CONTROL OF QUEUES WITH UNKNOWN COMPLETION TIME

by

Jing Tang

A Major Paper

Submitted to the Faculty of Graduate Studies and Research

through the Department of Mathematics and Statistics

in Partial Fulfillment of the Requirements for

the Degree of Master of Science at the

University of Windsor

Windsor, Ontario,Canada

2005

Abstract

We consider the problem of admission control to queues (M/M/1, M/E_2/1 and M/E_k/1) with unknown completion time. Customers arrive at a gatekeeper according to a Poisson process. The gatekeeper is not informed as to when services are completed. Upon a new arrival, the gatekeeper has to decide whether to admit or to block that customer. The gatekeeper will be informed if an admitted customer finds the server is busy, and that customer is rejected and leaves the system immediately without being served. We assume that a smaller cost c will be incurred if an arrival is blocked by the gatekeeper from entering the system and a larger cost K will be incurred if an admitted customer is rejected by the server. In the case of an M/M/1 queue, Lin and Ross (2003) prove that a threshold-type policy that blocks for a certain amount of time after an admitted arrival is optimal. They also show a relationship between the blocking costs and the optimal blocking time. We use an integral approach to obtain the same results. In the case of an M/E_2/1 queue, we assume a two-stage Erlang service time. A longer blocking time is used if an admitted customer is blocked at the first stage and a shorter blocking time is applied if that customer is blocked at the second stage. We analytically derive the cost function in terms of the two blocking times. For a given blocking cost, we use Maple to numerically solve those two blocking

times that minimize the total cost. We then propose a heuristic policy to handle the case when customer stage in service is unknown. Finally, we extend the model of an $M/E_2/1$ queue to the problem of admission control for an $M/E_k/1$ queue.

Acknowledgements

With a deep sense of gratitude, I wish to express my sincere thanks to my supervisor Dr. Myron Hlynka. I thank him for his patience and encouragement that carried me on through difficult times, and for his insights and suggestions that helped to shape my research skills. I am really glad to be his student.

My warm thanks are due to Dr. Severien Nkurunziza, who took effort in reading and providing me with valuable comments.

I am grateful to my former advisor Dr. Meiling Huang, who introduced and helped me to start my graduate student life in Statistics. The encouragement and motivation that was given to me to carry out my research work by her is remembered.

I also want to thank my parents, who taught me the value of hard work by their own examples. I would like to share this moment of happiness with them.

1 Introduction

According to Lin and Ross (2003), admission control has been widely used to improve the performance of a system which has problems occurring due to overload. Consider a web site consisting of web servers and HTTP GET requests sent by the users. When the arrival rate of new requests exceeds the servers' capacity, queues build up and the consequence is the long response times when the users are visiting the web site. There is a tendency for users to leave and never come back to sites that perform poorly. This could lead to profit loss if the site is commercial. Such performance problems can be solved by implementing admission control mechanisms in the sites. The main idea is to admit only a certain numbers of requests coming into the server and to block others whenever the arrival rate is too high.

To solve an admission problem, the fundamental thing to consider is the formulation of the objective function that associates a reward for serving a job, a cost when blocking customers in queue and a cost for rejecting a request. Optimal policies are considered in many different models. Under the assumption that the number of customers in the system is unknown, Altman and Koole (1995) considered a service control problem with noisy delayed information. Lin and Ross (2003) considered a multiple-server loss model. They showed that, in the case of a single server with exponential service,

a threshold-type policy is optimal and they proposed two types of heuristic policies when there are multiple servers. Cao and Nyberg (2004) considered the problem of admission control to an M/M/1 queue under periodic observations with average cost criterion.

In this paper, we consider a model similar to that of as Lin and Ross (2003) with only one server. However we allow Erlang service times rather than only allowing exponential service times. Customers arrive to a gatekeeper according to a Poisson process with rate λ. When a new customer arrives, the gatekeeper has to decide whether to admit or to block that customer. The gatekeeper will be informed if an admitted customer finds the server busy, and that customer leaves the system immediately without being served. However, the gatekeeper has no information on the departure time of the last accepted customer. We assume that a smaller cost c will be incurred if an arrival is blocked by the gatekeeper from entering the system and a larger cost K will be incurred if an admitted customer is rejected by the server. We also have a benefit occurring as a result of an accepted customer which is admitted by the server. By means of translation, we set the benefit to zero. Further, without loss of generality, we let $K = 1$ and $0 < c < 1$. In the next section, we consider the problem of admission control to an M/M/1 queue. We use an integral approach to show a relationship between the blocking cost and the optimal blocking time. In section 3, we consider the

case of an $M/E_2/1$ queue, and we assume a two-stage Erlang service time. A longer blocking time is used if an admitted customer is blocked at the first stage and a shorter blocking time is applied if that customer is blocked at the second stage. We analytically derive the cost function in terms of the two blocking times. For a given blocking cost, we use a local search method to numerically solve for those two blocking times that minimize the total cost. In section 4, we propose a heuristic policy to deal the case when the customer stage in service is unknown. Extension to an $M/E_k/1$ queue is presented in section 5. Finally conclusions are presented in the last section.

2 An M/M/1 Queue

In this section we consider the problem of admission control to an M/M/1 queue with unknown completion time. Lin and Ross (2003) show that a threshold-type policy that blocks for a certain amount of time after an admitted arrival is optimal. So in our problem, the gatekeeper blocks all the arrivals for a time units whenever an admission is occurred. Because of the memoryless property of the exponential service time, no matter when an admitted customer is accepted or is rejected, the gatekeeper blocks new arrivals for the next a time units. Upon acceptance of a customer, we want to find the expected time to the next acceptance. See Figure 1.

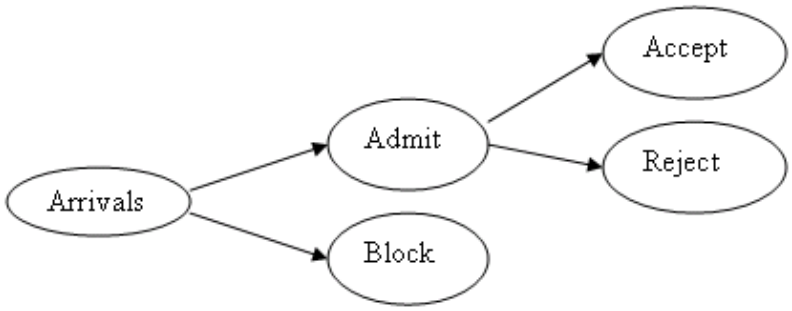

Figure 1: arrival rate = accept rate + reject rate + block rate.

Suppose that there is a customer which has just been accepted. Then after a time units, a customer arriving t time units later has two possible situations at time $t + a$. One possibility is that the last accepted customer has completed service and thus the new customer could be served immediately. The other situation is that the server is still busy so the new customer is rejected, has to leave the system and the gatekeeper blocks all the arrivals for another a time units. Then the next customer arriving t time units later will encounter the same problem. We define T_1, T_2 to be the interaccept and interadmit times respectively. Then

$$
\begin{aligned}
E(T_1) &= a + \int_0^\infty t\lambda e^{-\lambda t}\left(1 - e^{-\mu(a+t)}\right) dt + \int_0^\infty \left(t + E(T_1)\right) \lambda e^{-\lambda t} e^{-\mu(a+t)} dt \\
&= a + \int_0^\infty t\lambda e^{-\lambda t} dt + \int_0^\infty E(T_1)\lambda e^{-\lambda t} e^{-\mu(a+t)} dt \\
&= a + \frac{1}{\lambda} + \frac{E(T_1)\lambda e^{-\mu a}}{\lambda + \mu},
\end{aligned}
$$

Solving gives

$$
E(T_1) = \left(a + \frac{1}{\lambda}\right)\left(1 - \frac{\lambda e^{-\mu a}}{\lambda + \mu}\right)^{-1}.
$$

The expected interadmit time depends on the expected interarrival time. Thus

$$
\begin{aligned}
E(T_2) &= a + E\left(\text{interarrival time}\right) \\
&= a + \frac{1}{\lambda}
\end{aligned}
$$

The arrival rate consists of accept rate, reject rate and block rate. Let $\lambda_1, \lambda_2, \lambda_3$ be the accept, reject and block rates respectively, so

$$
\lambda = \lambda_1 + \lambda_2 + \lambda_3,
$$

where

$$\lambda_1 = \frac{1}{E(T_1)},$$
$$\lambda_2 = \frac{1}{E(T_2)} - \frac{1}{E(T_1)},$$
$$\lambda_3 = \lambda - \frac{1}{E(T_2)}.$$

Then the expected cost per unit time is

$$\begin{aligned}
&\lambda_1(0) + \lambda_2(1) + \lambda_3(c) \\
&= \frac{1}{a + \frac{1}{\lambda}} - \frac{\left(1 - \frac{\lambda e^{-\mu a}}{\lambda+\mu}\right)}{a + \frac{1}{\lambda}} + \lambda c - \frac{c}{a + \frac{1}{\lambda}} \\
&= \left(\frac{\lambda}{a\lambda + 1}\right) e^{-\mu a} \left(\frac{\lambda}{\lambda + \mu}\right) + \left(\lambda - \frac{\lambda}{a\lambda + 1}\right) c.
\end{aligned}$$

For fixed c, λ, μ, this is a function of a which we call it $f(a)$. It can be shown that a unique minimum exists. To minimize $f(a)$, we take the derivative w.r.t. a and set it to 0, yielding

$$\frac{-\lambda^2}{(a\lambda + 1)^2} e^{-\mu a} \frac{\lambda}{\lambda + \mu} + \frac{\lambda}{a\lambda + 1} \left(-\mu e^{-\mu a}\right) \frac{\lambda}{\lambda + \mu} + \frac{\lambda^2 c}{(a\lambda + 1)^2} = 0.$$

By examining the second derivative, we could check that we have a minimum. Thus

$$c = e^{-\mu a^*} \frac{\lambda}{\lambda + \mu} (\lambda + \mu + a^* \lambda \mu) = e^{-\mu a^*} \left(1 + \frac{a^* \mu \lambda}{\lambda + \mu} \right). \tag{2.1}$$

where a^* is an optimal blocking time. This matches the result in Lin and Ross (2003).

According to Lin and Ross (2003), we could see that if we fix the arrival rate λ then a^* will increase as blocking cost c and service rate μ decrease. On the other hand, a^* will increase as λ increases when the other two parameters are fixed. Intuitively, when the blocking cost is small, we are likely to increase our blocking time to avoid the possible rejection cost 1. When the service rate is high, the probability of an admitted customer being rejected is small, so we could reduce the blocking time to get a lower total cost.

3 An $M/E_2/1$ Queue

In this section we extend the model of Lin and Ross (2003) to the problem of admission control for an $M/E_2/1$ queue with unknown completion time. Motivated by the policy we used on an M/M/1 queue, we employ a similar strategy for an $M/E_2/1$ queue. First we assume that when an admitted

customer is rejected, the gatekeeper will be informed immediately at which stage the last accepted customer is being served. Our threshold policy states that upon rejection of an admitted customer,

1. The gatekeeper blocks any arrival for the next b time units if the last accepted customer is being served at the first stage,

2. The gatekeeper blocks any arrival for the next a time units if the last accepted customer is being served at the second stage,

3.1 The Cost Function

Recall that the arrival rate consists of accept rate, reject rate and block rate. We are going to find the expected interaccept time first. We define

$E_{AA} \equiv$ expected interaccept time,

$E_{1A} \equiv$ expected time from stage 1 rejection to the next acceptance,

$E_{2A} \equiv$ expected time from stage 2 rejection to the next acceptance.

Note $E_{AA} = E_{1A}$, because regardless of whether the last admitted customer is accepted or is rejected at the first stage, the gatekeeper has to block all

arrivals within the next b time units. The two-stage Erlang service time is the sum of two exponentials. To simplify the calculations, we could interpret the system as having two servers in series each with an exponential service time but with the condition that a new customer cannot enter the first stage if the second stage is nonempty. Thus at any time, the status of an accepted customer can be obtained by counting the number of service completions which occur according to a Poisson process, i.e. during a time interval, an accepted customer remains at first stage if there are no arrivals in the relevant time interval. So,

$$\begin{aligned}
E_{AA} &= b + \int_0^\infty t\lambda e^{-\lambda t}\left(1 - \frac{(\mu(t+b))^0 e^{-\mu(t+b)}}{0!} - \frac{(\mu(t+b))^1 e^{-\mu(t+b)}}{1!}\right)dt \\
&\quad + \int_0^\infty (t + E_{1A})\lambda e^{-\lambda t}\frac{(\mu(t+b))^0 e^{-\mu(t+b)}}{0!}dt \\
&\quad + \int_0^\infty (t + E_{2A})\lambda e^{-\lambda t}\frac{(\mu(t+b))^1 e^{-\mu(t+b)}}{1!}dt \\
&= b + \int_0^\infty t\lambda e^{-\lambda t}dt + \int_0^\infty E_{1A}\lambda e^{-\lambda t}e^{-\mu(t+b)}dt + \int_0^\infty E_{2A}\lambda e^{-\lambda t}\mu(t+b)e^{-\mu(t+b)}dt \\
&= b + \frac{1}{\lambda} + E_{AA}\frac{\lambda e^{-\mu b}}{\lambda+\mu} + E_{2A}\lambda\mu e^{-\mu b}\left[\frac{1+(\lambda+\mu)b}{(\lambda+\mu)^2}\right], \qquad (3.1)
\end{aligned}$$

where

$$
\begin{aligned}
E_{2A} &= a + \int_0^\infty (t + E_{2A})\, \lambda e^{-\lambda t} \frac{(\mu\,(t+a))^0\, e^{-\mu(t+a)}}{0!} dt + \int_0^\infty t\lambda e^{-\lambda t} \left(1 - e^{-\mu(t+a)}\right) dt \\
&= a + \int_0^\infty E_{2A} \lambda e^{-\lambda t} e^{-\mu(t+a)} dt + \int_0^\infty t\lambda e^{-\lambda t} dt \\
&= a + \frac{1}{\lambda} + E_{2A} \frac{\lambda e^{-\mu a}}{\lambda + \mu}.
\end{aligned}
$$

Solving for E_{2A} yields

$$
E_{2A} \left(1 - \frac{\lambda e^{-\mu a}}{\lambda + \mu}\right) = a + \frac{1}{\lambda},
$$

or

$$
E_{2A} = \left(a + \frac{1}{\lambda}\right) \left(1 - \frac{\lambda e^{-\mu a}}{\lambda + \mu}\right)^{-1}. \tag{3.2}
$$

Substituting(3.2) into (3.1), we have

$$
E_{AA} = b + \frac{1}{\lambda} + E_{AA} \frac{\lambda e^{-\mu b}}{\lambda + \mu} + \left(a + \frac{1}{\lambda}\right) \left(1 - \frac{\lambda e^{-\mu a}}{\lambda + \mu}\right)^{-1} \lambda \mu e^{-\mu b} \left[\frac{1 + (\lambda + \mu)\, b}{(\lambda + \mu)^2}\right],
$$

so

$$E_{AA} = \left(b + \frac{1}{\lambda}\right)\left(1 - \frac{\lambda e^{-\mu b}}{\lambda + \mu}\right)^{-1} + \frac{\left(a + \frac{1}{\lambda}\right)\left(\lambda\mu e^{-\mu b}\right)\left(1 + (\lambda + \mu)\, b\right)}{\left(\lambda + \mu - \lambda e^{-\mu a}\right)\left(\lambda + \mu - \lambda e^{-\mu b}\right)}.$$

Next we compute the expected time between two consecutive rejections which occur at the same stage. Let

$E_{11} \equiv$ expected time between two consecutive stage 1 rejections,

$E_{A1} \equiv$ expected time between an acceptance and the next stage 1 rejection,

$E_{21} \equiv$ expected time between a stage 2 rejection and the next stage 1 rejection,

$E_{22} \equiv$ expected time between two consecutive stage 2 rejections,

$E_{A2} \equiv$ expected time between an acceptance and the next stage 2 rejection,

$E_{12} \equiv$ expected time between a stage 1 rejection and the next stage 2 rejection.

Note that $E_{A1} = E_{11}, E_{A2} = E_{12}$. Then the expected time between two consecutive stage 1 rejections is given by

$$
\begin{aligned}
E_{11} &= b + \int_0^\infty t\lambda e^{-\lambda t} \frac{\left(\mu\left(t+b\right)\right)^0 e^{-\mu(t+b)}}{0!} dt + \int_0^\infty \left(t + E_{21}\right) \lambda e^{-\lambda t} \frac{\left(\mu\left(t+b\right)\right)^1 e^{-\mu(t+b)}}{1!} dt \\
&+ \int_0^\infty (t + E_{A1}) \lambda e^{-\lambda t} \left(1 - \frac{\left(\mu\left(t+b\right)\right)^0 e^{-\mu(t+b)}}{0!} - \frac{\left(\mu\left(t+b\right)\right)^1 e^{-\mu(t+b)}}{1!}\right) dt \\
&= b + \int_0^\infty E_{21} \lambda e^{-\lambda t} \mu\left(t+b\right) e^{-\mu(t+b)} dt + \int_0^\infty t\lambda e^{-\lambda t} dt + \int_0^\infty E_{A1} \lambda e^{-\lambda t} dt \\
&- \int_0^\infty E_{A1} \lambda e^{-\lambda t} e^{-\mu(t+b)} dt - \int_0^\infty E_{A1} \lambda e^{-\lambda t} \mu\left(t+b\right) e^{-\mu(t+b)} dt \\
&= b + \frac{1}{\lambda} + E_{11} + E_{21} \lambda\mu e^{-\mu b} \left[\frac{1 + (\lambda+\mu)b}{(\lambda+\mu)^2}\right] - E_{11} \frac{\lambda e^{-\mu b}}{\lambda+\mu} \qquad (3.3) \\
&- E_{11} \lambda\mu e^{-\mu b} \left[\frac{1 + (\lambda+\mu)b}{(\lambda+\mu)^2}\right],
\end{aligned}
$$

where

$$
\begin{aligned}
E_{21} &= a + \int_0^\infty \left(t + E_{21}\right) \lambda e^{-\lambda t} \frac{\left(\mu\left(t+a\right)\right)^0 e^{-\mu(t+a)}}{0!} dt + \int_0^\infty (t + E_{A1}) \lambda e^{-\lambda t} \left(1 - e^{-\mu(t+a)}\right) dt \\
&= a + \int_0^\infty E_{21} \lambda e^{-\lambda t} e^{-\mu(t+a)} dt + \int_0^\infty t\lambda e^{-\lambda t} dt + \int_0^\infty E_{A1} \lambda e^{-\lambda t} dt - \int_0^\infty E_{A1} \lambda e^{-\lambda t} e^{-\mu(t+a)} dt \\
&= a + \frac{1}{\lambda} + E_{11} + E_{21} \frac{\lambda e^{-\mu a}}{\lambda+\mu} - E_{11} \frac{\lambda e^{-\mu a}}{\lambda+\mu}.
\end{aligned}
$$

Solving for E_{21} yields

$$E_{21}\left(1-\frac{\lambda e^{-\mu a}}{\lambda+\mu}\right)=a+\frac{1}{\lambda}+E_{11}-a+\frac{1}{\lambda}+E_{11}$$

$$E_{21}=\left(a+\frac{1}{\lambda}\right)\left(1-\frac{\lambda e^{-\mu a}}{\lambda+\mu}\right)^{-1}+E_{11}. \tag{3.4}$$

Substituting (3.4) into (3.3), we get

$$\begin{aligned}E_{11}&=b+\frac{1}{\lambda}+E_{11}+\left(a+\frac{1}{\lambda}\right)\left(1-\frac{\lambda e^{-\mu a}}{\lambda+\mu}\right)^{-1}\lambda\mu e^{-\mu b}\left[\frac{1+(\lambda+\mu)b}{(\lambda+\mu)^2}\right]\\&\quad+E_{11}\lambda\mu e^{-\mu b}\left[\frac{1+(\lambda+\mu)b}{(\lambda+\mu)^2}\right]-E_{11}\frac{\lambda e^{-\mu b}}{\lambda+\mu}-E_{11}\lambda\mu e^{-\mu b}\left[\frac{1+(\lambda+\mu)b}{(\lambda+\mu)^2}\right]\\&=\left(b+\frac{1}{\lambda}\right)\left(\frac{\lambda+\mu}{\lambda e^{-\mu b}}\right)+\frac{\left(a+\frac{1}{\lambda}\right)\mu\left(1+(\lambda+\mu)b\right)}{\lambda+\mu-\lambda e^{-\mu a}}.\end{aligned}$$

The expected time between two consecutive stage 2 rejections is equal to

$$\begin{aligned}E_{22}&=a+\int_0^\infty t\lambda e^{-\lambda t}\frac{\left(\mu\left(t+a\right)\right)^0 e^{-\mu(t+a)}}{0!}dt+\int_0^\infty (t+E_{A2})\lambda e^{-\lambda t}\left(1-e^{-\mu(t+a)}\right)dt\\&=a+\int_0^\infty t\lambda e^{-\lambda t}dt+\int_0^\infty E_{12}\lambda e^{-\lambda t}dt-\int_0^\infty E_{12}\lambda e^{-\lambda t}e^{-\mu(t+a)}dt\\&=a+\frac{1}{\lambda}+E_{12}\left(1-\frac{\lambda e^{-\mu a}}{\lambda+\mu}\right),\end{aligned} \tag{3.5}$$

where

$$E_{12} = b + \int_0^\infty (t + E_{12})\, \lambda e^{-\lambda t} \frac{(\mu\,(t+b))^0\, e^{-\mu(t+b)}}{0!} dt + \int_0^\infty t\lambda e^{-\lambda t} \frac{(\mu\,(t+b))^1\, e^{-\mu(t+b)}}{1!} dt$$
$$+ \int_0^\infty (t + E_{A2})\lambda e^{-\lambda t} \left(1 - \frac{(\mu\,(t+b))^0\, e^{-\mu(t+b)}}{0!} - \frac{(\mu\,(t+b))^1\, e^{-\mu(t+b)}}{1!}\right) dt$$
$$= b + \frac{1}{\lambda} + E_{12} - E_{12}\lambda\mu e^{-\mu b} \left[\frac{1 + (\lambda+\mu)\, b}{(\lambda+\mu)^2}\right].$$

Rearranging the terms, we have

$$E_{12}\lambda\mu e^{-\mu b} \left[\frac{1 + (\lambda+\mu)\, b}{(\lambda+\mu)^2}\right] = b + \frac{1}{\lambda}.$$

Thus

$$E_{12} = \left(b + \frac{1}{\lambda}\right) \frac{(\lambda+\mu)^2}{\lambda\mu e^{-\mu b}\,(1 + (\lambda+\mu)\, b)}. \tag{3.6}$$

Substituting (3.6) into (3.5) yields

$$E_{22} = a + \frac{1}{\lambda} + \frac{\lambda + \mu - \lambda e^{-\mu a}}{\lambda + \mu} \left(b + \frac{1}{\lambda}\right) \frac{(\lambda+\mu)^2}{\lambda\mu e^{-\mu b}\,(1 + (\lambda+\mu)\, b)}$$
$$= a + \frac{1}{\lambda} + \left(b + \frac{1}{\lambda}\right) \frac{(\lambda+\mu)\,(\lambda + \mu - \lambda e^{-\mu a})}{\lambda\mu e^{-\mu b}\,(1 + (\lambda+\mu)\, b)}.$$

We denote $\lambda_a, \lambda_1, \lambda_2, \lambda_b$ as the accept rate, reject stage 1 rate, reject stage

2 rate and block rate respectively. Then

$$\lambda = \lambda_a + (\lambda_1 + \lambda_2) + \lambda_b.$$

where $\lambda_a = \frac{1}{E_{AA}}, \lambda_1 = \frac{1}{E_{11}}, \lambda_2 = \frac{1}{E_{22}}, \lambda_b = \lambda - \lambda_a - (\lambda_1 + \lambda_2)$.

The expected cost per unit time yields,

$$\lambda_a(0) + (\lambda_1 + \lambda_2)(1) + \lambda_b(c). \tag{3.7}$$

For a given blocking cost c, we are interested in finding a pair (a^*, b^*) that minimizes the function (3.7). Since it is not possible to analytically derive a formula for a^* and b^*, we run simulations to numerically calculate such pairs of values for each given value of c.

3.2 Minimization Results

We apply a local search method in which we randomly pick two initial values of a, b (we use point $(0, 0)$) and then make small random changes to get updated values of a, b. This was programmed in MAPLE. If new point is better, keep it, otherwise keep the original point and make some other small random changes. By changing our time units, without loss of generality, we can set $\lambda = 1$. Then (3.7) becomes a function of four variables: a, b, c, μ. In order to show that there is a unique pair of a, b which minimizes this

function, we plot a 3D graph of the expected cost function (3.7) in terms of a, b by setting $\mu = 1$, $c = 0.5$. See Figure (2). The result can be seen more

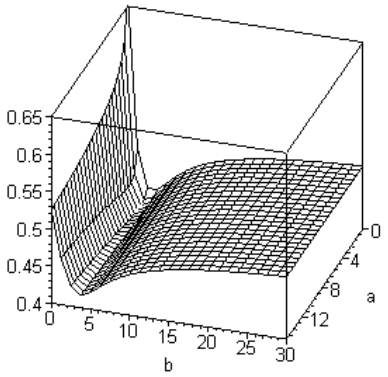

Figure 2: 3D-plot

clearly in Figure (3) and Figure (4), i.e. when $a^* = 1.43$ and $b^* = 2.63$, this

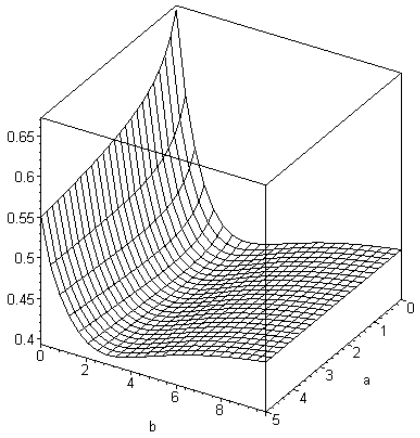

Figure 3: Another 3D-plot

function reaches the minimum at 0.399.

Then we plot 4 graphs for 4 values of μ (0.1,0.5,1,2), we choose 13 values

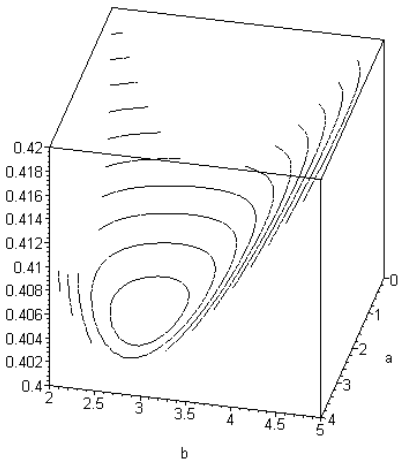

Figure 4: Contour plot

of c (0.01,0.05,0.l,0.2,...,0.8,0.9,0.95,0.99), and each of the 4 graphs is two dimensional with a and b for axes.

The results can be seen in Figure (5). In general, μ decreases in blocking times a, b. This makes sense, when μ is large, the gatekeeper would rarely block an arrival. As μ decreases, the number of customers that are blocked per unit time also increases. On the other hand, when the blocking cost c decreases, a and b are both increase. Note that, in the case of $c = 0.01$, its performance is close to that with no blocking cost, so the gatekeeper blocks a large number of arrivals. In the case of $c = 0.99$, its performance is close to that with blocking cost equals rejection cost, thus the gatekeeper would open the gate for almost all arrivals. Also note that when $\mu >> \lambda$, i.e. $\mu = 10$, its performance is close to that with no control and the gatekeeper admits

almost all the arrivals. In this case, we encounter numerical difficulties, so the graph is not shown.

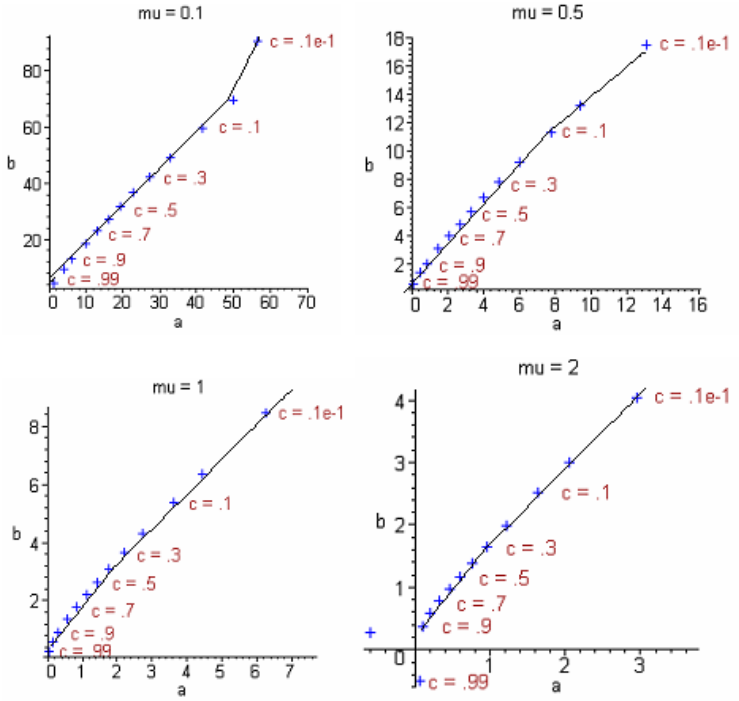

Figure 5: 4 graphs for 4 values of μ, each chooses 13 values of c.

4 Control When the Stage of Rejection is Not Communicated

In this section, we consider the case when the customer stage in service is not known to the gatekeeper. Our heuristic policy proposes that if last customer was accepted, wait time of length b. If last customer was rejected, measure time t, since the last accepted customer. At any time, there is only one customer in the system at either stage 1 or stage 2. We compute the probability that the customer remains at stage 1 and the probability that the customer is at stage 2 as follows:

$$\begin{aligned} \text{P(stage 1)} &= \text{P(0 completions in } t) \\ &= \frac{(\mu t)^0 e^{-\mu t}}{0!} = e^{-\mu t} = u(t), \\ \text{P(stage 2)} &= \text{P(1 completion in } t) \\ &= \frac{(\mu t)^1 e^{-\mu t}}{1!} = \mu t e^{-\mu t} = v(t). \end{aligned}$$

Then wait a weighted average

$$\frac{u(t)}{u(t)+v(t)} b + \frac{v(t)}{u(t)+v(t)} a.$$

before accepting a new customer.

5 An M/E$_k$/1 Queue

A two-stage Erlang service can be extended to E$_k$ service times. Let the stage of service be $1, 2, ..., k$. Let a_i be the gap time from a rejection at stage i until next allowable admission. Let A denote an acceptance. Let

E_{AA} = expected time between consecutive acceptances,

E_{iA} = expected time between rejection at stage i and next acceptance.

Then $E_{AA} = E_{1A}$.

$$E_{AA} = a_1 + \int_0^\infty t\lambda e^{-\lambda t}\left(1 - \sum_{j=0}^{k-1} \frac{(\mu\,(a_1+t))^j\, e^{-\mu(a_1+t)}}{j!}\right) dt$$
$$+ \int_0^\infty \lambda e^{-\lambda t} \sum_{j=0}^{k-1} (t + E_{j+1,A}) \frac{(\mu\,(a_1+t))^j\, e^{-\mu(a_1+t)}}{j!} dt$$

for $i = 1, ..., k$;

$$E_{iA} = a_i + \int_0^\infty t\lambda e^{-\lambda t}\left(1 - \sum_{j=0}^{k-i} \frac{(\mu\,(a_i+t))^j\, e^{-\mu(a_i+t)}}{j!}\right) dt$$
$$+ \int_0^\infty \lambda e^{-\lambda t} \sum_{j=0}^{k-i} (t + E_{j+i,A}) \frac{(\mu\,(a_i+t))^j\, e^{-\mu(a_i+t)}}{j!} dt.$$

These equations can be solved for E_{AA} and E_{iA}, $i = 1, ..., k$. We then proceed as in Section 3.1.

6 Conclusions

In this paper we studied the problem of admission control to an M/M/1, M/E_2/1 and M/E_k/1 queues with unknown completion time. In the case of an M/M/1 queue, we used an integral approach to analytically specify the relationship between the optimal blocking time and the blocking cost c, and obtained the same results as Lin and Ross (2003). In the case of an M/E_2/1 queue, we analytically derived the cost per unit time function and numerically obtained a pair (a^*, b^*) which minimizes the cost function for a given value of c.

A heuristic was presented in section 4 to handle the case when the stage of service is not communicated to the gatekeeper. The two-stage Erlang model was generalized in a fairly straight forward way to a k-stage Erlang for any arbitrary k in section 5.

References

[1] Altman,E. and Koole,G. 1995. Control of a random walk with noisy delayed information. *Systems Control Lett.* **24** 207-213

[2] Anders,R., Wittenmark, B. Kihl, M. 2003. Analysis and design of admission control in web-server systems. *The Proceeding of American Control Conference 2003.*

[3] Cao,J. and Nyberg, C. 2004. A monotonic property of the optimal admission control to an M/M/1 queue under periodic observations with average cost criterion. *17th Nordic Teletraffic Seminar (NTS).*

[4] Lin, K.Y. and Ross, S.M. 2003. Admission control with incomplete information of a queueing system. *Operations Research.* **51(4)** 645-654.