Available online at www.sciencedirect.com
SCIENCE DIRECT®
Operations Research Letters 34 (2006) 214–218
Operations Research Letters
www.elsevier.com/locate/orl

Optimal prices for finite capacity queueing systems

Serhan Ziya[a,*], Hayriye Ayhan[b], Robert D. Foley[b]

[a]*Department of Statistics and Operations Research, University of North Carolina, CB# 3260, 213 Smith Building, Chapel Hill, NC 27599, USA*
[b]*School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA*

Received 16 November 2004; accepted 4 April 2005
Available online 1 June 2005

Abstract

We prove a lower bound on the optimal price for a fairly large class of blocking systems with general arrival and service processes, determine optimal price expressions for M/M/1/m and M/GI/s/s systems, and investigate how optimal prices change with changes in the size of the waiting room and service capacity.

Keywords: Pricing in queueing systems; Finite capacity queues; Erlang loss system

1. Introduction

This paper investigates the best price to charge customers in blocking systems and the effect blocking has on these best prices. By "best," we mean the price that maximizes long-run average profit per unit time. We consider G/GI/s/m queues and its special cases and we restrict attention to the simplest pricing strategy: one fixed, static price for all customers. We assume that customers are not delay sensitive; however, $m < \infty$ ensures an upper bound on the mean waiting time. We assume that each arriving customer has her own cut-off point for how much she is willing to pay for service. If the price charged is more than the amount the customer is willing to pay, the customer will not purchase service. If the advertised price is less than or equal to a customer's cut-off point and if there are less than m in the system as the customer arrives, the customer joins the system.

For G/GI/s/m queues, we cannot determine optimal prices since we cannot determine the blocking probability. However, under fairly general conditions, Proposition 3.1 gives a lower bound on optimal prices. The lower bound can be thought of as the optimal price in the corresponding G/GI/∞ system. For M/M/1/m queues, Proposition 4.1 determines the optimal prices under a certain price elasticity assumption (increasing price elasticity, IPE). Proposition 4.2 shows that these optimal prices are monotonic in m. However, whether they are increasing or decreasing depends on whether the offered load is above or below a certain critical threshold. For the Erlang loss system M/GI/s/s,

* Corresponding author.
E-mail address: ziya@unc.edu (S. Ziya).

0167-6377/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.orl.2005.04.001

Proposition 5.1 gives the optimal prices under the same price elasticity assumption. Proposition 5.2 shows that the optimal price decreases as the number of servers s increases.

Although there is vast literature on pricing for queueing control, pricing of blocking systems has received relatively little attention. Courcoubetis and Reiman [1] analyze optimal static pricing decisions for a loss system in an asymptotic regime where the capacity and the potential load of the system go to infinity. For a similar loss model, Paschalidis and Tsitsiklis [9] show that (analytically and numerically) suitably chosen static pricing policies perform almost as well as optimal dynamic policies. Paschalidis and Liu [8] and Lin and Shroff [4] generalize some of these results in a network setting. Low [5] proves structural results on optimal dynamic policies for M/M/1/m systems. Miller and Buckman [7] investigate optimal transfer prices for a service department modelled as an M/M/s/s queueing system. Sumita et al. [10] analyze loss externalities and find optimal transfer prices for a finite buffer system.

2. Model description and notation

We consider a G/GI/s/m system where the arrival rate depends on the price. For the arrival process, define $N(t)$ to be the random number of customer arrivals during $(0, t]$ where $0 \leqslant N(t) < \infty$; we assume that $N(t)/t$ converges to a strictly positive finite number Λ, the *maximal arrival rate*. Our objective is to maximize the long-run average profit. We assume that the total cost to serve a group of customers is simply a linear function of the number of customers served, and we let x denote the variable cost per customer. Let y denote the *mark-up* over the variable cost x per customer; hence, the price is $x + y$, the sum of the mark-up y and the variable cost x. For simplicity in this paper and without loss of generality, we assume that $x = 0$ so that the mark-up y can also be called the price.

The amounts that customers are willing to pay for are assumed to be independent, identically distributed non-negative random variables. Let $F(y)$ denote the proportion of customers willing to pay a price of at most y. We call $F(\cdot)$ the *willingness-to-pay* distribution and we assume that it has a finite mean and is absolutely continuous with density $f(\cdot)$. Define $\alpha \equiv \inf\{t|F(t) > 0\}$ and $\beta \equiv \sup\{t|F(t) < 1\}$. Note that $0 \leqslant \alpha < \beta \leqslant \infty$ and $0 < F(y) < 1$ for all $y \in (\alpha, \beta)$. Let $\bar{F}(y) = 1 - F(y)$. Since there is no need to consider other prices, we restrict y to the interval $y \in [\alpha, \beta)$. The *hazard rate function* for $F(\cdot)$ will be denoted by $r(y) \equiv f(y)/\bar{F}(y)$ for $y \in [\alpha, \beta)$. Since F is continuous, the probability that an arriving customer is willing to join the system at an advertised price y is $\bar{F}(y)$. If we let $\lambda(y)$ denote the arrival rate of customers at a price y, then

$$\lambda(y) = \Lambda \bar{F}(y).$$

Note that $\lambda(0) = \Lambda$. The *price elasticity function* $e(y)$ is defined as

$$e(y) \equiv -\lim_{\Delta y \to 0} \frac{[\lambda(y + \Delta y) - \lambda(y)]/\lambda(y)}{\Delta y / y}.$$

For us, this simplifies to $e(y) = yr(y)$ for $y \in [\alpha, \beta)$. Often, we will be able to show the existence of a unique optimal price by assuming IPE, which is defined as $e(y)$ being strictly increasing over $[\inf\{y : e(y) \geqslant 1\}, \beta)$. As Lariviere and Porteus [3] indicate, IPE is satisfied by many continuous distributions (e.g., gamma and Weibull).

Service times are assumed to be i.i.d. random variables with c.d.f. $G(\cdot)$ and mean μ^{-1}, $0 < \mu < \infty$. For simplicity, assume that the queue discipline is first come, first served. The service process, arrival process, and willingness-to-pay amounts are assumed to be mutually independent.

For any given price $y \geqslant 0$, we assume that the long-run fraction of customers blocked, which we refer to as the blocking probability, is a constant $B(\lambda(y), s, m)$. More precisely, let $N(y, t)$ be the number of customers who are willing to pay at least y and arrive during $(0, t]$, and let $N^{\mathrm{B}}_{s,m}(y, t)$ be the number of blocked customers during $(0, t]$ among the customers who are willing to pay the price y. We assume that

$$\lim_{t \to \infty} N^{\mathrm{B}}_{s,m}(y, t)/N(y, t) = B(\lambda(y), s, m) \text{ a.s.}$$

When the arrival process is stationary and ergodic, a sufficient condition for the blocking probability $B(\lambda(y), s, m)$ to exist is that 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. [2].

Fix the arrival process N, willingness-to-pay distribution $F(\cdot)$, service time distribution $G(\cdot)$, and let $R(y, s, m)$ be the long-run average revenue per unit time for the system with price y, number of servers s, and system capacity m. Under the assumption that the blocking probability exists, it is straightforward to show that $R(y, s, m)$ exists and can be expressed as

$$R(y, s, m) = y\lambda(y)[1 - B(\lambda(y), s, m)]. \quad (1)$$

Let $Y^*(s, m)$ denote the set of optimal prices that maximize (1), and when there is a unique optimal price, let $y^*(s, m)$ denote this unique optimal price. The *total offered load* is $\Lambda/\mu = \lambda(0)/\mu$, and the *offered load at price* y is $\rho(y) \equiv \lambda(y)/\mu$, which is non-increasing in y. Let $y^o = \sup\{y : \rho(y) = 1\}$ be the highest price at which the offered load is one.

3. A lower bound on the optimal price in a G/GI/s/m system

In general, our objective function (1) for the G/GI/s/m system cannot be evaluated since we cannot compute the system's blocking probability $B(\cdot)$. However, $y^*_\infty \equiv \inf\{y : e(y) \geqslant 1\}$ will be a lower bound on optimal prices. In addition, y^*_∞ can be computed since the price elasticity function $e(\cdot)$ only depends on the willingness-to-pay distribution F—not on $B(\cdot)$. Note that, if the corresponding G/GI/∞ system has a unique optimal price, then that optimal price would be y^*_∞.

Proposition 3.1. *In the G/GI/s/m system, y^*_∞ is a lower bound on optimal prices; that is, $y^*_\infty \leqslant \inf Y^*(s, m)$. If the blocking probability $B(\lambda(y), s, m)$ is left-continuous in y, then optimal prices exist.*

Proof. To prove the lower bound, we need only consider the case $Y^*(s, m) \neq \emptyset$; otherwise, the result holds trivially since $\inf \emptyset = \infty$. Assume the opposite; that is, assume there exists $y^*(s, m) \in Y^*(s, m)$ such that $y^*(s, m) < y^*_\infty$. Using (1) and the optimality of $y^*(s, m)$, we have

$$\frac{1 - B(\lambda(y^*_\infty), s, m)}{1 - B(\lambda(y^*(s, m)), s, m)} \leqslant \frac{y^*(s, m)\lambda(y^*(s, m))}{y^*_\infty \lambda(y^*_\infty)}. \quad (2)$$

From [11], we know that $B(\lambda(y), s, m)$ is non-increasing in y, which means that the l.h.s. of (2) is greater than or equal to 1. Since $\alpha \leqslant y^*(s, m) < y^*_\infty$ and $y\lambda(y)$ is strictly increasing over $[\alpha, y^*_\infty)$, the r.h.s. is strictly less than one, which is a contradiction.

To prove that optimal prices exist when the blocking probability is left-continuous, note that the objective function $R(y, s, m)$ is a non-negative function of y and can only jump downwards at discontinuities. To complete the proof, we need only show that $\lim_{y\to\beta} R(y, s, m) = 0$. Note that $R(y, s, m) \leqslant y\lambda(y) \leqslant \Lambda \int_y^\beta z f(z)\,\mathrm{d}z$ and the last expression converges to zero as y goes to infinity since the mean of F is finite. □

4. Single server system with exponential interarrival and service times—M/M/1/m

When arrivals are Poisson and service times are exponentially distributed, we have an expression for the blocking probability. Under certain conditions, this allows us to derive expressions for the optimal prices and prove some structural results. (The proofs are in the appendix.)

Proposition 4.1. *Under IPE, the M/M/1/m system has a unique optimal price*

$$y^*(1, m) = \inf\{y : e(y)\gamma_m(y) \geqslant 1\},$$

where

$$\gamma_m(y) = \begin{cases} \dfrac{1 + m(\rho(y))^{m+1} - (m+1)(\rho(y))^m}{(1 - (\rho(y))^{m+1})(1 - (\rho(y))^m)} & \text{if } \rho(y) \neq 1, \\ \dfrac{1}{2} & \text{if } \rho(y) = 1. \end{cases} \quad (3)$$

It turns out that the optimal price $y^*(1, m)$ is monotone in the waiting room capacity m. Interestingly, whether the optimal prices are increasing or

decreasing in m depends on whether the total offered load Λ/μ is bigger or smaller than a certain critical load $\rho^c \equiv 1/\bar{F}(y^c)$ where $y^c \equiv \inf\{y : e(y) \geqslant 2\}$, i.e., it depends on whether $\rho(y^c)$ is bigger or smaller than 1. Clearly, $y^*_\infty \leqslant y^c$.

Proposition 4.2. *Under IPE, the optimal prices in an M/M/1/m system satisfy*:

$$\begin{aligned}
&\text{(i)} \quad y^c \leqslant y^*(1,m) \leqslant y^*(1,m+1) \leqslant y^o \\
&\qquad = \lim_{m\to\infty} y^*(1,m) \quad \textit{if } \Lambda/\mu \geqslant \rho^c, \\
&\text{(ii)} \quad y^c \geqslant y^*(1,m) \geqslant y^*(1,m+1) \geqslant \max(y^*_\infty, y^o) \\
&\qquad = \lim_{m\to\infty} y^*(1,m) \quad \textit{if } \Lambda/\mu \leqslant \rho^c.
\end{aligned}$$

Given the offered load, there are two factors that determine the utilization of the server: the waiting room capacity and the price. Pricing is a more effective tool when the waiting room capacity is large since there is less blocking. When the offered load is high, as the waiting room capacity increases, the optimal price also increases since the system will be admitting higher paying customers while still managing to keep a high utilization level. However, when the offered load is low, high utilization levels come at the expense of very low prices especially when the waiting room capacity is small. As the capacity gets larger, it becomes easier to increase the utilization levels and this results in lower prices.

We have counterexamples showing that the optimal prices are not monotonic in m for the M/GI/1/m and M/M/s/m systems.

5. Erlang loss system—M/GI/s/s

In this section, we give an expression for the optimal price in the M/GI/s/s system and prove that the optimal price decreases with the number of servers. The proof of Proposition 5.1 is similar to that of Proposition 4.1 and therefore is omitted. The proof of Proposition 5.2 is given in the Appendix.

Proposition 5.1. *Under IPE, the M/GI/s/s system has a unique optimal price*

$$y^*(s,s) = \inf[y : e(y)\delta_s(y) \geqslant 1],$$

where

$$\delta_s(y) = 1 + \rho(y)[(B(\lambda(y), s, s) - B(\lambda(y), s-1, s-1))]$$

and

$$B(\lambda(y), s, s) = \frac{(\rho(y))^s/s!}{\sum_{i=0}^{s} (\rho(y))^i/i!}.$$

Note the similarity of the expressions for $y^*(1,m)$ (given in Proposition 4.1) and $y^*(s,s)$. In both cases, we have a product of the price elasticity function and some function of the parameters of the queueing system that only depend on F and the price y through $\lambda(y)$.

Proposition 5.2. *Under IPE, the optimal prices in an M/GI/s/s system are ordered as follows*: $y^*(s,s) \geqslant y^*(s+1, s+1)$ *for* $s \geqslant 1$.

The fundamental difference between the monotonicity result given in Proposition 4.2 and the monotonicity result given in Proposition 5.2 is that while the former is about the effects of changes in the waiting room capacity alone, with no change in the service capacity, the latter concerns the changes in the service capacity. Increasing the service capacity increases the customer arrival rate that the system can efficiently handle, which results in a lower optimal price.

Acknowledgment

The authors would like to thank the referee for comments and suggestions that have significantly improved the paper.

Appendix A.

Lemma A.1. *The function* $\gamma_m(y)$ *is strictly decreasing in* $\rho(y)$ *and non-decreasing in* y. *Furthermore*,

$$\rho(y) > (<)(=) 1 \quad \Leftrightarrow \quad \gamma_m(y) > (<)(=)\gamma_{m+1}(y). \tag{A.1}$$

Proof. First, note that

$$g_m(z) = \frac{1 + mz^{m+1} - (m+1)z^m}{(1-z^m)(1-z^{m+1})}$$

is a strictly decreasing function of z for $z > 0$. Since $\gamma_m(y) = g_m(\rho(y))$, the first statement of the lemma

holds. The remaining results easily follow after establishing that $g_{m+1}(z) - g_m(z)$ is strictly positive for $0 < z < 1$, and strictly negative for $z > 1$. □

Proof of Proposition 4.1. The steady-state solution for the M/M/1/m queueing system gives an explicit expression for the blocking probability $B(\lambda(y), 1, m)$. Using the expression, we can show that

$$\frac{\mathrm{d}R(y,1,m)}{\mathrm{d}y} < (=)(>)0 \quad \Leftrightarrow \quad e(y)\gamma_m(y) > (=)(<)1.$$

We know from Proposition 3.1 that $y^*(1,m) \in [y^*_\infty, \beta)$. From IPE and Lemma A.1, $e(y)\gamma_m(y)$ is strictly increasing over $[y^*_\infty, \beta)$. Since there exist prices with positive rewards and since $\lim_{y\uparrow\beta} R(y,1,m) = 0$, $R(y,1,m)$ has a unique global maximum at $y^*(1,m) = \inf\{y : e(y)\gamma_m(y) \geqslant 1\}$. □

Proof of Proposition 4.2. Let $\gamma_0(y) \equiv 1/2$ for all y. Notice that (A.1) also holds for $m = 0$ and the new variable $y^*(1,0) = y^c$. From (3) and (A.1), we have

$$\begin{aligned}\gamma_{m+1}(y^*(1,m)) > (=)(<)\gamma_m(y^*(1,m)) \\ \Leftrightarrow \rho(y^*(1,m)) < (=)(>)1. \end{aligned} \tag{A.2}$$

Notice that $\gamma_m(y^o) = 1/2$ for $m = 0, 1, 2, \ldots$ and that

$$\begin{aligned}\Lambda/\mu \geqslant (\leqslant)\rho^c &\Leftrightarrow e(y^o) \geqslant (\leqslant)2 \\ &\Leftrightarrow e(y^o)\gamma_m(y^o) \geqslant (\leqslant)1 \\ &\Leftrightarrow y^o \geqslant (\leqslant)y^*(1,m) \\ &\Leftrightarrow \rho(y^*(1,m)) \geqslant (\leqslant)1, \end{aligned} \tag{A.3}$$

where we use the fact that IPE implies that $e(y)\gamma_m(y)$ is strictly increasing over $[y^*_\infty, \beta)$, the interval where $y^*(1,m)$ lies and that when $y^o > y^*_\infty$, y^o is the unique y for which $\rho(y) = 1$. Proposition 4.1, together with (A.2) and (A.3) establish monotonicity. To prove the limit, first note that for any $y > \max(y^*_\infty, y^o)$, we have $e(y) > 1$ and $\rho(y) < 1$. Then, it follows that $\lim_{m\to\infty} \gamma_m(y) = 1$ and therefore, there exists K such that $e(y)\gamma_m(y) > 1$ for all $m \geqslant K$. Hence,

$$\begin{aligned}\lim_{m\to\infty} y^*(1,m) &\leqslant \inf\{y : y > \max(y^*_\infty, y^o)\} \\ &= \max(y^*_\infty, y^o). \end{aligned} \tag{A.4}$$

Now, for any $y < \max(y^*_\infty, y^o)$, we have $e(y) < 1$ or $\rho(y) > 1$ (strict inequality follows since IPE implies that y^o is the unique y for which $\rho(y) = 1$ when $y^o > y^*_\infty$). If $\rho(y) > 1$, then $\lim_{m\to\infty} \gamma_m(y) = 0$ and therefore, there exists K such that $e(y)\gamma_m(y) < 1$ for all $m \geqslant K$. On the other hand, if $e(y) < 1$, then $e(y)\gamma_m(y) < 1$ for all m since $\gamma_m(y) < 1$ for all m. Hence, regardless of whether $e(y) < 1$ or $\rho(y) > 1$, we have

$$\begin{aligned}\lim_{m\to\infty} y^*(1,m) &\geqslant \sup\{y : y < \max(y^*_\infty, y^o)\} \\ &= \max(y^*_\infty, y^o). \end{aligned} \tag{A.5}$$

It follows from (A.4) and (A.5) that $\lim_{m\to\infty} y^*(1,m) = \max(y^*_\infty, y^o)$. Note that for the case of $\Lambda/\mu \geqslant \rho^c$, $\max(y^*_\infty, y^o) = y^o$ from Proposition 3.1 and the fact that $y^o \geqslant y^*(1,m)$ for all m. □

Proof of Proposition 5.2. It is known that $B(\lambda(y), s, s)$ is convex in s for fixed y (see [6]). From Proposition 5.1, this implies that $e(y)\delta_{s+1}(y) \geqslant e(y)\delta_s(y)$. Finally, it follows from Proposition 5.1 that $y^*(s+1, s+1) \leqslant y^*(s,s)$. □

References

[1] C.A. Courcoubetis, M.I. Reiman, Pricing in a large single link loss system, in: P. Key, D. Smith (Eds.), Teletraffic Engineering in a Competitive World, Elsevier, Amsterdam, 1999, pp. 737–746.

[2] P. Franken, D. Konig, U. Arndt, V. Schmidt, Queues and Point Processes, Akademie, Berlin, 1981, pp. 112.

[3] M.A. Lariviere, E.L. Porteus, Selling to the newsvendor: an analysis of price-only contracts, Manuf. Serv. Oper. Manag. 3 (2001) 293–305.

[4] X. Lin, N.B. Shroff, Simplification of Network Dynamics in Large Systems, Working Paper, School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA.

[5] D.W. Low, Optimal dynamic pricing policies for an M/M/s Queue, Oper. Res. 22 (1974) 545–561.

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

[7] B.L. Miller, A.G. Buckman, Cost allocation and opportunity costs, Manag. Sci. 33 (1987) 626–639.

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

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

[10] U. Sumita, Y. Masuda, S. Yamakawa, Optimal internal pricing and capacity planning for service facility with finite buffer, Euro. J. Oper. Res. 128 (2001) 192–205.

[11] S. Ziya, H. Ayhan, R.D. Foley, A monotonicity result for the blocking probability in a G/GI/c/m queueing system, under review, available at www.unc.edu/~ziya/monotoneblocking.pdf.