SERVICE CONTROL IN A FINITE BUFFERED QUEUE WITH MULTIPLE SERVICE RATES

Fikri Karaesmen[†]

Florin Avram[††]

and

Surendra M. Gupta[*,†]

(* Corresponding author)

† Department of Mechanical, Industrial and Manufacturing Engineering
334 Snell Engineering Center
Northeastern University
360 Huntington Avenue
Boston, Massachusetts 02115, USA
(617) 373-4846 (Phone)
(617) 373-2921 (Fax)

†† Department of Mathematics
Northeastern University
Boston, MA 02115

March 1997

SERVICE CONTROL IN A FINITE BUFFERED QUEUE WITH MULTIPLE SERVICE RATES

Fikri Karaesmen, Florin Avram and Surendra M. Gupta
Northeastern University

Abstract

We consider a finite buffered queue in which the queue length is controlled by dynamically selecting between two possible service rates (low and high). Using the faster service rate requires higher operating costs which may justify using the slower rate from time to time. Moreover, the system incurs holding costs for customers waiting to be processed and setup costs for each service rate change. When both service rates are close to the arrival rate, a heavy traffic (diffusion) approximation is valid for the expected queue length for this system. Furthermore, the approximating dynamic control problem has an explicit solution for a special case of the parameters which leads to an approximate solution for the general case. We present numerical examples that analyze the validity of the heavy traffic approach.

1 Introduction

Consider a finite buffered queue with two possible service rates (fast and slow). Using the faster server has the obvious advantage of keeping the queue lengths shorter. However, it may be less costly to use the slower server. This motivates a dynamic optimization problem of selecting between the two servers. In particular, we are interested in a case of this problem in which setup costs are incurred each time the service rate is switched.

Service control problems for queueing systems have been studied extensively for many particular cases. An important class of problems was introduced by Yadin and Naor [24] where the server may be controlled by turning it off. Yadin and Naor studied the queue lengths under an operating policy termed *N-policy.* Under this operating policy, the server is turned off when the queue length is zero and turned back on when the queue length reaches a threshold level N. Later, Heyman [11] and Sobel [21] proved that N-policy is the optimal operating policy for a variety of dynamic control problems.

For the problem with two positive service rates (viz. a fast server and a slow server), it is known that a single threshold type policy is optimal in the absence of setup costs. Using a threshold policy, it is optimal to switch to the faster server when the queue length reaches the threshold and switch back to the slower server when the queue length falls below the threshold. In particular, for the problem of selecting the optimal service rate in the case of

Poisson arrivals and exponential service times, Crabill [5] showed that the optimal service rate must be increasing in the queue length. On the other hand, when setup costs are taken into consideration it is likely that the optimal policy is of double band type, similar to the well known (s, S) inventory control policies. For our problem, in an (s, S) policy, the faster server is started when the queue length reaches S and is used until the queue length falls to s. Similarly, the slower server is used from the time the queue length falls to s until the queue length reaches S. Gebhard [7] considered a problem with two exponential servers and a Poisson arrival process and obtained the queue length distribution for a given (s, S) policy. Tijms [23] used the framework of Markov decision processes and proposed a policy iteration algorithm to compute the optimal values of the parameters (s, S). Federgruen and Tijms [6] studied the M/G/1 queue with two different service time distributions and gave a recursive algorithm to compute the performance of a given (s, S) policy.

The case of finite buffers has been given less attention than the infinite buffer counterpart. Hersh and Brosh, [10] studied the M/M/1/K queue with a removable server and showed that N-policy is still optimal if holding costs are negligible. Their results were later extended to the case of general service time distributions by Teghem [22]. Karaesmen and Gupta [13] studied the effect of non-negligible holding costs for the M/G/1/K case and reported an efficient computational procedure to compute the performance of a given (s, S) policy for removable servers. For the problem with two positive service rates, Neuts and Rao [16] employed matrix-geometric techniques to compute the performance of a given (s, S) policy for the M/G/1/K queue (with phase type service time distributions.) Gupta [8], studied both service and arrival control problems and showed that performance analysis for both problems are closely related.

The use of diffusion processes for storage/queueing systems in the operations research literature was pioneered by Bather [2], [3]. As well as being of interest in themselves, diffusion processes also arise as heavy traffic approximations of queueing systems. These approximations are particularly attractive as they are supported by formal limit theorems (see Iglehart and Whitt [12]). More recently, diffusion approximations of queueing models have been succesfully used to solve difficult dynamic optimization problems. Utilizing this approach, dynamic decision problems for queues are replaced by stochastic control problems for diffusion processes which may yield explicit or approximate solutions.

Rath [18] considered the service rate control problem for a queueing system. He showed that when both service rates are close to the arrival rate (i.e. in heavy traffic), the queue length process for the system converges to Brownian motion processes with two different sets of drift and variance parameters corresponding to each server. In addition to the convergence of queue lengths, it was shown that the costs in the queueing problem also converge to the costs of the Brownian motion control problem. Later, Rath [19] showed that the optimal policy for the above diffusion control problem is an (s, S) policy by using the corresponding controlled random walk and proving that the costs of the random walk converges to those of the diffusion process. Chernoff and Petkau [4] treated the same problem as a stochastic control problem and reproduced Rath's results without using the convergence from a random walk. Perry and Bar-Lev [17] considered the inventory control version of the same problem. In their case the contents of a bounded storage system is controlled by changing the drift of a diffusion process. In addition to holding and setup costs, Perry and Bar-Lev's model also includes penalties for hitting the boundaries. They compute the performance of this system for infinite horizon discounted costs using renewal arguments.

In a recent paper, Avram and Karaesmen [1] introduced a new method to solve the above drift control problems for the (long run) average cost optimization case. The method reveals the relationship between the problem without setup costs and the one with setup costs and is computationally very efficient.

In this paper, we analyze the validity of diffusion approximations as alternative performance analysis and optimization tools for a finite buffered queue with two different sets of arrival/service rates. Diffusion approximations have the attractive property that they capture the means and variances of the service and arrival processes which are critical for the performance of a system. This enables the approximate analysis of GI/G/1 type queues for which no efficient analytical techniques exist.

The arrangement of the paper is as follows. In section 2, we introduce some of the notations used and give a formal definition of the problem. Section 3 considers the Markovian case in detail. In section 4, we discuss diffusion approximations to compute the average queue length and the expected cost for a given (s, S) policy. We also summarize the method in [1] and describe how it can be used to compute the optimal average cost and the optimal values of the parameters (s, S). In section 5, we give numerical examples to demonstrate the behaviour of the proposed methodology for different cost and traffic parameters. Section 6 includes the conclusions and future research.

2 Definitions

We study a queueing system with two different sets of arrival and service time distributions. We denote by A_1 and B_1 the random variables denoting the times between arrival and service times respectively for the first set and the analogous random variables A_2 and B_2 for the second set. We will use λ_i and μ_i for the average arrival and service rates respectively when the ith $(i = 1, 2)$ set of distributions is used. Holding costs are incurred at rate h dollars per customer and operating the queue costs c_1 and c_2 dollars using the set of distributions 1 and 2 respectively. Switching from the first set to the second set costs k_1 dollars per switch and switching from the second set to the first set costs k_2 dollars per switch. Finally we let the total switching (or setup) cost, $k = k_1 + k_2$.

In order to compute the long run average cost, we denote the expected queue length for a given (s, S) policy by $L(s, S)$. We also denote by $\pi(i, s, S)$ the proportion of time the ith set of parameters is used. Finally, we denote by $\mathcal{K}(t)$ the total switching cost incurred until time t. Then the long run average cost per unit time, $\gamma(s, S)$, is:

$$\gamma(s, S) = hL(s, S) + c_1\pi(1, s, S) + c_2\pi(2, s, S) + \lim_{t->\infty} \frac{\mathcal{K}(t)}{t} \tag{1}$$

Our goal is to obtain the optimal cost γ^*, and the corresponding (s^*, S^*) pair.

3 The Markovian Case

In this section, we concentrate on the special case of Poisson arrival distributions and exponential service times. It will be seen that in this case, the computation of the performance of measures for given threshold levels (s, S) is straightforward. Thus, the results of this section are useful benchmarks for comparison purposes.

3.1 The Stationary Queue Length Distribution

Consider the finite (K) capacity Markovian queueing system with an arrival rate λ and two service rates μ_1 and μ_2 (where $\mu_1 < \mu_2$). An (s, S) policy is used to control the operation of the servers. Hence the service rate is switched to μ_1 whenever rate μ_2 is used and the queue length drops to s and is switched back to μ_2 when the queue length builds up to S while rate μ_1 is used.

In the next lemma, we obtain the stationary distribution of the service control problem described above in closed form.

Lemma 1 Consider the system subject to service rate control described above. If, $\rho_1 =$

λ/μ_1 and $\rho_2 = \lambda/\mu_2$, then:

$$\begin{array}{rcll} \pi_0^K & = & \kappa \left[\frac{1}{1-\rho_1} - \frac{(S-s)\rho_1^{S-1}(\rho_1-\rho_2)}{(1-\rho_1^{S-s})(1-\rho_2)} \right]^{-1} & \\ \pi_i^K & = & \rho_1^i \pi_0^K & \text{for } \; i = 1, 2, .., s-1 \\ \pi_i^K & = & \frac{\rho_1^s(1-\rho_1)}{(1-\rho_1^{(S-s)}} \pi_0^K \left[\frac{(\rho_1^{(i-s)} - \rho_1^{(S-s)}}{(1-\rho_1)} + \frac{\rho_1^{(S-s-1)}\rho_2(1-\rho_2^{(i-s)})}{(1-\rho_2)} \right] & \text{for } \; i = s, s+1, ..., S-1 \\ \pi_i^K & = & \frac{(1-\rho_2^{S-s})(1-\rho_1)\rho_1^{S-1}\rho_2^{i-S+1}}{(1-\rho_1^{S-s})(1-\rho_2)} \pi_0^K & \text{for } \; i = S, S+1, ...K \end{array} \tag{2}$$

where

$$\kappa = \frac{1}{1-\alpha(K+1)} \tag{3}$$

and

$$\alpha(K+1) = \frac{(1-\rho_2^{S-s})(1-\rho_1)\rho_1^{S-1}}{(1-\rho_1^{S-s})(1-\rho_2)} \left[\frac{1}{1-\rho_1} - \frac{(S-s)\rho_1^{S-1}(\rho_1-\rho_2)}{(1-\rho_1^{S-s})(1-\rho_2)} \right]^{-1} \frac{\rho_2^{2+K-S}}{1-\rho_2} \tag{4}$$

Proof: Let π_i^∞ denote the stationary probability of having i customers in the identical infinite buffered queue. The underlying Markov Chain for this system is reversible (see Kelly [15]) when $K \geq S$. Reversibility implies that the stationary distribution under a state space restriction is proportional to the stationary distribution for the unrestricted state space case. Hence:

$$\pi_i^K = \kappa \pi_i^\infty \quad \text{for } \; i = 0, 1, 2, ..., K \tag{5}$$

where κ is a normalizing constant given by:

$$\kappa = 1/(\sum_{i=0}^{K} \pi_i^\infty) \tag{6}$$

To obtain equation (3), we use the tail probability:

$$\alpha(K+1) = \sum_{i=K+1}^{\infty} \pi_i^\infty \tag{7}$$

along with the identity:

$$\sum_{i=0}^{K} \pi_i^\infty = 1 - \alpha(K+1) \tag{8}$$

Therefore, we are able to describe π_i^K's completely in terms of π_i^∞'s. Using the stationary distribution, π_i^∞ obtained in Gebhard [7], we immediately get equations (2) and (4).
□

Remarks: 1. The expression in (2) for the states from s to $S-1$ is slightly different than that in [7] which has a typographical error.
2. The stationary distribution in (2) is a function of ρ_1 and ρ_2 and not the service rates μ_1 and μ_2. This implies that lemma 1 also covers the case where both service and arrival rates vary.

3.2 Computation of the Cost Function

The cost function (1) for our problem consists of three components, viz., the holding cost, the operating cost and the setup cost. We have already dealt with the computation of the holding cost implicitly as we have obtained the stationary queue length distribution. The expected holding cost per unit time is:

$$hL = h\sum_{i=1}^{K} i\pi_i^K \tag{9}$$

Next, we obtain the operating and setup costs per unit time in closed form.

Lemma 2 The operating cost per unit time in the system subject to service rate controlled system descrived in Section 3.1 is:

$$(c_1 - c_2)\left[\kappa\left(1 - \frac{(S-s)\rho_1^{S-1}\rho_2(1-\rho_1)\pi_0^\infty}{\kappa(1-\rho_1^{S-s})(1-\rho_2)}\right)\right] + c_2 \tag{10}$$

Proof: Gebhard [7] gives the probability of using the high rate server in the infinite capacity queue as:

$$\frac{(S-s)\rho_1^{S-1}\rho_2(1-\rho_1)\pi_0^\infty}{(1-\rho_1^{S-s})(1-\rho_2)} \tag{11}$$

Thus, the proportion of time the low rate server is used (in the infinite capacity queue) is:

$$1 - \frac{(S-s)\rho_1^{S-1}\rho_2(1-\rho_1)\pi_0^\infty}{(1-\rho_1^{S-s})(1-\rho_2)} \tag{12}$$

which can also be written as:

$$\sum_{i=0}^{S-1} \pi_{i,1}^\infty. \tag{13}$$

By reversibility, the long run proportion of time that the low rate server is used in the corresponding finite capacity queue is:

$$
\begin{aligned}
\sum_{i=0}^{S-1} \pi_{i,1}^{K} &= \sum_{i=0}^{S-1} \kappa \pi_{i,1}^{\infty} && (14) \\
&= \kappa \sum_{i=0}^{S-1} \pi_{i,1}^{\infty} && (15) \\
&= \pi(1, s, S) && (16)
\end{aligned}
$$

where $\pi(1, s, S)$ was defined in Section 2. Noting that $\pi(2, s, S) = 1 - \pi(1, s, S)$, the operating cost per unit time of the queue subject to service control given in (10) follows.
□

Lemma 3 The setup cost per unit time in the system subject to service rate control described in Section 3.1 is:

$$k\lambda_1 \frac{(\rho_1^{S-1}(1-\rho_1)\pi_0^{\infty})}{(1-\rho_1^{S-s})} \quad (17)$$

Proof: The setup rate is the rate at which the process jumps from the state of low rate to the state of high rate or vice versa. Using the probabilities given in [7] for the infinite buffer case and by reversibility, the result in (17) is obtained.
□

As a consequence of lemmas 2 and 3, we can compute the performance of the M/M/1/K queue with two different traffic rates when (s, S) are given. However, computing the optimal (s, S) values still requires a search over the parameter space.

4 Diffusion Approximations

4.1 Diffusion Approximations for the GI/G/1 Queue

Performance analysis of queueing systems for arbitrary service and arrival time distributions usually defies exact analysis. Among the approximation methods, diffusion (heavy traffic) approximations are particularly attractive as their (limiting) convergence can be proved. Furthermore, diffusion approximations replace discrete state space queueing processes by continuous state space processes for which many analytical tools are available.

In this section, we present the performance evaluation tools obtained through diffusion approximations. Consider a GI/G/1 queue where customers arrive according to a renewal process with mean rate λ and the service times have a general distribution with mean $1/\mu$. We denote by $\mathrm{Var}(A)$ and $\mathrm{Var}(B)$, the variances of the interarrival times and service times respectively. As usual, $\rho = \lambda/\mu$ denotes the traffic intensity. To obtain the approximation, one needs a sequence of the above GI/G/1 queues indexed by a parameter n, called a heavy traffic scaling parameter. The queueing process converges to the diffusion process as $n \longrightarrow \infty$ and $\rho \longrightarrow 1$ under a certain scaling of time and space. In particular, let $L(t)$ denote the queue length at time t. The approximation is valid for the scaled queue length process, $Z(t)$, defined by:

$$Z(t) = \frac{L(nt)}{\sqrt{n}} \tag{18}$$

Although, the selection of the scaling parameter n appears to be critical, for performance analysis, a large integer value is sufficient. In the numerical examples, it will suffice to fix n to a value of 1000.

Iglehart and Whitt [12] have shown that (in the limit) the scaled queue length, $Z(t)$, is a reflected (or regulated) Brownian motion with drift coefficient -ϵ (where $\epsilon = \sqrt{n}(1-\rho)$) and variance, σ^2, given by:

$$\sigma^2 = \lambda^3 \mathrm{Var}(A) + \mu^3 \mathrm{Var}(B) \tag{19}$$

One can then evaluate the performance of the queueing system by evaluating the corresponding performance measures for the above regulated Brownian motion and going back to the original scale using (18).

For the GI/G/1 queue (with infinite buffers), one can write $Z(t)$ in terms of a standard Brownian motion $X(t)$ (with identical drift and variance) and another process, i.e:

$$Z(t) = X(t) + \mathcal{L}(t) \tag{20}$$

with the convention that $X(0) = 0$. $\mathcal{L}$ is a process with the following properties:
1. $\mathcal{L}$ is increasing and continuous with $\mathcal{L}(0) = 0$
2. $\mathcal{L}$ increases only when $Z = 0$
(This property is due to the regulation principle, see Harrison [9] for more details).

Finite buffer capacity of the queueing system can also be taken into account in the approximation. For the buffer capacity of K, we can denote the scaled buffer capacity by $\hat{K}$ $(= K/\sqrt{n})$. Then the scaled queue length, $Z(t)$, is regulated at $\hat{K}$ in addition to 0, i.e.:

$$Z(t) = X(t) + \mathcal{L}(t) - \mathcal{U}(t) \tag{21}$$

where $X(0) = 0$, $\mathcal{L}$ is the lower regulating process described above and $\mathcal{U}$ has the following properties:

1. $\mathcal{U}$ is increasing and continuous with $\mathcal{U}(0) = 0$
2. $\mathcal{U}$ increases only when $Z = \hat{K}$

Hence, most performance analysis questions for GI/G/1 queues in heavy traffic reduce to the performance analysis questions of one-dimensional regulated Brownian motion, a well studied process.

4.2 Expected Queue Lengths

Consider a queueing system with two sets of traffic parameters (A_i, B_i), $i = 1, 2$ where A_i and B_i are random variables denoting the interarrival and service times of set i. We also denote the mean arrival and service rates by (λ_i, μ_i) $(i = 1, 2)$. Parameter set 1 is employed from the time the queue length falls to s until the time it increases to S. Similarly, parameter set 2 is used from the time the queue length increases to S until the time it falls to s. Let $L(t)$ be the queue length process for this system. Rath [18], proves the following result:

Theorem 1 Under heavy traffic conditions (i.e. $\lambda_1 \approx \mu_1$ and $\lambda_2 \approx \mu_2$) and infinite buffer capacity, the scaled version of the queue length process $Z(t)$ converges to a controlled diffusion process that moves according to two Brownian motions with parameter sets $(\varepsilon_i, \sigma_i)$, $i = 1, 2$ (with drift ε_i and variance σ_i for set i) with a lower barrier (regulation) at zero. Thus,

$$\varepsilon_i = -\sqrt{n}(1 - \rho_i) \quad \text{for} \quad i = 1, 2 \tag{22}$$

for a large integer n and

$$\sigma_i^2 = \lambda_i^3 \mathrm{Var}(A_i) + \mu_i^3 \mathrm{Var}(B_i) \quad \text{for} \quad i = 1, 2 \tag{23}$$

The selection of the parameter set is performed dynamically according to a $(\hat{s}, \hat{S})$ policy where $\hat{s} = s/\sqrt{n}$ and $\hat{S} = S/\sqrt{n}$ are the scaled versions of the thresholds in the queueing system. Hence, $(\varepsilon_1, \sigma_1)$ will be used in the region $[0, \hat{S})$ and $(\varepsilon_2, \sigma_2)$ will be used in the region $[\hat{s}, \infty)$.

Note that, if the buffer capacity is finite (K), the process $Z(t)$ will have an upper barrier at $\hat{K}$ $(= K/\sqrt{n})$. Next, we show that, the expected value of the process $Z(t)$ can be computed in closed form.

Theorem 2 Let $\mathrm{E}[Z]$) be the expected value of the process $Z(t)$, then:

$$\mathrm{E}[Z] = \frac{f_1(\hat{s}) + f_2(\hat{S})}{t_1(\hat{s}) + t_2(\hat{S})} \tag{24}$$

where

$$f_1(\hat{s}) = \frac{\hat{S}^2 - \hat{s}^2}{2\mu_1} + \frac{\sigma_1^2(\hat{s} - \hat{S})}{2\varepsilon_1^2} + \frac{\sigma_1^4(e^{-2\varepsilon_1 \hat{s}/\sigma_1^2} - e^{-2\varepsilon_1 \hat{S}/\sigma_1^2})}{4\varepsilon_1^3} \tag{25}$$

$$t_1(\hat{s}) = \frac{\hat{S} - \hat{s}}{\varepsilon_1} + \frac{\sigma_1^2(e^{-2\varepsilon_1 \hat{S}/\sigma_1^2} - e^{-2\varepsilon_1 \hat{s}/\sigma_1^2})}{2\varepsilon_1^2} \tag{26}$$

$$\begin{aligned} f_2(\hat{S}) \quad = \quad & \frac{-\hat{S}^2}{2\varepsilon_2} + \frac{\sigma_2^2 \hat{S} + \hat{s}(\varepsilon_2 \hat{s} - \sigma_2^2)}{2\varepsilon_2^2} \\ & + \frac{\sigma_2^2(2\hat{K}\varepsilon_2 + 2\varepsilon_2^2 - \sigma_2^2)(e^{2\varepsilon_2(\hat{K}-\hat{s})/\sigma_2^2} - e^{2\varepsilon_2(\hat{K}-\hat{S})/\sigma_2^2})}{4\varepsilon_2^3} \end{aligned} \tag{27}$$

and

$$t_2(\hat{S}) = \frac{-\hat{S}}{\varepsilon_2} + \frac{2\varepsilon_2 \hat{s} + \sigma_2^2(e^{2\varepsilon_2(\hat{K}-\hat{s})/\sigma_2^2} - e^{2\varepsilon_2(\hat{K}-\hat{S})/\sigma_2^2})}{2\varepsilon_2^2} \tag{28}$$

Proof: By renewal theory, we can consider a regenerative cycle of the process that starts at the time parameter set 1 is initiated and ends the next time the buffer level drops to $\hat{s}$ when parameter set 2 is in use. Let $T_{\hat{S}}$ denote the first time after the start of the cycle the process reaches $\hat{S}$. Similarly, let $T_{\hat{s}}$ denote the time process reaches $\hat{s}$ starting from $\hat{S}$. Obviously, the cycle time is $T_{\hat{S}} + T_{\hat{s}}$. Furthermore, let

$$f_1(x) = \mathrm{E}\left[\int_0^{T_{\hat{S}}} Z(z)dz | Z(0) = x\right] \tag{29}$$

and

$$f_2(x) = \mathrm{E}\left[\int_0^{T_{\hat{s}}} Z(z)dz | Z(0) = x\right] \tag{30}$$

Hence, f_1 and f_2 denote the expected total inventory levels when the parameter sets 1 and 2 are used in a cycle, respectively. By renewal theory, the expected value of the inventory level in a cycle is the ratio of the expected total inventory level in a cycle to the expected duration of a cycle, i.e:

$$\mathrm{E}[Z] = \frac{f_1(\hat{s}) + f_2(\hat{S})}{\mathrm{E}[T_{\hat{s}}] + \mathrm{E}[T_{\hat{S}}]} \tag{31}$$

letting $t_1(\hat{s}) = \mathrm{E}[T_{\hat{S}}]$ and $t_2(\hat{S}) = \mathrm{E}[T_{\hat{s}}]$, we obtain (24). Furthermore, f_1, f_2, t_1 and t_2 are solutions of the systems of ordinary differential equations by Ito's lemma given below (see Karlin and Taylor [14]). Note that the boundary conditions $f_1'(0) = f_2'(\hat{K}) = t_1'(0) = t_2'(\hat{K}) = 0$ correspond to regulation at the boundaries 0 and $\hat{K}$ (see Harrison [9]).

$$\left\{ \begin{array}{ccc} \frac{\sigma_1^2 f_1''(x)}{2} + \varepsilon_1 f_1'(x) & = & -x \\ f_1'(0) & = & 0 \\ f_1(\hat{S}) & = & 0 \end{array} \right. \tag{32}$$

$$\begin{cases} \frac{\sigma_1^2 t_1''(x)}{2} + \varepsilon_1 t_1'(x) & = & -1 \\ t_1'(0) & = & 0 \\ t_1(\hat{S}) & = & 0 \end{cases} \tag{33}$$

$$\begin{cases} \frac{\sigma_2^2 f_2''(x)}{2} + \varepsilon_2 f_2'(x) & = & -x \\ f_2'(\hat{K}) & = & 0 \\ f_2(\hat{s}) & = & 0 \end{cases} \tag{34}$$

$$\begin{cases} \frac{\sigma_2^2 t_2''(x)}{2} + \varepsilon_2 t_2'(x) & = & -x \\ t_2'(\hat{K}) & = & 0 \\ t_2(\hat{s}) & = & 0 \end{cases} \tag{35}$$

□

4.3 Expected Performance of a Given Policy

An advantage of replacing the discrete queueing processes by diffusion processes is in computing cost functionals associated with the control of the original process. By Rath's results [18], it is known that the cost functionals defined in section 2 converge to the corresponding cost functions for the diffusion process.

As mentioned before, the diffusion problem is obtained from the queueing problem by a rescaling of time and space. For example, the holding cost of the system is proportional to the average buffer level which is L in the queueing system but $\hat{L} = L/\sqrt{n}$ in the diffusion system. To solve the original optimization problem, costs have to be on the same scale in the rescaled problem. To achieve this, we scale the operating cost such that $\hat{c_1} = c_1/\sqrt{n}$, $\hat{c_2} = c_2/\sqrt{n}$ and the (total) setup cost such that $\hat{k} = k/n$ but leave the holding cost, h, unchanged (see [20] for a similar argument). If boundary costs exist in the original queueing sytem (i.e. α_0 dollars for unit idle time and α_K dollars for the time the buffer is full), the approximating diffusion system will have the rescaled boundary costs $\hat{\alpha_0} = \alpha_0/\sqrt{n}$ and $\hat{\alpha_K} = \alpha_K/\sqrt{n}$.

With the rescaled costs, we can compute the performance of a given (s, S) policy for the queueing system as showm in the following corollary:

Corollary 3 Let $\hat{k}$ denote the total setup cost in a cycle (i.e. $k = k_1 + k_2$) and let $\hat{\alpha_0}$ and $\alpha_{\hat{K}}$ denote the boundary costs at the lower and upper boundaries respectively. Then, the optimal cost of an $(\hat{s}, \hat{S})$ policy, $\gamma(\hat{s}, \hat{S})$, is given by:

$$\gamma(\hat{s}, \hat{S}) = \frac{\varphi_1(\hat{s}) + \varphi_2(\hat{s}) + \hat{k}}{t_1(\hat{s}) + t_2(\hat{S})} \tag{36}$$

where

$$\begin{aligned}\varphi_1(\hat{s}) \;&=\; \frac{-h\hat{s}^2}{2\varepsilon_1} + \frac{\hat{s}(2\hat{c}_1\varepsilon_1 + h\sigma_1^2) - \hat{S}(2\hat{c}_1\varepsilon_1 - h\varepsilon_1\hat{S} + h\sigma_1^2)}{2\varepsilon_1^2} \\ &\quad + \frac{\sigma_1^2(2\hat{c}_1\varepsilon_1 + 2\hat{\alpha}_0\varepsilon_1^2 + h\sigma_1^2)(e^{-2\varepsilon_1\hat{s}/\sigma_1^2} - e^{-2\varepsilon_1\hat{S}/\sigma_1^2})}{4\varepsilon_1^3} \qquad (37)\end{aligned}$$

$$\begin{aligned}\varphi_2(\hat{S}) \;&=\; \frac{-h\hat{S}^2}{2\mu_2} + \frac{\hat{S}(2\hat{c}_2\varepsilon_2 + h\sigma_2^2) - \hat{s}(2\hat{c}_2\varepsilon_2 - h\varepsilon_2\hat{s} + h\sigma_2^2)}{2\varepsilon_2^2} \\ &\quad + \frac{\sigma_2^2(-2\hat{c}_2\varepsilon_2 + 2hK\varepsilon_2 + 2\hat{\alpha}_K\varepsilon_2^2 - h\sigma_2^2)(e^{2\varepsilon_2(K-\hat{s})/\sigma_2^2} - e^{2\varepsilon_2(K-\hat{S})/\sigma_2^2})}{4\varepsilon_2^3} \qquad (38)\end{aligned}$$

and $t_1(\hat{s})$, $t_2(\hat{S})$ are given in (26) and (28) respectively.

Proof: The proof is similar to that of theorem 2. We will define a cycle that starts at $\hat{s}$ (at an instance of switch from parameter set 2 to parameter set 1) and ends at level $\hat{s}$ at the next switch instance from parameter set 2 to set 1. Using $T_{\hat{S}}$ and $T_{\hat{s}}$ to denote the time the process takes from the start of the cycle to reach level $\hat{S}$ and from $T_{\hat{S}}$ to the end of the cycle respectively and letting C define the total cost incurred during this cycle, we obtain:

$$\gamma(\hat{s}, \hat{S}) = \frac{\mathrm{E}[C]}{\mathrm{E}[T_{\hat{s}}] + \mathrm{E}[T_{\hat{S}}]} \qquad (39)$$

The expected total cycle cost E$[C]$ can be written as the sum of the holding, boundary and setup costs in a cycle. Defining:

$$\varphi_1(x) = \mathrm{E}\left[\int_0^{T_{\hat{S}}} (hZ(z) + \hat{c}_1)dz + \hat{\alpha}_0\mathcal{L}(T_{\hat{S}})\right] \qquad (40)$$

and

$$\varphi_2(x) = \mathrm{E}\left[\int_0^{T_{\hat{s}}} (hZ(z) + \hat{c}_2)dz + \hat{\alpha}_K\mathcal{L}(T_{\hat{s}})\right], \qquad (41)$$

where x is any starting point between 0 and $\hat{K}$, we can write (36) for $\gamma(\hat{s}, \hat{S})$. Finally, φ_1 and φ_2 can be obtained as the solution of following ordinary differential equations:

$$\begin{cases} \frac{\sigma_1^2\varphi_1''(x)}{2} + \varepsilon_1\varphi_1'(x) &= -hx - \hat{c}_1 \\ \varphi_1'(0) &= -\hat{\alpha}_0 \\ \varphi_1(\hat{S}) &= 0 \end{cases} \qquad (42)$$

and

$$\begin{cases} \frac{s_2^2\varphi_2''(x)}{2} + \varepsilon_2\varphi_2'(x) &= -hx - \hat{c}_2 \\ \varphi_2'(\hat{K}) &= \hat{\alpha}_K \\ \varphi_2(\hat{s}) &= 0 \end{cases} \qquad (43)$$

□

4.4 An Algorithm for the Computing the Optimal (s, S)

In section 3.2, we mentioned that the optimal parameters s and S for the Markovian case can be computed by a search procedure. In fact, the same type of computation can be done for general arrival or service time distributions as in Neuts and Rao [16]. In general, numerical methods are necessary and the computation is lengthy. In this section, we summarize how to use a result from [1] to compute the optimal threshold values s and S when the queueing process in question is approximated by a diffusion process.

Consider a heavy traffic approximation, for the G/G/1/K queue with two traffic parameters used according to an $(\hat{s}, \hat{S})$ policy. As the main results in [1] is proved for the case of equal variances, in this section we assume that $\sigma_1 = \sigma_2 = \sigma$. The state of the buffer (for the diffusion process), $Z(t)$ can be denoted as (z, i), $(0 \leq z \leq \hat{K}, \ \ i = 1, 2)$. Let $\mathcal{P} = (\mathcal{P}_1, \mathcal{P}_2)$ be a partition of the state space that determines which set of parameters is to be used (if state $(z, i) \in \mathcal{P}_j$ then the jth set of parameters will be used). Furthermore, let $\mathcal{I}.$ denote the indicator function, i.e:

$$\mathcal{I}_\Delta(x) = \begin{cases} 1 & \text{if} \qquad x \in \Delta \\ 0 & \text{otherwise} \end{cases} \tag{44}$$

for any set Δ.

Consider the cost function for the diffusion control problem U which is dependent on the starting point x, the policy $\mathcal{P}$ and the time horizon t:

$$U_{x,\mathcal{P},t} = E_{x,\mathcal{P}} \left[\int_0^t \sum_{i=1}^{2} hZ(z) + c_i \mathcal{I}_{[Z(z) \in \mathcal{P}_i]} dz + \hat{\alpha_0} \mathcal{L}(t) + \hat{\alpha_K} \mathcal{U}(t) + kN(t) \right] \tag{45}$$

where $E_{x,\mathcal{P}}$ is the expectation operator which also depends on the starting point x and the policy $\mathcal{P}$ and $N(t)$ is the number of cycles upto time t. The objective is to minimize the long run average cost

$$\gamma = \sup_{\mathcal{P}} \lim_{t \to \infty} \frac{U_{x,\mathcal{P},t}}{t}. \tag{46}$$

where sup denotes supremum.

The solution of the above optimization problem requires a decomposition procedure. To outline the procedure, we first consider the case of a single diffusion with infinitesimal generator $\mathcal{G} = \frac{\sigma^2 d^2}{2\, dx^2} + \epsilon \frac{d}{dx}$. Using the approach pioneered by Bather [2], we can write:

$$U(x, \mathcal{P}, t) \approx \gamma t + V(x), \tag{47}$$

where γ is the long run average cost per time unit and $V(x)$ (called the differential cost starting at x) measures the relative differences between the various starting points after the common long run average γt is subtracted. It can then be shown by dynamic programming arguments that $V(x)$ satisfies the differential equation

$$\mathcal{G}V = -h + \gamma \tag{48}$$

Now, letting $v(x)$ denote the derivative of V we replace the second order equation (48) by the first order equation

$$gv = -h + \gamma, \tag{49}$$

where g denotes the operator $\frac{\sigma^2 d}{2dx} + \epsilon$. Finally, we decompose $v(x)$ as $v = \psi - \gamma\tau$, where τ and ψ are solutions of

$$\begin{aligned} g\,\psi &= -h \quad \text{and} && (50) \\ g\,\tau &= -1 && (51) \end{aligned}$$

respectively. Thus, (50) is a cost equation and (51) is a time equation and the problem is decomposed.

The above decomposition procedure can now be applied for the two diffusion processes, with generators $\mathcal{G}_i = \frac{\sigma^2 d^2}{2\,dx^2} + \epsilon_i \frac{d}{dx}$, $i = 1, 2$ when parameter set i is used. Bather's approach for each parameter set, now yields:

$$\begin{aligned} \mathcal{G}_i V_i &= -h_i + \gamma && (52) \\ V_i' &= (-1)^i \alpha_i && (53) \end{aligned}$$

By letting $v_i = V_i'$ denote the derivatives of Bather's differential costs, these become:

$$\begin{aligned} g_i v_i &= -h_i + \gamma && (54) \\ v_i &= (-1)^i \alpha_i && (55) \end{aligned}$$

where g_i denote the first order differential operators $\frac{\sigma^2 d}{2\,dx} + \epsilon_i$.

Let now ψ_i and τ_i be decompositions of v_i in the form $v_i = \gamma\tau_i - \psi_i$. We choose ψ_i, τ_i as solutions of the following differential equations:

$$\begin{cases} g_i\,\psi_i &= h_i \\ \psi_i(i) &= (-1)^{i+1}\alpha_i \end{cases} \tag{56}$$

and

$$\begin{cases} g_i\,\tau_i &= 1 \\ \tau_i(i) &= 0. \end{cases} \tag{57}$$

Then, $\gamma\tau_i - \psi_i$ satisfy the equation (54). Finally, let $\psi = \psi_0 - \psi_1$ and $\tau = \tau_0 - \tau_1$.

The following theorem characterizes the optimal switching levels. A detailed proof can be found in [1].

Theorem 4 a) The policy of switching between the two diffusions at a fixed point x achieves the long run average value of

$$\gamma(x) = \frac{\psi(x)}{\tau(x)}. \tag{58}$$

b) If a double band policy $(\hat{s}, \hat{S})$ is optimal for some transaction cost k, then the switch levels $(\hat{s}, \hat{S})$ and the corresponding long run average γ satisfy:

$$\frac{\psi(\hat{S})}{\tau(\hat{S})} = \frac{\psi(\hat{s})}{\tau(\hat{s})} = \gamma \quad \text{(conjugacy equation)} \tag{59}$$

and

$$-\int_{\hat{s}}^{\hat{S}} (\psi(x) - \gamma\tau(x))\, dx = k \quad \text{(cost equation)} \tag{60}$$

A graphical interpretation of theorem 4 is provided in Figure 1. $\gamma(a)$ is the average reward when switching at a in the absence of switching costs. Note that if $\gamma(a)$ is unimodal, for any value $s \in [\max[\gamma(0), \gamma(\hat{K})], \gamma(a)]$, theorem 4 provides a unique switching cost determined by (60) and a double band $[\hat{s}, \hat{S}]$ which is optimal for that $\hat{k}$.

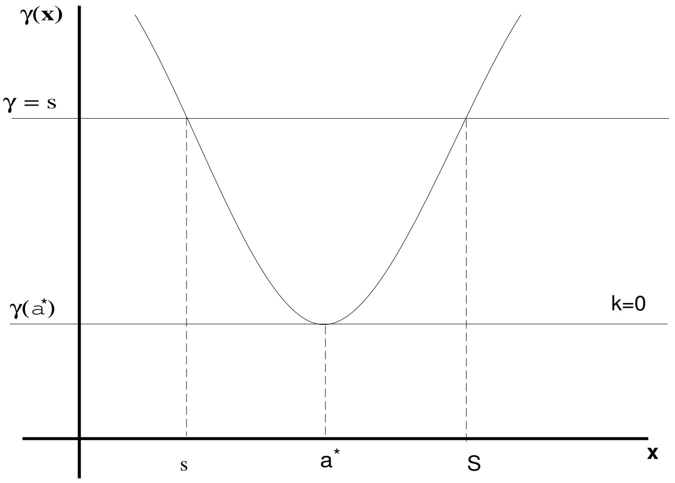

Figure 1: Graphical Interpretation of the Theorem

Theorem 4 suggests the following efficient search algorithm to find the optimal values of the thresholds $\hat{s}$ and $\hat{S}$:

Step 0. Set counter i to 0.

Step 1. Compute the optimal swithcing point in the absence of swithcing costs (a^*) by minimizing the function ψ/τ. If $a^* \notin [0, \hat{K}]$, stop a double band policy can not be optimal.

Step 2. Select $\hat{s}_i$ such that $0 < \hat{s} < a^*$

Step 3. Find $\hat{S}_i$ using the conjugacy equation (59).

Step 4. Find the corresponding switching cost k_i using the cost equation (60).

Step 5. If $|k_i - k| < \delta$ stop, else if $k_i < k$, select $\hat{s}_{i+1}$ such that $\hat{s}_i < \hat{s} < a^*$, else select $\hat{s}_{i+1}$ such that $0 < \hat{s} < \hat{s}_i$.

Step 6. Set $i = i + 1$, go back to step 3.

Remark: Let k_a denote the switching cost for which the corresponding long run average γ_a equals $\min[\gamma(0), \gamma(\hat{K})]$. For transaction costs larger than k_a the optimal policy is to always use the parameter set 1 or 2 depending on which of the costs $\gamma(0)$ or $\gamma(\hat{K})$ is smaller.

5 Numerical Examples

5.1 Approximating the Average Queue Length

Neuts and Rao [16] report a numerical study for the M/PH/1/K queue with two service rates operated according to a (s, S) policy. Matrix-geometric methods are used to compute a variety of exact performance measures. In this section, we compare the diffusion approximation with the results in [16] for the average queue lengths.

In [16] various performance measures are computed with the following setup: the buffer size, K is fixed at 50 and the service rate μ is fixed at 1, the traffic load ρ is adjusted by selecting two arrival rates λ_1 and λ_2. To translate these parameters to the diffusion approximation we select the heavy traffic scaling parameter n as 1000. We scale the parameters such that: $\hat{K} = 50/\sqrt{n}$, $\hat{s} = s/\sqrt{n}$ and $\hat{S} = S/\sqrt{n}$. Furthermore, the drift coefficient of diffusion i is given by $\epsilon_i = \sqrt{n}(1 - \rho_i)$ and the variances are given by: $\sigma_i = \lambda_i + Var(B)$ (where B is the random variable correspoding to service times). For each experiment, we report the absolute percentage error defined as:

$$\frac{|\text{Approximate Queue Length} - \text{Exact Queue Length}|}{\text{Exact Queue Length}} \tag{61}$$

Table 1 analyzes the quality of the approximation as the service distribution (and its coefficient of variation changes). It can be seen that the heavy traffic approximation is quite good when the coefficient of variation of the service process is less than 1. Table 2 displays the effect of varying the traffic load ratio ρ_1/ρ_2. Note that, the coefficient of variation of the selected distribution is low and the approximation is remarkably accurate even when the traffic rates are not close to 1. Table 3 compares the approximate and exact queue lengths when the thresholds (s, S) vary. Once again, the coefficient of variation of the service process is less than 1, and the response of the approximation to the changes in the threshold parameters is excellent.

5.2 Performance of the Optimization Algorithm

In this section, we provide numerical examples to analyze the performance of the optimization algorithm. Our benchmark is the M/M/1/K queue with two traffic rates for which the optimization can be performed by searching all the possible (s, S) pairs. We then compute threshold values by using the algorithm in section 4.4 and compute the cost of using the thresholds obtained by the algorithm. A relevant measure is the percent suboptimality of the policy obtained by the algorithm, defined as:

$$\%\text{suboptimality} = \frac{(\text{algorithm's cost} - \text{optimal cost})}{\text{optimal cost}}. \tag{62}$$

As a first example, consider the case where $\lambda_1 = 1.05$, $\mu_1 = 1$, $\lambda_2 = 1$, $\mu_2 = 1.05$, $h = 0.5$, $c_1 = 0$, $c_2 = 5$ and the buffer size $K = 50$. All cost parameters and the buffer size are scaled (as explained in section 4) to use the algorithm. We compare the exact (s, S) pairs with those found through the algorithm as the switching cost k varies from 0 to 10. Figure 2 displays the $\hat{\gamma}$ function for the given parameters.

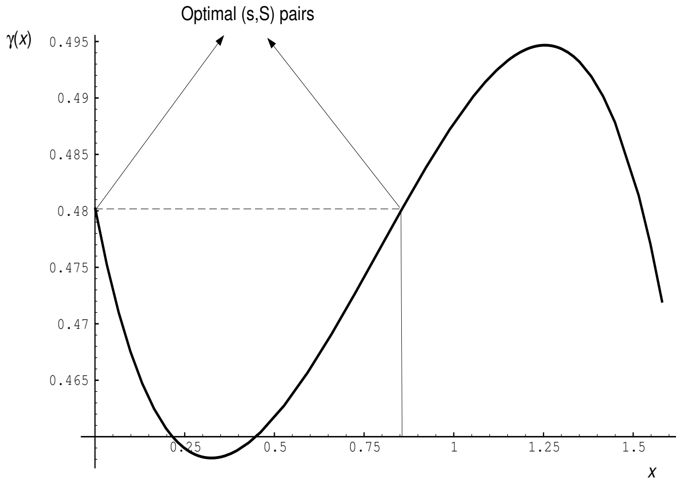

Figure 2: An Example of the Optimization Algorithm

Table 4 summarizes the results of the first example. The approximation is good for small values of the setup cost but starts to deteriorate as the setup cost increases. Also, the (s, S) pairs obtained by the algorithm are very sensitive to the setup cost whereas the exact (s, S) pairs do not vary much.

The second example has the parameters: $\lambda_1 = 1.2$, $\mu_1 = 1$, $\lambda_2 = 1$, $\mu_2 = 1.2$, $h = 0.5$, $c_1 = 0$, $c_2 = 10$ and the buffer size $K = 50$. Once again, we keep these parameters fixed and

λ_1	λ_2	CV	s	S	L (Exact)	L (App.)	% error
1	0.83	0.447	15	30	13.500	13.875	2.8
1	0.83	0.771	15	30	13.935	15.248	9.4
1	0.83	1	15	30	14.592	17.795	21.9
1	0.83	1.5	15	30	16.046	21.853	36.2
1	0.83	2	15	30	17.090	23.751	28.1
1	0.71	0.447	15	30	12.654	12.885	1.8
1	0.71	0.771	15	30	12.817	13.631	6.4
1	0.71	1	15	30	13.004	15.335	17.9
1	0.71	1.5	15	30	13.735	19.855	44.6
1	0.71	2	15	30	14.284	22.841	59.9
1.25	0.83	0.447	15	30	21.778	22.773	4.6
1.25	0.83	0.771	15	30	21.499	23.370	8.7
1.25	0.83	1	15	30	21.147	24.509	15.9
1.25	0.83	1.5	15	30	20.665	25.689	24.3
1.25	0.83	2	15	30	20.490	25.792	25.9
1.25	0.71	0.447	15	30	23.229	21.007	9.6
1.25	0.71	0.771	15	30	23.337	20.845	10.7
1.25	0.71	1	15	30	23.493	21.303	9.3
1.25	0.71	1.5	15	30	23.810	23.554	1.1
1.25	0.71	2	15	30	23.875	24.861	4.1
1.5	0.75	0.447	15	30	23.234	22.749	2.1
1.5	0.75	0.771	15	30	23.341	23.299	0.2
1.5	0.75	1	15	30	23.495	24.553	4.5
1.5	0.75	1.5	15	30	23.850	26.493	11.1
1.5	0.75	2	15	30	24.318	26.702	9.8

CV represents the coefficient of variation of the service process
L (Exact) is the exact average queue length
L (App.) is the approximate average queue length

Table 1: Effect of Service Time Distribution on the Approximation

λ_1	λ_2	CV	s	S	L (Exact)	L (App.)	% error
1	0.91	0.447	15	30	15.268	16.026	5.0
1	0.83	0.447	15	30	13.500	13.925	3.1
1	0.77	0.447	15	30	12.929	13.234	2.3
1	0.71	0.447	15	30	12.654	12.905	2.0
1	0.67	0.447	15	30	12.494	12.714	2.2
1	0.62	0.447	15	30	12.388	12.589	1.6
1	0.59	0.447	15	30	12.314	12.502	1.5
1	0.56	0.447	15	30	12.259	12.437	1.5
1	0.53	0.447	15	30	12.216	12.387	1.4
1	0.50	0.447	15	30	12.182	12.348	1.4
1.2	0.92	0.447	15	30	26.191	26.263	0.2
1.2	0.86	0.447	15	30	23.231	22.881	1.5
1.2	0.80	0.447	15	30	22.085	21.492	2.7
1.2	0.75	0.447	15	30	21.505	20.788	3.3
1.2	0.71	0.447	15	30	21.157	20.367	3.7
1.2	0.67	0.447	15	30	20.924	20.087	4.1
1.2	0.63	0.447	15	30	20.759	19.888	4.2
1.2	0.60	0.447	15	30	20.634	19.740	4.3
1.2	0.57	0.447	15	30	20.538	19.624	4.4
1.2	0.55	0.447	15	30	20.460	19.531	4.5
1.5	0.94	0.447	15	30	28.911	29.271	1.2
1.5	0.88	0.447	15	30	25.577	25.852	0.3
1.5	0.83	0.447	15	30	24.394	24.178	0.9
1.5	0.79	0.447	15	30	23.671	23.287	1.6
1.5	0.75	0.447	15	30	23.234	22.749	2.1
1.5	0.71	0.447	15	30	22.942	22.390	2.4
1.5	0.68	0.447	15	30	22.734	22.133	2.6
1.5	0.65	0.447	15	30	22.578	21.941	2.8
1.5	0.63	0.447	15	30	22.456	21.795	2.9
1.5	0.60	0.447	15	30	22.359	21.671	3.1
CV represents the coefficient of variation of the service process L (Exact) is the exact average queue length L (App.) is the approximate average queue length							

Table 2: Effect of the ratio ρ_1/ρ_2 on the Approximation

λ_1	λ_2	CV	s	S	L (Exact)	L (App.)	% error
1.25	0.83	0.771	15	20	18.388	18.666	1.5
1.25	0.83	0.771	20	25	23.337	23.329	0.03
1.25	0.83	0.771	25	30	28.292	28.055	0.84
1.25	0.83	0.771	30	35	33.197	32.717	1.4
1.25	0.83	0.771	35	40	37.961	37.147	2.1
1.25	0.83	0.771	10	20	15.967	16.466	3.1
1.25	0.83	0.771	15	25	20.863	21.011	0.71
1.25	0.83	0.771	20	30	25.811	25.693	0.46
1.25	0.83	0.771	25	35	30.736	30.373	1.2
1.25	0.83	0.771	30	40	35.555	34.893	1.9
1.25	0.83	0.771	10	25	18.433	18.795	2.0
1.25	0.83	0.771	15	30	23.337	23.370	0.14
1.25	0.83	0.771	20	35	28.262	28.018	0.86
1.25	0.83	0.771	25	40	33.110	32.571	1.6

CV the represents coefficient of variation of the service process
L (Exact) is the exact average queue length
L (App.) is the approximate average queue length

Table 3: Effect of the Control Limits on the Approximation

	Exact Solution			Approximate Solution			
k	s	S	γ	s	S	γ	% suboptimality
0	5	6	11.9848	10	11	12.3205	2.8
1	4	7	12.0098	4	18	12.5265	4.3
2	4	7	12.025	3	20	12.6504	5.2
3	3	7	12.0381	2	22	12.7842	6.2
4	3	7	12.0497	1	24	12.927	7.3
5	3	8	12.0601	1	25	13.0119	7.9
6	3	8	12.0692	1	26	13.0982	8.5
7	3	8	12.0783	no switch (2)		12.6904	5.1
8	3	8	12.0874	no switch (2)		12.6904	5.0
9	3	8	12.0966	no switch (2)		12.6904	4.9
10	3	8	12.1057	no switch (2)		12.6904	4.8
no switch(2) corresponds to always using set 2							

Table 4: Comparison of the Algorithm with Exact Results for Example 1

vary the setup cost k from 0 to 10. The results are summarized in Table 5. The algorithm does not fare well in this case, especially for large values of the switching cost. This can be explained by a combination of high coefficient of variation for the arrival and service processes and distance from heavy traffic conditions.

6 Conclusions

We studied performance evaluation and optimization problems for a finite buffered queueing system with two different types of traffic parameters.

We analyzed the performance of diffusion approximation for the problem as the approximating diffusion control problem can be solved exactly. The approximations give encouraging results for both performance evaluation and optimization. In particular, if the coefficient of variation of the arrival and service processes are small, the average queue length is approximated very well. As for optimization, the thresholds obtained by solving the approximating control problem seem to capture the trends of the exact solution but need more refinement to be useful.

Future research will investigate the applicability of similar ideas to problems with more than two types of trafic parameters and multiple classes of customers with class dependent

k	Exact Solution			Approximate Solution			% suboptimality
	s	S	γ	s	S	γ	
0	4	5	9.2215	6	7	9.43397	2.3
1	3	6	9.30643	2	12	9.8402	5.7
2	3	6	9.35315	1	14	10.1382	8.4
3	3	6	9.39988	1	15	10.3032	9.6
4	3	7	9.44324	1	16	10.4759	10.9
5	3	7	9.47752	no switch (2)		12.4977	31.8
6	2	7	9.50689	no switch (2)		12.4977	31.4
7	2	7	9.53511	no switch (2)		12.4977	31.1
8	2	7	9.565332	no switch (2)		12.4977	30.7
9	2	7	9.59153	no switch (2)		12.4977	30.3
10	2	7	9.61975	no switch (2)		12.4977	29.9
no switch(2) corresponds to always using set 2							

Table 5: Comparison of the Algorithm with Exact Results for Example 2

holding costs.

References

[1] Avram, F. and F. Karaesmen. "On the Optimal Time to Switch Between Two Diffusions". to appear in *Probability in the Engineering and Informational Sciences*, 1996.

[2] Bather J. "A Continuous Time Inventory Model". *Journal of Applied Probability*, 3:538–549, 1966.

[3] Bather J. "A Diffusion Model for the Optimal Control of a Dam". *Journal of Applied Probability*, 5:55–71, 1968.

[4] Chernoff H. and A. Petkau. "Optimal Control of Brownian Motion". *SIAM Journal on Applied Mathematics*, 34:717–731, 1978.

[5] Crabill, T.B. "Optimal Control of a Service Facility with Variable Exponential Service Times and Constant Arrival Rate". *Management Science*, 18:560–566, 1972.

[6] Federgruen A. and H.C. Tijms. "Computation of the Stationary Distribution of the Queue Size in an M/G/1 Queueing System with Variable Service Rate". *Journal of Applied Probability*, 17:515–522, 1980.

[7] Gebhard, R.F. "A Queuing Process with Bilevel Hysteretic Service Rate Control". *Naval Research Logistics Quarterly*, 14:55–67, 1967.

[8] Gupta, S.M. "Interrelationships Between Controlling Arrival and Service in Queueing Systems ". *Computers and Operations Research*, 22:1005–1014, 1995.

[9] Harrison, J.M. *"Brownian Motion and Stochastic Flow Systems"*. John Wiley, New York, 1985.

[10] Hersh, M. and I. Brosh. "The Optimal Strategy Structure of an Intermittently Operated Service Channel". *European Journal of Operational Research*, 5:133–141, 1980.

[11] Heyman, D. "Optimal Operating Policies for M/G/1 Queueing Systems". *Operations Research*, 16:362–382, 1968.

[12] Iglehart, D. and W. Whitt. "Multiple Channel Queues in Heavy Traffic I". *Advances in Applied Probability*, 2:150–177, 1970.

[13] Karaesmen, F. and S.M. Gupta. "Service Control in a Finite Buffered Queue with Holding and Setup Costs". Technical Report, Dept. of Mech., Ind. and Manuf. Eng, Northeastern University, 1996.

[14] Karlin, S. and H. Taylor. *"A Second Course in Stochastic Processes"*. Academic Press, New York, 1981.

[15] Kelly, F. *"Reversibility and Stochastic Networks"*. John Wiley and Sons, New York, 1979.

[16] Neuts, M.F. and B.M. Rao. "On the Design of a Finite Capacity Queue with Phase-type Service Times and Hysteretic Control". *European Journal of Operational Research*, 62:221–240, 1992.

[17] Perry, D. and S.K. Bar-Lev. "A Control of a Brownian Motion Storage System with Two Switchover Drifts". *Stochastic Analysis and Applications*, 7:103–115, 1989.

[18] Rath, J. "Controlled Queues in Heavy Traffic". *Advances in Applied Probability*, 7:656–671, 1975.

[19] Rath, J. "The Optimal Policy for a Controlled Brownian Motion Process". *SIAM Journal on Applied Mathematics*, 32:115–125, 1977.

[20] Reiman, M.I. and L. M. Wein. "Dynamic Scheduling of a Two-Class Queue with Setups". MIT, Alfred P. Sloan School of Management Working Paper, No: 3692-94-MSA, 1994.

[21] Sobel, M. "Optimal Average Cost Policy for a Queue with Start-up and Shut-down Costs". *Operations Research*, 17:145–162, 1969.

[22] Teghem Jr., J. "Optimal Control of a Removable Server in an M/G/1 Queue with Finite Capacity". *European Journal of Operational Research*, 31:358–367, 1987.

[23] Tijms, H.C. "An Algorithm for Average Cost Denumerable State Semi-Markov Decision Problems with Applications to Controlled Queuing Systems". In L.C. Thomas R. Hartley and D.J. White, editors, *"Recent Developments in Markov Decision Processes,"*, pages 143–179. Academic Press, New York, 1980.

[24] Yadin, M. and P. Naor. "Queueing Systems with a Removable Service Station". *Operational Research Quarterly*, 14:393–405, 1963.