Optimal Pricing for Finite Capacity Queueing Systems with Multiple Customer Classes

Serhan Ziya*
Department of Statistics and Operations Research
University of North Carolina
CB#3260, 213 Smith Building, Chapel Hill, NC 27599 U.S.A.
Email: ziya@unc.edu, Phone: +1 919 933 3899, Fax: +1 919 962 0391

Hayriye Ayhan and Robert D. Foley
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
765 Ferst Drive, Atlanta, GA 30318 U.S.A.
Email: hayhan@isye.gatech.edu, rfoley@isye.gatech.edu
Phone: +1 404 894 2308, +1 404 894 2309
Fax: +1 404 894 2301

Abstract

We determine optimal static prices for a finite capacity queueing system serving customers from different classes. We prove an upper bound for the optimal arrival rates for a fairly large class of queueing systems and provide sufficient conditions that ensure the existence of a unique optimal arrival rate vector. We show that these conditions hold for M/M/1/m and M/G/s/s systems and prove structural results on the relationships between the optimal arrival rates and system capacity.

*Corresponding author
I. INTRODUCTION

This paper investigates optimal static control policies for a finite capacity queueing system serving customers from different classes. The objective of the controller is to determine the arrival rate (or equivalently the price) for each customer class that maximizes the revenue for the service provider. There are no priorities; all customers are served according to the First-Come-First-Served discipline. Customers are not delay sensitive, however, since the system capacity is finite, there is an upper bound on the expected waiting time.

The main contribution of this paper is the analysis of optimal static control policies for finite capacity queues, which has received relatively little attention in the literature. Without any restrictions on the arrival and service processes, a complete characterization of optimal arrival rates is not possible even under reasonable assumptions on the customers’ reservation prices since the blocking probabilities are not known in general. For such general systems, we provide an upper bound for the optimal arrival rates. Under certain conditions on the blocking probabilities, we prove the existence and uniqueness of the optimal solution. We then show that these conditions hold for M/M/1/m and M/G/s/s systems and investigate the relationships between the optimal arrival rates and the system’s waiting room and service capacity. For the M/M/1/m system, the optimal arrival rates are shown to be either monotone increasing or decreasing in $m$. Whether they are increasing or decreasing depends on the demand on the system relative to the service capacity. For the M/G/s/s system, the optimal arrival rates are shown to be monotone increasing in $s$ regardless of the demand or the service capacity.

There is a rich literature on using pricing to control queues. Most of the existing work considers infinite capacity systems and deals with using pricing as a tool to induce delay sensitive self-optimizing customers to act in some socially optimal way. Two fundamental papers are Naor [13] and Mendelson [10]. Also see Stidham [16] for a review of early work in this area. On the other hand, pricing for systems with finite capacity has received relatively little attention. Courcoubetis and Reiman [3] analyze the pricing decisions for a loss system under an asymptotic regime where the capacity and the system load go to infinity. Paschalidis and Tsitsiklis [14] consider a similar loss model and show that static pricing policies (when suitably set) perform almost as well as optimal dynamic pricing policies. Similar results are established in a more general network setting by Paschalidis and Liu [15] and Lin and Shroff [8]. Low [9] considers dynamic pricing policies for an M/M/1/m system and shows the existence of a dynamic pricing policy.
that is monotone non-decreasing in the number of customers in the system. On the other hand, Miller [12], Lewis, Ayhan, and Foley [6], [7] study dynamic admission control decisions without pricing for finite capacity systems where customers’ valuations are deterministic given their class identities. These papers consider models where the decision is whether to accept or reject an incoming customer depending on the number of customers in the system and the reward that the customer will leave, which depends on the customer’s class.

The closest work to this paper are Ziya [18], Caro and Simchi-Levi [2], and Ziya, Ayhan, and Foley [20]. Ziya [18] develops optimal static pricing policies for a large class of queueing systems and proves structural results on the optimal prices. Motivated by a company selling phone cards, Caro and Simchi-Levi [2] prove structural results on optimal prices in an M/G/s/s system. On the other hand, Ziya, Ayhan, and Foley [20] are interested in optimal pricing policies for finite capacity systems with a single customer class. They give a lower bound on the optimal price for finite capacity systems with general arrival and service processes, prove that optimal prices are either monotone increasing or decreasing in \(m\) in an M/M/1/m system and monotone decreasing in \(s\) in an M/G/s/s system. This paper generalizes results of Ziya, Ayhan, and Foley [20] to systems with multiple customer classes.

In the next section, we describe our model, which assumes very mild conditions on the arrival and service processes. In Section III, we prove some structural results for this general model. In particular, we prove an upper bound on optimal arrival rates, and we give sufficient conditions for the existence of a unique optimal arrival rate vector. Sections IV and V show that these uniqueness conditions hold in M/M/1/m and M/G/s/s systems, respectively and provide structural results on the optimal arrival rates. Section VI contains our concluding remarks. Finally, proofs omitted in the text appear in the Appendix.

II. MODEL DESCRIPTION

Let \(N(t)\) denote the random number of customer arrivals during \((0, t]\). We assume that \(0 \leq N(t) < \infty\) for all \(t\) and \(N(t)/t\) converges to a strictly positive finite number \(\Lambda\), the maximal arrival rate. An arriving customer belongs to class \(i \in \mathcal{C} = \{1, 2, \ldots, I\}\) with probability \(p_i\) so that the maximal arrival rate for customer class \(i\) is \(\Lambda_i = \Lambda p_i\). Service times are i.i.d. random variables with c.d.f. \(G(\cdot)\) and mean \(\mu\), \(0 < \mu < \infty\). Hence, service times are independent of
customers’ class identities. There are \( s \) identical servers and the maximum allowable number of customers in the system at any given time is \( m \). Clearly, \( m \geq s \). No class has priority over the others. All customers are served in a First-Come-First-Served fashion. Customers who find the system full are lost.

We control the arrival rate for each customer class in the following manner. If the arrival rate of class \( i \) customers is set to be \( \lambda_i \), each arriving customer is allowed to enter the system with probability \( \lambda_i / \Lambda_i \) provided the system is not full. The admission probability \( \lambda_i / \Lambda_i \) can be interpreted as the probability that the arriving customer is willing to pay the price charged. Let \( y_i(\lambda) \) denote the price charged to achieve the arrival rate \( \lambda \) for class \( i \) customers. For example, if customers’ reservation prices are i.i.d. with a strictly increasing cumulative distribution function \( F_i(\cdot) \), we have \( y_i(\lambda) = F_i^{-1}(1 - \lambda / \Lambda_i) \) where \( F_i^{-1}(\cdot) \) is the inverse of \( F_i(\cdot) \). With this relationship between the prices and the arrival rates, setting the arrival rates is equivalent to setting the prices.

We define \( Q_i(\lambda) \) as the revenue rate function for class \( i \) customers for an arrival rate of \( \lambda \). To be more precise,

\[
Q_i(\lambda) = \lambda y_i(\lambda).
\]

Then, we make the following assumption:

**Assumption AII.1** For each \( i \in \mathcal{C} \), \( Q_i : [0, \Lambda_i] \to \mathbb{R} \) is continuously differentiable and concave. Furthermore, \( \lambda_i^{\infty} \) satisfies the first order condition for \( Q_i(\cdot) \) where \( \lambda_i^{\infty} \) is the unique maximizer for \( Q_i(\cdot) \).

The concavity assumption corresponds to having decreasing marginal revenue (with respect to demand) for each customer class and is a fairly common assumption. (For example, see Gallego and van Ryzin [5]. See also Ziya, Ayhan, and Foley [19] for a comparison of such common assumptions from the literature). The second part of the assumption is used to simplify the proofs and the presentation of the results. Obvious generalizations of our results can be proven without the second part of the assumption.

Define \( N_i(\lambda_i, t) \) as the number of class \( i \) arrivals who are admitted to the system by time \( t \) (but may be blocked if the system is full at the time of their arrival). Also, let \( N_i^B(\lambda, s, m, t) \) denote the number of class \( i \) customers who are allowed to enter but are blocked due to the

\(^1\)Service times can be allowed to be class dependent if \( m = s \) and the arrival process is Poisson.
limiting capacity $m$ during $[0, t]$. We assume that

$$
\lim_{t \to \infty} \frac{N_i^B(\lambda, s, m, t)}{N_i(\lambda_i, t)} = B_i(\lambda, s, m) \text{ a.s.} \tag{1}
$$

We refer to $B_i(\lambda, s, m)$ as the blocking probability for class $i$ customers. Since service requirements of the customers are independent of their class identities, it can easily be shown that the blocking probability for each class is the same for a given arrival rate vector $\lambda$ and depends on the arrival rates only through the total arrival rate $\sum_{i=1}^{I} \lambda_i$. Therefore, with a slight abuse of notation, we also define $B(\rho, s, m)$ as the blocking probability for all customer classes where $\rho = \sum_{i=1}^{I} \lambda_i / \mu$ is the traffic load. We choose to define $B(\cdot)$ as a function of the traffic load rather than the total arrival rate for mathematical convenience. A sufficient condition for (1) to exist is that the arrival process is stationary and ergodic, with probability 1 at most one customer departs at any time and the departure time of a served customer does not coincide with an arrival; see Franken et al. [4].

Clearly, the long-run average revenue $R(\lambda, s, m)$ has the form

$$
R(\lambda, s, m) = \sum_{i=1}^{I} Q_i(\lambda_i) \left[1 - B(\rho, s, m)\right]. \tag{2}
$$

Our objective is to maximize $R(\lambda, s, m)$ by choosing the arrival rates $\lambda_i$ from the interval $[0, \Lambda_i]$ for $i \in \mathcal{C}$.

### III. Structural Results

In this section, we prove some results for the general model described in Section II. In the following sections, we use these results to prove some structural properties for the M/M/1/m and M/G/s/s queues.

#### A. An Upper Bound on the Optimal Arrival Rates

Recall that $\lambda_i^\infty$ is the unique maximizer for $Q_i(\cdot)$. Hence, it is also the optimal arrival rate for class $i$ customers when the system waiting room capacity is infinite. The following proposition states that $\lambda_i^\infty$ is an upper bound on the arrival rate for class $i$ customers in any optimal solution to $R(\lambda, s, m)$.

**Proposition III.1** Let $\lambda^*(s, m)$ be an optimal solution to $R(\lambda, s, m)$. Then, $\lambda^*_i(s, m) \leq \lambda_i^\infty$ for all $i \in \mathcal{C}$. 
Proof: Let $i \in C$. We have $Q_i(\lambda_i^\infty) > Q_i(\xi)$ for any $\xi \in [0, \Lambda_i]$ and $\xi \neq \lambda_i^\infty$. From Proposition 2.1 of Ziya, Ayhan, and Foley [21], we also have

$$1 - B(\lambda_1, \lambda_2, \cdots, \lambda_i^\infty, \cdots, \lambda_I, s, m) \geq 1 - B(\lambda_1, \lambda_2, \cdots, \xi, \cdots, \lambda_I, s, m)$$

for $\xi > \lambda_i^\infty$ (\xi corresponds to the $i^{th}$ component in the $B(\cdot)$ term on the right hand side). Thus, we have

$$Q_i(\lambda_i^\infty)(1 - B(\lambda_1, \lambda_2, \cdots, \lambda_i^\infty, \cdots, \lambda_I, s, m)) > Q_i(\xi)(1 - B(\lambda_1, \lambda_2, \cdots, \xi, \cdots, \lambda_I, s, m))$$

for $\xi > \lambda_i^\infty$. Also, for any $j \in \{1, 2, \cdots, I\}$ and $j \neq i$,

$$Q_j(\lambda_j)(1 - B(\lambda_1, \lambda_2, \cdots, \lambda_i^\infty, \cdots, \lambda_I, s, m)) \geq Q_j(\lambda_j)(1 - B(\lambda_1, \lambda_2, \cdots, \xi, \cdots, \lambda_I, s, m))$$

for $\xi > \lambda_i^\infty$. Hence, the result follows. \qed

B. An Alternative Form for the Objective Function

In this section, we give an alternative expression for the objective function given in (2). This new expression will mainly help us prove some structural results in the following sections. However, it also suggests an alternative method to standard non-linear search algorithms if one is interested in finding optimal arrival rates.

Let $Q(\theta) : [0, \Lambda] \to \mathbb{R}$ denote the optimal value of the following problem:

$$\max \sum_{i=1}^{I} Q_i(\lambda_i)$$

subject to

$$\sum_{i=1}^{I} \lambda_i = \theta,$$

$$\lambda_i \geq 0 \text{ for } i \in \{1, 2, \cdots, I\}.$$

Then, $Q(\rho_0\mu>(1 - B(\rho_0, s, m))$ is the optimal revenue for a fixed value of $\rho = \rho_0$ (or a fixed total arrival rate of $\rho_0\mu$). Now, define a new function $R_0(\rho, s, m) : [0, \Lambda/\mu] \to \mathbb{R}$ as

$$R_0(\rho, s, m) = Q(\rho\mu)(1 - B(\rho, s, m)).$$

Then, we have

$$\max_{\rho} R_0(\rho, s, m) = \max_{\lambda} R(\lambda, s, m).$$
Hence, we can now express the objective as a single dimensional function. Although one still needs to deal with the \( I \)-dimensional problem (3) to evaluate \( R_0(\cdot) \), this alternative formulation will be very useful in proving our results. Note that this equivalence also suggests a method for finding optimal arrival rates assuming that there is an expression for the blocking probability \( B(\cdot) \). A line search over \( R_0(\cdot) \) would give \( \rho^* \), the optimal value for \( \rho \). Then, solving (3) for \( \theta = \mu \rho^* \) gives the optimal arrival rates. The problem (3) is a standard non-linear knapsack problem, which has received a lot of attention in the literature (e.g., see Zipkin [17] and Bitran and Hax [1]).

The following result immediately follows from Proposition III.1.

**Corollary III.1** Any optimal solution to \( R_0(\rho, s, m) \) lies in the interval \([0, \rho^\infty] \) where \( \rho^\infty = \sum_{i=1}^{I} \lambda^\infty_i / \mu \).

Hence, in order to find an optimal solution we can restrict ourselves to the interval \([0, \rho^\infty] \).

**C. Characterizing the Optimal Traffic Load**

In this section, we give some conditions under which there exists a unique value for the traffic load that maximizes \( R_0(\cdot) \) and characterize this optimal traffic load under these conditions. First, we make the following assumption on the blocking probability \( B(\cdot) \).

**Assumption AIII.1** \( B(\rho, s, m) \) is twice differentiable with respect to \( \rho \) for \( \rho > 0 \).

While it is not difficult to come up with examples where this assumption does not hold, it is satisfied for most of the standard queueing systems where the blocking probability can be computed (e.g. M/M/s/m queue).

Using the fact that \( Q_i(\cdot) \) is continuously differentiable and concave for all \( i \) (Assumption II.1), we can prove the following result.

**Corollary III.2** \( Q(\theta) \) is continuously differentiable and strictly concave for \( 0 < \theta \leq \mu \rho^\infty \). Furthermore, \( Q'(\theta) \) is differentiable a.e. for \( 0 < \theta \leq \mu \rho^\infty \) where \( Q'(\theta) \) denotes the first derivative function of \( Q(\theta) \).

The first part follows almost directly from the corollary on page 38 of Zipkin [17] and the second part immediately follows from the first part. Therefore, the proof is omitted.
Corollary III.2 implies that $R_0(\rho, s, m)$ is differentiable and twice differentiable a.e. in $\rho$ over the interval $[0, \rho^\infty]$. Note that this is the interval where any optimal solution of $R_0(\rho, s, m)$ lies (see Corollary III.1). Hence, any optimal solution for $R_0(\rho, s, m)$ has to satisfy the first order condition. Next, we give conditions that ensure that there is a unique solution to the first order condition. First, for ease of exposure, we define three new functions, $\Psi(\cdot)$, $\Gamma_1(\cdot)$, and $\Gamma_2(\cdot)$.

$\Psi(\rho) = \frac{\mu Q'(\mu\rho)}{Q(\mu)}$ for $0 < \rho \leq \rho^\infty$,

$\Gamma_1(\rho, s, m) = \frac{B'(\rho, s, m)}{1 - B(\rho, s, m)}$ for $\rho > 0$,

$\Gamma_2(\rho, s, m) = -\frac{B''(\rho, s, m)}{2B'(\rho, s, m)}$ for $\rho > 0$,

where $B'(\cdot)$ and $B''(\cdot)$ denote the first and second derivatives of $B(\cdot)$ with respect to $\rho$.

**Proposition III.2** Suppose that the following conditions are satisfied:

(i) $B'(\rho, s, m) > 0$ for $\rho > 0$,

(ii) $\Gamma_1(\rho, s, m) > \Gamma_2(\rho, s, m)$ for $\rho > 0$.

Then, $\Psi(\rho) = \Gamma_1(\rho, s, m)$ has a unique solution $\rho^*(s, m)$ and $\rho^*(s, m)$ is the unique maximizer for $R_0(\rho, s, m)$.

The first condition of Proposition III.2 requires that the blocking probability is strictly increasing in the traffic load. The condition can be shown to hold for M/M/s/m systems. Under more general conditions, Ziya, Ayhan, and Foley [21] proves the non-strict version of this result for a given service rate. The second condition is more technical in nature and it is more difficult to establish. In the following sections, we prove that it holds for M/M/1/m and M/G/s/s systems.

**IV. The M/M/1/m System**

We start by showing that for the M/M/1/m system, there is a unique value for the traffic load maximizing the revenue.

**Proposition IV.1** For the M/M/1/m system, there exists a unique optimal solution $\rho^*(1, m)$ that maximizes $R_0(\rho, 1, m)$ and $\rho^*(1, m)$ is the unique solution to $\Psi(\rho) = \Gamma_1(\rho, 1, m)$.

This result is established mainly by showing that the M/M/1/m system satisfies the conditions of Proposition III.2. One can also show that under Assumption II.1, there is a unique arrival rate.
vector for a given value of the traffic load. Hence, Proposition IV.1 also implies that there is a unique optimal arrival rate vector maximizing the long-run revenue rate.

We next investigate how the changes in the waiting room capacity affect the optimal traffic intensity and prices. It turns out that there is a monotonic structure, however, the direction of the monotonicity depends on the value of $\Psi(1)$ and $\rho^\infty$, which do not depend on the waiting room capacity $m$.

**Proposition IV.2** For the $M/M/1/m$ system, we have the following:

(i) If $\rho^\infty \geq 1$ and $\Psi(1) = 1/2$, then $\rho^*(1, m) = 1$ for all $m \geq 1$.

(ii) If $\rho^\infty \geq 1$ and $\Psi(1) > 1/2$, then $\rho^*(1, m + 1) \geq \rho^*(1, m) > 1$ for all $m \geq 1$.

(iii) If $\rho^\infty < 1$ or $\Psi(1) < 1/2$, then $1 < \rho^*(1, m + 1) \leq \rho^*(1, m)$ for all $m \geq 1$.

Both $\rho^\infty$ and $\Psi(1)$ (when $\rho^\infty \geq 1$) can be viewed as measures of sufficiency of the service capacity of the system relative to the demand. Recall that $\rho^\infty$ is the optimal traffic load when the service capacity is infinite and $\Psi(1) = \mu Q'(\mu)/Q(\mu)$ where $Q(\mu)$ is given by (3). Hence, higher values of $\rho^\infty$ and/or $\Psi(1)$ indicate that there is more to gain from increasing the service rate. Then, according to Proposition IV.2, when the service capacity is low relative to the demand, as the waiting room capacity increases, the optimal (offered) load on the system also increases. On the other hand, when the service capacity is relatively sufficient to utilize the potential demand, the optimal load decreases with the waiting room capacity. When the number of servers is more than one, there are examples that show that the optimal traffic load is not necessarily monotone in the waiting room capacity $m$.

Proposition IV.2 generalizes Proposition 4.2. of Ziya, Ayhan, and Foley [20] to multiple customer classes. When there is a single customer class, it can be shown that

$$\Psi(1) = (>)(<)1/2 \iff \Lambda/\mu = (<)(>)\rho^c;$$

where $\rho^c$ is a constant that only depends on the reservation price distribution of the customers. The equivalent conditions on the right hand side demonstrate the dependence on the demand level with respect to the service capacity more clearly.

Note that it is also easy to show that the optimal arrival rate for each customer class has the same monotonic structure as the optimal traffic load.
V. THE M/G/s/s SYSTEM

We first establish that there is a unique optimal traffic load for the M/G/s/s system.

**Proposition V.1** For the M/G/s/s system, there exists a unique optimal solution $\rho^*(s, s)$ that maximizes $R_0(\rho, s, s)$ and $\rho^*(s, s)$ is the unique solution to $\Psi(\rho) = \Gamma_1(\rho, s, s)$.

To prove this result, we show that for the M/G/s/s system, conditions of Proposition III.2 hold for $s \geq 2$. For $s = 1$, the result directly follows from Proposition IV.1.

We next investigate how the optimal load $\rho^*(s, s)$ changes with the number of servers $s$.

**Proposition V.2** For the M/G/s/s system, we have $\rho^*(s + 1, s + 1) \geq \rho^*(s, s)$ for $s \geq 1$.

Proposition V.2 simply states that the optimal load is monotone non-decreasing in the number of servers. This differs from the effect of the waiting room capacity on the M/M/1/m system, which depends on the level of the demand on the system relative to the service capacity. Proposition V.2 also implies that optimal arrival rate for each class is increasing in the number of servers.

For the M/G/s/s system, generalized versions of Propositions V.1 and V.2 can be proven if service times are allowed to depend on the class identities of the customers. Suppose that the service time for class $i$ customers are independent and identically distributed with mean $1/\mu_i$. In this case, it can be shown that the blocking probability depends on the arrival rates only through the expression $\sum_{i=1}^{I} \lambda_i/\mu_i$. Now, redefine $Q(\theta)$ as the optimal value of the following problem:

$$
\max \sum_{i=1}^{I} Q_i(\lambda_i)
$$

subject to

$$
\sum_{i=1}^{I} \lambda_i/\mu_i = \theta,
$$

$$
\lambda_i \geq 0 \text{ for } i \in \{1, 2, \ldots, I\}.
$$

Also redefine $R_0(\rho, s, m)$ as

$$
R_0(\theta, s, m) = Q(\theta)(1 - B(\theta, s, m))
$$

where $\theta = \sum_{i=1}^{I} \lambda_i/\mu_i$. Then, we get

$$
\max_{\theta} R_0(\theta, s, m) = \max_{\lambda} R(\lambda, s, m).
$$
Then, as before, we can show that Propositions V.1 and V.2 hold for this more general model by redefining $\Psi(\cdot)$ and $\Gamma_1(\cdot)$ accordingly and by replacing $\rho^*(s, s)$ by $\theta^*(s, s)$ where $\theta^*(s, s)$ represents the optimal value of $\sum_i \lambda_i/\mu_i$.

Note that, independently of this work, Caro and Simchi-Levi [2] have also proven Propositions V.1 and V.2. Same results also appeared in the dissertation of the first author of this paper (see [18]).

VI. SUMMARY AND CONCLUSIONS

In this paper, we investigated optimal static pricing policies for a finite capacity queueing system serving customers from different classes. Our results generalize the results of Ziya et.al. [20] to a multiple customer class setting. We first derived an upper bound on the optimal arrival rates for a large class of queueing systems and then developed sufficient conditions for the existence and uniqueness of an optimal solution. We showed that these conditions are satisfied for M/M/1/m and M/G/s/s systems and proved structural results on the relationships between the optimal arrival rates and the system capacity.

REFERENCES

APPENDIX

Proof of Proposition III.2: Using Corollary III.1, it is sufficient to consider the interval \((0, \rho^\infty]\) only. Hence, in the following, we restrict ourselves to this interval. The objective function \(R_0(\rho, s, m)\) is differentiable and twice differentiable a.e. in \(\rho\). Let \(R'_0(\cdot)\) and \(R''_0(\cdot)\) denote the first and the second derivative functions for (4), respectively. Then, we have

\[
R'_0(\rho, s, m) = \mu Q'(\mu \rho)(1 - B(\rho, s, m)) - Q(\mu \rho) B(\rho, s, m). \tag{5}
\]

From the fact that \(R_0(0, s, m) = 0\) and Corollary III.1, it follows that there exists \(\rho \in (0, \rho^\infty]\) such that \(R'_0(\rho, s, m) = 0\). For such \(\rho\), rewriting (5), we get the first order condition

\[
\Psi(\rho) = \Gamma_1(\rho, s, m). \tag{6}
\]

When it exists, \(R''_0(\rho, s, m)\) can be shown to be equal to

\[
R''_0(\rho, s, m) = \mu^2 Q''(\mu \rho)(1 - B(\rho, s, m)) - 2\mu Q'(\mu \rho) B'(\rho, s, m) - Q(\mu \rho) B''(\rho, s, m). \tag{7}
\]

We know from Corollary III.2 that \(Q''(\mu \rho) < 0\). Then, \(R''_0(\rho, s, m) < 0\) if \(-2\mu Q'(\mu \rho) B'(\rho, s, m) - Q(\mu \rho) B''(\rho, s, m) \leq 0\). Since \(Q(\mu \rho) > 0\) and \(B'(\rho, s, m) > 0\) for \(\rho > 0\), we can show that

June 9, 2006 DRAFT
\[ \Psi(\rho) \geq \Gamma_2(\rho, s, m) \Rightarrow R''_\rho(\rho, s, m) < 0. \] Then, the result follows from (6) and the assumption that \( \Gamma_1(\rho, s, m) > \Gamma_2(\rho, s, m) \) for \( \rho > 0. \)

**Proof of Proposition IV.1:** First, let \( m = 1. \) Then, the first order condition \( \Psi(\rho) = \Gamma_1(\rho, 1, 1) \) can also be written as \( \xi(\rho) = 0 \) where \( \xi(\rho) = Q(\mu\rho) - \mu(1 + \rho)Q'(\mu\rho). \) Let \( \xi'(\rho) \) denote the first derivative of \( \xi(\rho) \) with respect to \( \rho \) when it exists. Then, \( \xi'(\rho) = -\mu^2(1 + \rho)Q''(\mu\rho). \) Since \( Q(\cdot) \) is concave (by Corollary III.2), it follows that \( \xi'(\rho) > 0 \) for \( \rho > 0. \) Hence, \( \Psi(\rho) = \Gamma_1(\rho, 1, 1) \) has a unique solution.

Now, let \( m \geq 2. \) It is sufficient to prove that \( B(\rho, 1, m) \) is twice differentiable and the conditions of Proposition III.2 are satisfied. Since the system is an M/M/1/m system, the expression for \( B(\rho, 1, m) \) is known:

\[
B(\rho, 1, m) = \begin{cases} \frac{\rho^m - \rho^{m+1}}{1 - \rho^{m+1}} & \text{if } \rho \neq 1, \\ \frac{1}{m+1} & \text{if } \rho = 1. \end{cases}
\] (7)

The proof of twice differentiability is straightforward and thus is omitted. The first derivative is given by

\[
B'(\rho, 1, m) = \begin{cases} \frac{\rho^m - (m + 1)\rho^m + m}{(1 - \rho^{m+1})^2} & \text{if } \rho \neq 1, \\ \frac{m}{2m+2} & \text{if } \rho = 1. \end{cases}
\] (8)

In order to prove the first condition of Proposition III.2, from (8), it is sufficient to show that \( \rho^{m+1} - (m + 1)\rho + m > 0 \) for all \( \rho. \) Let \( \Phi(\rho) = \rho^{m+1} - (m + 1)\rho + m. \) Then, \( \Phi'(\rho) = (m + 1)\rho^m - (m + 1). \) Thus, \( \Phi(\rho) \) is decreasing for \( 0 \leq \rho < 1 \) and increasing for \( \rho > 1. \) Since we have \( \Phi(1) = 0, \) it follows that \( \rho^{m+1} - (m + 1)\rho + m > 0 \) for all \( \rho. \)

Finally, we prove the second condition of Proposition III.2. First, let \( \rho \neq 1. \) Then, using (7) and (8) and after some algebraic manipulations, we have

\[
\Gamma_1(\rho, 1, m) - \Gamma_2(\rho, 1, m) = \frac{-m(\sum_{n=1}^{m-1} nm - n^2)\rho^{m-n-1}(\rho^{m+1} - 1)^2(\rho - 1)^3}{2\rho(1 - \rho^{m+1})^2(1 - \rho^m)(m + \rho^m + 1 - (m + 1)\rho)}.
\]

Then, using the fact that \( (m + \rho^{m+1} - (m + 1)\rho) > 0, \) it follows that \( \Gamma_1(\rho, 1, m) - \Gamma_2(\rho, 1, m) > 0 \) for \( \rho > 0 \) and \( \rho \neq 1. \) For \( \rho = 1, \) we get \( \Gamma_1(1, 1, m) - \Gamma_2(1, 1, m) = \frac{m-1}{6} > 0 \) for \( m \geq 2. \) Hence, the result follows.

**Proof of Proposition IV.2**

(i) The result immediately follows from Proposition IV.1 and Lemma A.1 (i).

(ii) In the following, using Corollary III.1, we restrict attention to the interval \([0, \rho^\infty).\) Using (5), it can be shown that for any \( m \geq 1, \) \( \Psi(\rho) > \Gamma_1(\rho, 1, m) \) for \( 0 < \rho < \rho^*(1, m). \) From
Corollary III.2, we know that $\Psi(\rho)$ is decreasing in $\rho$ and from Proposition IV.1, we know that $\Psi(\rho) = \Gamma_1(\rho, 1, m)$ has a unique solution. Since $\Psi(1) > 1/2$ and $\Gamma_1(1, 1, m) = 1/2$ for $m \geq 1$, we conclude that $\rho^*(1, m) > 1$ for all $m \geq 1$. Then, from Lemma A.1 (ii), we conclude that $\rho^*(1, m)$ is decreasing in $m$.

(iii) In the following, using Corollary III.1, we restrict our attention to the interval $[0, \rho^\infty)$. Using (5), it can be shown that for any $m \geq 1$, $\Psi(\rho) > \Gamma_1(\rho, 1, m)$ for $0 < \rho < \rho^*(1, m)$. From Corollary III.2, we know that $\Psi(\rho)$ is decreasing in $\rho$ and from Proposition IV.1, we know that $\Psi(\rho) = \Gamma_1(\rho, 1, m)$ has a unique solution. First, if $\rho^\infty < 1$, it immediately follows from Corollary III.1 that $\rho^*(1, m) < 1$ for all $m \geq 1$. Now, suppose that $\rho^\infty \geq 1$ and $\Psi(1) < 1/2$. Since $\Psi(1) < 1/2$ and $\Gamma_1(1, 1, m) = 1/2$ for $m \geq 1$, we conclude that $\rho^*(1, m) < 1$ for all $m \geq 1$. Then, from Lemma A.1 (iii), we conclude that $\rho^*(1, m)$ is increasing in $m$.

**Proof of Proposition V.1**

For $s = 1$, the result immediately follows from Proposition IV.1 since the expression for the blocking probability is the same. Suppose that $s \geq 2$. It is sufficient to prove that $B(\rho, s, s)$ is twice differentiable and the conditions of Proposition III.2 are satisfied. Since the system is an M/G/s/s system, the expression for $B(\rho, s, s)$ is known to be $B(\rho, s, s) = \frac{\rho^s / s!}{\sum_{i=0}^{\infty} \rho^i / i!}$. Clearly, $B(\rho, s, s)$ is differentiable with respect to $\rho$. Then, we have from Lemma A.2 that $B'(\rho, s, s) = (1 - B(\rho, s, s))(B(\rho - 1, s - 1) - B(\rho, s, s))$. Hence, $B(\rho, s, s)$ is twice differentiable in $\rho$.

Next, we prove that the first condition of Proposition III.2 holds. For some fixed $\rho > 0$, we have $\lim_{s \to \infty} E(\rho, s, s) = 0$. Suppose for contradiction that $B'(\tilde{\rho}, \bar{s}, \bar{s}) \leq 0$ for some $\tilde{\rho} > 0$ and $\bar{s} \geq 1$. Then, using Lemma A.2, we have $B(\tilde{\rho}, \bar{s} - 1, \bar{s} - 1) - B(\tilde{\rho}, \bar{s}, \bar{s}) \leq 0$. Since $B(\rho, s, s)$ is known to be convex in $s$ (see, e.g., Messerli [11]), this implies that $B(\tilde{\rho}, \bar{s} - 1, \bar{s} - 1) - B(\tilde{\rho}, s, s) \leq 0$ for any $s \geq \bar{s}$. Then, we have $B(\tilde{\rho}, s, s) > B(\tilde{\rho}, \bar{s}, \bar{s})$ for any $s > \bar{s}$. This is a contradiction to the fact that $\lim_{s \to \infty} B(\tilde{\rho}, s, s) = 0$. Thus, $B'(\rho, s, s) > 0$ for $\rho > 0$ and $s \geq 1$.

Now, we prove that the second condition of Proposition III.2 holds. It can be shown that $\frac{d}{d\tilde{\rho}} \left(-\ln(1 - B(\rho, s, s))\right) = \frac{B'(\rho, s, s)}{1 - B(\rho, s, s)} = \Gamma_1(\rho, s, s)$ and $\frac{d}{d\tilde{\rho}} \left(-\frac{1}{2} \ln(B'(\rho, s, s))\right) = -\frac{B''(\rho, s, s)}{2B(\rho, s, s)} = \Gamma_2(\rho, s, s)$. Then, it follows that $\Gamma_1(\rho, s, s) - \Gamma_2(\rho, s, s) = \frac{1}{2} \frac{d}{d\tilde{\rho}} \left(\ln\left(\frac{B'(\rho, s, s)}{(1 - B(\rho, s, s))^2}\right)\right) > 0$. Thus, we have

$\Gamma_1(\rho, s, s) - \Gamma_2(\rho, s, s) > 0 \iff \frac{d}{d\tilde{\rho}} \left(\ln\left(\frac{B'(\rho, s, s)}{(1 - B(\rho, s, s))^2}\right)\right) > 0$. Since $\ln(\cdot)$ is a strictly increasing function, we also have $\Gamma_1(\rho, s, s) - \Gamma_2(\rho, s, s) > 0 \iff \frac{d}{d\tilde{\rho}} \left(\frac{B'(\rho, s, s)}{(1 - B(\rho, s, s))^2}\right) > 0$. Now, from Lemma A.2, we have $\frac{B'(\rho, s, s)}{(1 - B(\rho, s, s))^2} = \frac{B(\rho, s - 1, s - 1) - B(\rho, s, s)}{1 - B(\rho, s, s)}$. It follows that $\frac{B'(\rho, s, s)}{(1 - B(\rho, s, s))^2} = 1 - \frac{\Gamma_1(\rho, s, s) - \Gamma_2(\rho, s, s)}{1 - B(\rho, s, s)}$. Then, we have $\Gamma_1(\rho, s, s) - \Gamma_2(\rho, s, s) > 0 \iff \frac{d}{d\tilde{\rho}} \left(\frac{1 - B(\rho, s - 1, s - 1)}{1 - B(\rho, s, s)}\right) < 0$. Taking the
derivative, we get \( \frac{d}{d\rho} \left( \frac{1-B(\rho,s-1,s-1)}{1-B(\rho,s,s)} \right) = \frac{-B'(\rho,s-1,s-1)(1-B(\rho,s,s))}{(1-B(\rho,s,s))^2} + \frac{(1-B(\rho,s-1,s-1))B'(\rho,s,s)}{(1-B(\rho,s,s))^2} \). Then, 
\[ \frac{d}{d\rho} \left( \frac{1-B(\rho,s-1,s-1)}{1-B(\rho,s,s)} \right) < 0 \iff (1-B(\rho,s-1,s-1))B'(\rho,s,s) < B'(\rho,s-1,s-1)(1-B(\rho,s,s)) \iff \frac{B'(\rho,s,s)}{1-B(\rho,s,s)} < \frac{B'(\rho,s-1,s-1)}{1-B(\rho,s-1,s-1)} \iff B(\rho,s-1,s-1) - B(\rho,s,s) < B(\rho,s-2,s-2) - B(\rho,s-1,s-1).
\]

Note that the last equivalence follows from Lemma A.2. Finally, using the fact that \( B(\rho,s,s) \) is convex in \( s \) for fixed \( \rho \) (see, e.g., Messerli [11]), we conclude that \( \Gamma_1(\rho,s,s) > \Gamma_2(\rho,s,s) \) for \( \rho > 0 \).

**Proof of Proposition V.2:** The proof immediately follows from Proposition V.1, Lemma A.2 together with the fact that \( B(\rho,s,s) \) is convex in \( s \) for fixed \( \rho \) (see, e.g., Messerli [11])

**Lemma A.1** Suppose that interarrival and service times are exponentially distributed and let \( m \geq 1 \). Then:

(i) \( \Gamma_1(\rho,1,m) = 1/2 \) for \( \rho = 1 \).

(ii) \( \Gamma_1(\rho,1,m+1) > \Gamma_1(\rho,1,m) \) for \( \rho > 1 \).

(iii) \( \Gamma_1(\rho,1,m+1) < \Gamma_1(\rho,1,m) \) for \( \rho < 1 \).

**Proof:** (i) Using (7) and (8), it can be shown that \( \Gamma_1(1,1,m) = 1/2 \).

(ii) Using (7) and (8), after some algebra, it can be shown that
\[
\Gamma_1(\rho,1,m+1) - \Gamma_1(\rho,1,m) = -\rho^{m-1}(1-\rho)^4 \sum_{k=1}^{m} k(m-k+1)\rho^{m-k} \frac{1}{(1-\rho^m)(1-\rho^{m+1})(1-\rho^{m+2})}.
\]

Then we have \( \Gamma_1(\rho,1,m+1) - \Gamma_1(\rho,1,m) > 0 \) if \( \rho > 1 \).

(iii) From (9), it follows that \( \Gamma_1(\rho,1,m+1) - \Gamma_1(\rho,1,m) < 0 \) if \( \rho < 1 \).

**Lemma A.2** \( B'(\rho,s,s)/(1-B(\rho,s,s)) = B(\rho,s-1,s-1) - B(\rho,s,s) \) for \( s \geq 1 \).

**Proof:** Let \( S(\rho,c) \) be defined as \( S(\rho,s) = \sum_{i=0}^{s} \rho^i/i! \). Then, \( B(\rho,s,s) = \frac{\rho^s/s!}{S(\rho,s)} \), and \( B'(\rho,s,s) = \frac{(\rho^{s-1}/(s-1)!)(S(\rho,s) - (\rho^s/s!))S(\rho,s-1)}{(S(\rho,s))^2} \). Using \( 1-B(\rho,s,s) = S(\rho,s-1)/S(\rho,s) \), it follows that
\[
B'(\rho,s,s) = \frac{(\rho^{s-1}/(s-1)!)(S(\rho,s) - (\rho^s/s!))S(\rho,s-1)}{(S(\rho,s))^2} \frac{S(\rho,s)}{S(\rho,s-1)} = \frac{\rho^{s-1}/(s-1)!}{S(\rho,s-1)} - \frac{\rho^s/s!}{S(\rho,s)} = B(\rho,s-1,s-1) - B(\rho,s,s).
\]

\[ \square \]