5. [§]

CHAPTER 5

DYNAMIC BUFFERING OF A CAPACITY CONSTRAINED RESOURCE VIA THE THEORY OF CONSTRAINT

[§] A modified version of this first section of this chapter has been submitted to IEOM conference, a peer reviewed international conference holding at Kuala Lumpur, Malaysia in January 2011

5.1. PART A: BUFFERING WITH ZERO SHORTAGE COST

5.1.1. INTRODUCTION

The determination of the size of an inventory buffer placed ahead of the critical resource is one of the main issues deserving of attention in the application of the Theory of Constraints ($\boldsymbol{TOC}$). 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 (natural) statistical 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 the arguably one of the most important measures of the system. The definitions of terms such as throughput, inventory and operating expense are strictly in the context of Goldratt's Theory of Constraints.

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 ($\boldsymbol{DBR}$) system" for this Philosophy of Management, the 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 to address at the outset relates to the principal function of the buffer in this system. This question is important since it essentially relates to the buffer size, which has been dealt with extensively by the relevant literature on Inventory Control. The obvious answer is that it serves to protect the critical station which is either the Bottleneck ($\boldsymbol{BN}$) or the most Capacity Constrained Resource ($\boldsymbol{CCR}$) against loss of throughput.

While this answer seems adequate, further elucidation is required on the loss of throughput. The answer that does not seem to always be obvious, is whether the loss is due to the natural process variations that are inherent to the entire system as a result of the variation of the processing time of each work station, or the breakdown of any of the machines that are upstream to the critical station.

Another important issue is the relationship between the Work in Process ($\boldsymbol{WIP}$) 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 Hopp (2008, pp22-37) and Hopp and Spearman (2009, pp264-349).

A well known equation is the little's law that states that

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

This shows that the quantity of inventory ahead of the critical station cannot be determined as if being independent of the flow rate through the station, especially as the station works close to its full capacity. The effect of utilisation, termed as the curse of utilisation by some authors (Webster, 2008) is presented in figure 5.2.1. This diagram represents the behaviour of an $\boldsymbol{M/M/1/}\infty$ queue before it becomes a bottleneck

$(i.e. 0 < \rho < 1)$. It could be seen that the queue length grows exponentially as the resource transits from a Non-Bottleneck ($\boldsymbol{NBN}$), to a $\boldsymbol{CCR}$ and towards a $\boldsymbol{BN}$. The graph slopes up very quickly as the level of utilisation approaches full utilisation of the resource. This makes it imperative for every manager to place this effect in context as consideration is given to the loading of the system to cover more throughputs and balance the return from such increase in utilisation to have more system throughput against a possible "skyrocketing" cost of holding inventory in the system. That is about the main thrust of this chapter.

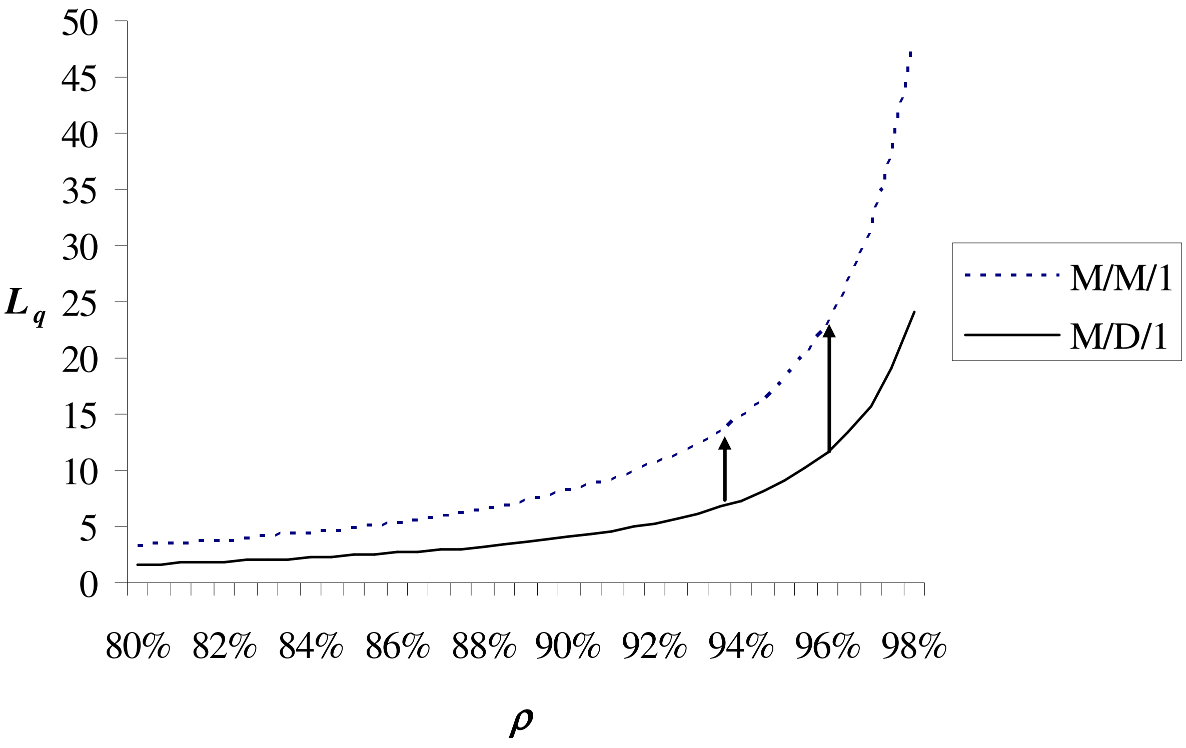


Figure 5.2.1: Curse of utilisation and variance (Webster S. 2008, pg 176)

5.1.2. Some Relevant Salient Features of the TOC

Ronen and Starr (1990) stated some outstanding features of the $\boldsymbol{OPT}$ technique (now commonly referred to as the $\boldsymbol{TOC}$). 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 problems of variability. These then dovetail into the effect of such on the $\boldsymbol{WIP}$ 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 without such stability, it will be difficult to designate a manufacturing resource as the critical one since its criticality will depend on the current production schedule of the company. This chapter, 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 $\boldsymbol{M/M/1}$ queuing environment as a reference.

The organisation of the remaining sections of this part of the chapter is as follows. First is a review of some pertinent literature in this area, while trying to identify the purpose of the buffers considered in such literature. Next is the presentation of the model. The next section presents some motivations for considering the process flow rate as an important variable when buffering decisions are being made. This is then followed by a section on numerical example, and then, the suggested areas for further research and conclusions.

5.1.3. Literature

Various authors have written about the applications of the $\boldsymbol{TOC}$ 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 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 (Spencer, 1991).

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 (2008) 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 (2004) used a $\boldsymbol{GI/G/m}$ queues and a queuing network analysis of 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 (1989,1997) 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 (1991) studied a balanced assembly line system being fed from storage buffers. Processing time is assumed deterministic. Battini et al (2009) 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 (1998) 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 $(\boldsymbol{MTTR})$ of each feeder machine, a mathematical relationship was formulated and the time buffer computed. Powel and Pyke (1996) 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 on machines affect placement of equal buffers between stations.

Not much authors appear to have focussed on buffering exclusively for the purpose of process variation and not resource failure, and to this author's knowledge, none considers, explicitly, managing flow in a $\boldsymbol{TOC}$ environment with considerations for the cost of keeping $\boldsymbol{WIP}$ inventory relative to the gain of achieving such level of utilisation. This directly affects the level of inventory, which is also supposed to be managed by the buffer size, in any system with stochastic input and processing time as typified in an $\boldsymbol{M/M/1}$ 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 (1998). This section seeks to build on Radovilsky's work, considering Radovilsky to be good for a $\boldsymbol{BN}$ system but not ideal for a $\boldsymbol{CCR}$ system.

5.1.4. 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 profit may always be realised from the sale of every extra unit of product, the cost that would have resulted from the $\boldsymbol{WIP}$ 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 chapter, therefore, seeks to contribute to how decisions about flow should be made in an $\boldsymbol{M/M/1}$ arrival and processing system in a $\boldsymbol{TOC}$ environment. This is then placed in the context of strategic buffer placement in such environment, bearing in mind the contributions the unit profit per product, unit holding cost per unit product per unit time, and the resource utilisation, ρ, 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 Radovilsky (1998). 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 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 a processing centre is involved. This is to simplify the analysis without loss of generalisation. The objective is the maximisation of the Net Profit function which is defined as

$$NP = TH - OE \tag{5.1.1}$$

$$TH = \mu(1 - P_0)C_{TH} \tag{5.1.2}$$

$$OE = L_S C_{OE} \tag{5.1.3}$$

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 on the 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 ${}^{D}/_{\mu}$.

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

$$P_0 = 1 - \rho \tag{5.1.4}$$

$$L_S = \frac{\rho}{1-\rho} \tag{5.1.5}$$

So, the net profit equation becomes

$$NP = \mu\rho C_{TH} - \frac{\rho C_{OE}}{1-\rho} \tag{5.1.6}$$

This makes the optimal ρ to be

$$\rho^* = 1 - \sqrt{\frac{C_{OE}}{\mu C_{TH}}} \tag{5.1.7}$$

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

$$L_S = \sqrt{\frac{\mu C_{TH}}{C_{OE}}} - 1 \qquad 5.1.8$$

And the optimal net profit, $\boldsymbol{NP}^*$, function becomes,

$$NP^* = \left(\sqrt{\mu C_{TH}} - \sqrt{C_{OE}}\right)^2 \qquad 5.1.9$$

Radovilsky (1998) had derived a similar equation for the optimal buffer size for considering the process to be an $\boldsymbol{M/M/1/K}$ for case $\rho = 1$. The results are that

$$K^* = \sqrt{\frac{2\mu C_{TH}}{C_{OE}}} - 1 \quad (\rho = 1) \qquad 5.1.10$$

and $$NP^* = \frac{1}{2}\left(\sqrt{2\mu C_{TH}} - \sqrt{C_{OE}}\right)^2 \quad (\rho = 1) \qquad 5.1.11$$

Radovilsky's assumptions connote the $\boldsymbol{BN}$ condition, hence, solving the case $\rho = 1$. He also did some numerical analysis for the case $\rho > 1$.

5.1.5. Benefits of optimising with respect to the 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 possible exponentially increasing 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 variability. 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 queuing 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 would be easier.

5.1.6. ANALYSIS AND DEDUCTIONS

From equation 5.1.7, 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 5.1.8 because the average queue length increases, meaning more inventory is allowed. The effect of C_{TH} is the reverse; increase in C_{TH} leads to increase in in both the flow rate and average queue length. Also, optimal buffer size increases with increase in service rate (or capacity) of the system. The effects of increase or decrease in C_{TH}, C_{OE} and μ are also apparent in equation 5.1.9; as either of μ and C_{TH} increases, net profit also increases, and as C_{OE} increases, net profit decreases as expected.

5.1.7. Numerical Analysis

The effect of using the dynamic buffering approach proposed is compared to the result from Radovilsky's model. This is done using a numerical example. But before the numerical analysis is done, an observation is raised.

In any $M/M/1$ queuing model, working at 100 percent utilisation is not theoretically unachievable because of the corrupting influence of variability on the build up of $\boldsymbol{WIP}$ ahead of the critical station. This has been explained with 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 ever visited is infinite. This implies that the queue would grow on perpetually. (An interested reader may refer to Hopp (2008, section 1.3 pg 15) and Cinlar (1975, Chapter 6, 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 except there is an infinite space in between the critical resource and the feeding resource. For there not to be blocking in the queue type considered at a specified probability level, the buffer size in equation 5.1.10 to be greater than kL_s, for most μ in equation 8. $k = 2$ for about 95 percent level. This means

$$\sqrt{\frac{\mu C_{TH}}{C_{OE}}} - 1 < \frac{1}{2}\sqrt{\frac{2\mu C_{TH}}{C_{OE}}} - 1 \qquad 5.1.12$$

The condition for this to happen is that

$$\mu < \frac{1}{2(3-2\sqrt{2})}\frac{C_{OE}}{C_{TH}} \qquad 5.1.13$$

This implies that the processing rate has to be quite 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 made from a unit product, divided by $1/[2(3-2\sqrt{2})]$. Very few products will probably fulfil this. This makes it imperative to seek to optimise ρ in the $\boldsymbol{CCR}$.

Figure 5.1.2 shows the behaviour of the system net profit before and after the optimal flow rate. This picture shows that the net profit increases somehow linearly until the maximum at the optimal flow rate, but declines very rapidly after the optimal flow rate. This shows that the curse of utilisation kicks in very strongly once the optimal flow rate is exceeded, and every marginal gain in profit is quickly eroded by the ballooning inventory cost. This indicates that it might be better not to meet all the customer demands that are between ρ^* and ρ_D. This gives a guide as to making trade off decisions in a $\boldsymbol{CCR}$ environment.

Next is presented the results of some numerical analysis in graphical form. Since Radovilsky's model uses $\rho = 1$, there is the need to scale the model so that an effective comparison can be made. It was noted earlier that full utilisation would perpetually build up finished goods inventory which, theoretically, could increase the buffer size to infinity. This would mean the cost also grows to infinity, thereby decreasing productivity accordingly (in line with $\boldsymbol{TOC}$'s technical definitions). This implies that the throughput in Radovilsky could have been overstated because it was assumed there that all output at $\rho = 1$ is throughput.

A benign alternative is to imagine that the full capacity of the station mentioned in Radovilsky is actually μ', a down-scaled portion of the actual μ, which is determined by $\mu' = \rho.\mu = \lambda$. It would also be assumed that this μ' is the production output that is guaranteed to be purchased by the market, and is the actual throughput in the context of $\boldsymbol{TOC}$. This means the constraint moves from the market to the production facility and the $\boldsymbol{CCR}$ "behaves" like the $\boldsymbol{BN}$ which now runs at 100 percent utilisation. The capacity then changes to $\mu\rho$, where ρ is what the new model determines as the actual feed rate to control the entire system to build the dynamic buffer ahead of the $\boldsymbol{CCR}$. This second scenario is, therefore, taken here as the upper bound for the Net Profit using Radovilsky's model. Based on this modification, the comparison was done.

For the purpose of this numerical illustration, arbitrary values were chosen as follows: Service rate = 50 items per time; Profit from unit sale = 50 units of money; Unit inventory holding cost = 20 units of money. For some dynamic analysis to track the behaviour of the model as a given parameter changes while others are kept fixed, an upper limit as set for the three variables that determine ρ, K and $\boldsymbol{NP}$ are as follows: Service rate = 100; Profit from unit sale = 150 units of money; Unit inventory holding cost = 100 units of money.

With all other variables held constant, figure 5.1.3 shows that optimal feed rate increases with increasing service rate; figure 5.1.4 shows that optimal buffer size increases with increasing service rate; figure 5.1.5 shows that optimal buffer size increases with increasing profit per unit sale; figure 5.1.6 shows that optimal buffer size decreases with increasing unit holding cost. It is worth mentioning that the effect of decreasing holding cost seems more drastic than those of other parameters on the optimal buffer size. This would be further buttressed when the graph of the Net Profit function is also interpreted. This is noticeable from the slopes of each of the curves. The same pattern is observed for the effect of each of the parameters on the average inventory and as such, the diagrams were not repeated.

The impact of the three key variables on Net Profit is examined in figures 5.1.7 to 5.1.10. Holding all other parameters constant, it can be seen from figure 5.1.7 that the net profit increases with increase in service rate; figure 5.1.9 shows that net profit increases with increasing profit per unit sale; figure 5.1.10 shows that net profit decreases with increase in unit holding cost. It can also be seen that the rate of decrease in net profit per unit increase in holding cost is more drastic, buttressing the initial observation with the buffer size. This is actually why the optimal buffer size drops sharply with every increase in unit holding cost.

One can also observe from the net profit function graphs that if adjustment is made for the fact that not all products made for full utilisation could be sold if the demand is less than the capacity, then, the profit margin for the proposed model seems higher than that of Radovilsky in the range $0 < \rho < 1$.

5.1.8. CONCLUSION

In conclusion, a model has been presented that has the potential for more profit in a $\boldsymbol{CCR}$ system than that which was done earlier. The focus of the model is on buffering a $\boldsymbol{DBR}$ system for statistical process fluctuations, without 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 in Radovilsky's model because it is not necessary to build up any inventory ahead of the $\boldsymbol{CCR}$ before regulating the feed rate of the $\boldsymbol{CCR}$ line. With the optimal ρ already determined, the system dynamically adjusts the optimal time buffer accordingly. Also, the optimal buffer size was retrieved indirectly from the optimum ρ. 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, which are some of the interesting areas of research to be explored after this work.

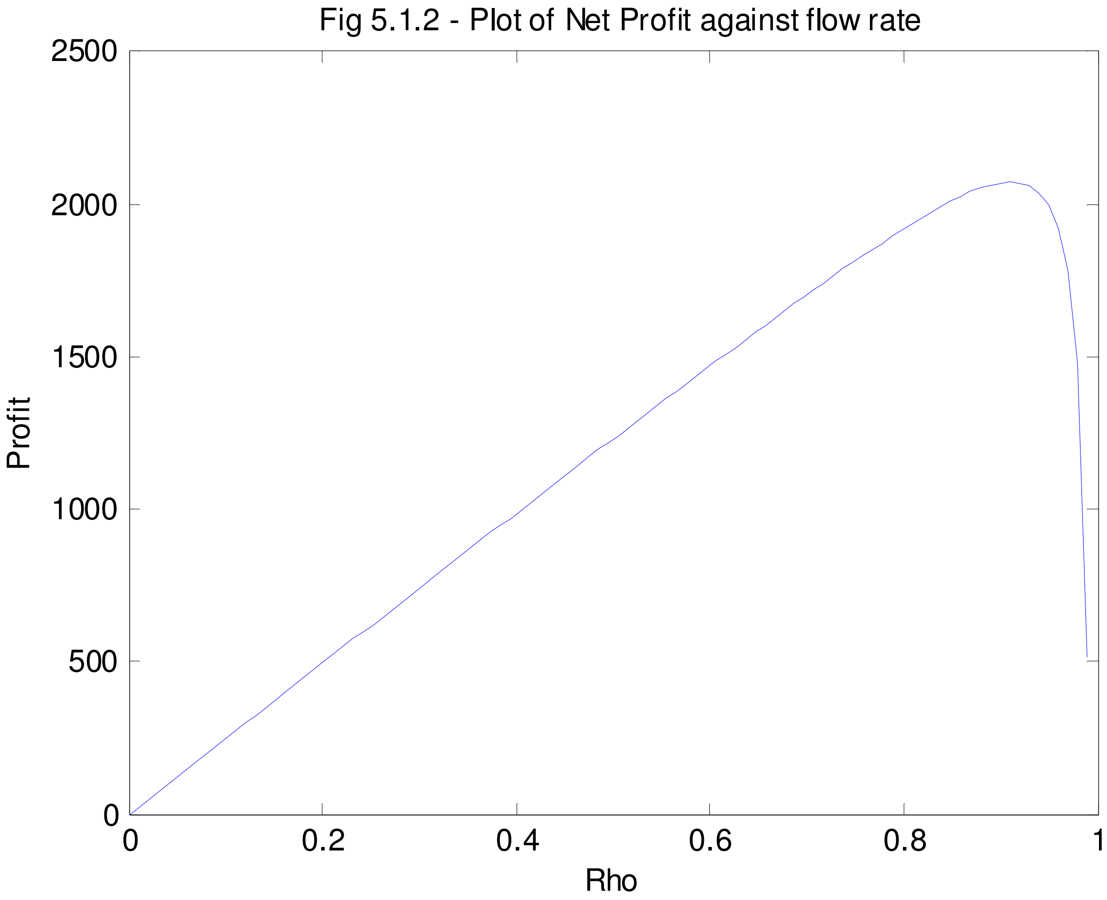


Figure 5.1.2: Net profit change with rho for $0 < \rho < 1$

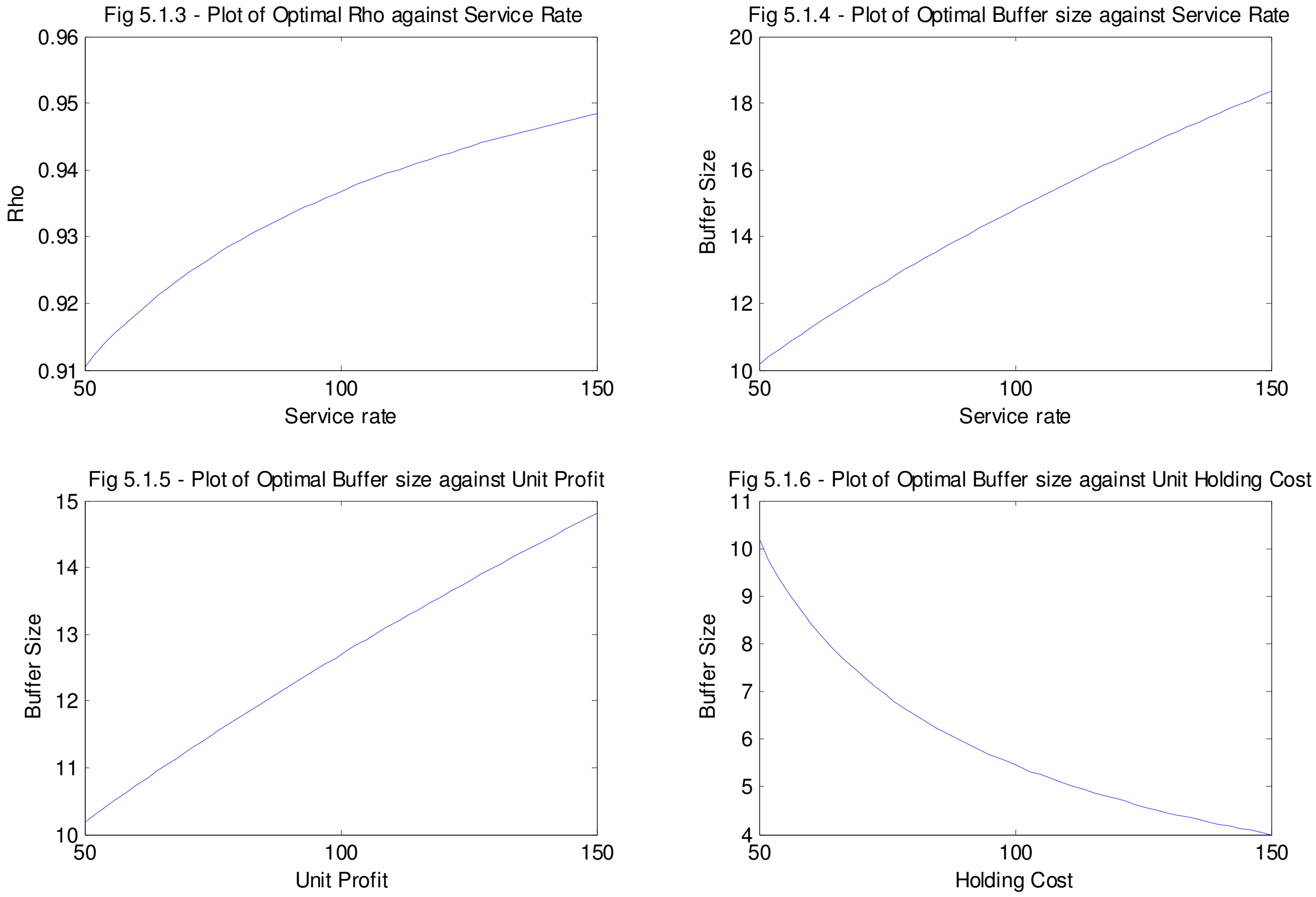
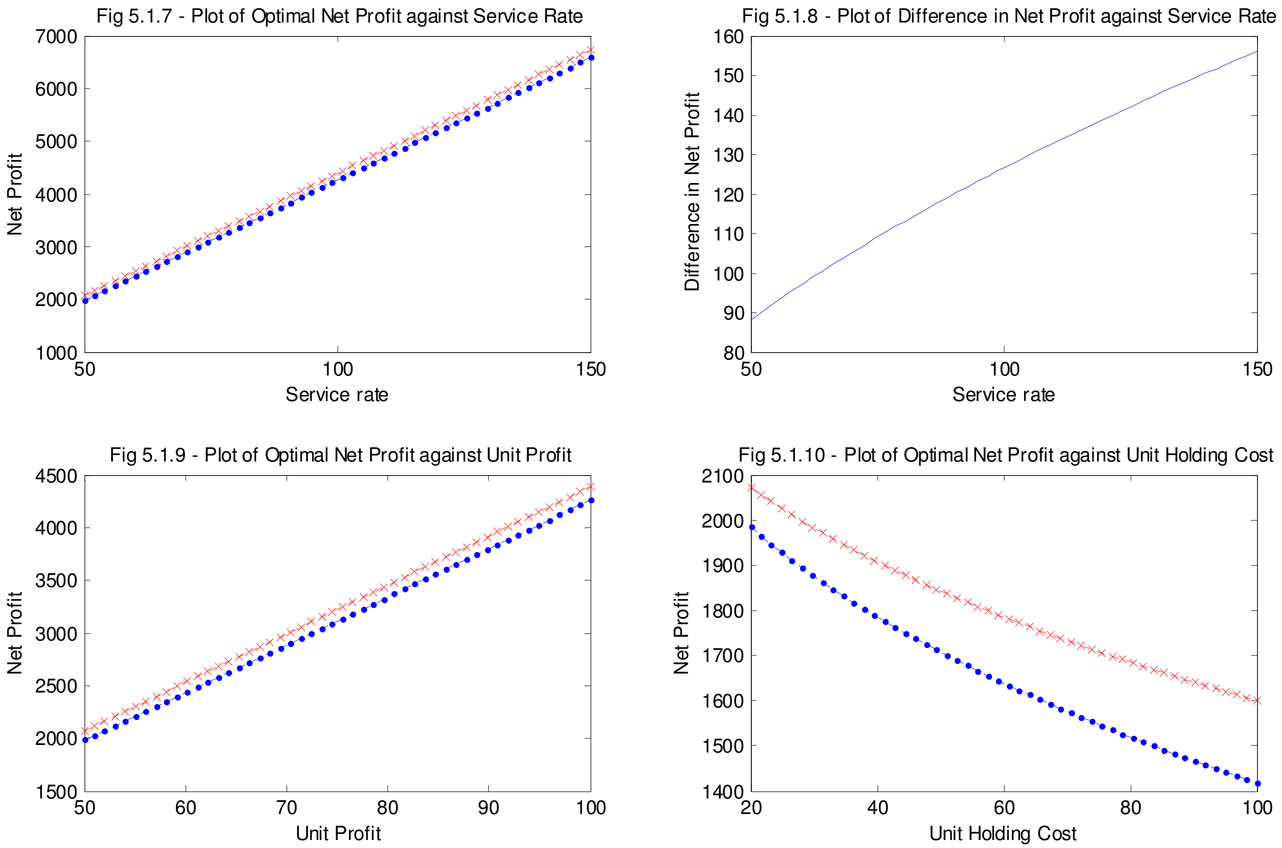


Figure 5.1.3 to 5.1.6: Changes in rho and buffer size with input parameters



Figures 5.1.7 – 5.1.10: Changes in net profit with input parameters

[**]

OPTIMISING FLOW IN AN M/M/1 SYSTEM WITH SHORTAGE COST: A THEORY OF CONSTRAINTS APPROACH

[**] A modified version of this part of this chapter has been accepted for presentation at the IASTED conference, a peer reviewed international conference, holding at Gaborone, Botswana in September, 2010.

5.2. PART B: BUFFERING WITH POSITIVE SHORTAGE COST

5.2.1. INTRODUCTION

Excessive build up of Inventory in a production system is one of the critical wastes that the Theory of Constraints seeks to attack. Based on this principle, the focus of a production system should be on maintaining flow rather than keeping inventory in the system. Inventory should only be kept ahead of the most critical work station and at some strategic points where the most critical line meet other lines in such a manner that other resources are scheduled to support this critical resource. The determination of the appropriate buffer size to place ahead of this critical resource and at the strategic points in the network is an area that has generated diverse interests, but most authors have not discussed issues of optimising flow through these lines.

In this section, the problem of the determination of the optimal rate of flow in a production system is being further considered. Such flow would automatically build up inventory ahead of the critical station, which in this case is a Capacity Constrained Resource ($\boldsymbol{CCR}$), in a production management environment utilising the Theory of Constraints ($\boldsymbol{TOC}$), and where every unit of lost production throughput has a stipulated cost. This seems plausible because, based on queuing theory, they are jointly determined, and the optimal value of one implies that of the other. Decision for any extra inventory may be made, however, based on marginal return of such extra inventory. In deriving this model, it was assumed the cost paid is once off, and not time dependent, for every throughput that is lost. This model is an extension of that derived in the previous section (5.1), and which was compared to that developed by Radovilsky (1998).

This second section, therefore, presents a more generalised model. The model in section 5.1 is a particular case of this extended model where it is implicitly taken that the unit shortage cost is zero.

5.2.2. LITERATURE REVIEW

Much work has been done on buffering in a manufacturing flow process. The majority appears to have focussed on integrated (automated and semi-automated) systems. This makes the focus of most such articles to be the solution to the design problem of the space to be allowed in-between processing centres in such an integrated environment which needs to be determined before construction, which is different from the problem of the management of the actual production process flow.

Some of the early contributions to this area include the paper by Hunt (1956), which was an analysis of a system where service is to be done in stages. This work was different from phase type process earlier done by Jackson (1954) in that simultaneity and blocking are allowed in the processes. Poisson input and exponential service time was assumed and the model is basically Markovian. Others include machine reliability approach by Enginarlar et al (2002) and Bartini et al (2009), and Production system with three unbalanced stations by Powell (1994) amongst others.

Something common to almost all these papers is that all the machines in the production network were being buffered. The approach, therefore, seems rather different from that being advocated by the Theory of Constraint ($\boldsymbol{TOC}$), where buffers are included only in strategic locations and not ahead of all machines/processing centres as in almost all the cases reported earlier. $\boldsymbol{TOC}$ advocates the presence of spare capacities in many areas of the production system but disapproves of holding inventories except where necessary.

Also, most of the works done seem to be buffering for the failure of feeder machines upstream to the critical station. Buffering for the purpose of the statistical fluctuations in the input and processing times seems not to be the main concern. Only Radovilsky (1998) appears to be quite applicable to buffering for the flow of the process, and it explicitly includes unit profit and unit holding cost in the model.

In summary, a review of literatures on the determination of an appropriate buffer size to place ahead of the critical resource in a production environment utilising the Theory of Constraint has been done by in a previous paper in section 5.1. The summary of the contributions of several authors like Faria et al (2006), Han and Ye (2008), Li and Tu (1998), Powell and Pyke (1996) and Radovilsky (1998) were discussed amongst others.

The effect of utilisation on the Work-in-Process ($\boldsymbol{WIP}$) inventory and its implication on the system cost appears not yet fully researched. Most authors that have written on buffering the relevant stations of the theory of constraints appear to have assumed that all the demands from the market should be met. But in order to meet these demands sometimes, the utilisation of the resources may need to be quite high. This has been discussed in section 5.1 and illustrated with figure 5.1.1.

Radovilsky (1998) has shown how the buffer size to support the bottle neck ($\boldsymbol{BN}$) station could be estimated using the capacitated queue M/M/1/K approach, where he found the derivative ofthe profit function relative to the queue capacity, K, and derived the optimal queue size.

While this is a good attempt, it has two key drawbacks. Firstly, it is difficult to extend this model to a case where other types of inputs (e.g. deteriorating inputs or balking inputs) could be considered. Secondly, it is difficult to include the range $0 < \rho < 1$ in the analysis. This has also been discussed in section 5.1, where it was shown that a solution to both drawbacks could be to optimise the flow rather than the buffer. The optimal buffer size can then be obtained from the steady state size of the queue once the maximum allowable shortage is specified (this can be a policy matter). Controlling the production system should also become easier since the feed rate controls the whole production line rather than just one machine. This makes the management of the system easier. This, actually, is in full sync with the philosophy of the $\boldsymbol{TOC}$, where the focus should be on the flow rather than the capacity of the system, and hence the Drum-Buffer-Rope approach.

The previous works did not consider the possibility of paying some cost for every throughput lost. A more realistic model will seem to be the one that accounts for the possibility of paying for every lost throughput. The least that could be paid is the opportunity cost of the revenue that should have been earned. In addition to this, there could be other penalties imposed on the company by its clients, especially in a case where it has one or more major client(s) that are responsible for the purchase of the bulk of its output. This scenario is not farfetched today where supply chain management ($\boldsymbol{SCM}$) is rife and many major global companies are implementing lean techniques and having their inputs delivered Just-in-Time and probably Just-in-Sequence.

The need to account for this cost of failure to deliver output as needed necessitates this extension. The shortage cost here is, however, assumed to be a fixed cost paid per unit product of output not supplied to the customer as and when needed and not increasing with the length of time for which the output was not available.

5.2.3. MODEL PRESENTATION

In this section, the net profit function is defined to include some cost of shortages. The net profit function then becomes

$$NP = TH - OE - SC \qquad 5.2.1$$

$$TH = \mu(1 - P_0)C_{TH} \qquad 5.2.2$$

$$OE = L_s C_{OE} \qquad 5.2.3$$

$$SC = \mu P_0 C_{SH} \qquad 5.2.4$$

where NP is the Net Profit,

TH is the throughput,

OE is the Operating Expense

SC is the Shortage Cost

μ is the rate of service at the station

$P_0$is the probability that waiting 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 on the resource

C_{OE} is the inventory cost per unit (product-time)

C_{SH} is the shortage cost for every unit throughput lost

An implicit assumption in the models in section 5.1 and Radovilsky (1998) is that this cost of shortages is actually zero. This can be seen by looking at equation 1. The new term introduced, $\boldsymbol{SC}$, as seen in equation 4, must be zero if we must have equation 1 appearing in the initial form. For this term to be zero, at least one of μ or P_0 or C_{SH} equals 0. Since it is not reasonable for either μ or P_0 to be zero, else the first term, $\boldsymbol{TH}$, would have also been zero or the resource becomes a bottleneck, so then, only C_{SH} could have been zero.

From the solution to $\boldsymbol{M}/\boldsymbol{M}/\mathbf{1}/\infty$, queue P_0 and L_S are:

$$P_0 = 1 - \rho \qquad 5.2.5$$

$$L_S = \frac{\rho}{1-\rho} \qquad 5.2.6$$

Having done this, the net profit equation becomes

$$NP = \mu\rho C_{TH} - \frac{\rho C_{TH}}{1-\rho} - \mu(1-\rho)C_{SH} \qquad 5.2.7$$

Differentiating equation 7 with respect to ρ and setting the derivative to zero to obtain the optimal ρ gives

$$\rho^* = 1 - \sqrt{\frac{C_{OE}}{\mu(C_{TH}+C_{SH})}} \qquad 5.2.8$$

The optimal buffer size can then be recovered from the optimal steady state queue length, L_S, corresponding to ρ*, and this is obtained by substituting equation 8 into equation 6 to obtain

$$L_S = \sqrt{\frac{\mu(C_{TH}+C_{SH})}{C_{OE}}} - 1 \qquad 5.2.9$$

Putting 5.2.8 and 5.2.9 into 5.2.7 and solving for NP*, the maximum profit function,

$$NP^* = \left(\sqrt{\mu(C_{TH}+C_{SH})} - \sqrt{C_{OE}}\right)^2 - \mu C_{SH} \qquad 5.2.10$$

One can see that this model is similar to the one obtained for the case where shortage cost was not considered in section 1 and reproduced here as equation 5.2.11.

$$NP^* = \left(\sqrt{\mu C_{TH}} - \sqrt{C_{OE}}\right)^2 \qquad 5.2.11$$

5.2.4. DEDUCTIONS FROM THE OPTIMAL NP EQUATION

One could easily see from equation 5.2.10 that if C_{TH} is zero, the solution is the same as that obtained in the previous section. But since the cost of shortages is hardly ever zero, then the model presented in this paper should give a more realistic profit estimate than equation 5.2.11.

The effects of C_{TH}, C_{OE} and μ are easily observed from the optimal ρ, optimal L_S and optimal $\boldsymbol{NP}$ equations. One could see that as C_{TH} increases, the optimal ρ, the optimal L_S as well as the Net Profit increase. One can also notice that as C_{OE} increases, the optimal ρ decreases, the optimal L_S decreases and the expected net profit decreases as well.

The effect of the unit shortage cost is easily seen for both the optimal ρ and optimal L_S. One can see that as C_{SH} increases, both the optimal ρ and the optimal L_S increase. But the effect of an increase in C_{SH} on the optimal $\boldsymbol{NP}$ is not so obvious from equation 5.2.10 since C_{SH} is in the two terms of the NP function, where its increase will tend to have an increasing effect due to the first one and a decreasing effect due to the other.

The effect of the unit shortage cost on the new profit function would be done in the section where numerical analysis is carried out, but it is worth exploring how the new variable affects the overall profit function. The effect of C_{SH} on the optimal profit function could be analytically studied by assuming one function is greater than the other and finding the condition under which that could be true. Intuitively, one can assume that including the shortage cost in the equation should reduce the profit function as shown in equation 5.2.12.

$$\left(\sqrt{\mu(C_{TH} + C_{SH})} - \sqrt{C_{OE}}\right)^2 - \mu C_{SH} < \left(\sqrt{\mu C_{TH}} - \sqrt{C_{OE}}\right)^2 \qquad 5.2.12$$

Solving the inequality and find the condition under which that could be true. This gives

$$\mu C_{SH} C_{OE} > 0 \qquad 5.2.13$$

Since it has been established that both μ and C_{OE} are not zero (actually positive), the condition in inequality 5.2.12 can only be true for C_{SH} greater than zero. The same conclusion could have been easily reached by simply looking at equation 5.2.7 and noting that $C_{SH} < 0$ increases the profit function, $C_{SH} > 0$ decreases the profit function, while $C_{SH} = 0$ makes the profit function to be equal to the model in equation 5.2.11.

This means that the expression in equation 5.2.10 is equal to the expression in equation 11 only when C_{SH} is zero. If C_{SH} is negative, then the expression in equation 5.2.10 is always greater than that equation 5.2.11 and if C_{SH} is positive, the expression in equation 5.2.10 is always less than that in equation 5.2.11. Since having negative C_{SH} is unreasonable, the value of C_{SH} can only range from zero to positive. This means the Net Profit function is of 5.2.10 always less than that in equation 5.2.11 for as long as there is cost of shortages, which makes intuitive sense.

The models derived in equations 5.2.8 and 5.2.9 therefore give guidance for how to select the optimal feed rate to optimise the net profit in a system that has a Capacity Constrained Resource but no Bottleneck when applying the Theory of Constraints in a production system, and/or where buffering is being made for statistical fluctuation in processing time and not for breakdown of the upstream stations to the critical resource.

5.2.5. NUMERICAL ANALYSIS

The effect of the inclusion of shortage cost in the model on the net profit is shown here. The net profit realised with shortage cost included is compared to that the dynamic buffering approach in section 1.

Figure 5.2.1 shows that as the flow rate moves towards the optimal rate, the difference between the model with and that without the shortage cost narrows. This shows that the effect of shortage cost becomes more pronounced as the system operates below the optimal level. But as the utilisation moves towards unity, the effect of shortage cost fizzles away. An explanation for this is that the possibility of shortage becomes almost zero as the queue length increases tremendously. This is because it is almost impossible to have shortages as a result of an idle resource as the probability of being idle goes towards zero. Also, the holding cost term dominates the profit function.

Next, the effect of changes in the various input parameters on the optimal utilisation (intensity), ρ, and the optimal average queue length, L_S, were graphically evaluated. For the purpose of our analysis, starting values were randomly chosen for the input variables. All of them were initialised to 50. With every other variable kept constant, the effect of each of the input variable on the optimal output values were observed by varying only the variable of interest.

Figures 5.2.2 to 5.2.5 show the effects of the changes in the values of the input variables on the optimal value of ρ. From these, optimal ρ increases with every of the input except the holding cost, and this is easily seen from equation 5.2.8. Also, both the shape and the slope of the curves of change in unit profit and change in unit shortage cost are the same. This can also be easily deduced from equation 5.2.8. It can also be seen that the effect of the service rate and the holding cost are more dramatic than those of unit profit and unit shortage costs. As each of the input variable quadruples from 50 to 200, one would notice that the rate of change in value of ρ for both the holding cost and the service rate are double those of unit profit and shortage cost. This is also apparent from equation 5.2.8. The effects of each of the input variables on the optimal average buffer build up is exactly the same as that noticed in ρ, and this is seen from figures 5.2.6 to 5.2.9.

Figures 5.2.10 to 5.3.13 show the effects of changes in the values of the input variables on the optimal net profit. It could be seen that while net profit increases with increasing service rate and unit profit, it decreases with increasing unit holding cost and unit shortage cost.

The net profit functions of the models with and without shortages have been plotted on the same axes. The diagram suggests that if the effect of shortage cost is neglected as done in the previous models, the changes in unit profit appears to have less effect on difference in profit predicted by the model with shortage cost and the one without it. But changes in holding cost appear to have the most dramatic effect. This can be explained by looking at equation 5.2.13.

In figure 5.2.7, the net profit function changes relative to changes in unit shortage cost is seen as a straight line for the model without shortages since C_{SH} has been taken as zero here. But the effect of increasing the holding cost on the net profit appears more drastic than that of the shortage cost.

Following the analyses of the effects of the various input variables on the computed output parameters, the holding cost appears to be the most important variable whose changes should be monitored to make the necessary flow adjustments to keep the system optimal.

5.2.6. CONCLUSION

The model of dynamic buffering of a $\boldsymbol{TOC}$ with shortage cost has been presented. It was assumed that the cost of shortage is a once off unit cost charged per unit product short. The previous model without shortages was shown to be a particular case of this model where the cost of shortages could be taken as zero. This model should be more realistic than a model without shortage cost included.

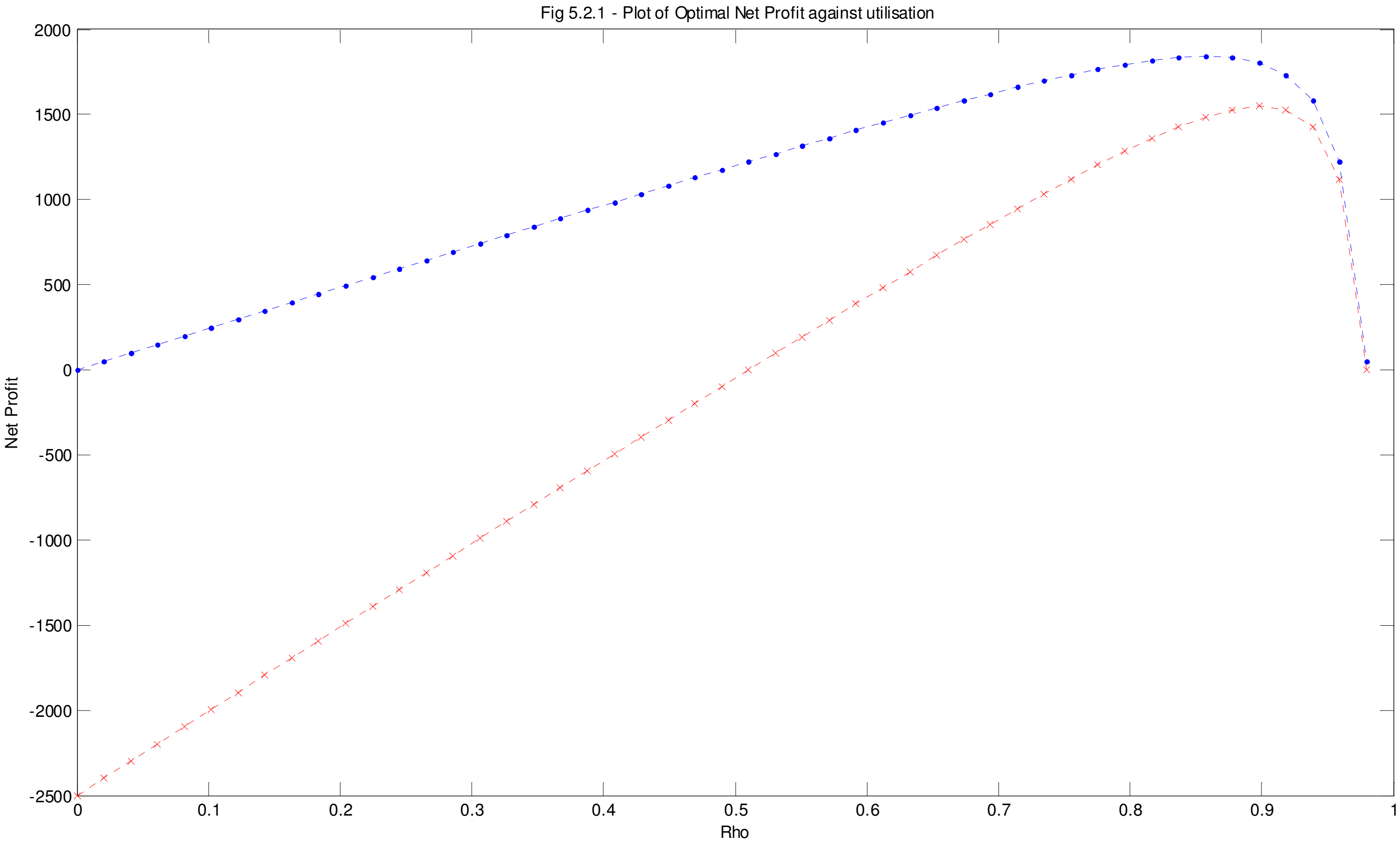


Figure 5.2.1: Changes in profit with rho $(0 \leq \rho \leq 1)$

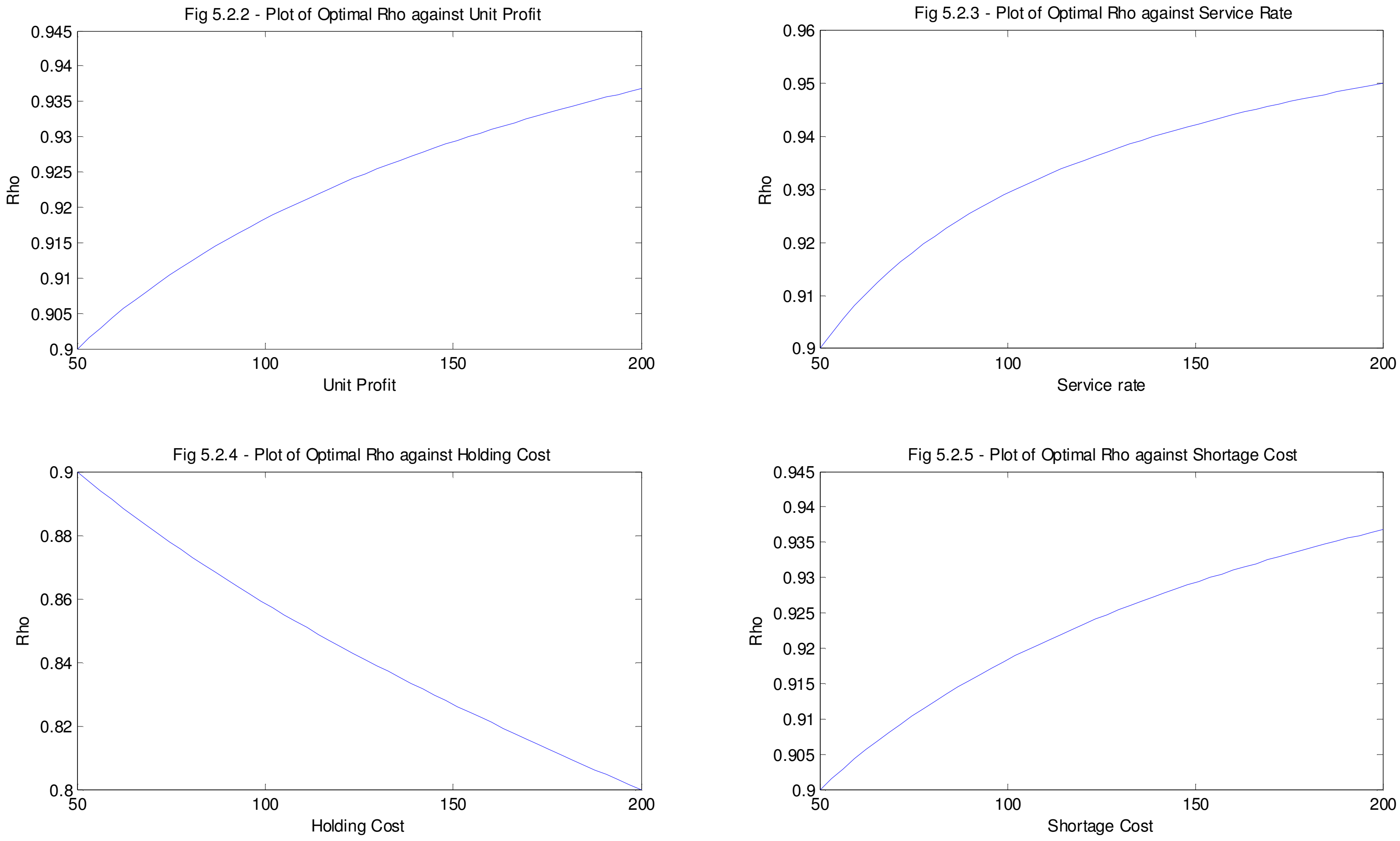


Fig 5.2.2 - Plot of Optimal Rho against Unit Profit

Fig 5.2.3 - Plot of Optimal Rho against Service Rate

Fig 5.2.4 - Plot of Optimal Rho against Holding Cost

Fig 5.2.5 - Plot of Optimal Rho against Shortage Cost

Figure 5.2.2 – 5.2.5: Changes in rho with input parameters

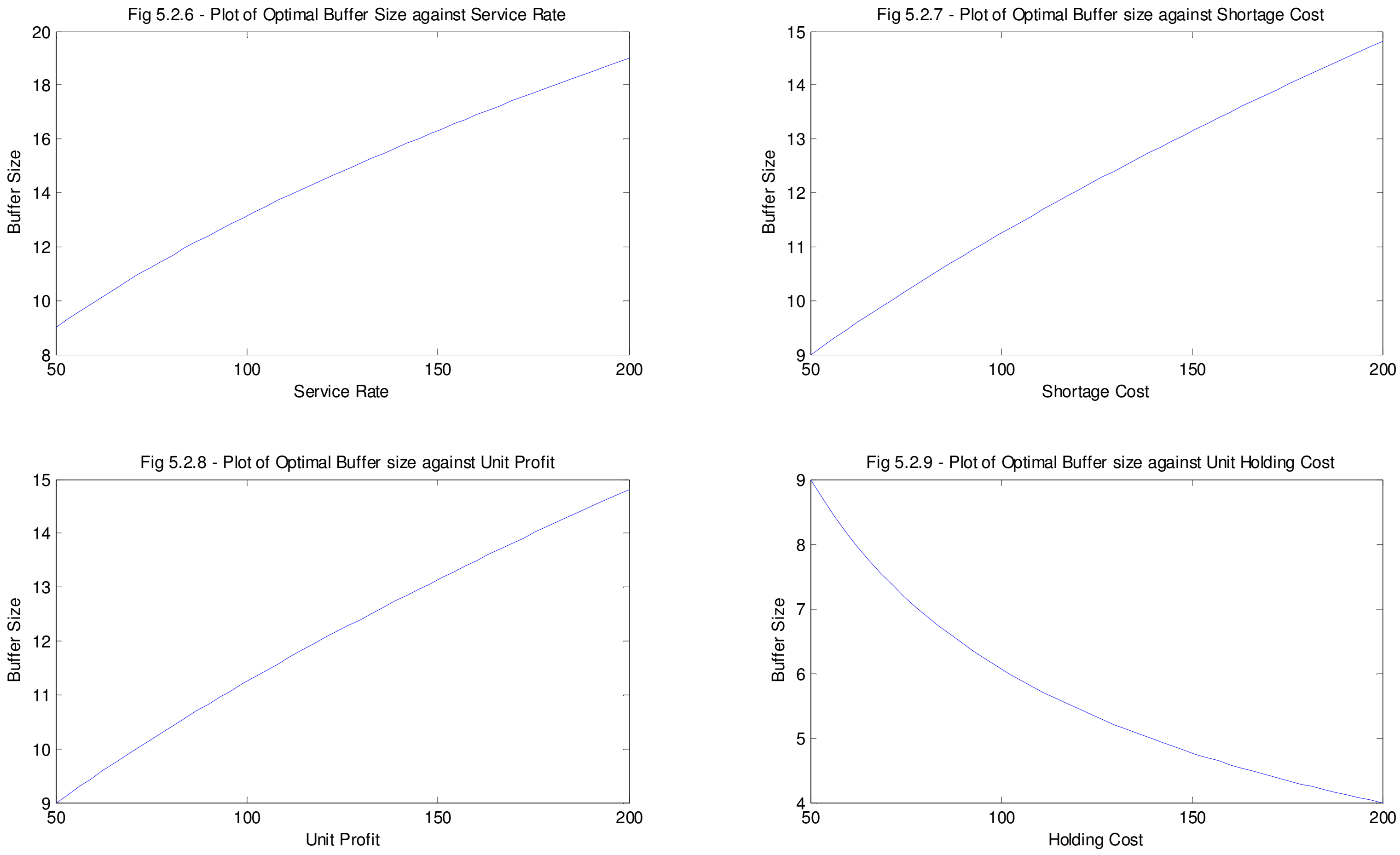


Fig 5.2.6 - Plot of Optimal Buffer Size against Service Rate

Fig 5.2.7 - Plot of Optimal Buffer size against Shortage Cost

Fig 5.2.8 - Plot of Optimal Buffer size against Unit Profit

Fig 5.2.9 - Plot of Optimal Buffer size against Unit Holding Cost

Figure 5.2.6 – 5.2.9: Changes in buffer size with input parameters

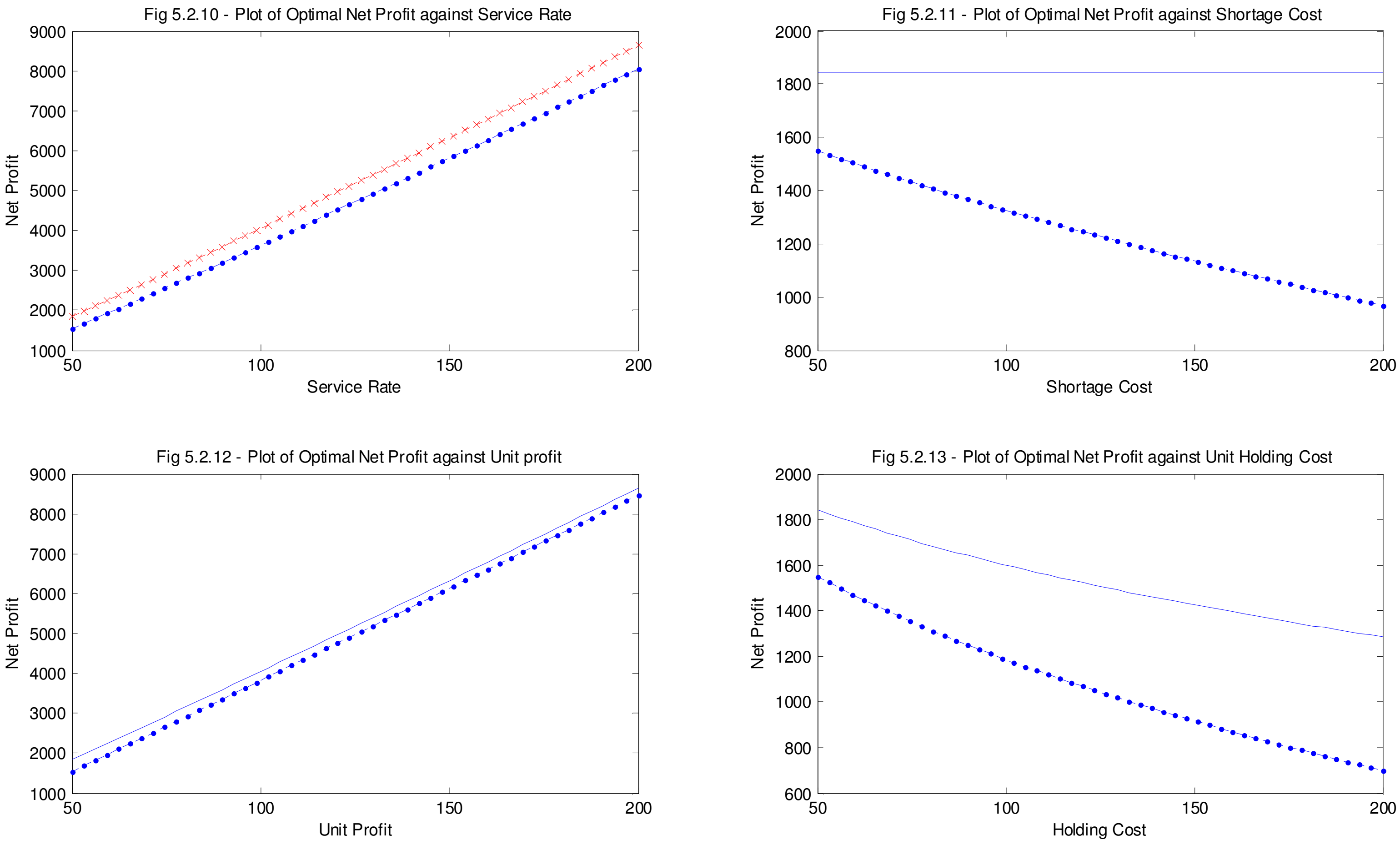


Fig 5.2.10 - Plot of Optimal Net Profit against Service Rate

Fig 5.2.11 - Plot of Optimal Net Profit against Shortage Cost

Fig 5.2.12 - Plot of Optimal Net Profit against Unit profit

Fig 5.2.13 - Plot of Optimal Net Profit against Unit Holding Cost

Figure 5.2.10 – 5.2.13: Changes in buffer size with input parameters

6.

CHAPTER 6

CONCLUSION

6.1. CONCLUDING OVERVIEW

Two common threads can be found in the compendium of works presented in this document. The first is that queuing principles with stochastic parameters have been used to analyse or applied to the various types of systems considered. The second is that the performance of the inventory management system has been studied directly or indirectly throughout. The focus and applications and/or contributions of each chapter can be summarised as follows. The work in the first three chapters have made particular use of the Markov Arrival Process $(\boldsymbol{MAP})$ that makes it possible to expand the basic Poisson input stream to various practical environments that have more complex input systems, but that could still take advantage of the memorylessness properties of the attendant exponential distribution to simplify the calculations.

6.2. SOME POSSIBLE APPLICATIONS OF DERIVED MODELS

Chapters 2 and 3 contain the analyses of systems where products are not delivered immediately in response to demands, but where some services are further done on the items to be delivered before actual delivery. Exponential distributions were assumed for the lead time between the order placement and actual delivery. These types of systems are currently pervasive. A common knowledge today is the need to decide if the production system is to be managed as a make-to-stock, make-to-order, or assemble-to-order (or even engineer-to-order) system. This decision is usually dependent on the level of trade off desirable between long supply lead time and explosive inventory level.

While making to stock generally guarantees high responsiveness, it usually implies carrying a large volume of inventory. On the other hand, making to order reduces the inventory level drastically but leads to high response (lead) time. A recent best practice is that of delayed differentiation of products, which is some form of assembling to order. This type of environment usually leads to some final services being done on the inventory stock before being delivered. This implies that inventory is depleted at the rate of the services performed on the stock rather than directly on the demand for such products. Such systems seek to find some form of compromise between managing explosive inventory levels and having a long supply lead time.

With the general shift in the production environment towards lean manufacturing and -assembling-to-order, models developed for such systems (as in this work) would start having more applications, as compared to the traditional queuing systems that implicitly assumes that items are produced to stock and orders are immediately fulfilled from stocks. Herein lays the importance of the first two models presented in this work. The distributions and steady state parameters of some such systems have been studied in chapters 2 and 3. These steady state parameter estimates could be used in further applied probability contexts in many systems. This will be further discussed briefly in section 3 of this chapter.

Chapter 4 is a contribution to the field of Joint Replenishment Planning $(\boldsymbol{JRP})$. Such systems are more practical in many real life instances than the typical assumptions around which some $\boldsymbol{ERP}$ systems are built. There are usually advantages in seeking how two or more products could be ordered together (usually from the same source) or produced together on the same machine. This may lead to savings in order (or set up) cost and thus overall reduction in the total production cost. Chapter 4 furthers the work done in this area.

While chapters 2, 3 and 4 are focused on the derivation of system parameters using queuing principles, chapter 5 is an application of the parameters derived in an $\boldsymbol{M/M/1}$ environment in the management of flows in a production system utilising the theory of constraint. The first part shows that determining the optimal buffer size indirectly by first determining the optimal flow rate, leads to further simplification of the application of optimisation techniques, and probably a more optimal profit function as compared to the previously documented approach of optimising the profit function directly with respect to the buffer size. This approach has been referred to in this book as dynamic buffering.

A more interesting observation made from this indirect approach is that it makes it easier to notice if it is actually necessary to seek to meet all customer demands in the first instance. It then makes it possible to obtain the optimal buffer size for more general systems other than the $\boldsymbol{M/M/1}$ because such can also be indirectly retrieved since the flow intensity is a more pervasive parameter in all queuing models, while models explicitly containing a buffer size parameter are limited. This makes it possible to generalise the model to other types of systems. This was illustrated with a simple modification of the $\boldsymbol{M/M/1}$ model initially presented to a case where there is shortage cost included.

6.3. POSSIBLE AREAS FOR FUTURE RESEARCH

The field of queuing theory is very popular and has enjoyed (and still enjoys) tremendous research focus, partially because if the ubiquity of queues, and therefore, the applicability of its theory. But it is possible to extend its applicability in many other ways, for instance, with the $\boldsymbol{MAP}$ input stream replacing the traditional Poisson input flow, and the $\boldsymbol{PH}$ service time model extending the traditional exponential model, as is currently being done by many authors, and in this work as well. The stochastic $\boldsymbol{JRP}$ system that has an $\boldsymbol{MAP}$ input like in chapters 2 and 3 are possible areas for further research. Models with input recovery system are another area that seems, for instance, yet to be explored. Such models would have another input stream recovered from the imperfections in removal of deteriorated items from wholesome stocks. This has generally not been considered in any work hitherto.

Also, the application of the steady state distributions and parameter estimates of the first three models considered in work, like many other such results by diverse authors, are fertile areas for improvement of the relevant areas in many production management philosophies. For instance, the application of some Phase distribution models like the Erlang, Hyper-exponential and Hyper-Erlang seems like possible candidates for resolving the issues of determining the transfer batch sizes in the Theory of Constraints environment. No application of stochastic processes appears to have been made in these areas. Others include management of system nervousness due to non-deterministic demand and lead times in the $\boldsymbol{MRP}$.

Steady state queue solutions, including those developed here, appear to have possible applications in such systems. While it is pertinent to state ahead that many such models may not have closed form solutions due to the nature of the solutions derived for the parameter estimates from many complex systems, it is anticipated that numerical iterative solutions would be useful tools in solving such problems. Such problems are being considered as part of the possible areas to explore by this author going forward from here.