Analysis of Capacity Constrained Resource with Parallel Processes

Lan Wang*
California State University - East Bay, Hayward, USA

Zinovy Radovilsky
California State University - Eat Bay, Hayward, USA

This paper analyses and provides insights into the time buffer size of a capacity constrained resource (CCR) with multiple parallel processes in the context of the Theory of Constraints (TOC). The appropriate capacity size would protect the CCR from becoming idle with a certain probability, denoted as the accepted idle rate. We formulate a constrained optimization model, where the key decision variable is the capacity size, and the objective is to maximize the net profit. For the in-depth understanding of the model's solutions, we conduct a sensitivity analysis of the optimal capacity size and net profit based on variations of the model parameters, i.e., accepted idle rate, arrival rate, processing rate, and cost characteristics. We prove the existence of important thresholds for these model parameters, and we explore the behavior of the optimal solutions depending on the values of these parameters.

* Corresponding Author. E-mail address: lan.wang@csueastbay.edu.

I. INTRODUCTION

The Theory of Constraints (TOC) has been widely recognized as a well-established theory and application of modern business management (Naor, Bernardes and Coman 2013, Spector 2011, Gupta and Boyd 2008, Inman, Sale and Green Jr. 2009, Mabin and Balderstone 2003). The fundamental work of Eliyahu Goldratt (Goldratt and Cox 1992, Goldratt 1990) paved the way for the development of the TOC framework. In the core of his theory is the important recognition of a system constraint that limits the system's ability to achieve its goal (Goldratt and Cox 1992). This constraint is typically described as a physical constraint, such as a machine or space with limited capacity, or a raw material. The constraint can be also identified as a company's policy or behaviour constraint. In this paper, we focus exclusively on a physical constraint, i.e., the capacity constrained resource (CCR). According to TOC, the system results including product throughput, operating expenses, and overall system profitability, are fully associated with a proper management and scheduling of CCRs (Goldratt 2010, Inman, Sale and Green Jr. 2009, Corbet and Csillag 2001).

The drum-buffer-rope (DBR) method is a cornerstone of the TOC constraints' management system; it is widely accepted by the research and business communities as a principal way to do effective production scheduling (Tseng and Wu 2006, Ye and Han 2008, Tukel, Rom and Eksioglu 2006). The DBR's centrepiece is a time buffer that establishes a protection of the CCR, and thus does not allow the entire production system to slow down or reduce its productivity despite the fluctuations of the resource capacity or production demand. The subject of managing the time buffer in front of a single-process CCR is discussed in various literature sources (Gupta and Boyd 2008, Hadas, Cyplik and Fertsch 2009, Hwang, Huang and Li 2011, Lee, et al. 2009, Radovilsky 1998).

The issue of CCR and its time buffer is relevant to more than just one particular resource or process. In many cases and for various processes there may be a capacity constrained resource with several processes (activities) working simultaneously on the same product, which may be defined as a CCR with parallel processes. Plentiful examples of parallel processes can be seen in a variety of business scenarios, e.g., check-out lines in a supermarket, parallel security check lines in an airport, parallel conveyors of goods production in a manufacturing company, etc. However, despite the present extensive research on DBR and its time buffers, optimization models and tools for identifying time buffers for a system with parallel processes are still not well established and utilized.

This paper expands the existing research of the DBR's time buffer by focusing on its optimization in CCRs with parallel processes. The paper structure incorporates the following sections. In the next Section 2 we analyse numerous literature sources on buffer size and its calculations, with the emphasis on the current research in buffer size optimization for parallel processes. In Section 3, we present our optimization model for a capacity (buffer) size in front of the CCR with parallel processes. An extensive numeric sensitivity analysis of the optimal solutions to the proposed model is given in Section 4, with conclusions presented in Section 5 of the paper.

II. LITERATURE REVIEW

A variety of methods and approaches have been created to address the DBR's time buffer size. They range from introducing relatively simple formulas for identifying time buffers to presenting sophisticated quantitative analytical tools for time buffer analysis. A group of researches (Hadas, Cyplik and Fertsch 2009) develop a set of empirically-driven formulae that define the time buffer as a ratio of a real lead time at a CCR and the production capacity excess of non-critical resources. Newbold (1998) designed a method of identifying a buffer size for a project CCR, which is based on each task's duration uncertainty. The author recommends that the buffer size should be empirically defined as a double value of the standard deviation for to the entire set of project tasks.

The more sophisticated research methods and tools are being used to optimize the time buffer size. Ye and Han (2008) describe an analytical approach based on reliability analysis that determines the size of a constraint buffer and assembly buffer in a DBR-controlled production system. They analyse the situation with one regular or assembly CCR in an in-line production process, but stop short of considering a production environment with a CCR for many parallel processes. Long and Ohsato (2008) analyse the DBR application for a project with possible activities' uncertainty and delays. Using the fuzzy numbers approach, they determine the buffer size as a square root of the sum of the squares of safety times estimated by fuzzy numbers. However, this fuzzy logic analysis does not realistically provide an optimized approach to identify the buffer size. Similarly to that, Bie, Cui and Zhang (2012) analyse the effects of the dependence between activities on project duration performances, and introduce a method for determining buffer sizes with the dependence assumption between activities.

Another approach for calculating the optimal time buffer size is based on formulating the problem as a queuing model. General ideas of identifying an optimal buffer size using the queuing theory are presented in the paper by Ghosh and Weerasinghe (2008). The proposed optimization algorithm is based on minimizing the queuing network costs only; it does not incorporate the important in TOC trade-off between the throughput revenue and operating expenses associated with the buffer size (Goldratt 2010). Using the latter approach, Radovilsky (1998) identify the optimal size of the time buffer by formulating the problem in terms of a single-server finite queue. The author defined the optimal number of units waiting in front of the CCR that would maximize the CCR's operational profits while protecting the constrained resource from becoming idle. However, the introduced optimization of buffer size (Radovilsky 1998) was based on a CCR with a single process. It did not take into consideration CCRs with multiple parallel processes.

A review of the existing literature reveals a limited number of papers that discuss identification of the time buffer with multiple processes. Shaaban and McNamara (2009) develop a simulation algorithm to analyse the operating performance of parallel production lines that contain unbalanced buffer storage sizes. The metrics that was used to analyse the simulated line behaviour included idle time and average buffer level output, for which a variety of statistical tools were employed. Louw and Page (2004) develop the calculation of a time buffer which is based on a queuing model with multiple servers utilizing non-Poisson arrival processes and non-exponential service time distributions (GI/G/m). Sirikrai and Yenradee (2006) propose a new scheduling method, the modified DBR, which applies a backward finite capacity scheduling technique, including machine loadings and detail scheduling, instead of the rope mechanism in DBR.

Overall, in the described research for the buffer size calculations in a multi-process environment, the time buffer was not introduced as a part of the DBR-based system analysis. Moreover, these papers do not consider the TOC trade-off between the throughput (revenue) and operating expenses as a way to identify the optimal buffer size. Radovilsky and Frankel (2013) attempted to address those points. Their model is formulated in terms of a finite multi-server queue, in which a CCR is defined as a part of the process with multiple parallel channels. The optimal buffer size for this multi-channel CCR is identified by maximizing the profit received as a trade-off between the throughput and operating expenses, the elements of financial measurements in TOC (Goldratt 2010). In this paper, the specified research (Radovilsky and Frankel 2013) is used as a basis for developing a new modelling approach in Section 3. We also provide in section 4 an extensive numeric analysis of the model's optimal results and their sensitivity to variations of the model's parameters.

III. MODEL DESCRIPTION AND ANALYSIS

In this section, we consider a production or service CCR with a multi-server parallel processing with s servers ($s > 1$), for which we need to identify an optimal time buffer size. This system will be described as an *M/M/s/K* queue (Gross et al. 2008). The assumptions of the Poisson arrival and exponential service time in this queue are frequently employed in literature sources for analysing production systems with finite multi-server queues (see, for example, Feng, Zheng and Li 2012, R. Inman 1999, Askin and Standridge 1993, Simon and Hopp 1991).The "arrival of a customer" simply means that a new incoming product unit is added to the time buffer, with the average of λ units per unit time. The service of those units is then done as an operation of utilizing products from the time buffer to the parallel processing with s servers, and the average service (processing) rate of μ units per server.

Since CCR is a bottleneck, we assume

that the arrival rate (λ) is greater than or equal to the total processing rate in this parallel system ($s\mu$), or the utilization factor is at least 1 ($\rho \geq 1$). If λ is less than $s\mu$ ($\rho < 1$), the capacity of the resource is greater than the incoming demand, in which case the resource is not a CCR (bottleneck) resource. Any idleness in the multi-server CCR would have an adverse effect on the entire process efficiency. To prevent such situations, the CCR is protected by a time buffer. We denote the *capacity size* of this CCR as K, which combines the size of the waiting line in front of the CCR, i.e., the time buffer, and the number of servers, s.

The model notations applied in this paper are summarized in Table 1.

TABLE 1. NOTATIONS.

Notations	*Explanations*
α	Accepted idle rate of the CCR
λ	Arrival rate
μ	Processing (service) rate at each server
γ	Occupancy rate per server, λ/μ
ρ	Utilization factor, $\frac{\lambda}{s\mu}$
p_o	Probability that the entire CCR system is idle
p_k	Probability that *k* items are in the system (including the servers)
L_q	Average size of the queue
α	Accepted idle rate
$\widehat{\alpha}$	Threshold of the accepted idle rate
C_{TH}	Throughput per unit of sale
C_{OE}	Carrying cost per unit
TUC	Throughput per unit of cost, $\frac{C_{TH}}{C_{OE}}$
K	Capacity size, combines the CCR's time buffer (waiting line) size and units processed in servers (*s*)
K^*	Optimal capacity size
s	Number of servers
NP	Net profit
NP*	Maximum (optimal) net profit
TH	Throughput
OE	Operating expenses

In an *M/M/s/K* system (Gross et al. 2008), the probability p_0 that the servers will be idle is:

$$p_0 = \left(\sum_{n=0}^{s} \frac{\gamma^n}{n!} + \frac{\gamma^s}{s!} \cdot \frac{1 - \rho^{K-s+1}}{1 - \rho}\right)^{-1} \quad (> 1) \qquad (1)$$

According to (1), this probability is a function of the number of units in the system K and number of servers (s). To protect the CCR from being idle, we need p_0 not to exceed a pre-defined "accepted idle rate" α that, evidently, needs to be very small:

$$p_0(K) \leq \alpha\,. \qquad (2)$$

Solving inequality (2) for *K*, we can obtain the following expression:

$$K \geq s - 1 + \frac{ln\left[\frac{(\rho-1)(1-\alpha D)+\alpha B}{\alpha B}\right]}{l\,n\,\rho} \quad (\rho \neq 1), \qquad (3)$$

where

$$D = \sum_{n=0}^{s-1} \frac{\gamma^n}{n!}$$

$$B = \frac{\gamma^s}{s!}\,.$$

The imposed accepted idle rate actually serves as a constraint of the capacity size *K*. Using that, we formulate the following model for identifying the optimal capacity size:

$$\max_{K} \quad NP = s\mu(1-p_0)C_{TH} - [L_q + \gamma(1-p_k)]C_{OE} \qquad (4)$$

subject to

$$K \geq \max\{s+1, \bar{K}\}, \qquad (5)$$

where

$$\bar{K} = s - 1 + \frac{\ln\left[\frac{(\rho-1)(1-\alpha D)+\alpha B}{\alpha B}\right]}{\ln\rho} \qquad (6)$$

$$p_0 = \left[\sum_{n=0}^{s} \frac{\gamma^n}{n!} + \frac{\gamma^s}{s!} \cdot \frac{1-\rho^{K-s+1}}{1-\rho}\right]^{-1} \qquad (7)$$

$$L_q = \frac{p_0 \gamma^s \rho}{s!\,(1-\rho)^2}[1 - \rho^{K-s+1} - (1-\rho)(K-s+1)\rho^{K-s}] \qquad (8)$$

$$p_k = \frac{\gamma^K}{s!\,s^{K-s}} \cdot p_0. \qquad (9)$$

We now explain the proposed optimization model. The model's objective in (4) is based on the maximization of the net profit associated with the CCR system with parallel processes. According to TOC (E. M. Goldratt 2010), a net profit (*NP*) can be identified as a difference between the throughput and operating expenses of the system's CCR. Throughput (*TH*) is defined as a difference between the overall return generated through sales and material costs of goods sold, i.e.:

$$TH = s\mu(1-p_0)C_{TH}\,, \qquad (10)$$

where C_{TH} is the throughput per unit of sale.

Operating expenses (*OE*) represent the cost of carrying a unit of stock associated with the time buffer. They are calculated as the average number of units in the system $[L_q + \gamma\,(1-p_k)]$ times the per-unit carrying cost (C_{OE}), i.e.:

$$OE = [L_q + \gamma(1-p_k)]C_{OE}\,. \qquad (11)$$

The difference between *TH* and *OE* defines the objective function *NP* in (4). Inequality (5) describes a constraint for the lower bound of the capacity size. It cannot be smaller than (*s+1)*, the number of servers plus one unit in front of the CCR. In addition, the capacity size is also constrained by the accepted idle rate: $K \geq \bar{K}$, where $\bar{K}$ in (6) is derived from inequality (3). The rest of the parameters in (4) — p_0, L_q, and p_k — are defined by the respective (7)-(9) of an *M/M/s/K* queue (see, for example, Gross et al. 2008).

To be able to identify the optimal capacity size K^*, we need to make sure that an optimal solution for this model does exist. The latter can be proven by demonstrating that the function $NP(K)$ from (4), where K is not constrained, has a unique optimal solution of Capacity Size K. See proof in Appendix A.

Given (5) and (6), the optimal capacity size is affected by the pre-defined value of accepted idle rate (α). The smaller the α value, the lower the probability $p_0(K)$ of the servers to be idle, which implies the larger capacity size. This leads to a larger throughput, but, at the same time, increases the operating cost of carrying inventory stock in front of the multi-server CCR due to the larger buffer size. The following statement establishes an estimate of how the value of α affects the optimal capacity size and associated net profit:

There exists a threshold of α , $\hat{\alpha} = \frac{\rho-1}{(\rho^2-1)\,B+(\rho-1)\,D}$, such that

- When $\alpha < \hat{\alpha}$, the optimal capacity size decreases as α goes up, and the optimal net profit grows as α increases;
- When $\alpha \geq \hat{\alpha}$, the optimal capacity size and net profit are not affected by α.

The statements' prove is presented in Appendix B. It provides a very important meaning in terms of establishing an analytical lower bound for the α value, above which the optimal capacity size and net profit become insensitive to variations of α. This will be discussed in more details in the next section.

Despite our ability to develop the optimal capacity size model, prove the existence of its optimal solution, and establish the α value threshold, an analytical solution for the optimal capacity size K is hard to derive from the model in (4)-(5). Therefore, we developed and presented in Section 4 an in-depth numeric analysis of the optimal capacity size and associated net profit, and their sensitivity to variations of the model parameters.

IV. NUMERIC ANALYSIS AND SENSITIVITY OF OPTIMAL CAPACITY SIZE AND NET PROFIT

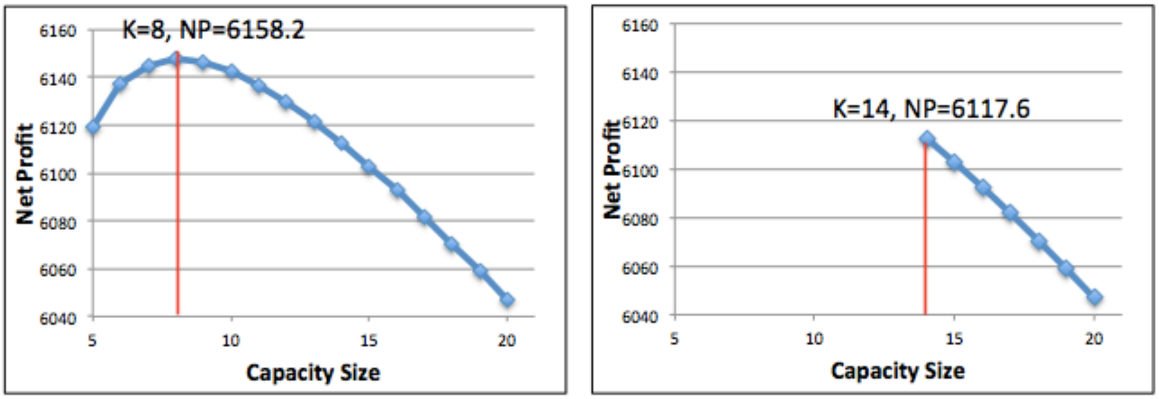
In this section, we first describe the algorithm used to identify numeric values of the optimal capacity size and associated net profit. Based on this algorithm, we calculate the optimal values and then numerically analyse them by varying the model parameters, i.e. accepted idle rate (α), arrival rate (λ), processing rate (μ), and throughput per unit cost (C_{TH}/C_{OE}). The optimal capacity size model is formulated in Matlab (version R2013a). We assign values for the parameters and exhaustively search for the optimal capacity size K^*, which generates the optimal (highest) net profit NP^*.

4.1. Impact of Accepted Idle Rate

The parameter α represents the accepted probability rate that the CCR with parallel processes may be empty. As described in the previous section, α would affect K and then the lower bound of K (LB_K) , $LB_K = max\,\{s+1, \bar{K}\}$ (see inequalities (3) and (5)).

To show changes of the optimal net profit (NP^*) with respect to the optimal capacity size (K^*), the values of other parameters have been fixed at: $\lambda = 250, \mu = 60, C_{TH} = 26.4, C_{OE} = 2\ 2$, and $s = 4$. We choose two values of α: 10%, a relatively large chance of a CCR being empty, and 0.5%, a relatively small one. The changes of NP^* with respect to K^* are shown in Fig. 1. At a larger α value, a lower K^* is observed ($K^* = 8$ when $\alpha = 10\%$). However, when $\alpha = 0.5\%$, the lower acceptance rate of p_0 results in a higher bound of $K^* = 14$. This prevents NP to reach the highest value since K needs to be at least equal or larger than 14.

Fig. 2 shows how both NP^* and K^* change with α . Their graphs clearly demonstrate that after a certain value of α , which is a threshold value, the optimal $NP^*$and K^* results are not affected by α. This provides an additional confirmation of the threshold for α that is defined and proven in Section 3.



High level of α $(\alpha = 10\%)$

Low level of α $(\alpha = 0.5\%)$

FIGURE 1. IMPACT OF CAPACITY SIZE ON NET PROFIT.

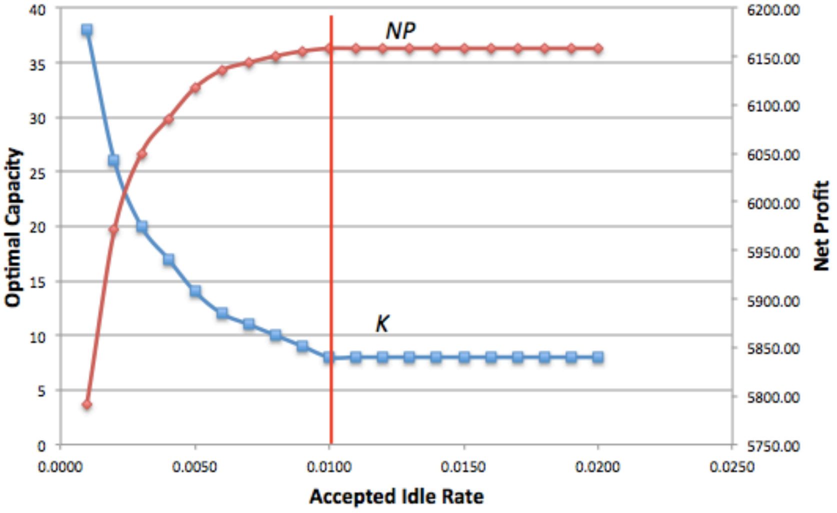


FIGURE 2. IMPACT OF ACCEPTED IDLE RATE ON OPTIMAL CAPACITY AND NET PROFIT.

The value of α below the threshold does affect the optimal values of capacity size and net profit (see Fig. 2). The larger the value of α (from 0 to threshold), the lower will be the K^* value, and the higher the NP^* value.

Overall, based on our sensitivity analysis and described threshold for α, we can derive several important observations:

- When the accepted idle rate (α) is set smaller to reduce a chance that the CCR becomes idle, the optimal capacity size needs to be set larger to hold more inventory in front of the CCR. However, this will lead to a lower optimal net profit as more is spent on the inventory carrying cost.
- The existence of the threshold for α means that, if α is larger than the threshold, its variations won't affect the optimal capacity size and net profit, with the latter reaching its highest value. Therefore, from the practical viewpoint of managing a CCR with parallel processes, it is always better to apply a greater than the threshold value of the accepted idle rate.

In the following sections we consider the case when α is large enough to exceed the respective threshold, and thus does not affect the sensitivity analysis of the optimal capacity size and associated net profit based on other model parameters.

4.2. Impact of Arrival Rate

In this section, we analyse the impact of changing arrival rate on the optimal capacity size and net profit. The following parameters are kept fixed: $\mu = 60, \alpha = 10\%, C_{TH} = 26.4,$ and $C_{OE} = 22$. The value of λ varies from 150 to 450 with an increment of 25. Then, we derive the optimal size of *K* and *NP* for each *s* (number of servers) varying from 3 to 5. Notice that λ also serves as the upper bound of *s*, since $\rho = \frac{\lambda}{\mu\, s} > 1$. Therefore at different values of λ, the range of feasible *s* may vary. For example, when $\lambda = 450$, the upper bound of s is 7. The variations of optimal capacity size depending on λ and for various s are presented in Fig. 3.

As can be observed from Fig. 3, if λ increases, the optimal capacity declines first (in most cases), and then stays practically unchanged. Take s=3 in Fig. 3 for example, the capacity firstly decreases with λ; once λ exceeds 350, the optimal capacity becomes constant at $s + 1 = 4$. It is not suppressing that that when λ is smaller, the operation manager tends to keep a larger buffer size (hence larger capacity) to prevent the system from being idle. Notice, the buffer size is not the actual queue in front. However, when λ becomes larger than a threshold, the systems are completely busy then there is no incentive to keep a larger buffer size as the operating expense is associated with the capacity K.

Overall, based on the graphs in Fig. 3, we can conclude that there exists a threshold of λ, such that:

- When λ is smaller then this threshold, the optimal capacity decreases with λ;
- When λ is greater or equal to the threshold, the optimal capacity is always $K^* = s + 1$.

Then we analyse NP^* variations for different values of λ (see Fig. 4). It is easy to see that as λ increases, given the same number of servers, the net profit would change insignificantly. For example, for s=4 and $\lambda = 250$, the optimal $NP^* = \$6158.24$; for *s*=4 and $\lambda = 450$, the optimal $NP^* = \$6230.70$, which shows that a major increase in λ leads to a very small change (increase in our case) in optimal *NP*.

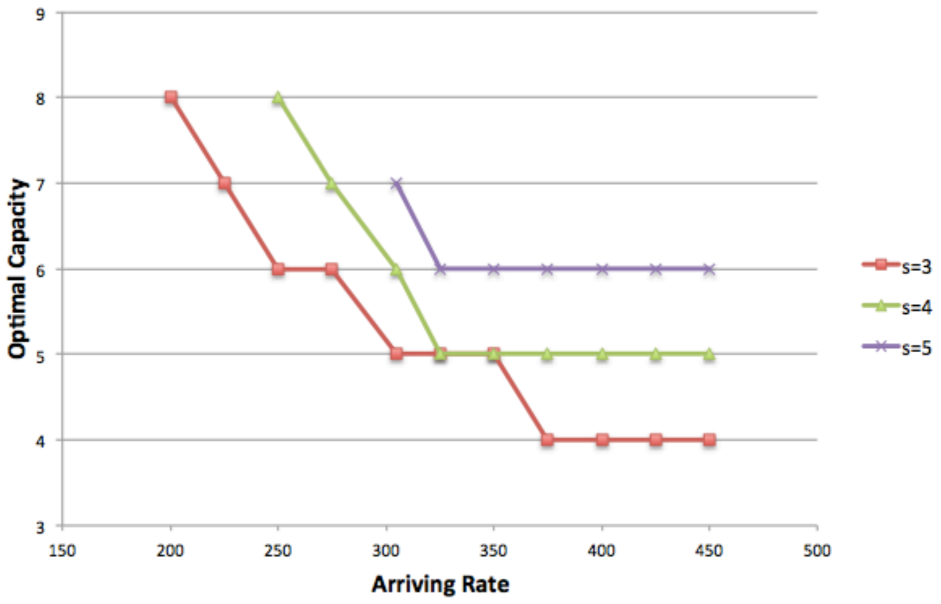


FIGURE 3. IMPACT OF ARRIVAL RATE ON OPTIMAL CAPACITY.

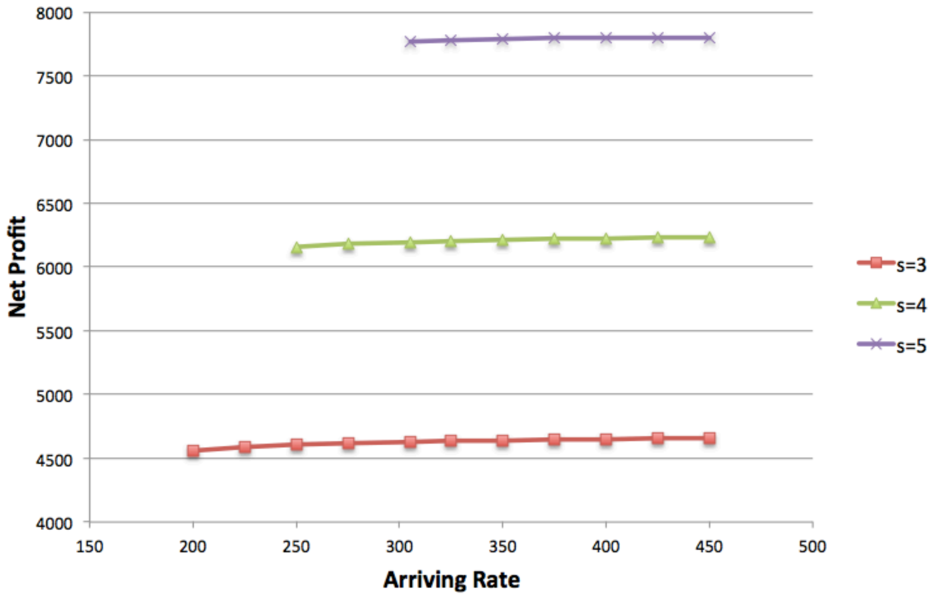


FIGURE 4. IMPACT OF ARRIVAL RATE ON OPTIMAL NET PROFIT.

The sensitivity analysis of the optimal solutions to variations of arrival rate can yield several important observations. When λ exceeds a respective threshold, it is the number of servers and associated processing rate that affect the optimal net profit. In this case, the optimal buffer size is always 1 as K needs to be at least $(s+1)$. When λ is relatively small and below the respective threshold, a larger buffer size should be maintained to guarantee that the CCR servers are fully utilized, and thus the net profit could be maximized. It is also interesting to see from Fig. 4 that the net profit is much more affected by the number of servers than it depends on the arrival rate. A major increase in λ leads to a very small change (increase in our case) in optimal *NP*. There are two reasons behind: first, in this CCR system, we assume that arrival rate (λ) is greater than the processing capacity ($s * \mu$) otherwise it is not a bottleneck. Second, the idle rate of the CCR system is controlled to be quite small. And it is cost-efficient to adjust the capacity (or buffer size) to make sure all the servers are fully utilized. Hence the impact of arrival rate on Net Profit is relative minor.

4.3. Impact of Processing Rate

In this section, we consider the sensitivity of the optimal capacity size and net profit to variations of the processing (service) rate μ. Values of the fixed parameters are: $\lambda = 250, \alpha = 10\%, C_{TH} = 26.4,$ and $C_{OE} = 22$. The value of μ is changed from 30 to 110 with an increment of 10. We derive K^* and NP^* for each value of s varying from 2 to 6 (see Fig. 5 and Fig. 6).

The charts in Fig. 5 clearly demonstrate that at each value of s, as μ increases, the optimal capacity would first stay unchanged and then increase (in most cases). Like in the previous case of sensitivity to variations of λ, we also observe here a threshold for μ, such that:

- When μ is smaller than the threshold, the optimal capacity size is always K = s + 1;
- When μ is greater than or equal to the threshold, the optimal capacity size increases with μ.

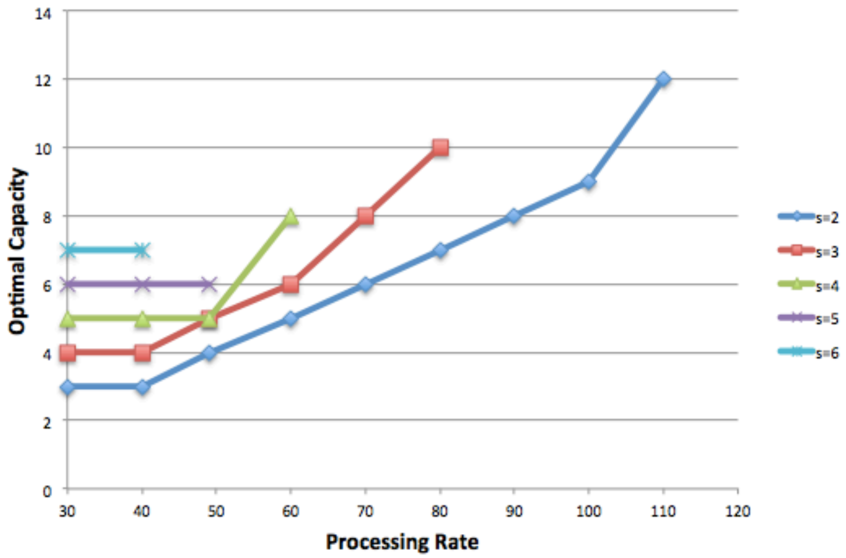


FIGURE 5. IMPACT OF PROCESSING RATE ON OPTIMAL CAPACITY.

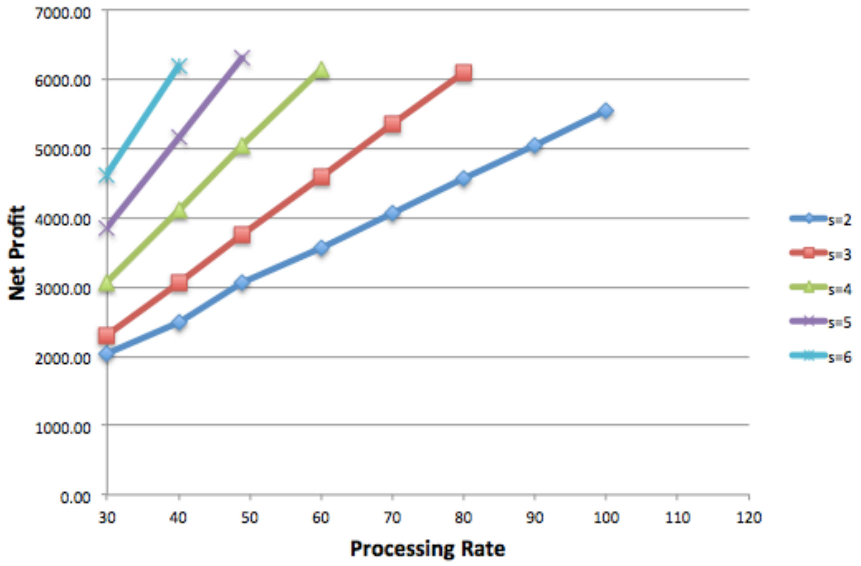


FIGURE 6. IMPACT OF PROCESSING RATE ON OPTIMAL NET PROFIT.

Then we analyse the sensitivity of the optimal *NP** to different values of μ. As can be seen from the charts in Fig. 6, the optimized net profit goes up with the increase of the processing rate. If the value of μ is fixed, then the optimal *NP** would go up as the number of servers increase.

Overall, based on the sensitivity of the optimal results to the processing rate μ, we can provide several important observations. First, when the processing rate is below a respective threshold, the optimal buffer size is always 1, i.e., $K = s + 1$, and, thus, there is no incentive to hold more buffer units, as the supply of units in the CCR's time buffer is always sufficient. Second, when the processing rate is increasing above the respective threshold, the CCR system should maintain a larger capacity size to guarantee that the servers are fully utilized and, as a result of such utilization, the optimal net profit will also go up. Finally, both s and μ would significantly impact the optimal net profit.

4.4. Impact of Throughput per Unit Cost

We define the ratio of throughput per unit (C_{TH}) to carrying cost per unit (C_{OE}), C_{TH}/C_{OE}, as the *throughput per unit cos*t (*TUC*). A larger *TUC* ratio represents a higher per-unit rate of return for each item processed in the CCR system.

For the sensitivity analysis of the optimal capacity size and net profit to variations of *TUC*, we fix values of the following parameters at: $\lambda = 250, \mu = 60, \alpha = 10\%$, and $C_{OE} = 22$. The value of C_{TH} varies from \$25 to \$175 with an increment of \$25, which produces a variation of *TUC* from a relatively low ratio of 1.14 to a relatively high ratio of 7.95. Then, we derive *K** and *NP** for a feasible value of *s* varying from 2 to 4 (see Fig. 7 and Fig. 8).

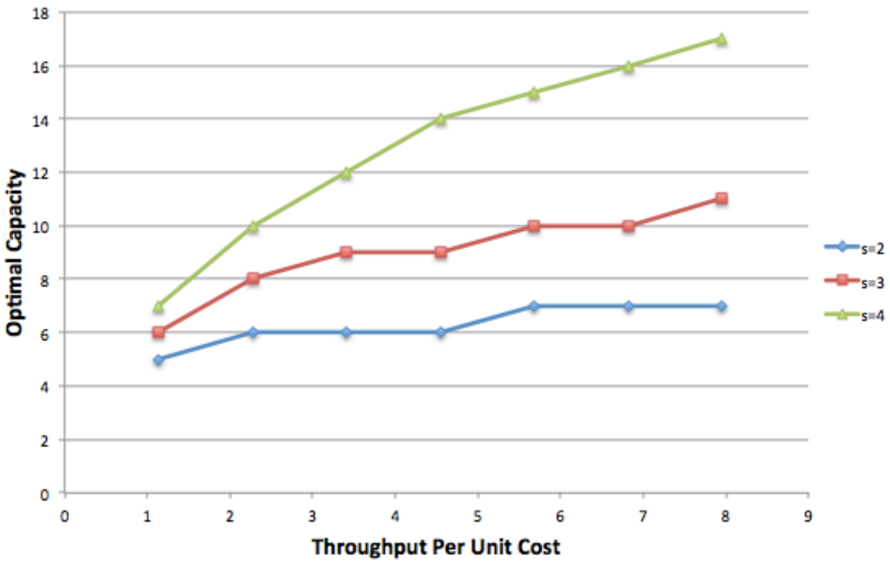


FIGURE 7. IMPACT OF TUC ON OPTIMAL CAPACITY.

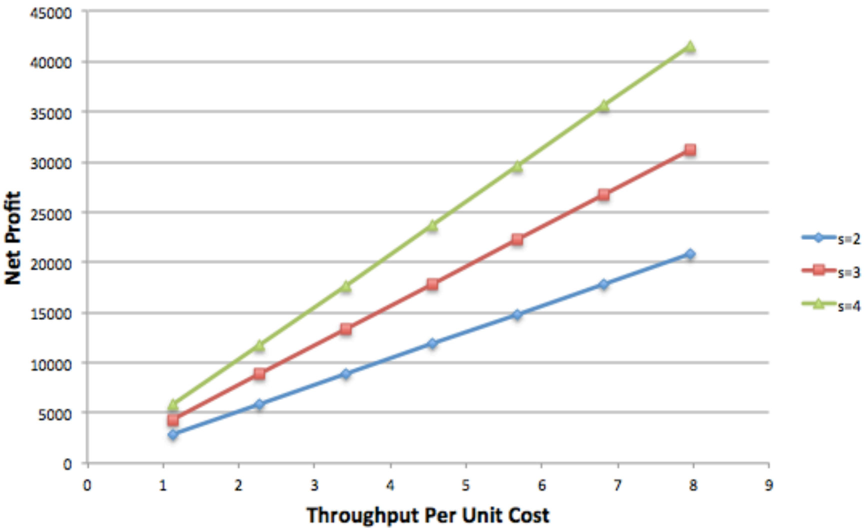


FIGURE 8. IMPACT OF TUC ON OPTIMAL NET PROFIT.

As can be seen from Fig. 7, when TUC increases with a fixed value of s, the optimal capacity size also goes up. For example, if $s = 4$, K^* increases from 7 to 17. In addition, the optimal capacity would go up as the value of TUC increases with more servers (s).

The graphs in Fig. 8 yield several important observations:

- When the per-unit throughput is relatively large comparing to the per-unit carrying cost (TUC is higher), it is more profitable to design a CCR with more servers; and corresponding optimal capacity size increases with TUC. The rate of increase in profit with respect to TUC is higher at a higher number of servers.
- The increase of TUC, given the same number of servers, would significantly affect the optimal NP growth. This is obvious since NP is linear in C_{TH} , assuming that all other parameters are fixed.

V. CONCLUSION

This paper provides analysis and insights into the time buffer size in a capacity constrained resource (CCR) with parallel processing. An appropriate size of the time buffer, which is, in combination with the number of servers, denotes the CCR's capacity size, would protect the CCR with certain probability from becoming idle. We formulate the optimal capacity size model as a constrained optimization problem in the setting of a CCR being a finite multi-sever queue. The objective of this model is to maximize the CCR's net profit in the context of the TOC. The major contributions and managerial insights stemming from this research are as follows.

First, we prove that the unconstrained optimization model has a unique optimal solution for capacity size. When the model is constrained, the optimal capacity is affected by the CCR's accepted idle rate. However, there exists a threshold for this accepted idle rate. When it is larger than the threshold, it would not affect the optimal capacity size, and, in this case, a higher optimal net profit can be achieved as compared to the case with a smaller than the threshold accepted idle rate. This finding suggests that a CCR system, in order to increase its net profit, may need to relax the accepted idle rate constraint.

Next, we provide an extensive analysis of the optimal capacity size and associated net profit sensitivity to variations of the model parameters.

Based on the numeric analysis of arrival rate, we observe a threshold for this parameter, with the values of arrival rate above the threshold leading to the optimal capacity size equal to the number of servers plus 1. Contrary to that, with the arrival rate being relatively small and below the threshold, the CCR is required to maintain a larger optimal capacity size to guarantee that the servers are fully utilized.

The results of numerical analysis of processing (service) rate reveal that there is also a threshold for this parameter. If the processing rate is below the threshold, the optimal buffer size is always 1, and the optimal capacity size is equal to the number of servers plus 1. This is due to the fact that the processing rate becomes lower than the arrival rate, and, therefore, there is no need to hold more buffer units in front of the CCR. When the processing rate is increasing above the threshold, the CCR is required to maintain a larger optimal capacity size to guarantee that the CCR servers are fully utilized.

Comparing the impact of arrival rate and processing rate, we find that the processing rate (μ) and processing capacity ($s * \mu$) have a much more significant impact on the net profit. It is because we control on the idle rate of the CCR system to be quite small. At the same time it is relatively cost-efficient to adjust the capacity (or buffer size) to make sure all the servers are fully utilized. Hence the impact of arrival rate on Net Profit is relative minor comparing with processing rate.

Finally, the sensitivity analysis to variations of the throughput per unit cost (*TUC*) shows that, when the *TUC* is relatively large, it is more profitable to design a CCR with more servers. In this case, the corresponding optimal capacity size increases with *TUC* growth, and the rate of increase in the optimal net profit with respect to *TUC* is higher with the larger number of servers.

The future research of the optimal time buffer (capacity) size in the CCR with parallel processes may cover several possible directions. For example, we may consider a multi-product CCR system, where the arrival rates dynamically change due to various product types and their arrival patterns. In addition, the processing (service) rate can potentially dynamically change due to the different processing requirement in a multi-product system. All this may necessitate an analysis of the optimal capacity size and net profit in a queuing system with a general arrival and service patterns.

REFERENCES

Askin, R., and Standridge, C., *Modeling and Analysis of Manufacturing Systems.* John Wiley and Sons, New York, 1993.

Bie, L., Cui, N. and Zhang, X., "Buffer sizing approach with dependence assumption between activities in critical", *International Journal of Production Research*, 50(24), 2012, 7343-7356.

Corbet, T. and Csillag, J.M., "Analysis of the effects of seven drum-buffer-rope implementations." *Production and Inventory Management Journal*, 42(3), 2001, 17-23.

Feng, W., Zheng, L. and Li, J., "The robustness of scheduling policies in multi-product manufacturing systems with sequence-dependent setup times and finite buffers", *Computers & Industrial Engineering*, 63(4), 2012, 1145-1153.

Ghosh , A. P. and Weerasinghe, A. P., "Optimal buffer size for a stochastic processing network", *Queuing Systems*, 55, 2007, 147-159.

Goldratt, E. M, *Theory Of Constraints Handbook.* Edited by James F. Cox III and John G. Schleier. McGraw Hill, Chicago, 2007.

Goldratt, E.M. *The Haystack Syndrome: Sifting Information Out of the Data Ocean.* North river Press, Croton-on-Hudson, NY, 1990.

Goldratt, E.M., and Cox, J., *The Goal* (2nd ed.), North River Press, Croton-on-Hudson, NY, 1992.

Gross , D., Shortle, J. F., Thompson, J. M. and Harris, C. M., *Fundamentals of Queuing Theory* (4th ed.), John Wiley and Sons, Hoboken, NJ, 2008.

Gupta, M. C., and Boyd, L. H., "Theory of constraints: a theory for operations management", *International Journal of Operations & Production Management*, 28(10), 2008, 991-1012.

Hadas, L., Cyplik, P. and Fertsch, M., "Method of buffering critical resources in make-to-order shop floor", *International Journal of Production Research*, 47(8), 2009, 2125–2139.

Hwang, Y.J., Huang, C.L. and Li, R.K., "Using Simplified Drum-Buffer-Rope To Rapidly Improve Operational Performance: A Case Study in China", *Production and Inventory Management Journal*, 47(1), 2001, 80-93.

Inman, R.A., Sale, M.L. and Green Jr., K.W., "Analysis of the relationships among TOC use, TOC outcomes, and organizational performance", *International Journal of Operations & Production Management*, 29(4), 2009, 341-356.

Inman, R.R., "Empirical Evaluation of Exponential and Independence Assumptions in Queuing Models of Manufactuing Systems", *Production and Operations Management*, 8(4), 1999, 409-432.

Lee, J.H., Chang, J.-G., Tsai, C.-H. and Li, R.-K., "Research on enhancement of TOC Simplified Drum-Buffer-Rope system using novel generic procedures", *Expert*

Systems and Applications, 37, 2010, 3747-3754.

Lee, J.H., Hwang, Y.J., Wang, M.T. and Li, R.K., "Why Is High Due-Date performance So Difficult to Achieve?--An Experimental Study", *Production and Inventory Management Journal*, 42(6), 2009, 30-43.

Long, L. D., and Ohsato, A., "Fuzzy critical chain method for project scheduling under resource constraints and uncertainty", *International Journal of Project Management* , 26(6), 2008, 688–698.

Louw, L., and Page, D.C., "Queuing network analysis approach for estimating the sizes of the time buffers in Theory of Constraints-controlled production system", *International Journal of Production Research*, 42(6), 2004, 1207-1226.

Mabin, V. and Balderstone, S., "The performance of the theory of constraints methodology." *International Journal of Operations & Production Management*, 23(6), 2003, 568-595.

Naor, M., Bernardes, E. S. and Coman, A., "Theory of constraints: is it a theory and a good one?" *International Journal of Production Research*, 51(2), 2013, 542-554.

Newbold, R. C. *Project management in the fast lane – applying the theory of constraints,* The St Lucie Press, Boca Raton, 1998.

Radovilsky, Z., "A quantitative approach to estimate the size of the time buffer in the theory of constraints", *International Journal of Production Economics*, 55(7), 1998, 113-119.

Radovilsky, Z. and Frankel, M., "Identifying buffer size in front of capacity-constrained resource with parallel processes", *International Jounral of Business Research*, 13(4), 2013 189-198.

Shaaban, S., and McNamara, T., "The Performance of Unpaced Lines with Unequal Buffer Sizes", *The Business Review, Cambridge*, 14(1), 2009, 105-110.

Simon, J., and Hopp, W., "Availability and Average Inventory of Balanced Assembly-Like Flow Systems", *IIE Transactions*, 23(2), 1991, 161-168.

Sirikrai, V., and Yenradee, P., "Modified drum-buffer-rope scheduling mechanism for a non-identical parallel machine flow shop with processing-time variations", *International Journal of Production Research*, 44(17), 2006, 3509-3531.

Spector, Y., "Theory of constraint methodology where the constraint is the business model", *International Journal of Production Research*, 49(11), 2011, 3387-3394.

Tseng, M.F., and Wu, H.H., "The Study of an Easy-to-Use DBR and BM System", *International Journal of Production Research*, 44(8), 2006, 1449-1478.

Tukel, O.I. , Rom, W.O. and Eksioglu, S. D., "An investigation of buffer sizing techniques", *European Journal of Operations Research*, 172, 2006, 401-416.

Ye, T., and Han, W., "Determination of buffer sizes for drum–buffer–rope", *International Journal of Production Research*, 46(10), 2008, 2827-2844.

Appendix A
Proof of Unique Optimal Solution Existing in Objective Function

Given the objective function in equation (4) (without the constraint):

$$\begin{aligned} NP(K) &= s\,\mu\,C_{TH} - \; s\,\mu\,C_{TH}\,p_0(K) - C_{OE}L_s(K) \\ &= \; s\,\mu\,C_{TH} - s\,\mu\,C_{TH}p_0(K) - C_{OE}L_q(K) + C_{OE}\gamma p_k(K) - C_{OE}\gamma, \end{aligned}$$

where

$$p_k = p_0\frac{\gamma^s}{s!}\rho^{K-s},$$

$$\begin{aligned} L_q &= \frac{p_0\gamma^s\rho}{s!\,(1-\rho)^2}[1-\rho^{K-s+1}-(1-\rho)(K-s+1)\rho^{K-s}] \\ &= \frac{p_0\gamma^s\rho}{s!\,(1-\rho)^2} - \frac{\rho^2}{(1-\rho)^2}p_k + \frac{\rho}{\rho-1}(K-s+1)p_k. \end{aligned}$$

In the equation for *NP(K)*, parameters $p_0(K)$, $L_q(K)$, and $p_k(K)$are dependent of *K*. Let us define $Z(K) = \; s\,\mu\,C_{TH}p_0(K) + C_{OE}L_q(K) - C_{OE}\gamma p_k(K)$, where members have an opposite sign to those in the equation for *NP(K)*. As long as $Z(K)$ is convex in *K* then it's direct that $NP(K)$ is concave.

$Z(K)$ could be rewritten as follows (note $\rho = \frac{\gamma}{s}$):

$$\begin{aligned} Z(K) &= s\,\mu\,C_{TH}p_0 + C_{OE}\left(\frac{p_0\gamma^s\rho}{s!\,(1-\rho)^2} - \frac{\rho^2}{(1-\rho)^2}p_0\frac{\gamma^s}{s!}\rho^{K-s} + \frac{\rho}{\rho-1}(K-s+1)p_0\frac{\gamma^s}{s!}\rho^{K-s}\right) \\ &\quad - C_{OE}\rho s p_0\frac{\gamma^s}{s!}\rho^{K-s} \\ &= p_0\left(s\,\mu\,C_{TH} + \frac{\gamma^s\rho}{s!\,(1-\rho)^2}C_{OE}\right) + C_{OE}p_0\frac{\gamma^s}{s!}\cdot\rho^{K-s}\left(\frac{\rho}{\rho-1}(K-s+1) - \frac{\rho^2}{(1-\rho)^2} - \rho s\right). \end{aligned}$$

Then we define $G(K) = \rho^{K-s}(\frac{\rho}{\rho-1}(K-s+1) - \frac{\rho^2}{(1-\rho)^2} - \rho s)$. If we can prove that $p_0(K)$ and $G(K)$ are both convex, then $Z(K)$ is convex.

Since $p_0(K) = \left(\sum_{n=0}^{s-1}\frac{\gamma^n}{n!} + \frac{\gamma^s}{s!}\,\frac{1-\rho^{K-s+1}}{1-\rho}\right)^{-1}$ $(\rho > 1)$. It has the same convexity with $\Lambda(K) = \rho^{K-s+1}$. And given that $\Lambda(K) = \rho^{K-s+1}$ is convex in K when $\rho > 1$, therefore, $p_0(K)$ is convex in *K*.

Then to prove the convexity of $G(K)$, we derive the second order condition as follows:

$$\begin{aligned} \frac{d^2G}{dK^2} &= \frac{\rho^{K-s+1}\log\rho}{(\rho-1)^2}[2(\rho-1) + \;\log\rho\,(2\,(\,\rho-1)\;+(-1+K\,(\rho-1)+s\,(\,\rho-2)(\,\rho-1) \\ &\qquad + 2\rho)]. \end{aligned}$$

As long as $\frac{d^2G}{dK^2} \geq 0$, then $G(K)$ must be convex. And the corresponding condition is $[2(\rho-1) + \;\log\rho\cdot(2\,(\rho-1)\;+(-1+K\,(\rho-1)+s\,(\,\rho-2)(\,\rho-1)+2\rho)] \geq 0$.

In sum, $NP(K)$ is concave in K when $[2(\rho-1) + \;\log\rho\cdot(2\,(\rho-1)\;+(-1+K\,(\rho-1)+s\,(\,\rho-2)(\,\rho-1)+2\rho)] \geq 0$. Further, there must exist unique optimal solution of K. Note that in the all of the numerical analysis in this paper, the parameter-settings guarantee the concavity of (K) .

Appendix B
Proof of Threshold for Accepted Idle Rate

From constraint (5), $K \geq (s+1)$. In addition, given the constraint (3) on the idle probability, $K \geq s - 1 + \frac{ln\left[\frac{(\rho-1)(1-\alpha D)+\alpha B}{\alpha B}\right]}{l\,n\,\rho}$. Then we denote $(s - 1 + \frac{ln\left[\frac{(\rho-1)(1-\alpha D)+\alpha B}{\alpha B}\right]}{l\,n\,\rho})$ as $\overline{K}$, and thus $LB_K = \max\{s+1, \overline{K}\}$, where LB_K is the lower bound of K.

The value of α affects the optimal capacity size when $\overline{K}$ is larger than $s+1$, and hence, $LB_K = \overline{K}$. $\overline{K} - (s+1) > 0$ when $s - 1 + \frac{\ln\left[\frac{(\rho-1)(1-\alpha D)+\alpha B}{\alpha B}\right]}{\ln \rho} - (s+1) = \frac{\ln\left[\frac{(\rho-1)(1-\alpha D)+\alpha B}{\alpha B}\right]}{\ln \rho} - 2 > 0$. Therefore, $\hat{\alpha} = \frac{\rho-1}{(\rho^2-1)\,B+(\rho-1)\,D}$.

In summary, we have the following:

- If $\alpha < \hat{\alpha}$, then $\overline{K} > (s+1)$ and $LB_K = \overline{K}$. The lower bound of K is larger than $(s+1)$.
- If $\alpha \geq \hat{\alpha}$, then $\overline{K} \leq (s+1)$ and $LB_K = s+1$. The lower bound of K is $(s+1)$.