A Note on Optimal Pricing for Finite Capacity Queueing Systems with Multiple Customer Classes

Serhan Ziya,[1] Hayriye Ayhan,[2] Robert D. Foley[2]

[1] *Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599*

[2] *H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30318*

Received 6 August 2007; revised 10 March 2008; accepted 15 March 2008
DOI 10.1002/nav.20295
Published online 5 May 2008 in Wiley InterScience (www.interscience.wiley.com).

Abstract: This article investigates optimal static prices for a finite capacity queueing system serving customers from different classes. We first show that the original multi-class formulation in which the price for each class is a decision variable can be reformulated as a single dimensional problem with the total load as the decision variable. Using this alternative formulation, 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. © 2008 Wiley Periodicals, Inc. Naval Research Logistics 55: 412–418, 2008

Keywords: pricing in queueing systems; finite capacity queues; admission control

1. INTRODUCTION

This article 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.

Pricing problem for finite capacity queues arises in different contexts including telecommunication networks (see [1, 3]) and on-line service providers (see [8]). Despite the increasing interest in Internet pricing, however, there has not appeared much work on pricing for finite capacity queues, in particular on static pricing policies. In theory, under ideal conditions, dynamic pricing policies would outperform static prices. In practice, however, dynamic pricing is not always preferred mainly because of its potential negative effects on the customer demand in the long run. Furthermore, static prices are easier to administer. Hence, for various practical concerns, static prices are widely used. In fact, within the queueing context, it has also been shown that even when the long-term negative effects of dynamic pricing are ignored, static prices perform quite well (see [8, 9]). Thus, the analysis of static prices, in particular within the queueing context, are highly relevant.

The objective of this article is to generalize the single-class results of Ziya, Ayhan, and Foley [14] to the multiple-class setting. Ziya et al. consider a similar queueing setting with a single customer class and prove several structural results on the relationships between the optimal price and system parameters. For the M/M/1/m queue, they show that under certain conditions on the customers' reservation price distribution, the optimal price is either increasing or decreasing in m, depending on whether the offered load is greater than a constant which is determined by the reservation price distribution alone. For the M/G/s/s queue, they show that the optimal price is decreasing in s. In this article, we show that both results can be generalized to the multiple class setting under slightly different conditions. While the generalized version of the M/G/s/s result of Ziya et al. is clear, the generalization of the M/M/1/m result is not since it is unclear how the condition that determines the direction of the monotonic behavior of the optimal price would generalize. In the following, we give a precise description of the generalized version of this condition.

Correspondence to: S. Ziya (ziya@unc.edu)
This article contains supplementary material available via the internet at http://www.interscience.wiley.com/jpages/0894-069X/suppmat.

To prove our results, we first show that our original multi-dimensional formulation that treats the arrival rate for each class as a decision variable (described in Section 2) can be reformulated as a one-dimensional problem with the total load as the decision variable. This alternative formulation is only one-dimensional on the surface since it requires solving a multi-dimensional knapsack problem. Nevertheless, this equivalence proves to be crucial in establishing our results. For this alternative formulation, we then provide sufficient conditions under which there is a unique value for the optimal load, and we show that these conditions are satisfied for the M/M/1/m and M/G/s/s queues. Finally, we establish the monotonicity properties. Note that the dimension reduction method that we are using in this paper has previously been used outside the queueing context, in the multi-product dynamic pricing literature. For example, see Maglaras and Meissner [6] and Talluri and van Ryzin [10]. Within the queueing context, Carrizosa, Conde, and Muñoz-Márquez [2] also use a dimension reduction method in order to analyze an admission control problem for an M/G/s/s queue, but technically their approach and formulation are quite different.

Although pricing for queueing systems has received significant attention, few articles have specifically addressed finite capacity queues. Apart from [14], the closest work to this article is Caro and Simchi-Levi [1]. Caro and Simchi-Levi are motivated by the pricing problem of a company selling phone cards. In their model, there are multiple call classes all sharing a common inbound link with a capacity of N calls and each class k has an outbound link with a capacity of N_k calls. The authors limit themselves to a particular type of admission policy where a call is admitted if there is an available inbound link and an available outbound link. This policy reduces to what the authors call the *complete sharing policy* when $N_k \geq N$ and to the *complete partitioning policy* when $\sum N_k \leq N$. Under either of the two policies, the system reduces to either an M/G/s/s queue or a finite number of independent M/G/s/s/ queues so that the pricing problem of [1] becomes essentially the pricing problem for an M/G/s/s queue. The authors carry out an analysis that is different from the one in this article, prove some structural results on the objective function as well as the optimal prices. One of their results state that the optimal price for each class decreases with the number of servers s, which we also establish in this article. Note that the same result also appeared in the dissertation of the first author of this article (see [12]). For a more extensive review of the existing literature on pricing for finite capacity queues, we refer the reader to [1, 12, 14].

2. MODEL DESCRIPTION

We consider the following multiple-class version of the model analyzed in Ziya et al. [14]. We have a queueing system with s identical servers. The maximum allowable number of customers in the system at any given time is $m \geq s$. Customers who find the system full are lost. We use Λ to denote the *maximal arrival rate*, the arrival rate of customers who may or may not join the system. Differing from Ziya et al. [14], an arriving customer belongs to class $i \in \mathcal{C} = \{1, 2, \ldots, I\}$ with probability α_i so that the *maximal arrival rate for customer class* i is $\Lambda_i = \Lambda\alpha_i$. Service times are independent of the customers' class identities. They are independent and identically distributed (i.i.d.) random variables with a common mean $1/\mu$, $0 < \mu < \infty$. No class has priority over the others. All customers are served in a First-Come-First-Served fashion.

We assume that each customer has a reservation price (the maximum amount the customer would be willing to pay) for the service provided and the reservation prices of the customers are i.i.d. with a strictly increasing, differentiable, cumulative distribution function $F_i(\cdot)$. Let p_i denote the price that is charged to class i customers. Then, the probability that an incoming class i customer will be willing to pay the price is $1 - F_i(p_i)$, and the arrival rate of class i customers who are willing to pay the price is $\lambda_i = \Lambda_i(1 - F_i(p_i))$. In this article, it will be convenient to let the arrival rates λ_j, $j \in \mathcal{C}$ be the decision variables as opposed to the prices p_j, $j \in \mathcal{C}$. Note that the inverse demand function, which simply returns the price that corresponds to a certain arrival rate λ for class i customers is given by $p_i(\lambda) = F_i^{-1}(1-\lambda/\Lambda_i)$ where $F_i^{-1}(\cdot)$ is the inverse of $F_i(\cdot)$.

We define $Q_i(\lambda)$ as the revenue rate function for class i customers for an arrival rate of λ. Thus,

$$Q_i(\lambda) = \lambda p_i(\lambda).$$

We make the following assumption:

ASSSUMPTION 2.1: For each $i \in \mathcal{C}$, $Q_i : [0, \Lambda_i] \to \mathbf{R}$ is continuously differentiable and concave. Furthermore λ_i^∞, the unique maximizer for $Q_i(\cdot)$, satisfies the first order condition for $Q_i(\cdot)$, i.e., $Q_i'(\lambda_i^\infty) = 0$ where $Q_i'(\cdot)$ is the first derivative of $Q_i(\cdot)$.

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 [4]. See also Ziya et al. [13] for a comparison of such common assumptions from the literature).

Since service requirements of the customers are assumed to be independent of their class identities, it can easily be shown that the blocking probability (when it exists)[1] for each class is the same for a given arrival rate vector $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_I)$ and depends on the arrival rates only through the total arrival rate $\sum_{i=1}^{I} \lambda_i$. We 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*.

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

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

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

3. AN ALTERNATIVE FORM FOR THE OBJECTIVE FUNCTION AND AN UPPER BOUND ON THE OPTIMAL SOLUTION

In this section, we first give an alternative expression for the objective function given in (1). This new expression will help us extend the results of Ziya et al. [14] to the multiple class setting.

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

$$\begin{aligned} &\max \sum_{i=1}^{I} Q_i(\lambda_i) \\ &\text{subject to} \\ &\sum_{i=1}^{I} \lambda_i = \theta, \\ &\Lambda_i \geq \lambda_i \geq 0 \text{ for } i \in \{1, 2, \ldots, I\}. \end{aligned} \tag{2}$$

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 $R_0(\rho, s, m) : [0, \Lambda/\mu] \to \mathbf{R}$ as

$$R_0(\rho, s, m) = Q(\rho\mu)(1 - B(\rho, s, m)). \tag{3}$$

Then, we have

$$\max_{\rho} R_0(\rho, s, m) = \max_{\boldsymbol{\lambda}} R(\boldsymbol{\lambda}, s, m).$$

Hence, we can now express the objective as a one-dimensional function. (Note that similar dimension reduction methods have previously been used outside the queueing context, see Maglaras and Meissner [6] and Talluri and van Ryzin [10].)

In the following proposition, we give an upper bound on the optimal arrival rates and the optimal solution to (3). These upper bounds are helpful since they do not depend on the blocking probability and thus apply to any finite capacity queue as long as the blocking probability exists. The proposition is also crucial in proving the results in the following sections since it implies that it is sufficient to limit the search for the optimal solution for (3) to the interval determined by the upper bound. (Proofs of all the results are given in the Online Appendix.)

PROPOSITION 3.1: Let $\boldsymbol{\lambda}^*(s, m) = (\lambda_1^*(s, m), \lambda_2^*(s, m), \ldots, \lambda_I^*(s, m))$ be a maximizer for $R(\boldsymbol{\lambda}, s, m)$. Then, $\lambda_i^*(s, m) \leq \lambda_i^\infty$ for all $i \in \mathcal{C}$. Consequently, any maximizer for $R_0(\rho, s, m)$ lies in the interval $[0, \rho^\infty]$ where $\rho^\infty = \sum_{i=1}^{I} \lambda_i^\infty/\mu$.

The first part of Proposition 3.1 generalizes the single class version of [14] to multiple classes. The result simply states that the maximizer for the system with infinite service and waiting room capacity is an upper bound on the maximizer for any finite capacity system. In a finite capacity system (compared with the infinite capacity system) having high arrival rates is not as valuable since some of the arriving customers will be lost. Therefore, the service provider is better off by charging higher prices (and thereby lowering the arrival rates) instead. The proof of the first part of Proposition 3.1 follows very similarly mainly utilizing Theorem 2.1. of Ziya et al. [15]. The second part of the proposition directly follows from the first part. One important implication of the proposition is that in order to find an optimal solution for (3) we can restrict ourselves to the interval $[0, \rho^\infty]$.

4. 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 $B(\cdot)$.

ASSSUMPTION 4.1: $B(\rho, s, m)$ is twice differentiable with respect to ρ 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).

[1] Blocking probability is the long-run fraction of customers who are blocked. A sufficient condition for the blocking probability 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. [5].

Using the fact that $Q_i(\cdot)$ is continuously differentiable and concave for all i (Assumption 2.1), we can prove the following result.

COROLLARY 4.1: $Q(\theta)$ is continuously differentiable and strictly concave for $0 < \theta \le \mu\rho^\infty$. Furthermore, $Q'(\theta)$ is differentiable a.e. for $0 < \theta \le \mu\rho^\infty$ where $Q'(\cdot)$ denotes the first derivative of $Q(\cdot)$.

The first part follows almost directly from the corollary on page 38 of Zipkin [11] and the second part immediately follows from the first part. Corollary 4.1 implies that $R_0(\rho, s, m)$ is differentiable and twice differentiable a.e. in ρ over the interval $[0, \rho^\infty]$. Note that this is the interval where any optimal solution of $R_0(\rho, s, m)$ lies (see Proposition 3.1). 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 $\Gamma_2(\cdot)$:

$$\Psi(\rho) = \frac{\mu Q'(\mu\rho)}{Q(\mu\rho)} \text{ for } 0 < \rho \le \rho^\infty,$$
$$\Gamma_1(\rho, s, m) = \frac{B'(\rho, s, m)}{1 - B(\rho, s, m)} \text{ for } \rho > 0,$$
$$\Gamma_2(\rho, s, m) = \frac{-B''(\rho, s, m)}{2B'(\rho, s, m)} \text{ for } \rho > 0,$$

where $B'(\cdot)$ and $B''(\cdot)$ denote the first and second derivatives of $B(\cdot)$ with respect to ρ. The following lemma is crucial in establishing our structural results for the M/M/1/m and M/G/s/s queues in the following sections.

LEMMA 4.1: 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, for fixed s and m, $R_0(\rho, s, m)$ is unimodal in ρ. Furthermore, $\Psi(\rho) = \Gamma_1(\rho, s, m)$ 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 Lemma 4.1 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. 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.

5. 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 5.1: For the M/M/1/m system, the objective function $R_0(\rho, 1, m)$ is unimodal in ρ. Furthermore, its maximizer $\rho^*(1, m)$ 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 Lemma 4.1. One can also show that under Assumption 2.1, there is a unique arrival rate vector for a given value of the traffic load. Hence, Proposition 5.1 also implies that there is a unique optimal arrival rate vector (and also unique optimal prices) maximizing the long-run revenue rate.

We next investigate how the changes in the waiting room capacity affect the optimal traffic load and prices. As in the single-class model of Ziya et al. [14], there is a monotonic structure and the direction of the monotonicity depends on the system parameters other than the waiting room capacity m. (Theorem 5.1 mainly follows from Lemma II.1 of Ziya [16]).

THEOREM 5.1: For the M/M/1/m system, we have the following:

(i) If $\rho^\infty \ge 1$ and $\Psi(1) = 1/2$, then $\rho^*(1, m) = 1$ for all $m \ge 1$.
(ii) If $\rho^\infty \ge 1$ and $\Psi(1) > 1/2$, then $1 < \rho^*(1, m+1) \le \rho^*(1, m)$ for all $m \ge 1$.
(iii) If $\rho^\infty < 1$ or $\Psi(1) < 1/2$, then $1 > \rho^*(1, m+1) \ge \rho^*(1, m)$ for all $m \ge 1$.

Using Theorem 5.1, 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. Consequently, the optimal prices are also monotone, but the direction of monotonicity is completely symmetric to that for the optimal arrival rates.

Both ρ^∞ and $\Psi(1)$ (when $\rho^\infty \ge 1$) can be viewed as measures of sufficiency of the service capacity of the system relative to the demand. Since ρ^∞ is the optimal traffic load when the service capacity is infinite, a higher value of ρ^∞ indicates a higher offered load on the system, and a more limited service capacity compared with the offered load. On the other hand, $\Psi(\rho)$ can be written as $\frac{d \ln(Q(\mu\rho))}{d\rho}$, and thus a higher value of $\Psi(1)$ similarly implies a higher marginal value for serving additional customers and indicates a relatively limited service capacity.[2] Then, according to Theorem

[2] For example, suppose that $I = 2$, and the relationship between the price p_i and the arrival rate $\lambda_i(p_i)$ for class $i = 1, 2$ is given by $\lambda_i(p_i) = \Lambda\alpha_i(1 - p_i)$. Then, one can show that $Q_i(\lambda) = \lambda(1 - \frac{\lambda}{\Lambda\alpha_i})$ for $i = 1, 2$. One can see that a higher value of Λ implies a higher load on the system and a more limited service capacity since higher Λ implies higher arrival rate for each fixed price. Now, for this simple example, it is straightforward to show that $\rho^\infty = \Lambda/2\mu$ and $\Psi(1) = (\Lambda - 2\mu)/(\Lambda - \mu)$. Then, for a fixed service capacity μ, a higher value of ρ^∞ or a higher value of $\Psi(1)$ implies a higher value of Λ, which in turn indicates a higher load on the system and a more limited service capacity.

5.1, when the service capacity is low relative to the demand, as the waiting room capacity increases, the optimal load on the system decreases. On the other hand, when the service capacity is relatively sufficient to utilize the potential demand, the optimal load increases with the waiting room capacity.

This interesting monotonic behavior characterized by Theorem 5.1 is caused by the tradeoff between throughput and revenue per customer. Obviously, the service provider would like to increase both throughput and revenue per customer, but these are two conflicting objectives since higher throughput is achieved at the expense of lower prices. The optimal prices essentially find the "right" balance between these two objectives. Now, keeping throughput high is equivalent to keeping utilization of the server high. In a system with relatively low service capacity, the server rarely idles, but this is especially the case when the waiting room capacity is large. If the waiting room capacity is small, the service provider keeps the load on the system high so as to keep the server occupied. As capacity increases, there is less incentive to keep the load high because there is more buffer space to help keep the server busy, and thus prices can be increased. However, this behavior is reversed when the service capacity is sufficiently large. In that case, high utilization rates come at the expense of very low prices, especially if the waiting room capacity is small. When the waiting room capacity is small, even with a high load on the system many customers are blocked and servers end up idling frequently, and thus influencing server utilization through pricing is difficult. As the waiting room capacity gets larger, utilization can be more easily controlled since less customers are blocked, and thus the service provider is more inclined towards increasing load by lowering prices. To summarize, the behavior of the optimal load as a function of the waiting room capacity is structurally different depending on whether the service capacity is high or low (or that potential demand is low or high). Theorem 5.1 provides a precise description of when a system would be regarded as low or high service capacity system.

Theorem 5.1 generalizes Proposition 4.2 of Ziya, Ayhan, and Foley [14] 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 ρ^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.

6. THE M/G/s/s SYSTEM

First, we establish that there is a unique optimal traffic load for the M/G/s/s system.

PROPOSITION 6.1: For the M/G/s/s system, the objective function $R_0(\rho, s, s)$ is unimodal in ρ. Furthermore, its maximizer $\rho^*(s, s)$ 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 Lemma 4.1 hold for $s \geq 2$. For $s = 1$, the result directly follows from Proposition 5.1. Second, we investigate how the optimal load $\rho^*(s, s)$ changes with the number of servers s.

THEOREM 6.1: For the M/G/s/s system, we have $\rho^*(s + 1, s + 1) \geq \rho^*(s, s)$ for $s \geq 1$.

Theorem 6.1 states that the optimal load is monotone increasing in the number of servers. This differs from the effect of the waiting room capacity on the optimal load for the M/M/1/m system. By increasing s and thereby increasing the service capacity, the system decreases the utilization of the servers. With this new idle capacity, the service provider can increase the throughput without significant changes in the prices. In a sense, additional service capacity makes the improvements in throughput less "costly". Thus, with the additional server, the service provider chooses to increase the offered load by reducing prices.

For the M/G/s/s system, generalized versions of Proposition 6.1 and Theorem 6.1 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 i.i.d. 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 $\sum_{i=1}^{I} \lambda_i/\mu_i$ using the fact that the system can be viewed as a single customer class system with service time requirement for each customer having mean $1/\mu_j$ with probability $\lambda_j / \sum_{i=1}^{I} \lambda_i$. Now, redefine $Q(\theta)$ as the optimal value of the following problem:

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

subject to

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

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

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

$$\max_{\theta} R_0(\theta, s, m) = \max_{\boldsymbol{\lambda}} R(\boldsymbol{\lambda}, s, m).$$

Table 1. Numerical results for the M/M/1/10 queue.

Case	a_1	b_1	a_2	b_2	$\rho_{\max}$	DP	SP	% Improvement with DP
1	2	10	4	20	0.4	0.300	0.300	0.00
2	1	10	3	10	0.4	0.250	0.250	0.00
3	4	10	8	20	0.8	1.200	1.200	0.00
4	2	10	6	10	0.8	1.000	1.000	0.00
5	6	10	12	20	1.2	2.695	2.694	0.06
6	3	10	9	10	1.2	2.246	2.245	0.07
7	10	10	20	20	2.0	7.193	7.089	1.46
8	5	10	15	10	2.0	6.012	5.921	1.53
9	20	10	40	20	4.0	22.077	21.238	3.95
10	10	10	30	10	4.0	18.812	18.182	3.47
11	50	10	100	20	10.0	76.553	73.852	3.66
12	25	10	75	10	10.0	62.773	61.054	2.82
13	100	10	200	20	20.0	175.236	170.940	2.51
14	50	10	150	10	20.0	137.396	134.674	2.02

As before, we can show that Proposition 6.1 and Theorem 6.1 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 $\sum_i \lambda_i/\mu_i$. Corollary 3.1 also generalizes. It is easy to show that $\theta^*(s,s)$ lies in the interval $[0,\theta^{\infty}]$ where $\theta^{\infty} = \sum_{i=1}^{I} \lambda_i^{\infty}/\mu_i$.

7. NUMERICAL STUDY: PERFORMANCE COMPARISON OF STATIC AND DYNAMIC PRICING POLICIES

In this section, our objective is to investigate how much more revenue would be realized in expectation if the service provider were to employ an optimal state-dependent dynamic pricing policy as opposed to an optimal static pricing policy. As in the numerical study carried out by Paschalidis and Tsitsiklis [8], we assume that there are two customer classes and we consider a number of scenarios each differing in their revenue function parameters. More specifically, we assume that the revenue functions are in the form of $Q_i(\lambda) = \lambda(a_i - b_i\lambda)$ for $\lambda \in [0, a_i/b_i]$ and for $i = 1, 2$, and pick different values for a_i and b_i so that the load on the system changes from very low to very high. We fix the service rate to $\mu = 1$. Table 1 reports our results for the $M/M/1/10$ queue. Note that in Tables 1 and 2, column $\rho_{\max} = (a_1/b_1 + a_2/b_2)/\mu$ gives the maximum possible load on the system, column DP gives the revenue under the optimal dynamic pricing policy, SP gives the revenue under the optimal static pricing policy, and the last column gives the percentage improvement in revenues that would be obtained by switching from optimal static policy to optimal dynamic policy.

According to Table 1, long-run average revenue under the optimal static policy is close to the long-run average revenue under the optimal dynamic policy for all of the fourteen cases – the smallest improvement being zero percent and the largest being 3.95 percent. The results also suggest that the improvement with dynamic pricing increases with increasing maximum load up to a certain point and then declines.

Table 2. Numerical results for the M/M/10/10 queue.

Case	a_1	b_1	a_2	b_2	$\rho_{\max}$	DP	SP	% Improvement with DP
1	10	10	20	20	2	7.500	7.500	0.00
2	5	10	15	10	2	6.250	6.250	0.00
3	20	10	40	20	4	29.999	29.999	0.00
4	10	10	30	10	4	24.999	24.999	0.00
5	30	10	60	20	6	67.452	67.446	0.01
6	15	10	45	10	6	56.211	56.205	0.01
7	50	10	100	20	10	184.881	184.453	0.23
8	25	10	75	10	10	154.131	153.742	0.25
9	100	10	200	20	20	646.046	637.830	1.29
10	50	10	150	10	20	542.800	534.971	1.46
11	150	10	300	20	30	1225.145	1204.417	1.72
12	75	10	225	10	30	1040.459	1022.194	1.79
13	250	10	500	20	50	2559.946	2505.896	2.16
14	125	10	375	10	50	2190.087	2162.139	1.29

Table 2 summarizes our findings for the M/M/10/10 queue. The maximum load on the system is set higher overall changing from 2 to 50 since the service capacity of the M/M/10/10 queue is ten times larger than that of the M/M/1/10 queue. From Table 2, we again observe that the percentage improvement with dynamic pricing is not significant. Especially when the load is low, there is either no improvement or a negligible improvement. As the load increases, improvements increase but in none of the cases, does the improvement exceed 2.16 percent. Note that these results are in agreement with those of Paschalidis and Tsitsiklis [8] who report similar improvements for a model that is a generalized version of the $M/M/s/s$ queue.

Our results suggest that switching from static pricing to dynamic pricing may not be worthwhile: the small increase in revenue may be offset by alienating customers and costs associated with implementing dynamic pricing.

REFERENCES

[1] F. Caro and D. Simchi-Levi, Static pricing for a network service provider, Working Paper, Anderson School of Management, University of California at Los Angeles, Los Angeles, CA, 2005.

[2] E. Carrizosa, E. Conde, and M. Muñoz-Márquez, Admission policies in loss queueing models with heterogeneous arrivals, Management Sci 44 (1998), 311–320.

[3] C.A. Courcoubetis and M.I. Reiman, "A Pricing in a large single link loss system," Teletraffic engineering in a competitive world, P. Key and D. Smith (Editors), Elsevier, Amsterdam, 1999, pp. 737–746.

[4] G. Gallego and G. van Ryzin, Optimal dynamic pricing of inventories with stochastic demand over finite horizons, Management Sci 40 (1994), 999–1020.

[5] P. Franken, D. Konig, U. Arndt, and V. Schmidt, Queues and point processes, Akademie-Verlag, Berlin, 1981.

[6] C. Maglaras and J. Meissner, Dynamic pricing strategies for multi-product revenue management problems, Manufact Service Oper Management 8 (2006), 136–148.

[7] E. Messerli, Proof of a convexity property of the Erlang B formula, Bell System Tech J 51 (1972), 951–953.

[8] I.C. Paschalidis and J.N. Tsitsiklis, Congestion-dependent pricing of network services, IEEE/ACM Trans Networking 8 (2000), 171–184.

[9] I.C. Paschalidis and Y. Liu, Pricing in multiservice loss networks: Static pricing, asymptotic optimality, and demand substitution effects, IEEE/ACM Trans Networking 10 (2002), 425–438.

[10] K.T. Talluri and G.J. van Ryzin, The theory and practice of revenue management, Springer, New York, 2005.

[11] P.H. Zipkin, Simple ranking methods for allocation of one resource, Management Sci 26 (1980), 34–43.

[12] S. Ziya, Uniform and precision pricing for a service facility, Ph.D. thesis, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 2003.

[13] S. Ziya, H. Ayhan, and R.D. Foley, Relationships among three assumptions in revenue management, Oper Res 52 (2004), 804–809.

[14] S. Ziya, H. Ayhan, and R.D. Foley, Optimal prices for finite capacity queueing systems, Oper Res Lett 34 (2006), 214–218.

[15] S. Ziya, H. Ayhan, R.D. Foley, and E. Pekoz, A monotonicity result for a G/GI/c queue with balking or reneging, J Appl Probab 43 (2006), 1201–1205.

[16] S. Ziya, On the relationships among traffic load, capacity, and throughput for the M/M/1/m, M/G/1/m-PS, and M/G/c/c Queues, Working Paper, University of North Carolina, Chapel Hill, NC, 2007, available at http://www.unc.edu/ziya/throughput-and-traffic-load.pdf.