

# On The Relationships among Traffic Load, Capacity, and Throughput for the $M/M/1/m$ , $M/G/1/m$ -PS, and $M/G/c/c$ Queues

Serhan Ziya

Department of Statistics and Operations Research

University of North Carolina

CB#3260, 213 Smith Building, Chapel Hill, NC 27599 U.S.A.

Email: [ziya@unc.edu](mailto:ziya@unc.edu), Phone: +1 919 933 3899, Fax: +1 919 962 0391

## Abstract

System throughput is one of the widely used performance measures in manufacturing, communication, and service networks. Although there are exceptions, throughput of such systems typically increases with additional capacity. We investigate how this improvement in throughput depends on the traffic load. More specifically, we consider  $M/M/1/m$ ,  $M/G/1/m - PS$ , and  $M/G/c/c$  queues, all of which arise in a variety of contexts. For the  $M/G/c/c$  queue, we show that throughput improvement (both nominal and relative) that would be obtained by adding an extra server is increasing in the traffic load. For the  $M/M/1/m$  and  $M/G/1/m - PS$  queues, we show that throughput improvement (both nominal and relative) that would be obtained by adding an extra buffer space is unimodal in traffic load. In particular, the relative improvement is maximized when the traffic load is one regardless of the buffer size. We also prove a new structural property for the blocking probability of the  $M/M/1/m$  and  $M/G/1/m - PS$  queues.

## I. INTRODUCTION

One of the widely used performance measures for both manufacturing and communication networks is the system *throughput*, which has also been well-studied in the academic literature. Many authors have used queueing models in order to generate insights on the relationship between throughput and system parameters, how networks should be designed so that throughput is maximized or kept at some reasonable level, and also how it can be estimated or approximated for relatively complex systems. For some examples of such work, see Buzacott and Shanthikumar [3], MacGregor Smith and Cruz [17], Duenyas and Hopp [7], and Chen and Jordan [6].

We are mainly concerned with understanding the relationships among system capacity (buffer or service capacity), throughput<sup>1</sup>, and traffic load. Typically, as intuition would suggest, throughput increases as the system capacity increases. (This can be easily proven for many standard single-station queueing models, but there are some exceptions.)<sup>2</sup> What is much less clear is how this increase in throughput depends on the system load. To be more precise, if we compute the throughput improvement that would be obtained as a result of an increase in the buffer or service capacity, would this improvement be higher when the system is lightly loaded or when it is heavily loaded, or perhaps when it is somewhere in between? Our objective here is to investigate this question using some standard queueing models. Our results provide insights on what type of systems would benefit most from an increase in capacity thereby also suggesting a way to aid in capacity expansion decisions concerning a group of independently operated stations.

More specifically, we consider  $M/M/1/m$ ,  $M/G/1/m - PS$ , and  $M/G/c/c$  queues. For the former two systems we increase the buffer space  $m$ , while for the last system, we increase the number of servers  $c$  and we observe how the system load effects the improvement obtained in throughput as a result of these changes. Single station queueing models are typically better fit for networks with simple structures, but their analysis can also provide valuable insights on complex systems since some of these simple models carry the essential characteristics of the

<sup>1</sup>The word “throughput” has been defined differently in different contexts (see, e.g., Chen and Jordan [6]). Here, we define it as the number of jobs served by the system per unit time.

<sup>2</sup>It is generally reasonable to expect the throughput to increase with the buffer and/or service capacity. This can be verified for many standard queueing systems such as the  $M/M/c/m$  queue (see, e.g., Köchel [15]) and in fact Sonderman [21] showed that the result holds under more general conditions. However, the result is not true in general. See, Whitt [22] for a counterexample.

original systems. The three queueing models we investigate in this paper have frequently been used in various settings.

**The  $M/M/1/m$  Queue:** Buzacott and Shanthikumar [3] extensively discuss how various queueing models can be used in the analysis of manufacturing systems. Among other queueing models, they use the  $M/M/1/m$  queue to model a single machine produce-to-stock system with lost sales. In this system, there are  $m$  “production authorization” (PA) tags in total and each product in store has a PA tag attached to it. When a product is sold, the attached PA tag is removed, and it is sent to the machine to authorize a new production. Hence, at any point in time, there are at most  $m$  products in stock. If there are no items in stock when a customer arrives, that customer is lost. In these systems, PA tags help keep production and inventory levels under control. One of the important decisions here would be how many PA tags to use, or equivalently how to set the maximum inventory level in the system. Such a decision would take various factors into account including inventory holding cost, system throughput, etc. (Buzacott and Shanthikumar consider one example where the objective is to maximize the long-run average net profit.) Therefore, it is important to understand how changes in the maximum inventory level would affect the system throughput, and under what conditions such changes would have the most impact on the throughput. We find that additional inventory space increases throughput by a larger amount when demand is neither too low nor too high compared with the production capacity.

**The  $M/G/1/m - PS$  Queue:** Processor sharing queueing models are frequently used in the networking literature to model systems where users share a fixed amount of bandwidth. For some examples, see Roberts [20], Berger and Kogan [2], Cao et al. [4], Chen and Jordan [6] and references therein. In particular, Cao et al. use the  $M/G/1/m - PS$  queue to model the jobs in a web server. The authors derive expressions for some of the web server performance measures including throughput, and show that the performance predicted by the model fits well with what the authors observe in practice in an experiment where a Web server is sent dummy requests from clients. On the other hand, Chen and Jordan specifically focus on throughput stating that the throughput “is the most common performance metric for data applications in the Internet”. The authors consider the  $M/G/1/m - PS$  queue, derive expressions for various performance measures related to throughput and average transmission rate, and provide several structural results on the relationships among these measures. In this context, the finite buffer capacity  $m$  stands for the queueing capacity for the server and thus our results provide insights

on how changes in this capacity affect the system throughput. According to our results, changes in the buffer capacity have the most significant effects if the load on the system is neither too low nor too high.

**The  $M/G/c/c$  Queue:** The Erlang loss system, i.e., the  $M/G/c/c$  queue is used in a variety of applications including telecommunication and service networks (see, e.g., Gross and Harris [9] and Gans, Koole, and Mandelbaum [8]). It is easy to show that as the service capacity of the system increases, the throughput also increases, but we investigate how this increase depends on the traffic load on the system. For telecommunication networks, our results provide insights on the benefit of having an additional line available for incoming calls and for service networks, they give insights on the benefits of having an additional server. It turns out that the benefit of having an additional line or a server is higher when system load is higher.

In short, the questions we investigate in this paper are relevant in the design and optimization of communication, manufacturing, and service networks. Numerous papers have addressed similar design and optimization questions, however, to our knowledge, there is no prior work that investigates the relationship between the traffic load and the effect of capacity changes on the throughput. There is significant amount of work that proves structural results on some performance measures including the system throughput and blocking probabilities. These papers mostly prove convexity of various performance measures and blocking probabilities with respect to certain system parameters.

The  $M/G/c/c$  queue in particular has received significant attention. For example, Messerli [18] shows that the blocking probability is convex in the number of servers. Jagers and Van Doorn [12] generalizes this result for the case where the parameter for the number of servers is not necessarily an integer. Jagerman [11] proves several structural properties including the log-convexity of the inverse of the blocking probability in the number of servers. Yao and Shanthikumar [23] show that the blocking rate is increasing convex in the arrival rate. On the other hand, Harel [10] and Krishnan [16] prove that the the blocking probability is jointly convex and throughput is jointly concave in the arrival and service rates.

On the other hand, with some exceptions, not much has appeared on the structural properties of the blocking probability and related performance measures for queueing systems with delays. Chang et al. [5] prove that the blocking probability is decreasing convex in the number of servers and in the service rate for the  $G/M/c/m$  queue, Pacheco [19] considers the  $M/M/c/m$

queue and shows that the blocking probability is convex in  $m$ , convex in the traffic intensity for low traffic load and concave for high traffic load. Ziya et al. [24] prove some monotonicity results for the  $G/GI/c$  queue. One implication of their results is that the blocking probability is non-decreasing in the arrival rate for the  $G/GI/c/m$  queue.

It appears that there are mainly two reasons why blocking probabilities have received such attention. First, blocking probability, like throughput is an important measure by itself (see Altman and Jean-Marie [1]). Second, structural properties of blocking probabilities can be quite useful in the optimal design of queueing systems since any optimal design problem that concerns a loss system is highly likely to include the blocking probability in the objective function. In this paper, although our main focus is on system throughput, we also prove a new structural property for the blocking probability for the  $M/M/1/m$  and  $M/G/1/m - PS$  queues, which we use in proving some of our main results, but the result is also of independent interest because of this importance of blocking probabilities.

The rest of the paper is organized as follows. In Section II, we investigate the relationship between the traffic load and the throughput improvement that would be obtained with an additional buffer space in  $M/M/1/m$  and  $M/G/1/m - PS$  queues. We prove that both nominal and relative improvement in throughput are unimodal functions of traffic load and that relative improvement is maximized when traffic load is 1. In Section III, we study the  $M/G/c/c$  queue and investigate the effect of changes in  $c$  on the throughput. We find that both nominal and percentage improvement in throughput are increasing functions of the traffic load on the system.

## II. INCREASING THE BUFFER SIZE OF THE $M/M/1/m$ AND $M/G/1/m - PS$ QUEUES

Consider an  $M/M/1/m$  queue with  $1 \leq m < \infty$  where jobs arrive with rate  $\lambda$  and the service rate  $\mu = 1$  without loss of generality. We use  $\rho = \lambda/\mu = \lambda$  to denote the traffic intensity and  $B(\rho, m)$  to denote the blocking probability, i.e. the long-run proportion of customers who find the system full and balk. It is well-known that  $B(\rho, m)$  can be expressed as follows (see, e.g., Gross and Harris 1998):

$$B(\rho, m) = \begin{cases} \frac{(1-\rho)\rho^m}{1-\rho^{m+1}}, & \rho \neq 1 \\ \frac{1}{m+1}, & \rho = 1 \end{cases}. \quad (1)$$

Now, consider an  $M/G/1/m - PS$  queue, i.e.,  $M/G/1/m$  queue working according to the processor sharing service discipline where jobs arrive with rate  $\lambda$  and the mean service time

requirement for each job is  $1/\mu = 1$ . Then, it is known that the limiting distribution of the number of jobs in the system is the same as the limiting distribution of the number of jobs in the  $M/M/1/m$  queue (see, e.g., Kleinrock [14]). Therefore, the blocking probability for the  $M/G/1/m - PS$  queue is also as given in (1) and both the  $M/M/1/m$  queue and the  $M/G/1/m - PS$  queue have the common expression for throughput given by

$$T(\rho, m) = \rho(1 - B(\rho, m)). \quad (2)$$

In the rest of this section, we prove several results on the blocking probability as well as the throughput given by (1) and (2), respectively. Obviously, our results apply to both the  $M/M/1/m$  and the  $M/G/1/m - PS$  queues.

We start with a technical result for the blocking probability, which is needed for the proof of one of the main results that follow. First, let  $B'(\cdot)$  denote the first derivative of  $B(\cdot)$  with respect to  $\rho$ . Then, we can show the following:

**Lemma II.1** *Let  $\Theta(\rho, m) = B'(\rho, m)/(1 - B(\rho, m))$  for  $m \geq 1$  and  $\rho \geq 0$ . Then*

- (i)  $\Theta(\rho, m) = 1/2$  for  $\rho = 1$ .
- (ii)  $\Theta(\rho, m + 1) > \Theta(\rho, m)$  for  $\rho > 1$ .
- (iii)  $\Theta(\rho, m + 1) < \Theta(\rho, m)$  for  $\rho < 1$ .

*Proof:* (i) The derivative of (1) with respect to  $\rho$  gives

$$B'(\rho, m) = \begin{cases} \frac{\rho^{m-1}(\rho^{m+1} - (m+1)\rho + m)}{(1 - \rho^{m+1})^2}, & \rho \neq 1 \\ \frac{m}{2m+2}, & \rho = 1 \end{cases}. \quad (3)$$

Then, it immediately follows that  $\Theta(1, m) = 1/2$ .

(ii) Using (1) and (3), after some algebra, it can be shown that

$$\Theta(\rho, m + 1) - \Theta(\rho, m) = \frac{\rho^{m-1}(1 - \rho)((m + 2)\rho(1 - \rho^m) - m(1 - \rho^{m+2}))}{(1 - \rho^m)(1 - \rho^{m+1})(1 - \rho^{m+2})}$$

which further simplifies to

$$\Theta(\rho, m + 1) - \Theta(\rho, m) = \frac{-\rho^{m-1}(1 - \rho)^4 \sum_{k=1}^m k(m - k + 1)\rho^{m-k}}{(1 - \rho^m)(1 - \rho^{m+1})(1 - \rho^{m+2})}. \quad (4)$$

Then we have  $\Theta(\rho, m + 1) - \Theta(\rho, m) > 0$  if  $\rho > 1$ .

(iii) From (4), it follows that  $\Theta(\rho, m + 1) - \Theta(\rho, m) < 0$  if  $\rho < 1$ . □

Now, define  $Q(\rho, m)$  to be

$$Q(\rho, m) = T(\rho, m + 1) - T(\rho, m) \text{ for } \rho > 0. \quad (5)$$

Then,  $Q(\rho, m)$  is the nominal improvement in throughput that would be obtained by adding one additional buffer space to the queue. We are interested in determining how  $Q(\rho, m)$  behaves as the traffic load  $\rho$  changes. Theorem II.1 characterizes this behavior.

**Theorem II.1** *For any fixed  $m \geq 1$ ,  $Q(\rho, m)$  is unimodal in  $\rho$  and it has a unique maximizer  $\rho^*(m) > 0$ . More precisely, there exists  $0 < \rho^*(m) < \infty$  such that*

$$\frac{d}{d\rho}Q(\rho, m) > (=)(<)0 \text{ for } \rho < (=)(>)\rho^*(m).$$

*Proof:* Suppose in the following that  $\rho \neq 1$ . (Note that explicitly considering the point  $\rho = 1$  does not change the proof significantly since it easy to show that  $\rho^*(m) \neq 1$  for any  $m$  and that  $Q(\rho, m)$  is continuous and differentiable in  $\rho$ .) Using (1), it can be shown that

$$Q(\rho, m) = \frac{\rho^{m+1}(1 - \rho)^2}{(1 - \rho^{m+1})(1 - \rho^{m+2})}.$$

Let  $Q'(\rho, m)$  denote the first derivative of  $Q(\rho, m)$  with respect to  $\rho$ . Then, after some algebraic simplifications, we obtain

$$Q'(\rho, m) = \frac{(1 - \rho)^2(1 - \rho^{m+1})(1 - \rho^{m+1})\rho^m\Gamma(\rho, m)}{(1 - \rho^{m+1})^2(1 - \rho^{m+2})^2} \quad (6)$$

where

$$\Gamma(\rho, m) = \frac{m + 1}{1 - \rho^{m+1}} + \frac{\rho^{m+2}(m + 2)}{1 - \rho^{m+2}} - \frac{2\rho}{1 - \rho}. \quad (7)$$

Clearly then the sign of  $Q'(\rho, m)$  is determined by the sign of  $\Gamma(\rho, m)$ . Now,  $\Gamma(0, m) = m + 1$ . Therefore, it is sufficient to show that  $\Gamma'(\rho, m) < 0$  for all  $\rho \geq 0$  and that there exists  $\rho^*(m) < \infty$  such that  $\Gamma'(\rho^*(m), m) = 0$  where  $\Gamma'(\rho, m)$  denotes the first derivative of  $\Gamma(\rho, m)$  with respect to  $\rho$ . Taking the derivative, we find that  $\Gamma'(\rho, m) < 0$  if and only if

$$(m + 1)^2 \frac{\rho^m}{(1 - \rho^{m+1})^2} + (m + 2)^2 \frac{\rho^{m+1}}{(1 - \rho^{m+2})^2} - \frac{2}{(1 - \rho)^2} < 0.$$

Then, it is sufficient to show that for all  $m \geq 1$  and  $\rho \geq 0$ ,

$$\alpha(\rho, m) - \frac{1}{(1 - \rho)^2} < 0$$

where

$$\alpha(\rho, m) = \frac{(m + 1)^2 \rho^m}{(1 - \rho^{m+1})^2}.$$

It is easy to establish that

$$\alpha(\rho, 1) = \frac{-(1-\rho)^2}{(1-\rho^2)^2} < 0.$$

We will now show that  $\alpha(\rho, m) \geq \alpha(\rho, m+1)$ , which will complete the proof. Now,  $\alpha(\rho, m) \geq \alpha(\rho, m+1)$  if and only if

$$\frac{(m+2)^2}{(m+1)^2} \leq \frac{(1-\rho^{m+2})^2}{\rho(1-\rho^{m+1})^2}. \quad (8)$$

Now, taking the derivative of  $\frac{(1-\rho^{m+2})^2}{\rho(1-\rho^{m+1})^2}$ , we find that  $\frac{\partial}{\partial \rho} \left( \frac{(1-\rho^{m+2})^2}{\rho(1-\rho^{m+1})^2} \right) > (<)0$  if and only if

$$(2m+2)\rho^{m+1}(1-\rho) - (1-\rho^{m+1})(1+\rho^{m+2}) > (<)0.$$

The left hand side can be simplified as

$$\begin{aligned} & (2m+2)\rho^{m+1}(1-\rho) - (1-\rho^{m+1})(1+\rho^{m+2}) \\ &= (1-\rho)((2m+2)\rho^{m+1} - (1+\rho+\dots+\rho^m)(1+\rho^{m+2})) \\ &= (1-\rho)((2m+2)\rho^{m+1} - (1+\rho+\dots+\rho^m) - (\rho^{m+2} + \rho^{m+3} + \dots + \rho^{2m+2})) \\ &= (1-\rho)(\rho^{m+1}(1-\rho) + \rho^{m+1}(1-\rho^2 + \dots + \rho^{m+1}(1-\rho^{m+1})) \\ &\quad - (1-\rho^{m+1}) - (\rho - \rho^{m+1}) - \dots - (\rho^m - \rho^{m+1})) \\ &= (1-\rho) \left( (1-\rho)(\rho^{m+1} - \rho^m) + (1-\rho^2)(\rho^{m+1} - \rho^{m-1}) + \dots + (1-\rho^{m+1})(\rho^{m+1} - 1) \right) \end{aligned}$$

which is positive for  $\rho > 1$  and negative for  $\rho < 1$  since each product in the second parenthesis is negative for  $\rho > 1$  and positive for  $\rho < 1$ . Clearly, then the function  $\frac{(1-\rho^{m+2})^2}{\rho(1-\rho^{m+1})^2}$  is minimized at  $\rho = 1$ . Using L'Hospital's rule gives  $\lim_{\rho \rightarrow 1} \frac{(1-\rho^{m+2})^2}{\rho(1-\rho^{m+1})^2} = \frac{(m+2)^2}{(m+1)^2}$ , which implies that (8) holds. Hence, we proved that  $\Gamma'(\rho, m) < 0$ , which implies that there is at most one solution to  $Q'(\rho, m) = 0$ . Finally, it follows from the proof of Theorem II.2 that there is indeed a solution  $\rho^*(m) < \infty$ .  $\square$

According to Theorem II.1, for any buffer capacity  $m$ , the nominal improvement in throughput is first increasing then decreasing in the traffic load  $\rho$  having a unique maximizer  $\rho^*(m)$ . When  $\rho$  is small (less than  $\rho^*(m)$ ), the buffer capacity does not have much impact on the throughput since few customers are blocked, however as  $\rho$  increases (getting closer to  $\rho^*(m)$ ), the effect of buffer space also increases and thus gains from an additional buffer space are also higher. However, when  $\rho$  is large (greater than  $\rho^*(m)$ ), additional buffer space does not have much use since service capacity (rather than the buffer capacity) becomes more of a bottleneck. As

$\rho$  increases over the interval  $(\rho^*(m), \infty)$ , the steady-state probability of the server being idle gets smaller for any fixed buffer capacity and thus an additional buffer space has diminishing marginal value.

The following theorem describes how  $\rho^*(m)$  changes with  $m$ .

**Theorem II.2** *The maximizer for  $Q(\rho, m)$ ,  $\rho^*(m)$  is monotone decreasing in  $m$ , bounded above by 1.8 and below by 1. More precisely, we have*

$$\rho^*(1) > \rho^*(2) > \cdots > \rho^*(m) > \rho^*(m+1) > \cdots \lim_{n \rightarrow \infty} \rho^*(n) = 1.$$

Furthermore,  $\rho^*(1)$  is the unique solution to

$$\rho^3 - 2\rho - 2 = 0, \rho > 1$$

and  $\rho^*(1) < 1.8$ .

*Proof:* First, note that  $\rho^*(m)$  is bounded below by 1 since  $Q'(1, m) = 1/(2(m+1)(m+2)) > 0$  where  $Q'(\rho, m)$  denotes the first derivative of  $Q(\rho, m)$  with respect to  $\rho$ . Then, in order to establish the monotonicity property, it is sufficient to show that  $\Gamma(\rho, m+1) < \Gamma(\rho, m)$  for  $\rho > 1$  where  $\Gamma(\rho, m)$  is as given in (7). After a few algebraic simplifications, we can show that  $\Gamma(\rho, m+1) - \Gamma(\rho, m) < 0$  if and only if

$$(\rho^{m+1} - 1)((m+1)\rho^m - (m+1)\rho^{m+2} + \rho^{2m+2} - 1) > 0$$

Thus, we need to show that  $(m+1)\rho^m - (m+1)\rho^{m+2} + \rho^{2m+2} - 1 > 0$  for  $\rho > 1$ . We establish that by first observing that  $(m+1)\rho^m - (m+1)\rho^{m+2} + \rho^{2m+2} - 1 = 0$  for  $\rho = 1$  and showing that its derivative is positive for  $\rho > 1$ . Now the derivative can be shown to be positive if and only if  $m - (m+2)\rho^2 + 2\rho^{m+2} > 0$ , which can be shown to be correct using a simple induction argument. Thus,  $\rho^*(m)$  is decreasing in  $m$ , bounded by 1 and therefore it has a limit. Then,  $\lim_{n \rightarrow \infty} \rho^*(n) = 1$  can be shown using the fact that  $Q'(1, m) = 1/(2(m+1)(m+2))$  and  $\lim_{n \rightarrow \infty} \Gamma(\rho, n) < 0$  for any  $\rho > 1$ . The rest of the result can easily be shown by solving for  $\Gamma(\rho, 1) = 0$ .  $\square$

According to Theorem II.2, the value of traffic load that maximizes the throughput improvement is between 1 and 1.8 and is decreasing in  $m$ . Theorems II.1 and II.2 together provide useful insights on the effect of traffic load on the improvement in throughput as a result of buffer size

additions. Our results imply that the throughput improvement as a result of a queueing capacity increase would be largest for servers with loads that are neither too large nor too small (more precisely when the load is between 1 and 2). In the case of produce-to-stock systems, this implies that adding another PA card will be most beneficial when the customer demand is neither too high nor too low compared with the production capacity. Similarly, for a web server, increasing the queueing capacity will have the largest benefit on throughput for systems that are neither too heavily nor lightly loaded.

Suppose now that, among a group of independently operated queues with identical service and buffer capacities, we would like to determine the one that would benefit most from having an additional buffer space. Then, based on our results, we know the following: For any two servers for which the traffic load is less than 1, throughput improvement will be higher if the additional buffer space is given to the queue with the higher traffic load. On the other hand, for any two queues for which the traffic load is greater than 1.8, throughput improvement will be higher if the additional buffer capacity is given to the queue with the lower traffic load. On the other hand when buffer size is very large, the additional buffer space should be given to the queue for which the traffic load is approximately 1, if there are any.

In the proofs of Theorems II.1 and II.2, we also obtain some useful expressions, which can be used for optimization purposes. For example, suppose that we would like to determine the potential throughput improvement that would be obtained by adding extra buffer spaces to a system of  $K$  parallel queues, by possibly redistributing the total system load to the queues but perhaps under some constraints. Adding subscripts to our original notation, let  $Q_i(\rho_i, m_i)$  denote the throughput improvement that would be obtained by an additional buffer space in queue  $i$ . Now, consider the following optimization problem:

$$\begin{aligned} \max_{\rho_1, \rho_2, \dots, \rho_K} \quad & \sum_{i=1}^K Q_i(\rho_i, m_i) \\ \text{s.t.} \quad & \sum_{i=1}^K \rho_i = \rho^0, \\ & \rho_i \in I_i, \quad i = 1, 2, \dots, K \end{aligned} \tag{9}$$

where  $\rho^0$  is the total load that needs to be distributed to  $K$  queues, and  $I_i$  is the constraint set for  $\rho_i$ , the load on queue  $i$ . Then, the solution to Problem (9) is an upper bound on the maximum improvement that would be obtained in throughput as a result of changes in the buffer capacities and in obtaining this solution, the derivative expression (6), which we derived in the proof of Theorem II.1 can be readily used.

In assessing the value of an additional buffer space, it is also of interest to look at the relative improvement in the throughput as well. Define  $R(\rho, m)$  to be

$$R(\rho, m) = \frac{T(\rho, m+1)}{T(\rho, m)} \text{ for } \rho > 0. \quad (10)$$

Hence,  $R(\rho, m)$  gives the relative improvement in throughput that would be obtained by adding one extra buffer space to the queue. Then, we can prove the following.

**Theorem II.3** *For any fixed  $m \geq 1$ ,  $R(\rho, m)$  is unimodal in  $\rho$  and  $\rho^* = 1$  is its unique maximizer. More precisely,*

$$\frac{d}{d\rho} R(\rho, m) > (=)(<)0 \text{ for } \rho < (=)(>)1.$$

*Proof:* First note that since  $\ln(\cdot)$  is a strictly increasing function, it is sufficient to show that

$$\frac{d}{d\rho} \ln \left( \frac{1 - B(\rho, m+1)}{1 - B(\rho, m)} \right) > (=)(<)0 \text{ for } \rho < (=)(>)1.$$

Now, it is easy to show that

$$\frac{d}{d\rho} \ln \left( \frac{1 - B(\rho, m+1)}{1 - B(\rho, m)} \right) = \frac{B'(\rho, m)}{1 - B(\rho, m)} - \frac{B'(\rho, m+1)}{1 - B(\rho, m+1)}.$$

Then, the result immediately follows from Lemma II.1.  $\square$

The fact that  $R(\rho, m)$  is unimodal in  $\rho$  may not be surprising given Theorem II.1. However, it is interesting that  $R(\rho, m)$  is maximized at  $\rho = 1$  for any value of  $m$ . Regardless of what the buffer space is, relative improvement in throughput is maximized when the job arrival rate to the system is the same as the service rate.

### III. INCREASING THE SERVICE CAPACITY OF THE M/G/C/C QUEUE

Consider an  $M/G/c/c$  queue with  $1 \leq c < \infty$ . For convenience, we use the same notation as in Section II. Customers arrive with rate  $\lambda$ , and the mean service time is  $1/\mu = 1$ . We use  $\rho = \lambda/\mu = \lambda$  to denote the traffic load. With a slight abuse of notation, we also use  $B(\rho, c)$  and  $T(\rho, c) = \rho(1 - B(\rho, c))$  to denote the blocking probability and throughput, respectively.

It is well-known that  $B(\rho, c)$  can be expressed as follows (see, e.g., Gross and Harris 1998):

$$B(\rho, c) = \frac{\frac{\rho^c}{c!}}{\sum_{i=0}^c \frac{\rho^i}{i!}} \text{ for } \rho > 0. \quad (11)$$

Before we give our main result of this section, we first give the following lemma, which is due to Kelly [13], and is crucial in establishing our main results of this section. Again, we use  $B'(\rho, c)$  to denote the first derivative of  $B(\rho, c)$  with respect to  $\rho$ .

**Lemma III.1** *We have*

$$\frac{B'(\rho, c)}{1 - B(\rho, c)} = B(\rho, c - 1) - B(\rho, c) \text{ for } c \geq 1 \text{ and } \rho > 0.$$

Now, define  $S(\rho, c)$  to be

$$S(\rho, c) = T(\rho, c + 1) - T(\rho, c) \text{ for } \rho > 0. \quad (12)$$

Hence,  $S(\rho, c)$  is the nominal improvement in throughput that would be obtained by adding one extra server to the  $M/G/c/c$  system. Also define  $V(\rho, c)$  to be

$$V(\rho, c) = \frac{T(\rho, c + 1)}{T(\rho, c)} \text{ for } \rho > 0. \quad (13)$$

Then,  $V(\rho, c)$  gives the relative improvement in throughput that would be obtained by adding one extra server to the  $M/G/c/c$  system.

We can prove the following result for  $S(\rho, c)$  and  $V(\rho, c)$ .

**Theorem III.1** *Both  $S(\rho, c)$  and  $V(\rho, c)$  are strictly increasing in  $\rho$ , i.e.,  $\frac{d}{d\rho}S(\rho, c) > 0$  and  $\frac{d}{d\rho}V(\rho, c) > 0$  for  $\rho > 0$ .*

*Proof:* (i) First, we prove the first part of the result:  $S'(\rho, c) > 0$  where  $S'(\rho, c)$  denotes the first derivative of  $S(\rho, c)$  with respect to  $\rho$ .

Let  $D(\rho, c) = B(\rho, c) - B(\rho, c + 1)$  and  $D'(\rho, c)$  denote the first derivative of  $D(\rho, c)$  with respect to  $\rho$ . From Lemma III.1, we have

$$D(\rho, c) = \frac{B'(\rho, c + 1)}{1 - B(\rho, c + 1)}.$$

Using this, we get

$$D'(\rho, c) = B'(\rho, c) - B'(\rho, c + 1) = D(\rho, c - 1)(1 - B(\rho, c)) - D(\rho, c)(1 - B(\rho, c + 1)). \quad (14)$$

where  $B'(\rho, m)$  denotes the first derivative of  $B(\rho, m)$  with respect to  $\rho$ . Using (14), we get

$$\begin{aligned} \frac{D'(\rho, c)}{D(\rho, c)} &= \frac{D(\rho, c - 1)}{D(\rho, c)}(1 - B(\rho, c)) - (1 - B(\rho, c + 1)) \\ &> (1 - B(\rho, c)) - (1 - B(\rho, c + 1)) = -D(\rho, c) \end{aligned}$$

where the inequality follows from the convexity result of Messerli (1972). Thus,

$$\frac{D'(\rho, c)}{D(\rho, c)} > -D(\rho, c). \quad (15)$$

Now, suppose for contradiction that there exists  $\rho^0 > 0$  such that  $S'(\rho^0, c) \leq 0$ . Then, it can easily be shown that

$$\frac{D'(\rho^0, c)}{D(\rho^0, c)} \leq -\frac{1}{\rho^0}.$$

Using (15), we get

$$D(\rho^0, c) > \frac{1}{\rho^0}. \quad (16)$$

Since throughput cannot increase more than the service capacity of a single server (which can be shown to be correct using the convexity of  $B(\rho, c)$  with respect to  $c$ ), for any  $\rho > 0$ , we have  $S(\rho, c) < \mu$ . Dividing both sides of this inequality by  $\lambda$ , we get  $D(\rho, c) < \frac{1}{\rho}$ , which is a contradiction to (16). Hence the result follows.

(ii) We now prove the second part of the result:  $V'(\rho, c) > 0$  where  $V'(\rho, c)$  denotes the first derivative of  $V(\rho, c)$  with respect to  $\rho$ .

It can be shown that

$$V'(\rho, c) = \frac{-B'(\rho, c+1)(1-B(\rho, c)) + (1-B(\rho, c+1))B'(\rho, c)}{(1-B(\rho, c))^2}.$$

It follows that  $V'(\rho, c) > 0$  if and only if

$$B'(\rho, c)(1-B(\rho, c+1)) - B'(\rho, c+1)(1-B(\rho, c)) > 0. \quad (17)$$

Dividing both sides of (17) by  $(1-B(\rho, c+1))$ , we get the equivalent condition

$$B'(\rho, c) - \frac{B'(\rho, c+1)(1-B(\rho, c))}{1-B(\rho, c+1)} > 0. \quad (18)$$

From Lemma III.1, we know that  $B'(\rho, c+1) = (1-B(\rho, c+1))(B(\rho, c) - B(\rho, c+1))$ . Using this property twice in (18) first for the term  $B'(\rho, c+1)/(1-B(\rho, c+1))$  and then for  $B'(\rho, c)$ , we find that (18) simplifies to

$$(1-B(\rho, c))[(B(\rho, c-1) - B(\rho, c)) - (B(\rho, c) - B(\rho, c+1))] > 0. \quad (19)$$

Finally, since  $B(\rho, c)$  is known to be convex in  $c$  for fixed  $\rho$  (see, e.g., Messerli 1972), it follows that (19) holds and thus  $V'(\rho, c) > 0$ .  $\square$

According to Theorem III.1, throughput improvement (both nominally and percentage-wise) that would be gained by an additional server is higher for higher values of  $\rho$ . An additional server would be most useful if its utilization can be kept at a high level and server utilization can be kept high if the traffic load on the queue is high. Therefore, the throughput improvement brought by additional server will be higher the higher the traffic load on the queue.

As in the case of the models we analyzed in Section II, some of the expressions we have derived in this section (e.g., Equation (14)) can also be useful in solving throughput optimization problems involving  $M/G/c/c$  queues (similar to Problem (9) of Section II).

#### IV. CONCLUSIONS

We have investigated how throughput improvements as a result of capacity increments are affected by the traffic load on the system. For the sake of analytical tractability, we focused on queues with relatively simple structures, however, interpretation of our results for these simple queues provides insights that can be useful for systems that are much more complicated. For example, based on our results for the  $M/M/1/m$  and  $M/G/1/m - PS$  queues, it appears that a lightly or very heavily loaded system with a finite space for incoming jobs will not benefit much from increasing its buffer size while systems with mid-levels of traffic load will benefit the most. Additional buffer space for a lightly loaded system is not of much use because unless the buffer space is extremely small, jobs are rarely blocked. On the other hand, for a heavily loaded system, the real problem is the limited service capacity. Unless the buffer space is extremely small, the server will rarely idle because of the high load on the system. Thus, there is no strong incentive to increase buffer space when the load is low or when it is high. Our analysis of the  $M/G/c/c$  system suggests that, at least for systems with no delays, the benefit of increasing the service capacity will be higher the higher the load on the system. Higher benefits will be realized if the utilization can be kept at high levels, which will be the case if the traffic load is high.

#### REFERENCES

- [1] E. Altman, A. Jean-Marie, The loss process of messages in an  $M/M/1/K$  queue, in Proceedings of INFOCOM '94. Networking for Global Communications, pp. 1191-1198, 1994.
- [2] A. W. Berger, Y. Kogan, Dimensioning bandwidth for elastic traffic in high-speed data networks, IEEE/ACM Transactions on Networking vol. 8 pp. 643-654, 2000.

- [3] J. A. Buzacott, J. G. Shanthikumar, *Stochastic Models of Manufacturing Systems*, Prentice-Hall Inc., 1993.
- [4] J. Cao, M. Andersson, C. Nyberg, M. Kihl, Web server performance modeling using an  $M/G/1/K^*PS$  queue, in *Proceedings of International Conference on Telecommunication*, 2003.
- [5] C. S. Chang, X. Chao, M. Pinedo, J. G. Shanthikumar, Stochastic convexity for multidimensional processes and its applications, *IEEE Transactions on Automatic Control* vol. 36 pp. 1347–1355, 1991.
- [6] N. Chen, S. Jordan, Throughput in Processor-Sharing Queues, *IEEE Transactions on Automatic Control* vol. 52 pp. 299–305, 2007.
- [7] I. Duenyas, W. Hopp, Estimating the throughput of an exponential CONWIP assembly system, *Queueing Systems* vol. 14 pp. 135–157, 1993.
- [8] N. Gans, G. Koole, A. Mandelbaum, Telephone Call Centers: Tutorial, Review, and Research Prospects, *Manufacturing and Service Operations Management* vol. 5 pp. 79–141, 2003.
- [9] D. Gross, C. M. Harris, *Fundamentals of Queueing Theory*, Wiley, 1998.
- [10] A. Harel, Convexity properties of the Erlang loss formula, *Operations Research* vol. 38 pp. 499–505, 1990.
- [11] D. L. Jagerman, Some properties of the Erlang loss function, *The Bell System Technical Journal* vol. 53 pp. 525–551, 1974.
- [12] A. A. Jagers, E. A. Van Doorn, On the continued Erlang loss function, *Operations Research Letters* vol. 5 pp. 43–47, 1986.
- [13] F. P. Kelly, Routing in circuit-switched networks: Optimization, shadow prices and decentralization, *Advances in Applied Probability* vol. 20 pp. 112–144, 1988.
- [14] L. Kleinrock, *Queueing Systems Volume II: Computer Applications*, Wiley, 1976.
- [15] P. Köchel, Finite queueing systems - structural investigations and optimal design, *International Journal of Production Economics* vol. 88 pp. 157–171, 2004.
- [16] K. R. Krishnan, The convexity of loss rate in an Erlang loss system and sojourn in an Erlang delay system with respect to arrival and service rates, *IEEE Transactions on Communications* vol. 38 pp. 1314–1316, 1990.
- [17] J. MacGregor Smith, F. R. B. Cruz, The buffer allocation problem for general finite buffer queueing networks, *IIE Transactions* vol. 37 pp. 343–365, 2005.
- [18] E. Messerli, Proof of a convexity property of the Erlang B formula, *The Bell System Technical Journal* vol. 51 pp. 951–953, 1972.
- [19] A. Pacheco, Second-order properties of the loss probability in  $M/M/s/s+c$  systems, *Queueing Systems* vol. 15 pp. 289–308, 1994.
- [20] J. W. Roberts, A survey on statistical bandwidth sharing, *Computer Networks* vol. 45 pp. 319–332, 2004.
- [21] D. Sonderman, Comparing multi-server queues with finite waiting rooms, I: Same number of servers, *Advances in Applied Probability* vol. 11 pp. 439–447, 1979.
- [22] W. Whitt, Counterexamples for comparisons of queues with finite waiting rooms, *Queueing Systems* vol. 10 pp. 271–278, 1992.
- [23] D. D. Yao, J. G. Shanthikumar, The optimal input rates to a system of manufacturing cells, *Information Systems and Operational Research* vol. 25 pp. 57–65, 1987.
- [24] S. Ziya, H. Ayhan, R. D. Foley, E. Pekoz, A monotonicity result for a  $G/GI/c$  queue with balking or reneging, *Journal of Applied Probability* vol. 43 pp. 1201–1205, 2006.