Proceedings of the 2011 International Conference on Industrial Engineering and Operations Management
Kuala Lumpur, Malaysia, January 22 – 24, 2011

Dynamic Buffering of a Capacity Constrained Resource via the Theory of Constraints

Olufemi Adetunji, Kris Adendorff and VSS Yadavalli*
Department of Industrial and Systems Engineering
University of Pretoria, Hatfield, Pretoria 0002, South Africa

Abstract

A model is proposed to optimise flow through a production system. Such flow dynamically provides the optimal level of Work-in-Process (WIP) inventory in the system, giving due consideration for the impact of system utilisation on the level of WIP. The production environment assumed is an M/M/1/∞ queue and the range of utilisation $0 < \rho < 1$ is considered. In a steady state, this model achieves the result of production pull environment (system state dependence) without the need to continuously monitor the job completion status of the resources. The model is also compared to a buffer-centred model.

Keywords
Flow, Buffer, Theory of Constraints, Markov process

1. Introduction

The determination of the size of an inventory buffer ahead of a critical resource appears to be one of the main issues deserving of attention in the application of the Theory of Constraints (TOC). This seems justified since excess inventory is a perennial problem that the technique is meant to address. Such production systems of interest have some level of statistical (natural) fluctuations in the processing time such that if the resource has an unplanned idle time, planned throughput may be lost. Since it is almost impossible to completely eliminate all forms of uncertainty, there is always a need to accommodate some slack in a system of the nature under consideration. A slack is usually either in the form of reserve capacity or inventory. System slack serves to ameliorate the effects of natural variations that could otherwise lead to the loss of system throughput.

The Theory of Constraints opts to employ the slack of excess capacity to respond to system contingencies that arise due to the natural variations in its processes. It is, however, still impossible to eliminate buffer inventory completely from such systems. It is essential to have a level of inventory necessary to decouple the system in some critical areas of the production network. Such critical stations are allowed time-buffers to maintain throughput, which is arguably one of the most important features of the system. The implication of the foregoing is that the level of inventory held in strategic positions is very important in the achievement of the system profit goal. This may explain why a lot of effort in improving the practical potency of the Theory of Constraints has been devoted to managing this type of inventory. The importance is emphasised by the use of the synonym "Drum-Buffer-Rope (DBR) system" for this Philosophy of Management, where the drum is essentially the critical station, and the buffer ahead of it is used to construct a name together with the third word, the rope, which also indicates how the entire system's production is scheduled.

An important question is the relationship between the Work in Process (WIP) Inventory and the flow rate of the system. The amount of inventory that is present ahead of any workstation is not only a function of the strategic buffer placed ahead of such station, but also of the rate of flow of the products through that station. The effect of resource utilisation on the average throughput time and consequently the average number of inventory in the system is well documented in literatures. Some good references are [1] pp22-37 and [2], pp264-349. A well known equation is Little's law that states that

$$Work - In - Process\ Inventory\ =\ Throughput\ time\ X\ Throughput\ rate$$

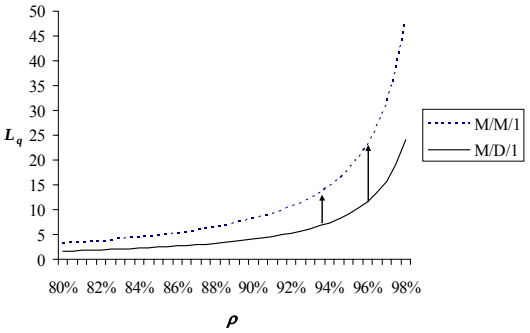

Figure 1: Curse of utilisation and variance (Webster 2008, pg 176)

This shows that the quantity of inventory ahead of the critical station cannot be determined as if it is independent of the flow rate through the station, especially as the station works close to its full capacity. The effect of utilisation, also termed the curse of utilisation by some authors, (e.g. [3]) is presented in figure 1. This diagram represents the behaviour of an $M/M/1/\infty$ queue before it becomes a bottleneck (BN) ($i.e. the\ range\ 0 < \rho < 1$). It could be seen that the queue length grows non-linearly as the resource transits from a Non-Bottleneck (NBN), to a Capacity Constrained Resource (CCR) and towards a BN. The graph slopes up very quickly as the utilisation approaches full utilisation. This makes it imperative for every manager to place this effect in context as considerations are made about the loading of the system to cover more throughputs, thereby balancing such throughput returns against the possibility of a skyrocketing inventory cost. Only few authors appeared to have highlighted the possible negative effect of high capacity utilisation on system performance in a TOC environment. A known example is Chakravorty and Atwater [4], and even at that, the implication of the holding cost of the WIP has not been discussed in any known article. This relationship between utilisation, WIP and systems inventory cost is the main thrust of this paper.

1.1. Some Relevant Features of TOC

Ronen and Starr [5] stated some outstanding features of the OPT technique (now commonly referred to as the TOC). Two of these are the "unavoidable" statistical fluctuation of the input arrival and service times; and the dependence of processes, one on the other, which further worsens the variabilities. This then dovetails into the effect of such on the WIP discussed earlier.

Another important feature is that this technique can work only in an environment that has a stable schedule, i.e. the product mix (volume and variety) have been stabilised. This is apparent because it will be difficult to designate a manufacturing resource as the critical one since its criticality will depend on the production schedule and the attendant bill of routing for the products involved. This paper, therefore, assumes a stable production environment and chooses the simplest of such case, perhaps where only one product is produced, and uses that to illustrate how the flow and the buffer in such systems are jointly determined, in tandem with a previous work done assuming a typical $M/M/1$ queuing environment as a reference.

The organisation of the remaining portion of this paper is as follows. First is a review of some pertinent literature in this area. Next is a review of Radovilsky's model and its key results. Next follows a critique of Radovilsky and changes of some parameters and inputs. Next, the model for this work is presented. This is then followed by a numerical example, suggested areas for further research and conclusions.

2. Literature

Various authors have written about the applications of the TOC in diverse contexts. But the review here would be limited to those applications that have focused on the determination of the buffer size to be used in the management of the network or the critical station of the system, especially in a quantitative manner.

Many other researchers have proposed various heuristics ranging from using the work equivalence of half the manufacturing lead time, a quarter of total lead time or even stating that initial estimation is unnecessary since it is an ongoing improvement process [6].

Most authors that estimated buffer size quantitatively have been motivated by the failure of the upstream section of the critical resource. Among such papers are Han and Ye [7] that used the reliability theory to model the machines in the system as having two states of up and down to construct a relationship between the feeder and the fed machines. Page and Louw [8] used a $GI/G/m$ queues and a queuing network analysis (multiproduct open queuing network modelling method) together with the assumption of normality of flow times and a chosen service level to determine the buffer size. So [9,10] reports an approximation scheme to determine buffer capacities required to achieve the target performance level in a general flexible manufacturing system with multiple products and another on the optimal buffer allocation problem of minimizing the average work-in-process subject to a minimum required throughput and a constraint on the total buffer space. Simon and Hopp [11] studied a balanced assembly line system being fed from storage buffers. Processing time is assumed deterministic. Battini et al [12] developed efficiency simulative study for the allocation of storage capacity in serial production lines and an experimental cross matrix was provided as a tool to determine the optimal buffer size. Li and Tu [13] presented a constraint time buffer determination model. The model first proposes a machine-view's bill of routing representing a structure that serves as a fundamental structure for formulating and computing the maximum time buffer. By incorporating the Mean-Time-To-Repair (MTTR) of each feeder machine, a mathematical relationship was formulated and the time buffer computed. Powel and Pyke [14] studied the problem of buffering serial lines with moderate variability and a single bottleneck. The focus was essentially on how large variations in mean processing times of machines affect placement of equal buffers between stations. Some more recent contributors in the management of buffers and flows in a TOC environment include 15, 16 and 17.

Not much authors appear to have focussed on buffering exclusively for the purpose of process variation with the exclusion of resource failure; and to our knowledge, none considers managing flow in a TOC environment with considerations for the cost of keeping WIP inventory relative to the gain of achieving high level of utilisation. Utilisation directly affects the level of in process inventory, and by extension the buffer size, in any system with stochastic input and processing time as typified in an $M/M/1$ queue. The work that appears to have focused exclusively on the critical work station only and in a stochastic processing time environment seems to be that of Radovilsky [18]. Our paper seeks to use the model in [18] as leverage for flow a process, which seems to be a good way to manage the case $\rho < 1$ not included in his model.

3. Model Presentation

In the models presented in the literature survey, the goal, generally, seems to be to determine the optimal size of the buffers (constraint or others). These models presuppose that covering the throughputs to meet the market demand to the best of the capacity of the constraint resource would always generate profit for the company. But this may not always be true. While more profit may always be realised from the sale of every extra unit of product, the cost that would have resulted from the WIP inventory held in the system as a result of the curse of utilisation might have contributed more expense that the profit realised. This is an often ignored reality in most models. The goal here is to rather seek to determine the optimal flow rate and study how the system profit goal behaves as a result of this flow.

This paper, therefore, seeks to contribute to how decisions about flow should be made in an $M/M/1$ arrival and processing system in a TOC environment. This is then placed in the context of strategic buffer placement in such an environment, bearing in mind the contributions the profit per unit product, holding cost per unit product per unit time, and the resource utilisation, ρ, have on the profit goal of the organisation. The implication of the Markovian environment is that the holding cost may indirectly be an exponential function, since it is affected by the rate of growth of the queue size ahead of the critical station.

The variables and notations adopted in this paper are consistent with the ones used in [18]. This is to allow for ease of comparison. So, an optimal flow rate is being sought to maximise the profit function of the system. From this, the average queue size (WIP) is to be retrieved. Other decisions about what size of buffer to allow would then be made based on these functions. It is also assumed that only one product is being produced in this system, and only a processing centre is involved. This is to simplify the analysis. The objective is the maximisation of the Net Profit function which is defined as

$$NP = TH - OE \quad (1)$$

$$TH = \mu(1 - P_0)C_{TH} \quad (2)$$

$$OE = L_S C_{OE} \tag{3}$$

where
NP is the Net Profit,
TH is the throughput rate,
OE is the Operating Expense (incurred during the same time window as the throughput, and is assumed here to be made up of only the holding cost)
μ is the rate of service at the resource over a stated time interval
$P_0$is the probability that constraint buffer of the resource is empty
C_{TH} is the profit earned from selling a unit of output
L_S is the average queue length ahead of the critical resource
C_{OE} is the inventory cost per unit (product-time)
K is the buffer size
D is the demand rate from the market
ρ_D is the level of utilisation based on D defined as the ratio D/μ.

The process is assumed to follow the $M/M/1/\infty$ queue and so, P_0 and L_S are substituted with the following in the NP equation:

$$P_0 = 1 - \rho \tag{4}$$

$$L_S = \frac{\rho}{1-\rho} \tag{5}$$

So, the net profit equation becomes

$$NP = \mu\rho C_{TH} - \frac{\rho C_{OE}}{1-\rho} \tag{6}$$

This makes the optimal ρ to be

$$\rho^* = 1 - \sqrt{\frac{C_{OE}}{\mu C_{TH}}} \tag{7}$$

Recovering the optimal buffer size simply becomes the steady state queue length, L_S, corresponding to ρ^*, and this is

$$L_S = \sqrt{\frac{\mu C_{TH}}{C_{OE}}} - 1 \tag{8}$$

And the optimal net profit function, NP^*, becomes,

$$NP^* = \left(\sqrt{\mu C_{TH}} - \sqrt{C_{OE}}\right)^2 \tag{9}$$

3.1. Motivations for optimising with respect to ρ

Before analysing and making deductions from the model proposed in this paper, some benefits of optimising the profit with respect to the flow rather than the buffer size would be pointed out. Firstly, the effect of exponential queuing time on the system profit as the flow rate gets closer to the full utilisation of the resource capacity is more easily observed. It may be more profitable to allow lost throughput than to buffer for process variabilities. This will be further discussed. Secondly, it is easier to extend the model to other queuing cases. This is because ρ is a more pervasive variable than K. While K is found in capacitated queues only, ρ is the main variable of interest of all queue types. This will make it possible to utilise other types of queues, e.g. queues with balking, perishable input, etc. Thirdly, controlling the buffer may be simply reduced to controlling the flow rate rather than monitoring the position of the buffer. The former should be easier.

4. Deductions from the model

From equation 7, one could notice that as C_{OE} decreases, other things being equal, ρ edges closer to unity indicating higher utilisation of resource. The corresponding effect is seen in L_S in equation 8 because the average queue length increases, meaning more inventory is allowed. The effect of decreasing C_{TH} is the reverse. Also, optimal buffer size increases with increase in service rate (or Bottleneck rate) of the system. The effects of an increase or a decrease in

C_{TH}, C_{OE} and μ are also apparent in equation 9. As either μ or C_{TH} increases, net profit also increases, and as C_{OE} increases, net profit decreases as expected. Figure 2 shows that the profit function declines very rapidly after the optimal flow rate is reached. This makes it unprofitable to target 100 percent utilisation of the critical resource.

5. Comparison to an earlier (Buffer-centered) model

Radovilsky [18] has derived a similar equation for the optimal buffer size by considering the process to be an $M/M/1/K$ for case $\rho = 1$. The results are that

$$K^* = \sqrt{\frac{2\mu C_{TH}}{C_{OE}}} - 1 \quad (\rho = 1) \tag{10}$$

$$\text{and} \quad NP^* = \frac{1}{2}\left(\sqrt{2\mu C_{TH}} - \sqrt{C_{OE}}\right)^2 \quad (\rho = 1) \tag{11}$$

Radovilsky's assumptions connote the BN condition, so he solved the case $\rho = 1$. He also did some numerical analysis for the case $\rho > 1$.The dynamic buffering approach proposed can be compared to the result from Radovilsky model. Numerical examples will not be used for want of space, but an attempt is made to seek a point of indifference between the two models.

Our analysis is limited to $0 < \rho < 1$ because a flow rate beyond this range is not practically feasible. In any $M/M/1$ queuing model, working at 100 percent utilisation is not possible because of the corrupting influence of variability on the build up of WIP ahead of the critical station. This has been explained in the curse of utilisation, and the implication is that inventory could theoretically build up ahead of the critical station infinitely. With $\rho = 1 \equiv \lambda = \mu$, a Markov chain in which all the states are recurrent null results, and the expected time of return to any of the states it has previously visited (i.e. returning to $L_S = n$ at any future time, given current $L_S > n$) is infinite. This implies that the queue could grow on perpetually. (An interested reader may refer to [1] pg 15 and [19] Lemma 5.33 pg 176).

There will be periods of blocking for as long as $\rho \geq 1$ in a series system that includes the critical resource somewhere along its line. The condition under which blocking will not happen is for the buffer size in equation 10 to be greater than $kL_s, k \geq 2$ for up to 95 percent coverage for most μ in equation 8. This means

$$\sqrt{\frac{\mu C_{TH}}{C_{OE}}} - 1 < \frac{1}{2}\sqrt{\frac{2\mu C_{TH}}{C_{OE}}} - 1 \tag{12}$$

The condition for this to happen is that

$$\mu < \frac{1}{2(3 - 2\sqrt{2})}\frac{C_{OE}}{C_{TH}} \tag{13}$$

This implies that the Bottleneck rate has to be very small compared to the cost of inventory relative to the unit profit. It should be noted that the unit of μ is $1/time$, the unit of C_{TH} is $money$ while that of C_{OE} is $1/(money.time)$. This means that the flow rate per time must be less than the ratio of the inventory cost per unit product per time to the profit from made from a unit product, divided by $1/[2(3 - 2\sqrt{2})]$. Very few products will probably fulfil this. This makes us to seek to optimise ρ in the CCR.

6. Conclusion

A model has been presented for the management of flow in a CCR. The focus of the model is to allow the optimal flow rate to dynamically buffer a DBR system for statistical process fluctuations, with no breakdown of upstream stations. More so, it is easier to control such system with the dynamic buffering approach through ρ than it would likely be through monitoring of buffer size because it is not necessary to build up any inventory ahead of the CCR before regulating the feed rate of the CCR line. With the optimal ρ already determined, the system dynamically adjusts the optimal time buffer accordingly. Also, the optimal buffer size could be retrieved indirectly from the optimum ρ, given a critical value (probability) for coverage. The elimination of the need to have the optimal buffer length involved in the derivation of the optimal Net Profit function makes it easy to extend the model to other more interesting areas like deteriorating inventory and network buffer balancing. This model incorporates the joint effect of management of input flow, system utilisation level, cost of keeping WIP and the profit from increased throughput in synergy to make better production management decisions. The result of this model can be utilised in other areas like managing NBN resources and production scheduling, which are also other possible areas for further research.

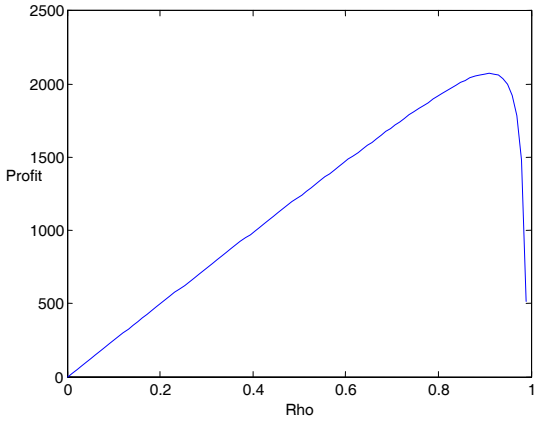

Figure 2: Plot of Net Profit against flow rate, ρ

Acknowledgements

Thanks are due to NRF for their financial support

References

1. Hopp W.J., 2008, "Supply Chain Science", McGraw-Hill/Irwin.
2. Hopp W.J. and Spearman M.L., 2008, Factory physics, McGraw-Hill International, New York.
3. Webster S, 2008, Principles and tools for supply chain management, McGraw-Hill International, New York
4. Chakravorty S.S. and Atwater J.B., 2006, "Bottleneck management: Theory and practice", Production Planning and Control, 17(5), 441 - 447
5. Ronen B. and Starr M.K., 1990, "Syncronised manufacturing as in OPT: from theory to practice", Computers and Industrial Engineering, 4(18), 585 – 600.
6. Spencer M., 1991, "Using The Goal in an MRP", Production and Inventory Management Journal 32(4), 22-28.
7. Han W. and Ye T., 2008, "Determination of buffer sizes for drum–buffer–rope (DBR) -controlled production systems", International Journal of Production Research, 46(10), 2827–2844.
8. Page D.C. and Louw L., 2004, "Queuing network analysis approach for estimating the sizes of the time buffers in Theory of Constraints-controlled production systems", International Journal of Production Research, 42(6), 1207–1226
9. So K.C., 1989, "Allocating Buffer Storages in a Flexible Manufacturing System", International Journal of Flexible Manufacturing Systems, 1(3), 223 - 237
10. So K.C., 1997, "Optimal buffer allocation strategy for minimizing work-in- process inventory in un-paced production", IIE Transactions Vol.29(1), 81 - 88.
11. Simon J.T. and Hopp W.J., 1991, "Availability and Average Inventory of Balanced Assembly-Like Flow Systems", IIE Transactions, 23(2). 161 - 168.
12. Battini D., Persona A. and Ragattieri A., 2009, "Buffer size design linked to reliability performance: A simulative study", Computers & Industrial Engineering 56 1633–1641.
13. Li R.K. and Tu Y.M., 1998, "Constraint time buffer determination model", International Journal of Production Research, 36(4), 1091 – 1103.
14. Powell S. G and Pyke D. F., 1996, "Allocation of buffers to serial production lines with bottlenecks", IIE transactions, 28(1), 18-29.
15. Woo S., Park S., and Fujimura S., 2009, "Real-time buffer management method for DBR scheduling", International Journal of Manufacturing Technology and Management, 16(1/2), 42 – 57
16. Woo S., Park S., and Fujimura S., 2007, "Buffer size setting method for DBR scheduling", IEEJ Transactions on Electronics, Information and Systems, 127(3), 416 – 424
17. Ribeiro M.A., Silveira J.L., Qassim R.Y., 2007, "Joint optimization of maintenance and buffer size in a manufacturing system", European Journal of Operational Research, 176, 405 – 413.
18. Radovilsky Z.Y., 1998, "A quantitative approach to estimate the size of the time buffer in the theory of constraints", International Journal of Production Economics, 55, 113- 119.
19. Cinlar E., 1975, Introduction to stochastic processes, Englewood Cliffs NJ, Prentice hall.