Convexity, Buffer Allocation, and Finite Buffer Blocking Probabilities

Extended Abstract

J. MacGregor Smith *
email: jmsmith@ecs.umass.edu

February 2, 1999

Abstract — **The convexity properties of blocking probabilities of finite buffer queueing formulas are examined and analyzed. From the simplest formula for an M/M/1/K system, a closed form expression is developed interrelating the minimum buffer size and the parameters of the blocking probabilities and system utilization. The closed form expression for the M/M/1/K system is shown to be essentially a convex function in the range of the critical parameters. Also, for M/M/C/K systems, with a fixed value of C, a convex function for the buffer size as a function the parameters of the blocking probabilities and system utilization is developed. It is also shown that for M/D/1/K and M/G/1/K systems the approximation formula of Gelenbe is convex for these systems. Finally, the closed form expressions of M/M/1/K and M/M/C/K systems can be utilized for the Buffer Allocation Problem (BAP) in series, merge, and splitting topologies of finite buffer queueing networks.**

Keywords — Blocking Probabilities, Buffer Allocation

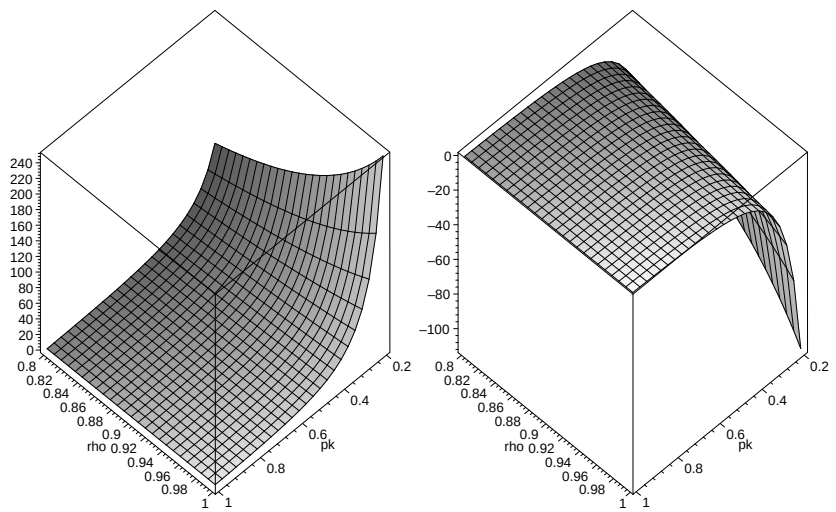

Figure 1: Properties of Blocking Probabilities

*Department of Mechanical and Industrial Engineering , University of Massachusetts Amherst Massachusetts 01003

1 INTRODUCTION

Convexity properties of the blocking probabilities of finite buffer queueing formulas are examined and analyzed in this paper with the idea of characterizing these properties and their eventual use in the buffer optimization problem.

From the simplest formula for an M/M/1/K system, a closed-form expression is developed interrelating the minimum buffer size and the parameters of the blocking probabilities and system utilization. The closed form expression for the M/M/1/K system is shown to be essentially a convex function in the range of critical parameters. Also, when one examines an M/M/C/K system, for fixed value of C, one can derive also a convex function for the buffer size as a function the parameters of the blocking probabilities and system utilization. Given the properties found for these Markovian systems, one wonders whether this convexity property would extend to other service time distributions.

Since closed form expressions of the blocking probabilities of M/G/1/K systems are not easily available, one can use approximations for the M/G/1/K systems to see if the convexity properties hold. In particular, it is shown that for M/D/1/K systems and M/G/1/K systems the approximation formula of Gelenbe is convex for these systems. Thus, for an effective approximation for M/G/1/K systems, it appears that the blocking probability should be a convex function of the system parameters. Finally, the closed form expressions are utilized for the Buffer Allocation Problem (BAP) in series, merge, and split topologies of finite buffer queueing networks.

A mathematical programming problem involving the closed form expression as the objective function to minimize the buffers at each workstation subject to a throughput constraint is shown to be an effective tool for explorations of the optimal buffer patterns in M/M/1/K and M/M/C/K systems. Extensions of the closed form expressions for M/D/1/K systems and M/G/1/K queueing network systems are reserved for another paper.

1.1 Motivation

In almost all real physical systems, finite buffers exist. In manufacturing systems there are assembly lines, material handling systems, and cellular manufacturing cells all with finite buffers. In telecommunication systems, finite capacity telephone lines, computers, and capacitated ATM switches are the norm. Finally, in service systems such as facilities, limited circulation systems (elevators, stairways and corridors), capacitated activity spaces, and storage areas are a reality. In all these system applications, it is crucial to compute the performance measures of these systems with finite buffers. It is also important to determine the optimal configuration of these systems with these finite buffers in mind. Any way in which this optimization process can be facilitated would be a great boost to the applications.

1.2 Outline of Paper

In Section 2 of this paper, a brief overview of the literature in this area is presented. In Section 3, the convexity properties of Markovian finite buffer systems are examined, while in Section, 4 the convexity of buffer properties of M/G/1/K and M/D/1/K systems are briefly examined. In Section 5, a mathematical programming formulation utilizing the convexity formulas for M/M/1/K and M/M/C/K systems is examined and Section 6 demonstrates the results for series, merge, and splitting topologies. Finally, Section 7 concludes the paper.

2 Literature Review

If one examines the literature on the convexity of queueing systems, one will come across the work of Harel [2], Harel and Zipkin[3], and Harel[4]. However, his examination of the convexity properties of queueing systems was generally focused on infinite systems except for their examination of Erlang loss systems. It appears, to the author's knowledge, that no one has really examined the convexity properties of the blocking probabilities of finite buffer systems largely because of the assumed integrality of k. There are certain flow systems such as bulk solids handling problems (fine particles, sand, powders, etc.) that non-integral buffers would make sense. If one relaxes this integrality as we shall see, then a closed form expressions for k can be realized. Also, when examining this closed form expression, the convexity properties will be examined.

2.1 Buffer Allocation Problem (BAP)

The Buffer Allocation Problem (BAP) is a notorious $\mathcal{NP}-$Hard combinatorial optimization problem. In general, there is no closed form expression interrelating the decision variable k with the throughput Θ and the other parameters of the queueing network: ρ, λ, μ etc. What we shall see, in this paper however, is that we will be able to generate a closed-form expression for the buffer sizes at each station as a function of ρ and p_k.

3 Markovian Systems

We will assume Poisson arrivals and Exponential service in these systems where finite buffers precede all servers. All routing probabilities will be assumed to be known apriori.

3.1 Notation

The following present some of the notation we need for the paper.

Λ := External Poisson arrival rate to the network.

λ_j := Poisson arrival rate to node j.

$\tilde{\lambda}_j$:= Effective arrival rate to node j.

μ_j := Exponential mean service rate at node j.

$\tilde{\mu_j}$:= Effective service rate at node j due to blocking.

c :=number of servers

$\rho = \frac{\lambda}{\mu c}$:= the traffic intensity

k_j := Buffer capacity at node j.

h := The holding node established in the Expansion method.

p_k := Blocking probability of finite queue of size K.

p'_k :=Feedback blocking probability in the Expansion method.

p_0^j :=Unconditional probability that there is no customer in the service channel at node j (either being served or being held after service).

Θ := Mean throughput rate.

3.2 M/M/1/K Systems

The blocking probability for an M/M/1/K system is well-known:

$$\frac{(1-\rho)\,\rho^k}{1-\rho^{k+1}} = p_k$$

If we relax the integrality of k, then we can arrive at a closed-form expression for the buffer size which is the largest integer as follows:

$$k = \left\lceil \frac{\ln\left(\frac{p_k}{1-\rho+p_k\rho}\right)}{\ln(\rho)} \right\rceil$$

This is a very useful formula for determining the buffer size k of an individual queue. What also is interesting about the formula is illustrated in Figure 2 which shows the smooth convex nature of the function over the range of ρ and p_k. The general range of the function for all values of ρ, p_k is on the left in Figure 2 and the specific function values for k in the range of most practical interest is on the right in Figure 2.

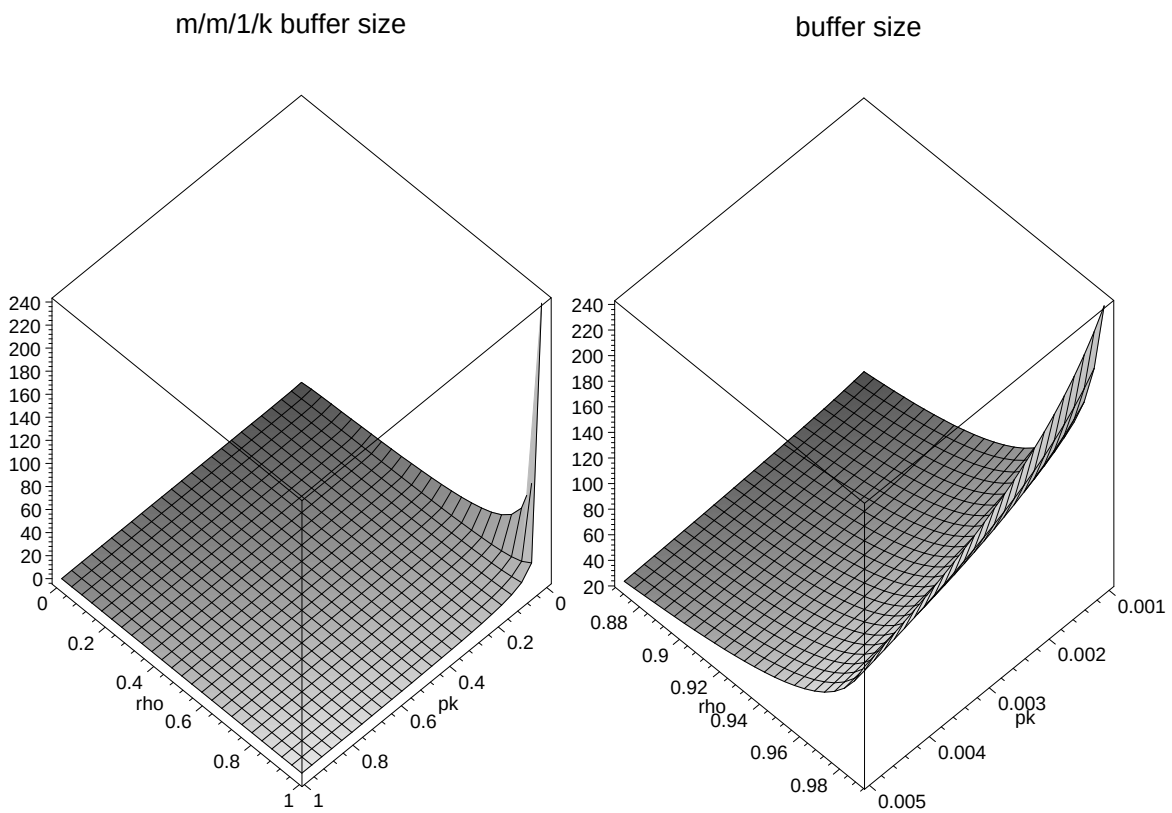

Figure 2: M/M/1/K Convex Function

3.3 Convexity of buffer formula

As is standard practice, in order to examine the convex nature of the closed-form expression, we shall form the Hessian matrix of the closed-form expression, differentiating with respect to ρ and

p_k. For the sake of the argument and compactness of the presentation, the terms of the Hessian are equal to:

$$H = \begin{bmatrix} H_{11}, H_{12} \\ H_{21}, H_{22} \end{bmatrix}$$

Let's examine each of the Hessian components, term-by-term. The first diagonal term H_{11} is:

$$H_{11} = \frac{(-1+\rho)\,(2\,p_k\,\rho + 1 - \rho)}{(1-\rho+p_k\,\rho)^2\,{p_k}^2\ln(\rho)}$$

If one examines the first term in the matrix, we shall see that it is positive everywhere in the range of $0 < \rho < 1$ and $0 < p_k < 1$. A graph of this function is indicated in Figure 3

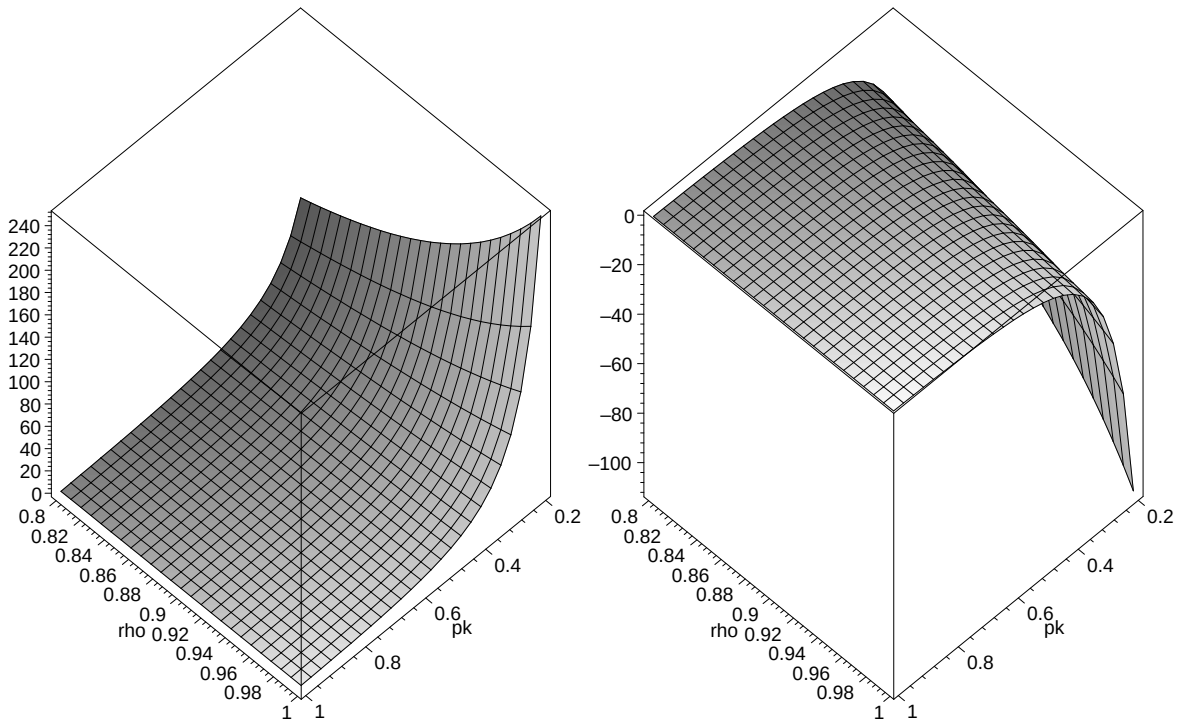

Figure 3: First diagonal H_{11}(left) and off-diagonal Hessian terms (right) H_{12}, H_{21}

For the off-diagonal terms $H_{12} = H_{21}$ we have:

$$H_{12} \equiv H_{21} = -\frac{\ln(\rho)p_k\,\rho - 1 + 2\,\rho - p_k\,\rho - \rho^2 + p_k\,\rho^2}{(1-\rho+p_k\,\rho)^2\,(\ln(\rho))^2\,p_k\,\rho}$$

This term is negative everywhere in the range $0 < \rho < 1$ and $0 < p_k < 1$. A graph of this function is indicated in Figure 3.

The second diagonal term H_{22} is given as:

$$H_{22} = \frac{(-1+p_k)^2}{(1-\rho+p_k\,\rho)^2\ln(\rho)} + 2\,\frac{-1+p_k}{(1-\rho+p_k\,\rho)\,(\ln(\rho))^2\,\rho}$$
$$+2\,\ln(\frac{p_k}{1-\rho+p_k\,\rho})\,(\ln(\rho))^{-3}\,\rho^{-2} + \ln(\frac{p_k}{1-\rho+p_k\,\rho})\,(\ln(\rho))^{-2}\,\rho^{-2}$$

This second diagonal term H_{22} unfortunately is not positive everywhere in the range of $0 < \rho < 1$ and $0 < p_k < 1$. A graph of this function is indicated in the range where it is actually positive which is a most practical situation.

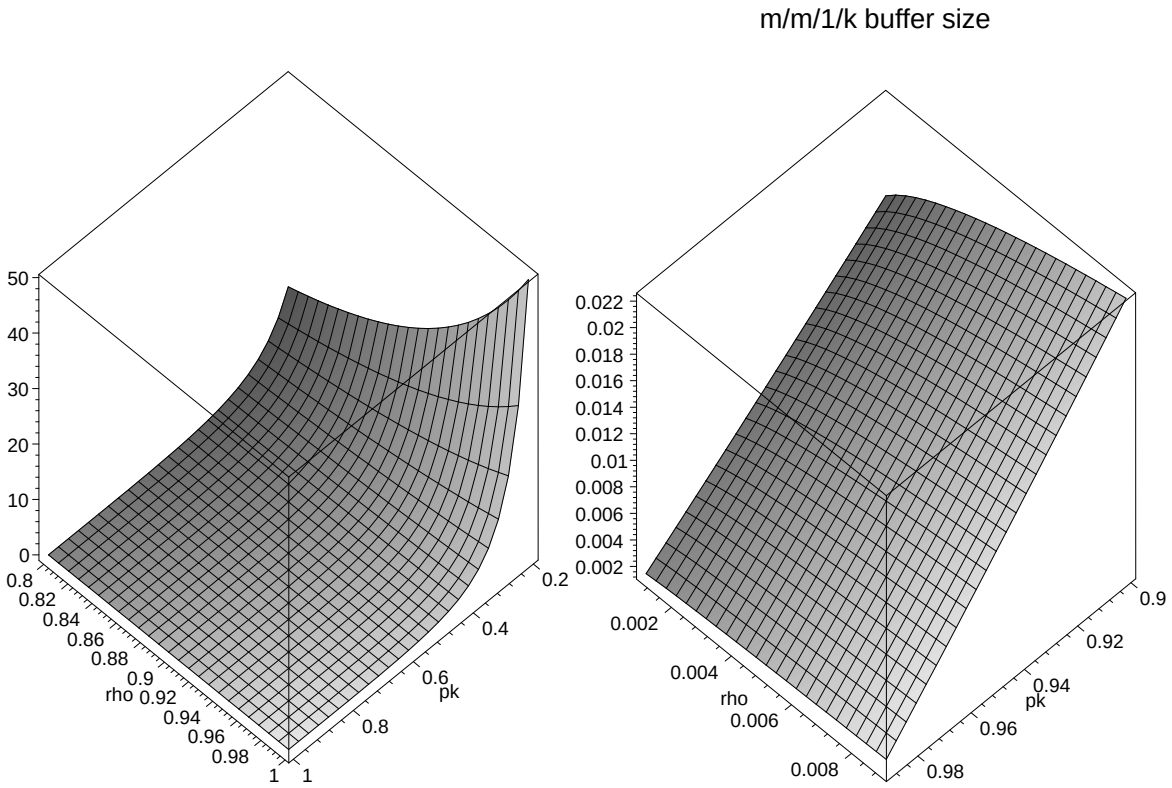

Figure 4: Hessian term H_{22} on left Non-convex region on right

Since the first term in the Hessian is positive everywhere, the Hessian will be positive definite if the magnitude of the product of the diagonal terms is strictly larger than the magnitude of the product of the off-diagonal terms, *i.e.*

$$H_{11} * H_{22} > H_{12} * H_{21}$$

For most of the values of the function, except for those values where this is not true in the range of low ρ and high p_k, the function is convex. Examination of the eigenvalues of the Hessian matrix also supports the positive definite nature of H in the restricted range.

For an example in the region where the function is not convex, Figure4 on the right suffices to indicate the non-convex nature of the function. Note the low vales of ρ and high values of p_k which makes little sense in a practical situation, but could occur. It may be argued, that where the buffer size is below $k = 1$ is of no importance, yet there may be certain applications where a buffer size $k < 1$ may be meaningful, so in this range, the function is not convex, in fact H it is not even positive-semidefinte.

3.4 M/M/C/K

Now, let's examine a multi-server Markovian system. We shall limit ourselves to those systems where $(\lambda/c\mu \neq 1)$.

3.5 General Formula for blocking probability

Analytical results from the $M/M/C/K$ model provide the following expression for p_k.

$$p_k = \frac{1}{c^{k-c}c!}\left(\frac{\lambda}{\mu}\right)^k p_0$$

where for $(\lambda/c\mu \neq 1)$

$$p_0 = \left[\sum_{n=0}^{c-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{(\lambda/\mu)^c}{c!}\frac{1-(\lambda/c\mu)^{k-c+1}}{1-\lambda/c\mu}\right]^{-1}$$

M/M/C/K composite formula for the blocking probability:

$$\frac{\left(\frac{\lambda}{\mu}\right)^k (c!)^{-1}\left(c^{k-c}\right)^{-1}}{\left(e^{\frac{\lambda}{\mu}}\Gamma(c,\frac{\lambda}{\mu})\left(\Gamma(c)\right)^{-1} + \frac{\left(\frac{\lambda}{\mu}\right)^c\left(1-\left(\frac{\lambda}{c\mu}\right)^{k-c+1}\right)}{(c!)^{-1}\left(1-\frac{\lambda}{c\mu}\right)^{-1}}\right)} = p_k$$

Since the blocking probability function is very complex, one cannot express the value of k without fixing C. If one fixes $C = 1$, the function becomes the same as the one which was derived from the previous formula for the M/M/1/K system.

3.6 C=2

If one fixes C=2, the following closed form expression occurs:

$$k = -\ln\left(1/2\,\frac{p_k\,(2\,\mu+\lambda)}{2\,\mu-\lambda+p_k\,\lambda}\right)\left(\ln(2\,\frac{\mu}{\lambda})\right)^{-1}$$

Figure 5 illustrates the convex nature of the C=2 expression and also the significantly reduced

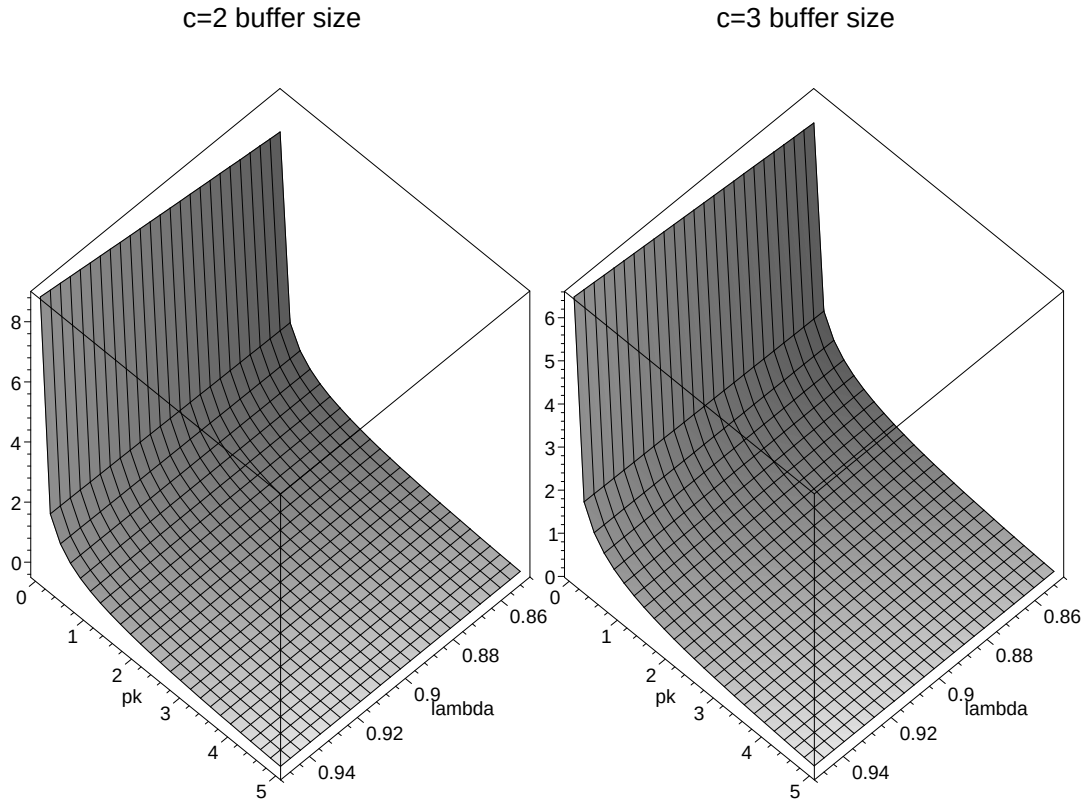

Figure 5: C=2,C=3 convex closed form expressions

buffer size as a result of the increased number of servers in the system over the M/M/1/K system. A similar analysis of the Hessian as was done for the M/M/1/K case of this expression could also be carried out to show the convexity of the C=2 function.

3.7 C=3

If one fixes C=3, the following closed form expression occurs:

$$k = -\ln\left(1/9\,\frac{p_k\,(6\,\mu^2+4\,\lambda\,\mu+\lambda^2)}{\mu\,(3\,\mu-\lambda+p_k\,\lambda)}\right)\left(\ln(3\,\frac{\mu}{\lambda})\right)^{-1}$$

Figure 5 also illustrates the convex nature of the C=3 expression and the marginally reduced buffer size as a result of the increased number of servers in the system. Obviously, we could continue this process, but for the sake of the argument will not do so at the present stage.

4 Non-Markovian Systems

In principle, for the M/G/1/K system, one could develop the blocking probability for fixed k and all the other probabilities, but this would be tedious and not as useful when it comes to modelling the design of queueing networks with varying buffer sizes.

4.1 Different approximations

There are numerous approximations for the blocking probability possible for M/G/1/K systems, One survey article by Springer and Makens [10] analyzes in some detail five approximation formula and concludes that the formula by Gelenbe is the most accurate and robust. In the following analysis, we shall restrict ourselves to this approximation formula, and as in the previous analysis in this paper, examine the convex nature of the blocking probability as provided by Gelenbe's formula. There are other approximations [11] that could be analyzed, and this may be pursued for the expanded part of the paper.

4.2 Gelenbe's Formula

Gelenbe's formula is based on approximating the discrete queueing process as a continuous diffusion process. The blocking probability from Gelenbe's equation with squared arrival and service process coefficient of variation is given by the following expression [1] where a^2 and s^2 are respectively the squared coefficient of variations of the arrival and service processes:

$$\frac{\lambda\,(1-\lambda)\,e^{2\,\frac{(\lambda-1)(k-1)}{\lambda\,a^2+s^2}}}{\left(1-\lambda^2 e^{2\,\frac{(\lambda-1)(k-1)}{\lambda\,a^2+s^2}}\right)} = p_k$$

The closed-form expression deriveable for k from Gelenbe's formula is the following:

$$k = 1/2\,\frac{2\,\lambda-2\,\mu+\ln(\frac{p\,k\,\mu^2}{\lambda\,(-\lambda+\mu+p\,k\,\lambda)})\lambda\,a^2+\ln(\frac{p\,k\,\mu^2}{\lambda\,(-\lambda+\mu+p\,k\,\lambda)})\mu\,s^2}{(\lambda-\mu)}$$

If we assume we have a Markovian system, then $a^2 = s^2 = 1$ and we get a function whose graph is illustrated on the left in Figure 6 and if we set $s^2 = 0$ we get a convex function that is indicated on the right in Figure 6. The Markovian approximation is very accurate, while the M/D/1/K is not as accurate an approximation, even though the order of magnitude reduction in the buffer size which is roughly 2 is quite accurate. This assessment of Gelenbe's formula essentially agrees with the assessment of Springer.

The convex nature of the M/G/1/K approximation is an important property that should be shared by all these approximations. In a future paper, the study of other M/G/1/K approximations will be carried out along with their use in a BAP program analysis.

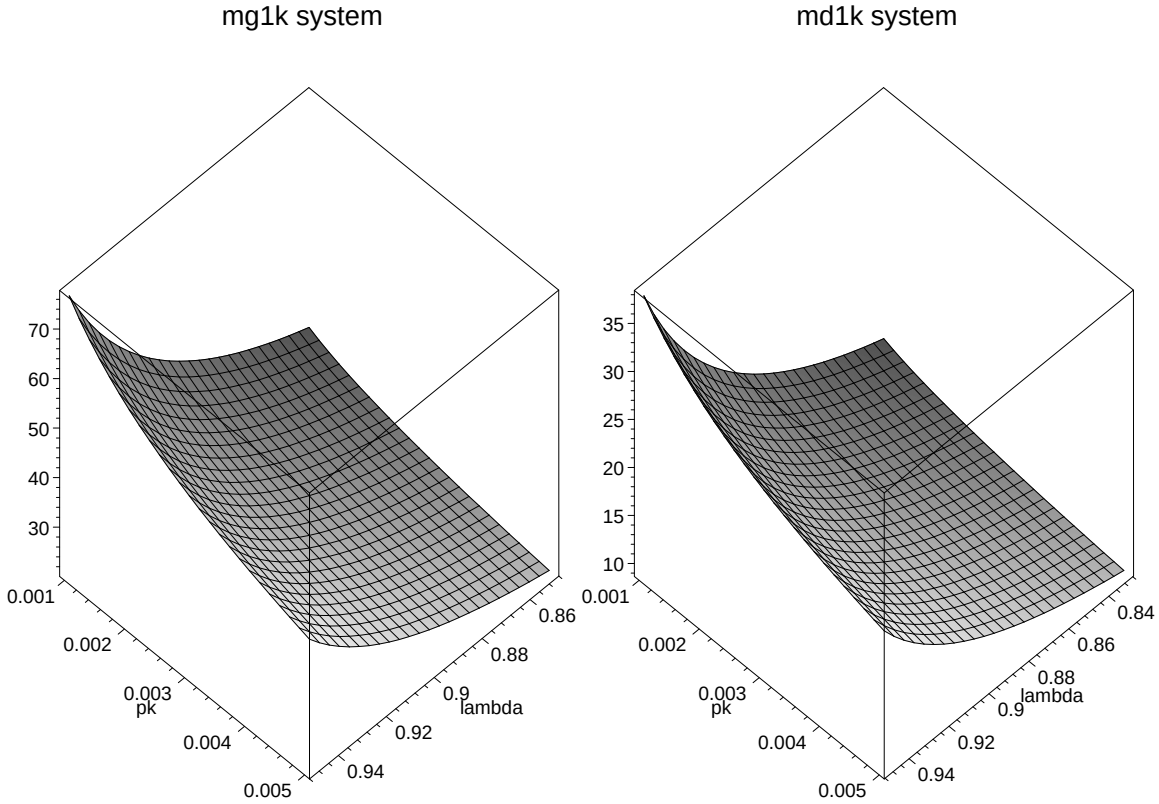

Figure 6: Gelenbe's M/G/1/K and M/D/1/K Closed-Form Expressions

5 BAP Optimization

The optimization problem with $M/M/1/K$ and $M/M/C/K$ systems that will be exanined here is essentially the following (shown for the $M/M/1/K$ system:

$$Minimize\ Z = \sum_{j=1}^{N} \left\lceil \frac{\ln(\frac{P_k^j}{1-\rho^j+P_k^j\rho^j})}{\ln(\rho^j)} \right\rceil$$

s.t.

$$\Theta \geq \Theta^\tau$$

$$k_j \geq 0 \forall j$$

where Θ^τ is a threshold throughput. Θ^τ can be prespecified and then serve as the input Λ to an approximate performance measure program such as the Expansion method program [5]. In the particular formulation of the problem the $k_j \forall j$ become the decision variables under optimization control. While these are essentially integer variables, they can be reasonably approximated by round off from the nonlinear programming solver.

One way to incorporate the throughput constraint is through a penalty function approach. The penalty function optimization problem is essentially the following:

$$Minimize\ Z = \sum_{j=1}^{N} \left\lceil \frac{\ln(\frac{P_k^j}{1-\rho^j+P_k^j\rho^j})}{\ln(\rho^j)} \right\rceil - \mu(\Theta - \Theta^\tau)^2$$

$$s.t.$$

$$k_j \geq 0\ \forall j, \mu \geq 0$$

The nice thing about this model is the Expansion Method can deliver the estimate of the $\Theta$and we can use the penalty function as a variable parameter to solve for the optimal k_j values.

In the remainder of the present paper, the computational results of employing the closed form expressions of M/M/1/K and M/M/C/K systems for the (BAP) in series, merge, and split topologies of finite buffer queueing networks will be presented.

Acknowledgements: I want to thank Peter Kubat for reading over the manuscript and Tayfur ALtiok for a discussion of ideas in this paper.

References

[1] Gelenbe, E., 1975. "On approximate Computer System Models," *JACM* **22** (2), 261-269.

[2] Harel, A. and P. Zipkin, 1987. "Strong Convexity Results for Queueing Systems," *Operations Research.* **35**(3), 405-418.

[3] Harel, A., 1988. "Sharp Bounds and Simple Approximations for the Erlang Delay and Loss Formulas," *Management Science,* **34**(8), 959-972.

[4] Harel, A., 1990. "The Convexity of the Erlang Delay Formula with Respect to the Number of Servers," *Working Paper #90-24.* Graduate School of Management, Rutgers, November 1990.

[5] Kerbache, L. and J. MacGregor Smith, 1988. "Asymptotic Behaviour of the Expansion Method for Open Finite Queueing Networks." *Computers and Operations Research,* **15** (2), 157-169.

[6] Kimura, T., 1996."A Transform-Free Approximation for the Finite Capacity M/G/s Queue,"*Operations Research.* **44**(6), 984-988.

[7] Kimura, T., 1996. "Optimal Buffer Design of an M/G/s Queue with Finite Capacity," *Commun. Statist.-Stocahstic Models,* **12**(1), 165-180.

[8] Singh, Ajay and J. MacGregor Smith 1997, "Buffer Allocation for an Integer Nonlinear Network Design Problem,"*Computers and Operations Research* **24** (5), 453-472.

[9] Smith, J. MacGregor and S. Daskalaki. "Buffer Space Allocation in Automated Assembly Lines." *Operations Research* **36***(2)*,343-358.

[10] Springer, M. and P. Makens, 1991. "Queueing Models for performance analysis and Selection of Single Station Models," *EJOR***58**, 123-145.

[11] Yao, D. and Buzacott, J.A., 1985. "Queueing Models for a Flexible Machining Station Part 1: The Diffusion Approximation." *EJOR***19**, 233-240.