Theory of constraints- Lessons for academicians and practicing managers from “The Goal”

1 Ajay Gupta 2 Dr Arvind Bhardwaj 3 Dr Arun Kanda

1. Sr. Lecturer, Department of Industrial Engineering, National Institute of Technology Jalandhar (India)
2. Professor, Department of Industrial Engineering, National Institute of Technology Jalandhar (India)
3. Professor, Department of Mechanical Engineering, Indian Institute of Technology, Delhi (India)

Abstract

Theory of Constraints has been emerging as an important tool for optimization of manufacturing/service systems. Although, the traditional approaches for optimization of systems had been yielding highly optimal results but it has been felt that those solutions were not found to give competitive advantage to the practicing managers in the real life situations. Present work highlights the conceptual framework as put forward by Goldratt and subsequently applied by different other researchers under the actual practicing environment. A comparative study of issues raised by different researchers and problems faced by them in the application of Theory of Constraints has also been presented. The results indicate that performance of the organizations improves, if the organizations are managed according to the throughput world as proposed by TOC, compared to managing according to the cost world alone. It has also been observed that under certain situations, certain combination based algorithms like Tabu – Simulated annealing may yield better results as compared to TOC. The review also indicates that DBR mechanism for material release in to the production system performs better than other release mechanisms.

Introduction

Theory of constraints is the result of pioneering work of Dr Eliyahu Goldratt [1]. Initially, Goldratt presented his rules of production scheduling in the form of software called Optimized Production Technology (OPT). He later on presented his ideas of production Management in the form of a book “ The Goal”. Goldratt [1] highlights the difficulties faced by most of the production managers in their day-to-day work. He presents some of the unorthodox methods to solve these problems. In this paper, an attempt has been made to capture the salient points from “The Goal [1]”. These points can act as a ready reference for the academicians and the practicing managers. Most of the production managers believe that quality, employee turnover

rate and unreliable suppliers are their biggest problems. Goldratt [2], through the use of computer simulations highlights that even if these problems have been eliminated, the output of the plant does not reach an optimal level as long as it is managed according to the cost world and the policy constraints are not eliminated.

Most of the researchers have used Simulation studies to evaluate the performance of DBR mechanism in comparison to other mechanisms. The findings of these researchers are contradictory. For example R Verma [5] concludes that TOC/OPT performs better than local optimization technique but poorer than Management Science approach. Daniel & Guide [6] concludes that DBR can be used to improve the output of recoverable manufacturing system. Chakravorty & Atwater[7] finds that the performance of DBR is very sensitive to the changes in the level of free goods (Goods not passing through the bottlenecks) released into the system. So, the schedulers using DBR need to be cognizant of how free goods are accepted and released. Duclos & Spencer [8] finds that the constraint buffer is very useful to improve the output of the production systems. Chakravorty [9] concludes that DBR release mechanism outperforms modified infinite loading and immediate release mechanisms.

The following paragraphs highlights the importance of measurement systems, drawbacks of conventional measurement systems, measurement system suggested by TOC, reasons for build up of and drawbacks of WIP and finished goods inventory, methods to identify and utilize the bottlenecks and the guidelines for operational planning and control. The paper ends with the concluding remarks.

Measurement systems: -

Measurements are needed for the following purposes: -

a. For control i.e. to know to what extent a company is achieving it's goal of making money.
b. Measurements should induce everyone to do what is good for the organization

Conventional measurement parameters like profit margins, capital investment, direct labour content, scrap factor, parts per shift etc. can be misleading. Generally these measurement yardsticks are faulty in most of the organizations and stress on local optima at the cost of global optima. These measurements make workers & management do what is bad for the organization. Efficiency is one of the most commonly used yardstick to measure the performance of a production system. The following paragraph explains the problems associated with this measure

Efficiency: - It is the ratio of actual output of a workstation to the maximum output obtainable from it. Due to this measure, every workstation in the system tries to maximize it's individual output and in this attempt forgets the systems viewpoint and many a time efforts of these individual workstations are not in line with the overall organizational objectives. TOC strongly advocates that a production manager should not attempt to utilize every resource 100%. An efficiency level at which the system is able to meet the customer requirements should be the aim of each sub unit. Increasing

the efficiency of a machine. beyond the rate of demand or rate of output of bottleneck is counterproductive for the organization.

In The Goal, Jonah further states that, a plant that is running at very high efficiency level is not necessarily making more money. In fact a plant in which every thing is working every time is very inefficient. As it is impossible to have a perfectly balanced combination of all the resources required in a production system. So, if everything is working every time, some of the machines must be overproducing some of the things.

A balanced plant is one, which has a capacity exactly equal to the market demand. The production mangers spend a lot of time and energy to try to balance their resources exactly to the market demand. TOC questions this approach by saying that closer a production system gets to a balanced plant the closer it gets to bankruptcy. When the capacity of each resource is trimmed exactly to the market demand, the throughput rate goes down and inventory goes up. It happens because of simultaneous existence of following two factors in any production system

a. Dependant events
b. Statistical fluctuations

Dependant events: - An event or a series of events that must take place before another can begin. The subsequent event depends upon the ones prior to it.

Statistical fluctuations: - The timing of many activities varies from cycle to cycle around some mean value. This random variation is called statistical fluctuations. People believe that since, these variations are on either side of the mean and should average out in the long run and a worker should be able to meet the production targets set on the basis of average cycle time. It would be true if an activity exists in isolation and does not depend upon any other activity. But this is not true for most of the activities in a production system. Therefore, these fluctuations do not average out and the maximum negative fluctuation of predecessor activities becomes the minimum fluctuation of a subsequent dependent activity e.g. when a workstation finishes the job in less than average time, it has to wait for the material being processed on a previous machine. Similarly, if the predecessor workstation takes more than the average time, the succeeding workstation may again be starved, this time is lost forever and the resulting average output of the workstation is less than the target output. The amount by which this actual output falls short of average depends upon the position of a workstation in the chain. More the distance of a workstation from the first workstation more is the deviation and vice – versa. Also the fluctuations are more for manual operations as compared to the automatic/machine-controlled operations. Generally a production system is composed of a combination of these two types of operations. This factor further reduces the throughput as compared to a totally manual system or an automatic system built to same average capacity. If the operation sequence can be arranged in the descending order of speed, then no workstation will be forced to remain idle due to starvation. But it will result in high work in process inventory. The true capacity of a resource depends upon its design capacity and it’s position in the manufacturing chain. The efforts of managers to level the capacity of each resource according to demand without taking into account statistical fluctuations is not good and such systems generally fall short of their target

output levels. So, some workstations need to have extra capacity particularly the workstations towards the end of the chain. The production managers should not just keep on producing during the periods of low demand just to maintain efficiency and should not choose any arbitrary efficiency target to measure the performance of various machines. It is to be determined on the basis of system constraints.

Appropriate measurement yardstick: -

The ultimate goal of every organization is to make money. Based upon this goal, Goldratt has given guidelines to distinguish between productive and non-productive actions.

Every action of the organization that takes it closer to its goal is productive and the action that does not is non-productive.

Achieving high quality levels, recruiting and retaining right employees, acquiring appropriate and cost effective technology are all very important but they are not the goals in themselves. They are the means to achieve the goal of making money. If a plant is not making money, all the above means cannot make a plant survive. There have been examples of many companies that were good in these factors but even then they were not profitable and were closed down.

Making people work and making money are not synonymous.

The three measures (at organizational level), which can be used to determine whether a plant is making money now and will continue to do so in the future, are

a. Return on investment
b. Net profit
c. Liquidity

But any one of the above measures can be improved at the cost of the other. So, the objective is to improve all the above measures simultaneously

TOC suggests three measurement yardsticks at the operational level. These are

a. **Throughput** It is defined as the rate at which a system generates money by the sale of goods and services which it produces
b. **Inventory** It is the money that a system has invested in purchasing the things that it intends to sell.
c. **Operational expenses** It is the money, which a system spends to convert inventory into throughput.

Throughput is the money coming into the system; inventory is the money currently inside the system and operational expenses is the money we have to pay to make the throughput happen. Therefore all the parameters can be defined in terms of money. Any activity that effects favorably at least one of the above three measures is productive. Otherwise it is unproductive

People trained in the conventional cost accounting methods use to doubt the sufficiency of these three measures to cover all the expenses and the functional areas of the organization. But Alex explains that every expenditure in the organization can be classified to fall in one of the above three categories e.g. investment in assets is inventory, annual depreciation is operational expense, Similarly for scrap, Loss in value is operational expense and the money that we can realize by selling the scrap is inventory.

Improvement is not just cost saving. It is improving at least two parameters out of inventory, throughput and operating expenses simultaneously. When people are

asked to reduce costs, they concentrate on cutting those costs that does not reduce operating expenses. e.g. saving a few set ups on the non- bottleneck operations does not save any operating expense.

Increasing throughput should be the first priority followed by reduction in inventory and reduction of operating expenses should be done at the end. People generally follow reverse order.

Reasons for the build up of WIP & finished goods inventory

Inventory is generally built up due to manufacturing something beyond immediate requirement. This act leads to WIP and finished goods inventory of some of the components/finished products and also the shortage of others at the same time. Secondly, in order to maintain high efficiency figures during the periods of low demand, the production systems just keeps on working and the result is build up of inventory. Bottlenecks not only account for the WIP in front of them. They also account for the inventory waiting at the final assembly stage because of the non-availability of components that require processing on bottlenecks. If we analyze the finished goods inventory in any system. Mostly, it is found to be composed of either obsolete items or of those items that do not pass through the bottlenecks. Inventory is shown on the assets side of the balance sheet and is valued at cost plus value addition. In such measurement systems, reduction of inventory will be shown as a loss in the financial statement that is a wrong thing and encourages building of inventory. The managers should understand that inventory is a liability and not an asset. With decrease in inventory, quality problems gets highlighted very quickly, manufacturing lead-time goes down thereby increasing the on time delivery performance of an organization. During the periods of low demand bottlenecks should produce according to the market demand and need not be loaded to 100% capacity because that will result in finished goods inventory only. But TOC does not advocate reducing the inventory to zero at all the points in the production system Rather it encourages maintaining certain pre-determined inventory levels before bottleneck operations. This will ensure 100% utilization of bottleneck resource even during breakdown at a workstation prior to bottleneck operation. After repair, the prior operation has to work to meet the current demand and also rebuild the buffer stock in front of the bottleneck. The amount of buffer required at any workstation is inversely proportional to the spare capacity available at the previous workstations.

Identification and management of bottlenecks:-

Bottleneck resource: - It is the resource whose capacity is less than or equal to the demand placed on it. In production management, a manger should not balance the capacity to the demand. Rather they should balance the flow of products through the system to the demand. This flow rate should be equal to the rate of output of bottleneck. Since the output rate of the bottleneck is less than the demand so efforts are to be concentrated upon increasing the rate of output of the bottleneck. Bottlenecks are neither good nor bad but they are a reality that always exists in a

production system. A production system is composed of a few bottleneck and a large number of non-bottleneck machines. These equipments are connected to each other by routing of the components.

Capacity Constrained resources:- These are the resources that have the capacity slightly more than that of bottlenecks & can sometimes be responsible for the shipment delays if not managed effectively. So, one has to concentrate on planning, control and efficiency improvement of such resources as well.

Fictitious bottlenecks: - These are the resources with capacity almost equal to the capacity of their pre-decessors and occurring towards the end of the chain.

To study the effect of bottlenecks on the total output of a system, the following four simple situations can be drawn

1 Y (Non-bottleneck) ⟶ X (bottleneck)

If a non-bottleneck operation feeds a bottleneck operation as shown above, then increasing the rate of production of Y beyond the rate of production of X results in increase of the WIP inventory in front of X and not the throughput. So, schedule the output of Y according to the output of X and let it remain idle for part of the time.

2 X (bottleneck) ⟶ Y (Non-bottleneck)

If a bottleneck operation feeds a non-bottleneck as shown above the rate of output of X will limit the rate of output of Y. The equipment Y will be forced to remain idle for part of the time depending upon the difference in the rate of output of X and Y.

3 X (bottleneck) ⟶ Assembly

If a bottleneck feeds the assembly directly, the rate of output at the assembly line will be limited by the rate at which the material is fed to it from X (Bottleneck)

4 Y (Non-bottleneck) ⟶ Assembly

If a non-bottleneck directly feeds the assembly. Since it is a non-bottleneck it's maximum rate of output will be more than the market demand. So, the market demand will be the constraint in this case and will limit the rate of output from Y.

Therefore we see that in all the four cases, bottleneck determines the rate of throughput of the system. Therefore, it is very important to manage the bottleneck properly and to improve it's output.

Guidelines to identify bottleneck: - The production managers generally face difficulty in identifying the bottlenecks. Variation in the product-mix and production volume over a period of time further adds to this difficulty. Some of the guidelines that can be used for the identification of bottlenecks are given below: -

a. Search through the production database

b. Prepare a list of items that generally reaches late at assembly line. These are the parts that usually pass through the bottleneck operations. Study their process sheets to trace the bottleneck.
c. The operation in front of which there is always a heap of unprocessed inventory is a most likely candidate for bottleneck. One should identify the operations with high WIP visually and then check it from the database whether these operations are actually bottlenecks or not. i.e. they account for the assembly delay or not. Generally these resources with high WIP are found to be bottleneck operations.
d. The most efficient equipment in the system is usually found to be bottlenecks. The reason for this happening is that it is a general human tendency to load the most efficient machine/resource to such an extent that it becomes bottleneck.
e. If the buffer stock at any machine starts increasing it is an indication that it is going to become bottleneck. On the other hand, if there is a decrease in the buffer stock at any machine. Then probably some of the previous operations is going to become a bottleneck
f. Once the production problems are rectified, the marketing may become the next bottleneck. Thus the bottleneck shifts from being inside the system to outside the system.
g. Bottlenecks may change with change in product or volume mix. So, one should not think that a particular resource will always remain a bottleneck.

Some methods to increase the capacity of bottlenecks: -

a. Use the bottlenecks to full load during every production cycle.
b. Take some load from the bottleneck operation and give it on some other machines, if possible, even at the loss of some efficiency.
c. Reduce the set up time at the bottleneck. SMED approach can be used to do it.
d. If possible, increase the rate of output at the bottlenecks.
e. Try to see if the components scheduled to route through bottlenecks actually require that operation or not. There may be some components that can function well without being processed through the bottlenecks. Eliminate bottleneck processing on such components.
f. Inspect the components before processing by the bottlenecks. That will save the bottlenecks from processing those components that are already defective and will ultimately be rejected. In this way, TOC helps us in inspection planning.
g. Implement strict process control at bottleneck and post bottleneck operations. So that there is absolutely minimum rejection of components during processing on bottleneck and post bottleneck operations. In this way, TOC helps us in identifying the operations where statistical process control is most beneficial. Identify and rectify the quality problems at the bottleneck and the post bottleneck operations and then at the remaining operations.

h. If we do not use control mechanisms then the variations in the cycle time are more in manual elements. Just introduction of some control mechanism reduces this variability substantially. So, introduce process control mechanisms on such operations.
i. Do not load the bottlenecks with the components required for future anticipated demand till they have not processed all the components required for the existing demand.
j. If by wasting some resources, like providing extra laborers, we can increase the efficiency of bottlenecks, it should be done without hesitation.
k. When a bottleneck is processing one lot. Prepare the second lot so that it can be immediately taken up for processing after completion of first lot e.g. interchangeable worktables for work piece holding reduce the machine stoppage time due to job set-ups.
l. If we reduce the batch size at the non-bottleneck operations, the movement from one workstation to next workstation becomes quicker. It helps in reducing the occasional idleness of workstations because of non-arrival of material from the previous workstation.
m. Provide suitable Time buffer between material release point and the bottleneck.

Some rules to guide the operational planning & control: -

Some of the rules for operational planning and control, mentioned in The Goal are given below. As indicated by Blackstone & Cox [3] many of these rules are a part of nine OPT rules.

a. An hour lost at bottleneck is the hour lost for the entire plant as plant looses the throughput for one hour and if the other resources are working during this time they will be building WIP inventory only there by increasing the carrying cost and not the throughput.
b. An hour saved at the non-bottleneck operation is a mirage only. By saving time at the non- bottleneck we do not make any real savings i.e. the operational expenses remain the same.
c. Since there is idle time at non-bottleneck operations, we can reduce the batch size at these operations. People resist such moves by saying that it will reduce the efficiency. This argument is not valid as the non-bottlenecks have spare capacity. We will be utilizing that time only. By doing so, WIP definitely goes down and response rate goes up.
d. Production shop floor people generally manufacture more components of a type than required for immediate orders just to reduce the number of set ups. This tendency results in excess of some components and shortage of others. There is no need to use EOQ formula to calculate the economic order quantity for non-bottleneck operations.
e. Balance the rate of flow to the demand and not the capacity.
f. Many a time, only one or two components get delayed in purchase or production. But this non-availability of one or two components delays the whole assembly and fulfillment of the customer order in time. So, the

Pareto 80-20 rule is to be modified to 99.9 – 0.1 rule. This point further highlights the importance of proper management of bottleneck operations.

g. A method is needed to communicate the shop floor people about the components that are late or will pass through the bottlenecks and thus require immediate processing by pre and post bottleneck operations. Color-coding can be used for this purpose.
h. The workers are given incentive for the total output given by them during wage period (day, week or a month). This method does not take into account the variations in the production rate during the production period. Reduction in this variation is very important for the increasing the throughput of an organization. So, the incentive system needs to be modified in such a way that incentives are given for consistent performance also. The output of an individual worker is constrained by the previous operations. Therefore, individual incentive system is not effective.
i. Determining the cost of operation of bottleneck in isolation is wrong. As it does not help in knowing the true importance of a bottleneck. To highlight this importance, the hourly cost of a bottleneck operation should be taken as the cost of operation of entire plant.
j. Production scheduling at the bottlenecks can be used to predict the timing of dispatch of a customer order with fair degree of accuracy.
k. Activating a resource and utilizing a resource are not synonymous. Activating means just keeping it busy while utilizing means making it work towards achievement of the goal of the organization.
l. Use the timing of processing at bottlenecks to signal the release of next lot of materials for processing in the system. This can be done by online production data entry at least at the bottlenecks.
m. Information is not the pile of reports generated periodically. It is the answer to the questions asked by the users.

n. Socratic methods are a very powerful approach to guide the people to develop strong commonsense solutions.
o. Lead-time for a new customer order should be based upon the volume of work in hand at the bottleneck.
p. Smaller the batch size, smaller the production cycle time and smaller the production cycle time faster the delivery/response rate.
q. Splitting a bigger order into many smaller orders is goof for the buyer as well as the supplier.
r. Time that material spends inside the production system is composed of four components as given below: -

I. **Set up Time:-** The time taken to reset the machine before the production of a new lot on it.
II. **Processing Time:-** The time taken by the machine to actually process a component.

III. **Queue Time:-** The time during which a components waits in front of a machine for it's turn to get processed on the machine.
IV. **Wait Time: -** It is the time when a part is waiting for another part to join at subassembly or final assembly stage.

For the parts passing through the bottlenecks, queue time is the major time factor. For the parts not passing through the bottlenecks, wait time is the major time factor. To take care of both of these factors, material release into the system has to be controlled. TOC advocates the use of **Drum- Buffer- Rope** mechanism for this purpose.

Procedure for ongoing improvements: -

The book also lists the five-step process for ongoing improvement. These steps are

a. Identify the system bottlenecks
b. Decide how to exploit these bottlenecks
c. Sub-ordinate everything else to the above decision
d. Elevate the system bottlenecks
e. If in the above process a bottleneck has been broken, go back to the first step. But do not allow inertia to cause a system constraint.

These steps have been explained in greater details by Goldratt [3] in his book "Theory of constraints"

Once a bottleneck has been improved/elevated, a new bottleneck will appear. It can be an erroneous policy or low market demand. When the nature of bottleneck changes, we have to change the way we operate non-bottlenecks as well. When we release the material according to the capacity of the bottlenecks, then everything becomes important. In this situation, every operation should process the material on first come first served basis.

A hole means shortage of some material. If, in a production system, the number of holes increases beyond certain limit then it is an indication of insufficient capacity.

Conclusion

Most of the Production Managers frequently face a situation in which they have piles of inventory of the items not needed and a shortage of the products required by the market. Goldratt through his novel " The Goal" tries to teach us that most of these problems are the result of wrong paradigms that the managers develop and use to manage the production systems. If the managers are seriously interested to improve their production systems, they must come out of the cost world and should learn to manage the system according to the throughput world. The managers should stop bothering about each and every link of the business system and should concentrate their efforts on the identification and improvement of the bottleneck operations. Managing the system according to the bottleneck and TOC will show a slight negative downturn in the value of conventional measurement parameters like efficiency value. But it will increase the cash flow and the due date performance of the organization. The book also talks about five-step generic process for on going improvements. Although the book gives the example of a job shop production system

but the process of ongoing improvement suggested in the book is generic in nature and can be applied to any business function and type of production system. The book talks about new shop floor procedures, measurement methods, quality control/improvement methods.

References

1 Goldratt, E. M and Cox, J., *The Goal: A Process of Ongoing Improvement.* North River Press, Croton-on-Hudson, New York, 2nd revised edition 1992.

2 Goldratt, E. M., Production The TOC way. North River Press, Croton-on-Hudon, New York, 1st revised edition 2003.

3 Goldratt, E. M., *Theory of Constraints.* North River Press, Croton-on-Hudon, New York, 1990.

4 Fry, T.D., Blackstone, J.H. and Cox, J.F., 1992. An analysis and discussion of the optimized production technology software and its use. Prod. Oper. Mgmt., 1: 229-242.

5 R Verma, 1997 Management Science, Theory of Constraints/Optimized Production Technology and Local Optimization *Omega, Int, J. Mgmt Sei.* Vol. 25, No. 2, pp. 189-200

6 Daniel, R. Guide Jr. Scheduling with priority dispatching rules and drum-buffer-rope in a recoverable manufacturing system Int. J. Production Economics 53 (1997) 101--11

7 Satya S. Chakravorty, J. Brian Atwater The impact of free goods on the performance of drum-buffer-rope scheduling systems Int. J. Production Economics 95 (2005) 347–357

8 Leslie K. Duclos, Michael S. Spencer The impact of a constraint buffer in a flow shop Int. J. Production Economics 42 (1995) 175-185 E

9 Satya S. Chakravorty An evaluation of the DBR control mechanism in a job shop environment Omega 29 (2001) 335–342