted capacity*. If the defective item reaches the constraint undetected, and is processed by the constraint, throughput potential is lost forever. Constraint capacity is *wasted*, but there is no commonly accepted term that treats this situation in the TOC literature. If the defective item is detected and removed before it reaches the constraint, this may increase the likelihood of constraint starvation. The unwanted upstream production of defective items means that additional non-constraint capacity must be exercised to replenish lost (defective) units. The unwanted downstream production of defective items means that throughput has been destroyed.

Scrap and rework are fundamentally different than other "statistical fluctuations" because they not only threaten, but destroy throughput directly or indirectly. No surge or sprint capacity can make up for the fact that the constraint, or a resource downstream, has just produced a defective item. Wasted capacity, and its corresponding wasted inventory, are things to be avoided. To better do so, we need to give them a name and focus on them. We need to understand *wasted inventory* and *wasted capacity* in the context of more familiar TOC capacity terminology.

THE IMPACT OF DEFECTIVE WORK

Spoilage, defects, and rework are common occurrences in manufacturing. An Industry Week survey of 884 responding manufacturing plants found that scrap and rework costs were estimated as two percent of sales [11]. In some industries this internal failure cost is much greater. For example, it is estimated that defective work costs aerospace/defense companies an average of six percent of their total sales [12].

While spoilage, defects, and rework are often encountered, the impact of this poor quality on the production system is not well understood [1]. Producing defective items results in lower yield. Adding more inspection or testing to find defectives adds time and money to the process. Rework reduces line yield losses but increases cycle time and process costs.

Much of this impact is manifested by increases in variability. Hopp and Spearman [8, p. 390] claimed that *quality problems are one of the largest and most common causes of variability* and warned that such problems can have extreme consequences on operations. They described some of the direct and indirect effects of rework and scrap using a simple serial asynchronous four-stage constant work-in- progress (CONWIP) line with deterministic processing times. Using simulation, they showed that poor quality lowers throughput (yield) and increases cycle time for any given level of work-in-progress (WIP) inventory, because rework causes increases in both variability of process times and average utilization of affected workstations. Impact of poor quality on the overall line differs depending on the workstation where the defective item is produced. High WIP levels tend to amplify financial losses due to poor quality because of the larger time between processing the defective item and detecting the defect.

Scrap, which can be seen as the most extreme form of rework, is often treated as a deterministic quantity (e.g., 10%) when in almost all real situations the scrap rate for a given process is a random quantity [8]. For example, the *mean* scrap rate may be 10% but scrap may actually vary from 0% to 20% in different parts of the production cycle or at different workstations within the line. Most manufacturing systems use some form of job size inflation to compensate for average or expected yield losses and to protect customer due dates. The consequences for a specific order for a specific customer may be felt as increased lead time, waiting to fill out a lot due to low yield, or increased finished goods inventory resulting from excess goods caused by high yield. The more variable the yields, the greater the cost and disruption.

THE CORRESPONDENCE OF CAPACITY AND INVENTORY

Betterton [4] proposed a taxonomy of capacity types for TOC, and since every type of capacity has a corresponding type of inventory, that taxonomy can be extended to include inventory types for TOC.

Little's Law [10] says that the average number of items in a queuing system, denoted L, equals the average arrival rate of items to the system, λ, multiplied by the average waiting time of an item in the system, W. Thus, $L = \lambda W$.

For a system in which no losses occur the arrival rate must equal the departure rate, which is the throughput rate. The waiting time corresponds to the time period from entry to departure, so is equivalent to flow time. The number of items in the system corresponds to the system inventory. Thus by Little's Law in general is Mean Throughput Rate equals Mean Inventory divided by Mean Flow Time.

Since throughput is a rate it can be expressed as production units per time period. Likewise the right side of the equation is inventory units per time period. Thus there is a one-to-one relation between production capacity and resulting inventory; every use or type of capacity has a corresponding result or type of inventory.

Consistent with the above, Anupindi et al. [2] define theoretical capacity of a resource as the maximum sustainable flow rate if the resource was fully utilized. They discuss flow time and break it into two parts: theoretical flow time and waiting time. The theoretical flow time is the combined time of all operations on the critical path of the process, and for a serial line is just the total of all operations making up the process. Thus, for theoretical flow time there is no waiting (no variability), and accordingly, no buffer inventory. Theoretical flow time corresponds to theoretical inventory and theoretical capacity.

Inventory discussed by Anupindi et al. [2] is referred to as buffer input, buffer output, and station in-process inventory. These inventory categories can be simplified to in-station inventory and in-transit inventory (inventory waiting or moving between stations). According to Anupindi et al [2] the minimum amount of inventory necessary to maintain a resource's full throughput rate can be called theoretical inventory and expressed as:

Theoretical inventory = (Max throughput rate) (Theoretical flow time)

This corresponds to the Hopp and Spearman [8] definition of Critical WIP as the WIP level at which a line, with no variability in process times, achieves maximum throughput with minimum cycle time. That is:

Critical WIP = (Bottleneck rate) (Raw process time)

where Raw process time equals total flow time with no variability.

Thus while different terms are used from author to author, the literature is consistent in recognizing the relationship of capacity and inventory at the global level. It is at the disaggregated level where a lack of clarity creeps in. As an example, Blackstone and Cox [7] define productive inventory as

> *…the amount of WIP inventory (measured in time units of the constraint work station) needed to support the constraint work station until material can get from the gating or first operation to the constraint work station … In a deterministic world, this WIP level would maintain throughput until a unit can get from the gating operation to the constraint work station. The example line is*

> *provided again in figure 4a. In figure 4b, 22 minutes are required for WIP to be processed through stations A (five units per hour represents 12 minutes per unit) and B (six units per hour represents 10 minutes per unit) therefore the* ***protective inventory*** *should equal this amount of constraint work station time. If 2 units of WIP are in the line at the constraint work station X (which processes parts at the rate of 15 minutes per unit), then this WIP equals 30 minutes of processing on the constraint work station. This material is the productive WIP required to support the constraint work station until another unit released from raw material could reach the constraint station X ... This productive WIP may be located anywhere between the material release point and the constraint work station; the closer to the constraint work station the better.* (p. 419, emphasis added).

This statement defines productive inventory as an amount or quantity of inventory rather than a kind or type of inventory. Worse for this reader, the definition sounds like protective inventory when it says "productive WIP is that necessary to maintain uninterrupted production at the constraint until material can get from the gating or first operation to the constraint work station." This statement implies an interruption of production at stations preceding the constraint - which would require protective inventory. Notice that within the same cited paragraph the article actually refers to the two units of inventory as "protective inventory" (see bold), but later switches back to calling it productive WIP. The statement that *productive WIP may be located anywhere between the material release point and the constraint work station* can be correct only if one presumes that protective WIP is also on hand. For example, say the constraint just finished processing a unit. If the two productive WIP units mentioned as being "anywhere" were located upstream of the constraint and the constraint had no units in its feeding buffer then the constraint would starve.

If we define *productive inventory* simply as that inventory created by a resource using its productive capacity then there is a direct correspondence between productive capacity and productive inventory. With this definition, there is no confusing productive WIP with protective WIP. With this definition the statement is true that productive WIP may be located anywhere between the gate and the constraint, including at a station being processed. This definition holds in a deterministic or stochastic world. With this definition one can still quantify productive WIP in time units of constraint production. With this definition, the idea of imperfect quality translates perfectly - we get "Wasted Inventory" (scrap) corresponding to "Wasted Capacity." All in all, there is a one-to-one correspondence between the capacity relationships and the inventory relationships, as it should be according to Little's Law. The remainder of this paper illustrates that correspondence and defines the related terms.

THROUGHPUT

Throughput is system output converted to sales (or other goal units). Throughput is *productive inventory* in its final form, and is created by utilizing a portion of the system's installed or theoretical capacity. The portion of theoretical capacity so utilized is *productive capacity*.

SYSTEM CAPACITY AND INVENTORY

System capacity is equal to the productive capacity of the system's busiest (most highly used) resource over the time horizon of interest. Such a resource is a constraint, and the constraint's *productive capacity* is equal to its *available capacity* less its *wasted capacity*. Good output of the constraint is productive inventory; defective output is wasted inventory. Thus the capacity of a constraint is *utilized* as productive capacity or it is lost as wasted productive capacity. Once a resource becomes a constraint, it has no

protective capacity and no excess capacity, and it can produce no protective inventory and no (potentially) excess inventory.

THEORETICAL CAPACITY AND THEORETICAL INVENTORY

Theoretical capacity refers to the upper limit of output (quantity per time period) that an individual resource or system is capable of producing under normal conditions on a sustained basis. Theoretical capacity is equivalent to design capacity and is the practical or full capacity ordinarily achievable. Theoretical capacity may be less than surge capacity achievable for short durations or under operational conditions that shorten the life of the resource, or which are not sustainable over the long run. We define *theoretical capacity* as the normal full capacity of a resource (worker or machine), having no failure or other downtime (unavailability), and producing perfect results. For a production line this would be the constraint capacity. Since all inventory must be produced with some form of capacity, theoretical inventory is the counterpart of theoretical capacity. Theoretical inventory is the category or amount of inventory that would be produced with theoretical capacity.

In reality, of course, all resources have some downtime and suffer from some imperfections. So theoretical capacity may be seen as being composed of two components, *available capacity* and *unavailable capacity*. Likewise, theoretical inventory may be seen as being composed of two corresponding components, *available inventory* and *unavailable inventory*.

Theoretical Capacity = Available Capacity + Unavailable Capacity ***(1)***

Inventory from Theoretical Capacity = Inventory from Available Capacity + Inventory from Unavailable Capacity ***(2)***

"Inventory from unavailable capacity" may seem like a contradiction in terms, but what it refers to is inventory that could be obtained if (for example) breakdowns could be reduced/eliminated.

AVAILABLE CAPACITY AND INVENTORY FROM AVAILABLE CAPACITY

Available capacity consists of *effective capacity*, *wasted capacity*, and *excess capacity*. Available capacity excludes periods when the resource is not operating, such as maintenance, repair, setup, or other downtime. We can use the widely accepted reliability definition for "availability" as:

Availability = MTBF / (MTBF + MTTR) ***(3)***

where, MTBF is mean time between failures, and MTTR is mean time to repair. Or more broadly, the definition given in the APICS Dictionary [6] can be used, according to which, availability is the portion of time that a resource is capable of working:

Availability = (S-B) / S ***(4)***

where, S is scheduled time, and B is downtime. Making use of (1) and (3) above, we define *available capacity* as the capacity remaining in a resource beyond any downtime, or:

Available Capacity = Availability x Theoretical Capacity ***(5)***
= (MTBF / (MTBF + MTTR)) x Theoretical Capacity = ((S-B) / S) x Theoretical Capacity

The available capacity of a resource can be directed toward productive ends (utilized), expended toward unproductive ends (activated), wasted, or remain idle and unused. Inventory potential from Available Capacity will be in the categories of Inventory from Effective Capacity, Inventory from Wasted Capacity, and Inventory from Excess Capacity.

By *utilized* is meant capacity used in support of the constraint and by *activated* is meant capacity used that is not needed to support the constraint, for example by making excess inventory. In the theory of constraints, *utilization* is the ratio of utilized capacity to available capacity, however, more commonly, *utilization* is the ratio of used capacity to theoretical (full) capacity [9], or the percentage of time a resource is busy [3]. Capacity that is wasted represents a special form of activation, in which productive ends are intended but defective work (wasted inventory) is produced. By *idle* is meant that a nonconstraint resource is in standby mode until it is needed.

EFFECTIVE CAPACITY AND INVENTORY - PRODUCTIVE AND PROTECTIVE

Effective capacity is composed of productive capacity and protective capacity. Inventory from Effective Capacity is composed of Inventory from Productive Capacity and Inventory from Protective Capacity. By *productive* is meant that the resource is creating throughput (good results), or is subordinating to the constraint in support of throughput production. The capacity used when a resource is engaged in throughput production, or in support of the constraint for throughput production, is termed *productive capacity*. Inventory created by a resource using its productive capacity is *productive inventory*.

A nonconstraint resource in idle or standby mode may be needed to protect constraint operation, for example, by "speeding up" for a period after it has experienced downtime, to replenish buffer inventory feeding the constraint, or to avoid blockage of the constraint by removing completed items from the constraint's downstream space buffer. The capacity used when the normally idle capacity of a nonconstraint resource is exercised effectively in a protective manner is termed *protective capacity*, and any inventory produced thereby is *protective inventory*. While protective capacity may remain idle for a while, eventually it is called upon to support throughput. Excess capacity by its nature is always idle and so never produces any inventory.

WASTED CAPACITY AND WASTED INVENTORY

It is possible that defective units can be produced by human error or by a machine that drifts out of control, or even through random chance in a capable process under statistical control. Defective units are not throughput, and they are not intended. The available capacity used to produce defective units is *wasted inventory*, and is accordingly termed *wasted capacity*.

Both productive and protective capacity may be wasted. A nonconstraint resource with less that 100 percent yield may invoke unused capacity in an attempt to compensate for some system fluctuation, i.e., to protect the constraint. To the extent that this resource produces good output (throughput) in support of the constraint, the capacity so exercised is protective capacity. To the extent that this resource produces defective output, even though intended to support the constraint, the capacity employed is wasted capacity and the inventory produced is *wasted inventory*.

In a serially linked set (chain) of resources with yield loss, additional wasted capacity and wasted inventory are produced when a downstream workstation converts good units created by an upstream workstation into defective units. This secondary conversion of the upstream resource's productive

capacity into wasted capacity may be thought of as a "system effect," and is more pronounced with longer resource chains.

EXCESS CAPACITY AND EXCESS INVENTORY

Excess capacity is that portion of available capacity not needed, or so rarely needed that its benefits are outweighed by the costs of keeping it. By definition, excess capacity never exercised will produce no inventory; the idea of *excess inventory* corresponds to the inventory that would be produced if the excess capacity were used. Deciding what capacity is excess means deciding how much protective capacity is sufficient. This is no trivial decision. As Cox and Blackstone [7] have pointed out, there is currently no mathematical approach for defining protective capacity.

By the nature and definition of a constraint, it has no excess capacity and no protective capacity. (It does not protect itself.) Excess and protective capacity appear only at nonconstraints. A constraint resource will experience some downtime in the form of failure, maintenance, setup, etc. This aggregate downtime is the constraint's unavailable capacity, and is subtracted from the constraint's theoretical capacity to arrive at its available capacity.

Since the constraint has no excess or protective capacity, all of its available capacity would ideally be viewed as productive capacity and scheduled to support throughput by producing *productive inventory*. However, this is not possible if there are yield losses at the constraint. The effective (productive) capacity of the constraint will be degraded to the extent that yield loss is present. Productive capacity for a non-constraint resource is equal to that of the constraint; the same quantity of productive inventory must flow through both constraint and nonconstraint resources. Any additional capacity exercised beyond this level by a non-constraint, in order to maintain uninterrupted production at the constraint, is protective capacity and is engaged in producing protective inventory.

CAPACITY AND INVENTORY – TWO SIDES OF THE SAME COIN

Capacity exists before inventory – inventory is the manifestation of capacity. When inventory is created capacity is consumed. Capacity is potential inventory; inventory is actualized capacity. Capacity can be consumed with either positive or negative results; intended good output can be created or defective output can result. Scrap, rework, and other forms of non-optimum production are fundamentally different than other “statistical fluctuations” in the TOC world because they destroy throughput directly or indirectly.

Some threshold amount of protective inventory is needed to transmit protective capacity and vice versa. The presence or absence of strategically located work-in-progress (WIP) inventory serves to protect the constraint from being idled. Located upstream of the constraint, such WIP constitutes a buffer that helps prevent starvation of that critical resource. (A space buffer downstream from the constraint helps avoid blockage of the constraint.) Like capacity, WIP can be classified and defined based on its characteristics and use - productive, protective, wasted, and excess. When a resource produces a defective item, that resource is actually removing from the flow stream a (wasted) WIP unit that must then be replaced by a good one. Capacity and WIP are related in that it takes inventory to transmit capacity from resource to resource. Producing wasted WIP converts potentially productive capacity to wasted capacity.

This inventory-capacity taxonomy attempts to clarify the relationship of inventory and capacity within the TOC framework, and to address the problem that there is no commonly accepted term in the TOC literature that deals with the reality of yield loss and the resulting wasted capacity and wasted inventory. The Appendix provides a summary of the inventory and capacity relationships discussed.

APPENDIX - SUMMARY OF TOC INVENTORY AND CAPACITY RELATIONSHIPS

Part 1 - Summary of TOC Capacity Relationships

The capacity relationships defined and discussed above are summarized below:

Total System Capacity Potential = Theoretical Capacity

Theoretical Capacity = Available Capacity + Unavailable Capacity

Available Capacity = Effective Capacity + Wasted Capacity + Excess Capacity

Effective Capacity = Productive Capacity + Protective Capacity

Theoretical Capacity = Productive Capacity + Protective Capacity + Excess Capacity
+ Wasted Capacity + Unavailable Capacity

Idle Capacity = Protective Capacity + Excess Capacity

Theoretical Capacity = Productive Capacity + Idle Capacity + Wasted Capacity + Unavailable Capacity

Part 2 - Summary of TOC Complementary Inventory Relationships

The inventory relationships defined and discussed above are summarized below:

Total System Inventory Potential = Inventory from Theoretical Capacity

Inventory from Theoretical Capacity = Inventory from Available Capacity + Inventory from Unavailable Capacity

Inventory from Available Capacity = Inventory from Effective Capacity + Inventory from Wasted Capacity + Inventory from Excess Capacity

Inventory from Effective Capacity = Inventory from Productive Capacity + Inventory from Protective Capacity

Inventory from Theoretical Capacity = Inventory from Productive Capacity + Inventory from Protective Capacity + Inventory from Excess Capacity + Inventory from Wasted Capacity + Inventory from Unavailable Capacity

Inventory from Idle Capacity = Inventory from Protective Capacity + Inventory from Excess Capacity

Inventory from Theoretical Capacity = Inventory from Productive Capacity + Inventory from Idle Capacity + Inventory from Wasted Capacity + Inventory from Unavailable Capacity

Notes:

1. “Inventory from unavailable capacity" may seem like a contradiction in terms, but what it refers to is inventory that could be obtained if (for example) breakdowns could be reduced/eliminated.

2. "Inventory from Idle Capacity" corresponds with the APICS definition of idle inventory — The inventory generally not needed in a system of linked resources. Idle inventory generally consists of protective inventory and excess inventory.

Part 3 - Summary of TOC Physical Inventory Relationships

Total System Inventory = Throughput (Finished Goods, good product) + Scrap (defective product) + Total WIP

Total WIP = Effective WIP + Excess WIP

Effective WIP = Productive WIP (line WIP) + Protective WIP (buffer WIP)

Effective Production = Throughput Production (Finished Goods without defects) + Effective WIP

Note 1: "Line WIP" refers to the productive WIP between the gate and the constraint, and "buffer WIP" refers to WIP that buffers the constraint against interruption.

REFERENCES

[1] Agnihothri, S. R. and Kenett, R. S. "The Impact of Defects on a Process with Rework." *European Journal of Operational Research*, 1995, 80: 308-327.

[2] Anupindi, R., Chopra, S., DeshMukh, S. D., Van Mieghem, J. A. and Zemel, E. *Managing Business Process Flows: Principles of Operations Management*, 2nd Ed. Pearson Prentice Hall, New Jersey, USA. 2006.

[3] Barbacci, M., Klein, M. H., Longstaff, T. H. and Weinstock, C. B. *Quality Attributes* (CMU/SEI-95-TR-021). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, 1995.

[4] Betterton, C. E. "A Proposed Capacity Taxonomy For TOC Incorporating Yield Loss." Southeast Decision Sciences Institute Conference, February 21-23, 2007, Savannah Marriott Riverfront Hotel, Savannah, Georgia. 2007.

[5] Betterton. C. E. and Cox, J. F. III. "Espoused Drum-Buffer-Rope Flow Control in Serial Lines: A Comparative Study of Simulation Models." *International Journal of Production Economics*, 2009, 117, 66-79.

[6] Blackstone, J. H. Jr. (Ed.). *APICS Dictionary* (Thirteenth Ed.). Chicago: APICS - The Association for Operations Management. 2010.

[7] Blackstone, J. H. Jr. and Cox, J. F. III. Designing Unbalanced Lines - Understanding Protective Capacity and Protective Inventory. *Production Planning & Control*, 13(4): 416-423. 2002.

[8] Hopp, W. J. and Spearman, M. L. *Factory Physics* (Second Ed.). Boston: Irwin McGraw-Hill. 2001.

[9] Krajewski, L. J., Ritzman, L. P., and Malhotra, M. K. *Operations Management: Processes and Value Chains* (Eighth Ed.). Upper Saddle River: Pearson Prentice Hall. 2007.

[10] Little, J. D. C. "A Proof for the Queuing Formula: L = λW." *Operations Research*, 1961, 9(3) 383–387.

[11] Taninetz, G. "Faster But Not Better." *Industry Week*, 2004, April, 47-49.

[12] Velocci, A. L. Jr. "Cost of Quality An Industry Challenge." *Aviation Week & Space Technology*, 1998, 149(58).