

Optimal Internal Pricing and Capacity Planning for Service Facility with Finite Buffer

**Ushio Sumita
Yasushi Masuda
and
Shigetaka Yamakawa**

**Working Paper No. 1
May, 1998**

Ushio Sumita is Dean of Graduate School of International Management, International University of Japan, Niigata, Japan.

Yasushi Masuda is an Associate Professor of Administration Engineering, Faculty of Science and Technology, Keio University, Japan.

Shigetaka Yamakawa is a Senior Research Fellow at IUJ ARIS-F Software R&D Center (IAC) in International University of Japan, Niigata, Japan.

Abstract

A general microeconomic model is developed for exploring optimal internal pricing and capacity planning for service facility with finite buffer capacity. Because of the limited buffer capacity, jobs finding buffer full upon their arrival would be rejected. Such rejections create a gap between the value collectively perceived by users and the actual achievement of the organizational value. This gap, called the loss externality, has never been studied before and plays an important role for designing optimal pricing scheme. In general, the underlying economic structure may involve multiple equilibria and it is unclear whether or not the system can be controlled through internal pricing. In this regard, a sufficient condition is given under which the system administrator can and two separate prices for accepted and rejected users at any demand level to be desired so that the desired demand level becomes the unique equilibrium of the system. For a short-run problem, it is shown that the optimal pricing scheme can be expressed as the sum of the usual congestion externality and the loss externality. For a long-run problem, the optimal pricing scheme is expressed in a unified manner so that the structural relationship between the short-run problem and the long-run problem at optimality can be readily observed. A necessary and sufficient condition is also given for the marginal capacity pricing to be optimal, i.e. the optimal long-run pricing consists of the marginal cost for processing capacity and the marginal cost for buffer capacity without involving any externality at all. (PRICING ; OPTIMAL CAPACITY ; SERVICE FACILITY ; FINITE BUFFER CAPACITY ; LOSS EXTERNALITY)

1 Introduction

When an organization has an internal service department which provides services to several other user departments, there exists an implicit supply-demand structure for the services within the organization. Since the effects of the services provided by the service department on the performances of user departments may not be necessarily observed in visible measures, the question of how to allocate the resources of the service department internally and how to justify its capacity investment imposes difficult managerial control problems. When we see the society as a huge organization, the resource allocation problem described above can be seen as the resource allocation problem of the Internet. Every year, we have a sequence of conferences on management issues of the Internet, including the problem of how to set pricing for optimal resource allocation, see, e.g., Public Access to the Internet [7], Internet Economics Workshop [11], INET 97 [6], and OECD's report [16]. There is no doubt about the importance of the pricing mechanism in administration of the future Internet and further research is needed.

For the use of pricing schemes for controlling the queue size, early models were largely centered on the optimal balking policy as a function of the number of jobs in system where the system capacity was fixed, see e.g. Naor [15], Yechiali [20], Knudsen [8], Edelson and Hildebrand [5], and Lippman and Stidham [9] among others. For investigating the tradeoff between delay cost and capacity and utilization of the system, a microeconomic model was first studied by Mendelson[12]. For a short-run problem with fixed system capacity, the paper showed that the optimal internal price should be equal to the negative congestion externality in order to maximize the overall organizational net-value attained through the services provided by the service department. For a long-run problem, more restrictive assumptions are imposed. With $M/M/1$ service system and linear delay and capacity costs, the optimal price should be equal to the linear capacity cost coefficient, i.e. the marginal capacity cost. Subsequently, Mendelson's model has been extended to several different directions. The impact of different accounting rules was analyzed in Whang [18] through a two-stage game theoretic approach, demonstrating that a full cost allocation scheme over-

comes possible information and incentive problems. Systems with multiple user classes were studied in Mendelson and Whang [13], yielding optimal incentive-compatible priority-pricing schemes. Dewan and Mendelson [3] extended the original model by analyzing the long-run problem with more general service systems and weaker assumptions. Stidham [17] incorporated $GI/GI/1$ system with linear delay cost for the long-run problem. He also proposed an iterative algorithm for computing the equilibrium arrival rate, system processing capacity and price for $M/M/1$ system.

Recently, these models and their variations were applied for the analysis of the Internet pricing. For instance, Cocchi, Shenker, Estin and Zhang [2] studied a priority pricing for congestion management with multiple types of applications. Afeche and Mendelson [1] examined the providers pricing strategies in the context of market segmentation in the Internet.

All of the models described above are restricted to systems with infinite buffer capacity, where all jobs arriving at a system are always accepted. To the authors best knowledge, only a paper by Miller and Buckman [14] treated possible lost jobs due to finite buffer capacity in analyzing cost allocation problems. They combined queueing analysis and dynamic programming approach to capture different opportunity costs depending on the number of idle servers. Their analysis relied on $M/M/s/s$ queueing system explicitly.

The thrust of this paper is to incorporate systems with finite buffer capacity in a general context, thereby providing fundamental tools for the analysis of economic control of information systems and other service facilities with finite buffers. New results specific to finite buffer systems include the followings.

1. There always exists a negative externality arising from the rejection, named the loss externality. Several properties of the loss externality are examined.
2. The optimal pricing strategy is dependent not only on the performance of the server but also on how the rejected jobs are treated.
3. Finite buffer systems are in general more complex than systems with infinite capacity, and it is intuitively unclear whether the finite buffer systems can be controlled by price.

It is shown, however, that the demand for a finite system can be controlled by choosing prices for the rejected jobs as well as the accepted ones.

Some results are naturally extended from those of the infinite buffer case as summarized below.

1. The optimal short-run price is decomposed into the sum of the congestion externality and the loss externality.
2. A common pricing practice that $\text{price} = \text{marginal capacity cost}$ is not always optimal. A necessary and sufficient condition is obtained under which the common pricing practice leads to optimality.

Indeed, these results are readily reduced to those of Mendelson [12] and Dewan and Mendelson [3] by letting the better capacity go to infinity.

The organization of this paper is as follows. In Section 2, a few examples of the finite buffer case are shown and the importance and urgency for conducting economic analysis of such systems is demonstrated. In Section 3, a model with finite buffer capacity is formally described and notation is introduced. Given prices for accepted jobs and rejected jobs, incremental users at arrival rate λ would independently determine whether or not they should try to use the system by comparing their marginal value against their marginal cost. The controllability of demand based on the price mechanism is discussed in Section 4. This controllability is closely related to the uniqueness of the equilibrium. For the system administrator to control the system by price, he/she has to know which equilibrium is achieved for a given price. Hence the conditions for guaranteeing the uniqueness of the equilibrium are useful to the administrator. Section 5 explores some properties of the loss externality, including negativity of the externality and its limiting behaviors for both light and heavy traffic cases. An interesting result is that the loss externality disappears in both cases. This fact implies that the loss externality is an externality arising from the uncertainty. When users know that all jobs are either accepted or rejected, there is no gap between the perception of users' value and the achieved social value, and hence there is no externality

caused by the uncertainty. A short-run problem is analyzed in Section 6 where the optimal pricing scheme is explicitly derived. Section 7 is devoted to analysis of a long-run problem. In particular, the optimal solution structure is expressed in terms of total derivatives of the underlying functions, providing an economic interpretation of the long-run optimal solution. Some concluding remarks are given in Section 8.

2 Finite Buffer Systems and Their Applications

Finite buffer systems are everywhere in the world; client server computing systems, the Internet and Intranet, banking systems, hospital service systems, etc. A few concrete examples are given below. The Internal Revenue Service (IRS) in the United States started an electronic filing system in which taxpayers dial-up an IRS site to file tax forms electronically. The IRS has a web site with FTP servers where the tax forms are stored in PDF¹ format and can be downloaded and printed by taxpayers. These services are expected to reduce a variety of costs such as human resource cost for handling tax forms and postage costs of both IRS and taxpayers. However, the electronic service may create unfilled tax forms due to busy signals resulting from the limited server capacity. Around the due date of filing, the IRS experiences a huge number of access requests to the server causing a fear of overflows². Since the maximum number of FTP sessions for downloading tax forms is much less than that of the web, many taxpayers experience heavy congestion and their attempts for establishing the session are often rejected³. Fortunately, the servers were upgraded in 1997 to handle up to six million visitors a day. Accordingly, there have been no major complaints from taxpayers despite the rapid growth of the traffic by 280% since the previous year. However, it is questionable that the IRS can afford to continue this luxurious over-provision strategy forever. Since the overflow problem can occur only around the due date, it is not necessarily a good idea to test the capacity to the peak load, leaving the servers idle for the rest of year. It might be better to set some price during the peak period and to control the demand accordingly.

¹ Adobe's document format

² IRS Web site weathers April 15, Computerworld Online News, <http://www.computerworld.com/>, April 15, 1997.

³ IRS under online onslaught C|Net News.com, <http://www.news.com/>, March 31, 1997.

How to control the peak load is an urgent issue in the next few years.

Another example showing the relationships among the pricing, the demand, and the rejection can be found in the case of the America Online (AOL). In 1996, the AOL employed a new flat-rate pricing in which users would pay not for usage but for membership. The AOL's busy signal rate jumped up dramatically after it moved to the flat-rate pricing and the users stayed online much longer than they did before. A statistics⁴ shows that AOL's call failure was 60.3% in January, 1997, whereas CompuServe had only 6.5%. This AOL example indicates that the demand for networking is elastic in price and the mismanagement of pricing strategy can lead to deadly severe capacity deficiency.

The busy-signal problem is also observed in corporate information systems, especially in SOHO (Small Office and Home Office) applications where employees or agents working at home dial-up the corporate server to share information via telephone or ISDN. Since the number of the modem ports is limited, many users get frustrated with busy-signals at peak times such as 9 AM when people rush to e-mail and news servers. Again, the pricing will be effective to manage the situation. In this case, the rejected users may try other modem port with old and slow functions. The dial alternation of this sort can be done automatically by a piece of software. Since the alternated jobs are treated with slower transfer rate, it should be priced differently. Otherwise, users will never agree to use an old modem port.

These are real examples of finite buffer systems. In this paper, we attempt to analyze a mathematical model which is an abstraction of these situations, and to find the optimal pricing strategy for management of the limited resources. In the next section, we formally introduce the model, and explain how the model incorporates the examples shown above.

3 Model Description and Notation

We consider a stationary service system which has a system processing capacity μ and a finite buffer with system capacity K (buffer capacity plus the maximum number of jobs that can be in service simultaneously). For the interpretation of μ , we follow Mendelson and Dewan

⁴ AOL leads in busy signals, study says, C|Net News.com, <http://www.news.com>, April 28, 1997

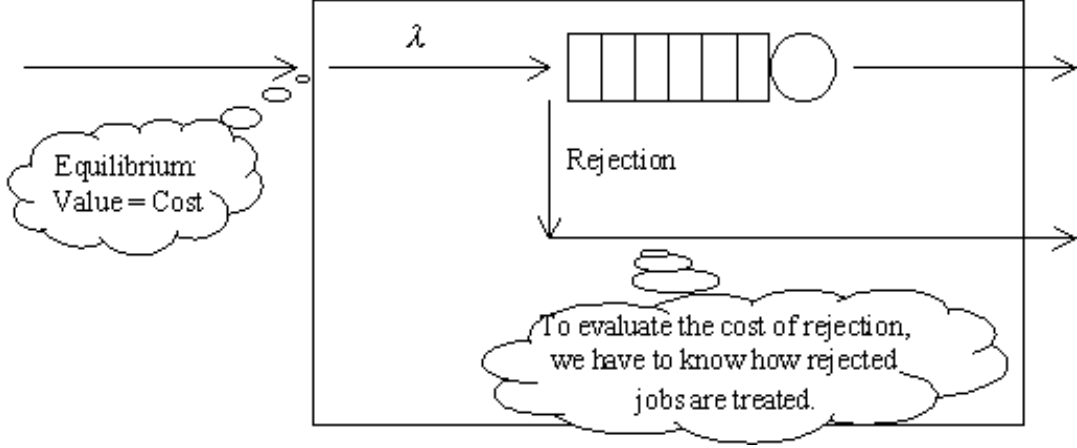


Figure 1:

[3] where μ represents the expected output rate, measured in jobs (or transactions) per unit time, when infinitely many jobs are available. Jobs arrive at system according to a Poisson process with stationary arrival rate of λ jobs per unit time. Because of finite buffer capacity, jobs finding the system full upon their arrival would not be accepted. The probability that a job is accepted to system at stationarity is a function of λ , μ , and K which we denote by $\alpha(\lambda, \mu, K)$. Similarly, the corresponding rejection probability is denoted by $\beta(\lambda, \mu, K)$, so that

$$\alpha(\lambda, \mu, K) + \beta(\lambda, \mu, K) = 1, \quad 0 \leq \alpha, \beta \leq 1, \quad (3.1)$$

where arguments of functions are omitted when no ambiguity is present. We assume that α is continuously differentiable with respect to λ and μ ,

$$\alpha_\lambda < 0, \quad \lambda > 0 \quad ; \quad \lim_{\lambda \rightarrow 0} \alpha = 1, \quad \lim_{\lambda \rightarrow \infty} \alpha = 0, \quad (3.2)$$

where we write $f_x = \partial f / \partial x$.

Although the buffer capacity parameter K may not be a continuous variable, we assume the differentiability with respect to K in order to avoid mathematical complexities. It should be noted that (3.2) holds true for most of queueing systems with finite buffer capacity. Because of rejection of jobs arriving at system when it is full, a usual stability condition $\lambda < \mu$ is not necessary here.

Let $a(\lambda)$ be the (inverse) demand function at arrival rate λ . That is, at arrival rate λ , individual jobs observe the value of their jobs as $a(\lambda)$ if they are accepted to the system. Similarly let $r(\lambda)$ be the marginal value of jobs at arrival rate λ when they are rejected. $r(\lambda)$ may be zero if rejected jobs produce no value, or may be positive if rejected jobs are processed by some other system. In the latter case, $r(\lambda)$ may be less than or equal to $a(\lambda)$ depending on different functionalities that different service systems offer. We assume that $a(\lambda)$ and $r(\lambda)$ are continuously differentiable, $a(\lambda)$ is strictly decreasing, and $r(\lambda)$ is nonincreasing. Furthermore, in order to avoid infinite demand, it is assumed that $a(\lambda)$ and $r(\lambda)$ become zero as λ goes to infinity. In summary, we assume :

$$a_\lambda < 0, r_\lambda \leq 0 \quad ; \quad a(0) > 0, r(0) \geq 0 \quad ; \quad r(\lambda) \leq a(\lambda), \quad (3.3)$$

and

$$\lim_{\lambda \rightarrow \infty} a(\lambda) = 0, \quad \lim_{\lambda \rightarrow \infty} r(\lambda) = 0. \quad (3.4)$$

The effective demand function for jobs at arrival rate λ is the expected marginal value of jobs perceived by individual users, given by

$$MV^i(\lambda, \mu, K) = \alpha(\lambda, \mu, K) a(\lambda) + \beta(\lambda, \mu, K) r(\lambda). \quad (3.5)$$

Let $A(\lambda)$ and $R(\lambda)$ be defined by

$$A(\lambda) = \int_0^\lambda a(y) dy \quad ; \quad R(\lambda) = \int_0^\lambda r(y) dy. \quad (3.6)$$

We note that $A(\lambda)$ represents the overall organizational value (per unit time) achieved by the service system at arrival rate λ if all jobs are accepted. Similarly, $R(\lambda)$ is the organizational value (per unit time) at arrival rate λ if all jobs are rejected. Then, the expected organizational value (per unit time) actually achieved by the service system at arrival rate λ , denoted by $TV^o(\lambda, \mu, K)$, is given by

$$TV^o(\lambda, \mu, K) = \alpha(\lambda, \mu, K) A(\lambda) + \beta(\lambda, \mu, K) R(\lambda). \quad (3.7)$$

From (3.5) and (3.7), one can see that there can exist a gap between the cumulative value perceived by individual jobs and the expected organizational value actually achieved by the

service system at arrival rate λ . More specifically, the *loss externality* LE is defined as

$$LE(\lambda, \mu, K) = MV^i(\lambda, \mu, K) - MV^o(\lambda, \mu, K) = -(\alpha_\lambda A(\lambda) + \beta_\lambda R(\lambda)) \quad (3.8)$$

where $MV^0 = \partial TV^o / \partial \lambda$. It should be noted that LE is the reduction in the total value due to the effect of one additional job submission per unit time at arrival rate λ . The loss externality is a new concept in the study of economics of queues, and is studied for the first time in this paper. As we will see, the loss externality plays an important role for determining optimal prices for both short-run and long-run problems.

Defined next are cost functions for accepted jobs and rejected jobs:

$$G(\lambda, \mu, K) = E[\text{cost per job} \mid \text{accepted}] \quad ; \quad M(\lambda, \mu, K) = E[\text{cost per job} \mid \text{rejected}]. \quad (3.9)$$

Typically, G corresponds to the delay cost of jobs accepted to system, which is a function of the expected time spent in system by individual jobs, see e.g. Mendelson [12]. It is assumed that G is continuously differentiable and

$$G_\lambda \geq 0, \quad G_\mu < 0 \quad ; \quad \lim_{\lambda \rightarrow 0} G(\lambda, \mu, K) > 0, \quad (3.10)$$

which again holds true for most of queueing systems. The cost for a rejected job would depend on how that job is treated. It may represent the procedural cost for rejection, or the price of the service at a different system if it is outsourced to an alternative server. It may be the delay cost due to the response time from the other server. In this paper, the original server in the system is called *the primary server* and the server processing rejected jobs is called *the alternative server*. When rejected jobs are simply lost, the system can be considered as a system with an alternative server which just passes the rejected jobs through.

We note that M is a function of system parameters μ and K of the primary system as well as λ because the effective arrival rate to the alternative server is given by $\lambda \beta(\lambda, \mu, K)$. Throughout the paper, we assume that M is differentiable and

$$M_\lambda(\lambda, \mu, K) \geq 0, \quad \lambda > 0. \quad (3.11)$$

Furthermore, in order to avoid zero demand for computing, we assume that for given μ and K ,

$$\lim_{\lambda \rightarrow 0} G(\lambda, \mu, K) < a(0). \quad (3.12)$$

Let the expected cost per job for individual users and the expected organizational cost at arrival rate λ be denoted by $MC^i(\lambda, \mu, K)$ and $TC^o(\lambda, \mu, K)$, respectively. From (3.1) and (3.9), it follows that

$$MC^i(\lambda, \mu, K) = \alpha(\lambda, \mu, K) G(\lambda, \mu, K) + \beta(\lambda, \mu, K) M(\lambda, \mu, K) \quad (3.13)$$

and

$$TC^o(\lambda, \mu, K) = \lambda MC^i(\lambda, \mu, K) = \lambda (\alpha G + \beta M). \quad (3.14)$$

Clearly, $MC^i(\lambda, \mu, K) \neq MC^o(\lambda, \mu, K) \equiv \partial TC^o / \partial \lambda$. Mendelson [12] and Dewan and Mendelson [3] articulately pointed out the gap between (3.13) and (3.14) for the infinite buffer case ($\beta = 0$), and related it to *the congestion externality*. In our model, the gap arises not only from the congestion externality but also from the loss externality in (3.8).

For the management, of interest is how to determine two prices p_A and p_R where

$$\begin{cases} p_A & : \text{ the price per job for those jobs accepted,} \\ p_R & : \text{ the price per job for those jobs rejected.} \end{cases} \quad (3.15)$$

At arrival rate λ , individual jobs should assume the price (3.15) plus the expected cost (3.13) as the total cost to be incurred. Thus the effective job cost for individual users is given by

$$MC^i(\lambda, \mu, K) + \alpha(\lambda, \mu, K) p_A + \beta(\lambda, \mu, K) p_R. \quad (3.16)$$

We next show three examples of applications of our model.

Example 1 (Complete Loss System) A first application of our model is the complete loss system where there exists no alternative server and rejected jobs are simply lost. Without any alternative server, the congestion cost function M will not depend on λ , and is zero or some positive number. In the case of IRS, the taxpayers will face a big problem when he/she cannot file the tax forms. Hence interpreting M as the cost for unfilling, one can see that the system with fairly large M is appropriate. Since the lost jobs do not contribute any to

the firm, one may assume that $r(\lambda) = 0$. □

Example 2 (Remote Mirror Server as Alternative Server) In a recent computing environment with world wide networking, a system with proxy servers and TM-monitors can resubmit over-owing jobs to another remote server which may be located at the opposite side of the globe. The alternative server is usually slower than the primary server due to communication delay and so on. Thus, it is natural to assume that $G(\lambda, \mu, K) \leq M(\lambda, \mu, K)$. If the alternative server is a mirror server with a comparable processing capability, then users will not lose any value or information of jobs. Therefore, one may assume that $a(\lambda) = r(\lambda)$. However, if the alternative machine has less capability of handling information or different presentations than the primary system does (e.g., it handles only displays texts while the original server is a multimedia server), some partial information and/or usability of the system can be lost. Then it is possible that $a(\lambda) \geq r(\lambda)$. □

Example 3 (Old System as Alternative) The progress of information technology is so fast that newly purchased computers can become out-dated within a few years. Those machines that are thought to be slow but still functioning properly can be used as supplemental systems for backup service. Then, similarly to the previous example, we have $G(\lambda, \mu, K) \leq M(\lambda, \mu, K)$, and $a(\lambda) \geq r(\lambda)$. The SOHO application shown in the previous section is one good example of this situation. □

4 Uniqueness of Equilibrium in Finite Buffer System and Price Controllability

Given a price vector (p_A, p_R) , individual users of the service system independently determine whether or not they should try to use the system. These decisions then give rise to a demand level expressed in terms of arrival rate λ . Incremental users make this decision by comparing

the marginal value of the service, $MV^i = \alpha a + \beta r$ from (3.5), with the job cost in (3.16). Hence the key equation determining the relationship between (p_A, p_R) and λ would be

$$MV^i(\lambda, \mu, K) = MC^i(\lambda, \mu, K) + \alpha(\lambda, \mu, K) p_A + \beta(\lambda, \mu, K) p_R. \quad (4.1)$$

If λ satisfies (4.1), it is then called an *equilibrium* of the economy in the system for price vector (p_A, p_R) .

In general, equilibrium determined by (4.1) is not necessarily unique. As we will see in Section 5, the uniqueness of the equilibrium is essential for system management where users' behavior is controlled through internal prices. In this regard, we define the price controllability of the system as follows.

Definition 1 *The system described in Section 3 is said to be controlled by price if for any $\hat{\lambda} > 0$, there exists (\hat{p}_A, \hat{p}_R) such that $\hat{\lambda}$ is the unique equilibrium in the system.*

Here we do not assume non-negativity of the prices since the negative price can be considered as a *bribe* for promoting usage of the system.

Before analyzing the price controllability, we first discuss the users' behavior when a price vector is given and provide sufficient conditions for which the uniqueness of the equilibrium is guaranteed. A preliminary lemma is needed, which involves the following conditions.

Condition 1 *For fixed $\mu, K > 0$, $G(\lambda, \mu, K) \leq M(\lambda, \mu, K)$ for all $\lambda > 0$.*

Condition 2 *For fixed $\mu, K > 0$, (p_A, p_R) satisfies that $p_A + G(\lambda, \mu, K) \leq p_R + M(\lambda, \mu, K)$ for all $\lambda > 0$.*

Condition 1 implies the cost advantage of processing with the primary server over rejection. Condition 2 also implies the advantage of processing with the primary server for cost plus price. We note that (p_A, p_R) is an internal pricing within an organization and is observed as a cost by users. Hence Condition 1 involves the actual cost, and Condition 2 is related to the virtual cost seen by the users.

If Condition 2 holds, then it follows from (3.3) that for fixed $\mu, K > 0$, one has $a(\lambda) - p_A - G(\lambda, \mu, K) \geq r(\lambda) - p_R - M(\lambda, \mu, K)$ for all $\lambda > 0$. Hence the price convinces all users

that the primary server is better than the alternative server. In other words, the users have no incentive to cheat and submit directly to the alternative server without first submitting their jobs to the primary one. The incentive control of this sort is one of major roles of price mechanism in designing a service system with multiple facilities. Furthermore, one will see the importance of having two variables p_A and p_R instead of a single price p . As is shown later, this flexibility of pricing is crucial to guarantee the price controllability.

Lemma 1

- (a) MV^i is strictly decreasing in $\lambda > 0$.
- (b) Under Condition 1, MC^i is nondecreasing in λ ,
- (c) Under Condition 2, the effective job cost (3.16) for individual users is nondecreasing in $\lambda > 0$.

Proof. For part (a), we have

$$MV_\lambda^i = \alpha_\lambda (a - r) + \alpha a_\lambda + \beta r_\lambda. \quad (4.2)$$

Then, it follows that $MV_\lambda^i < 0$ from (3.3) and (3.4). The strict inequality comes from the assumption that for any $\lambda > 0$, one has $\alpha > 0$ and $a_\lambda < 0$.

From (3.13) one sees that

$$MC_\lambda^i = \alpha_\lambda (G - M) + \alpha G_\lambda + \beta M_\lambda. \quad (4.3)$$

Since $\alpha_\lambda < 0$ from (3.2), $G - M \leq 0$ from Condition 1, $G_\lambda \geq 0$ from (3.10), and $M_\lambda \geq 0$ from (3.11), one has $MC_\lambda^i \geq 0$, proving part (b). Part (c) follows in a similar manner. \square

Theorem 3

- (a) Under Condition 1, there exists a unique $\tilde{\lambda} > 0$ such that $MV^i(\tilde{\lambda}, \mu, K) = MC^i(\tilde{\lambda}, \mu, K)$.
- (b) Under Condition 2, given a price vector (p_A, p_R) with $p_A < a(0) - \lim_{\lambda \rightarrow 0} G$, there exists a unique equilibrium $\hat{\lambda} > 0$ such that $MV^i(\hat{\lambda}, \mu, K) = MC^i(\hat{\lambda}, \mu, K)$.

Proof. From Lemma 1, MV^i is strictly decreasing while MC^i is nondecreasing in $\lambda > 0$. Furthermore, from (3.4) and (3.12), one has

$$\lim_{\lambda \rightarrow 0} MV^i = a(0) > \lim_{\lambda \rightarrow 0} G = \lim_{\lambda \rightarrow 0} MC^i \quad ; \quad \lim_{\lambda \rightarrow \infty} MV^i = 0 < \lim_{\lambda \rightarrow 0} G = \lim_{\lambda \rightarrow 0} MC^i. \quad (4.4)$$

Hence $MV^i = MC^i$ has a unique solution $\tilde{\lambda}$, proving (a). Similarly, $MV^i = MC^i + \alpha p_A + \beta p_R$ has a unique solution $\hat{\lambda}$ when $p_A < a(0) - \lim_{\lambda \rightarrow 0} G$ under Condition 2. \square

The equilibrium without prices is often said to be *individually optimal* since it is the best individual users can achieve without presence of price mechanisms. Thus, Theorem 3 (a) is a uniqueness condition for the individually optimal equilibrium. Theorem 3 (b) provides a sufficient condition for the uniqueness of the equilibrium with a price vector (p_A, p_R) . The uniqueness condition in Theorem 3 will also play an important role in analyzing the net-value of the system in a short-run problem in Section 6. We are now in a position to prove the next theorem regarding the price controllability.

Theorem 4 *For given μ and K , let $G(\lambda, \mu, K)$ be bounded above by $\gamma > 0$ and let $M(\lambda, \mu, K)$ be bounded below by $\delta \geq 0$. Then, the system can be controlled by price.*

Proof. By assumption,

$$G \leq \gamma \quad ; \quad M \geq \delta \text{ for any } \lambda > 0. \quad (4.5)$$

For arbitrarily given λ , let

$$\mathcal{S}_1 = \{ (p_A, p_R) : p_A \leq \delta, \gamma \leq p_R \}. \quad (4.6)$$

Then, for any $(p_A, p_R) \in \mathcal{S}_1$ it follows from (4.5) that $G \leq p_R$, $-M \leq -p_A$. Hence one has $G + p_A \leq M + p_R$, and Condition 2 is satisfied.

Similarly, for arbitrarily given λ , let

$$\mathcal{S}_2 = \{ (p_A, p_R) : \alpha p_A + \beta p_R = MV^i - MC^i \}. \quad (4.7)$$

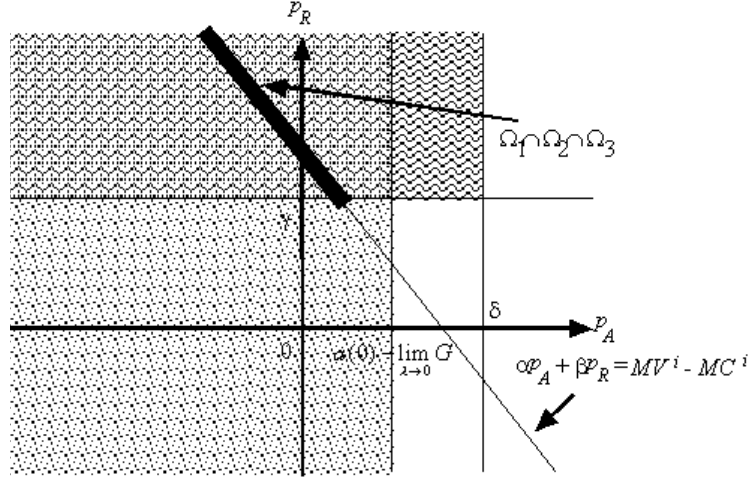


Figure 2:

We note that for $(p_A, p_R) \in \Omega_2$, the corresponding λ is an equilibrium in the system. Finally, for the same λ , let

$$\Omega_3 = \{ (p_A, p_R) : p_A < a(0) - \lim_{\lambda \rightarrow 0} G \}. \quad (4.8)$$

We show that $\bigcap_{j=1}^3 \Omega_j$ is nonempty for any λ . One easily sees that $\Omega_1 \cap \Omega_3$ is always nonempty. Since $\alpha > 0$ and $\beta \geq 0$ for any $\lambda > 0$, it is again obvious that the set of (p_A, p_R) satisfying $\alpha p_A + \beta p_R = c$ has an intersection with $\bigcap_{j=1}^3 \Omega_j$ for some constant c . Hence $\bigcap_{j=1}^3 \Omega_j$ is nonempty.

It is clear that any $(p_A, p_R) \in \bigcap_{j=1}^3 \Omega_j$ satisfies the conditions in Proposition 3 (b), and consequently the corresponding λ is the unique equilibrium. Therefore, for a target λ , the system administrator should pick a price bundle (p_A, p_R) in $\bigcap_{j=1}^3 \Omega_j$, which is not difficult, see Figure 2. \square

In many finite buffer systems, the cost function G for accepted jobs is bounded above, and the cost function M for rejected jobs is obviously bounded below by 0. Therefore, the finite buffer systems are usually controllable.

The assumption $a(\lambda) \geq r(\lambda)$ is not needed to prove the theorem above. Thus, the controllability is guaranteed in a larger class of finite buffer systems beyond our concern in

this paper.

The stability of the equilibrium is also important for system management. Stidham [17] derived stability conditions for $M/M/1$ system with linear capacity cost. Recently, Masuda and Whang [10] discussed a dynamic adoptive pricing for a queueing network model and derived conditions for local stability of the economy. For the finite buffer case, the required analysis involves dynamics of the economy with much more complexity and will be reported elsewhere.

5 Properties of Loss Externality

The loss externality defined by (3.8) plays a key role in determining the optimal price vector (p_A^*, p_R^*) and the corresponding optimal demand level λ^* . For this reason, we next investigate some properties of the loss externality.

Proposition 1 *For any μ and K , it holds true that*

$$LE(\lambda, \mu, K) \geq 0, \quad \lambda > 0. \quad (5.1)$$

Equality holds for λ' if and only if $a(\lambda) = r(\lambda)$ for all $\lambda \leq \lambda'$.

Proof. It is obvious from (3.3) and (3.6) that $A \leq R$, $\lambda > 0$. Equality holds for λ' if and only if $a(\lambda) = r(\lambda)$ for all $\lambda \leq \lambda'$. The result then follows from (3.8). \square

For the complete loss system described in Example 3, the loss externality is always positive, while it is zero in the mirror server case in Example 3 regardless of the degree of congestion. If the alternative server is more attractive than the primary server, i.e., $a(\lambda) \leq r(\lambda)$, then inequality in (5.1) is reversed which we state as a corollary below.

Corollary 1 *If $a(\lambda) \leq r(\lambda)$ and $\lambda > 0$, then*

$$LE(\lambda, \mu, K) \leq 0, \quad \lambda > 0. \quad (5.2)$$

It is nontrivial to see how the loss externality behaves as the system demand λ changes. By simple manipulation, one has $LE = \beta_\lambda(A - R)$, and therefore $LE_\lambda = \beta_{\lambda\lambda}(A - R) + \beta_\lambda(a - r)$. In many cases, the rejection probability is increasing and concave, i.e. $\beta_\lambda \geq 0$ and $\beta_{\lambda\lambda} \leq 0$, but the sign of LE_λ is not easily determined. In fact, the loss externality is not monotone: it grows for small λ but will vanish as $\lambda \rightarrow \infty$. In the next proposition, we see some interesting properties of the loss externality.

Proposition 2

(a) *Given μ and K , suppose that a, r, A , and β_λ are bounded in $\lambda \geq 0$. Then,*

$$\lim_{\lambda \rightarrow 0} LE = 0 \quad ; \quad \lim_{\lambda \rightarrow \infty} LE = 0, \quad (5.3)$$

and

$$\lim_{\lambda \rightarrow 0} LE_\lambda \geq 0. \quad (5.4)$$

(b) *Given $\lambda > 0$,*

$$\text{if } \lim_{\mu \rightarrow \infty} \alpha = 1, \text{ then } \lim_{\mu \rightarrow \infty} LE = 0, \text{ and} \quad (5.5)$$

$$\text{if } \lim_{K \rightarrow \infty} \alpha = 1, \text{ then } \lim_{\mu \rightarrow \infty} LE = 0. \quad (5.6)$$

Proof. By assumption, $\lim_{\lambda \rightarrow 0}(A - R) = 0$, and hence one sees that $LE = \beta_\lambda(A - R) \rightarrow 0$ and therefore $\lim_{\lambda \rightarrow 0} LE = 0$. One has $\lim_{\lambda \rightarrow 0} \beta = 1$ from (3.2) and $\beta_\lambda > 0$, so that $\lim_{\lambda \rightarrow \infty} \beta_\lambda = 0$. Since $A - R$ is bounded by assumption, it follows that $\lim_{\lambda \rightarrow \infty} LE = 0$. Inequality in (5.4) is immediate from Proposition 1 since we have a contradiction if $LE_\lambda < 0$. For Part (b), if $\lim_{\mu \rightarrow \infty} \alpha = 1$ or $\lim_{K \rightarrow \infty} \alpha = 1$, then one has $\lim_{\mu(K) \rightarrow \infty} \beta_\lambda = 0$, completing the proof. \square

Proposition 2 shows some interesting features of the loss externality. Part (a) claims that the loss externality vanishes for both $\lambda \rightarrow 0$ and $\lambda \rightarrow \infty$, indicating that the loss externality is closely related to the uncertainty of the rejection / acceptance. For small λ , almost all jobs are accepted and for large λ , virtually all jobs are rejected. In both cases, there is little uncertainty and the loss externality approaches zero. Part (b) shows that if capacities

become sufficiently large, virtually all jobs are accepted and the loss externality goes to zero again. This result is consistent with the corresponding result of the congestion externality for the infinite buffer case.

6 Short Run Problem

In a short-run problem, the processing capacity μ and the system capacity K of the service department are fixed. Hence the service department manager would face the decision problem of how to determine a price vector (p_A^*, p_R^*) and a desirable demand level λ^* based on (4.1), so that the expected net-value of the service department's services to the entire organization would be maximized. Here, the net-value NV is defined by

$$NV = TV^o - TC^o. \quad (6.1)$$

This decision problem is formulated more formally below.

Short-Run Problem

$$\begin{aligned} & \underset{\lambda, p_A, p_R}{\text{maximize}} && NV \\ & \text{subject to} && MV^i = MC^i + \alpha p_A + \beta p_R. \end{aligned}$$

In general, the individually optimal equilibrium obtained as a solution of $MV^i = MC^i$ is not socially optimal. Without price mechanisms, self-interested users tend to consume resources beyond the socially optimal level. We expect that prices appropriately set will suppress the unnecessary access to the system. In infinite buffer systems, for instance, Naor [15] and Lippman and Stidham [9] showed that the individually optimal usage rate would be always more than the socially optimal one. The next proposition shows that users in the loss system also over-consume the system resources without price mechanisms.

Proposition 3 *Suppose that Condition 1 holds. Let λ^* be the optimal arrival rate to Short-Run Problem, and let $\tilde{\lambda}$ be the individually optimal arrival rate, i.e., the solution to $MV^i = MC^i$. Then, $\lambda^* \leq \tilde{\lambda}$.*

Proof. From (3.14), one has

$$MC^o = TC_\lambda^o = MC^i + \lambda MC_\lambda^i. \quad (6.2)$$

Since $NV = TV^o - TC^o$, substitution of $MV^i - MV^o = LE$ and (6.2) into this equation yields

$$NV_\lambda = -LE - \lambda MC^i + (MV^i - MC^i). \quad (6.3)$$

We note that for any $\lambda > 0$, $LE \geq 0$ from Proposition 1 and $MC^i \geq 0$ from Lemma 1. Furthermore, from Lemma 1, $f(\lambda) = MV^i - MC^i$ is strictly decreasing and $f(\tilde{\lambda}) = 0$ so that $f(\lambda) < 0$ for $\lambda > \tilde{\lambda}$. Hence $NV_\lambda(\lambda^*) = 0$ implies $f(\lambda^*) = LE + \lambda MC_\lambda^i \geq 0 = f(\lambda^i)$ and therefore $\lambda^* \leq \tilde{\lambda}$, proving the proposition. \square

We now prove the main theorem of this section, providing a necessary condition for the optimality of Short Run Problem.

Theorem 5 *If $(\lambda^*, p_A^*, p_R^*)$ is optimal for Short-Run Problem, then*

$$\alpha^* p_A^* + \beta^* p_R^* = \lambda^* MC_\lambda^{i*} + LE^*, \quad (6.4)$$

where $f^* = f|_{(\lambda, p_A, p_R) = (\lambda^*, p_A^*, p_R^*)}$.

Proof. From the first order necessary condition for the optimality of Short-Run Problem, one has $NV_\lambda = 0$, i.e.,

$$MV^o - MC^o = MV^o - MC^i - \lambda MC_\lambda^i = 0. \quad (6.5)$$

From (6.5) and the constraint $MV^i = MC^i + \alpha p_A + \beta p_R$, one has $\alpha p_A + \beta p_R = (MV^i - MV^o) + \lambda MC_\lambda^i$. Since $LE = MV^i - MV^o$ by definition, the theorem follows. \square

The left hand side of (6.4) is the expected price of one job submission seen by an arriving user. Thus, Theorem 5 indicates that the expected price at optimality is equal to the sum of the congestion externality and the loss externality. In the following corollary, this sum is further decomposed into the externalities for those accepted and those rejected. Other types of decomposition of externalities are found in, for example, Yamakawa [19].

Corollary 2 *If $(\lambda^*, p_A^*, p_R^*)$ is optimal for Short-Run Problem, then*

$$\alpha^* p_A^* + \beta p_R^* = \lambda^* \alpha^* (G_\lambda^* + \beta_\lambda^* \theta_A^*) + \lambda^* \beta^* (M_\lambda^* + \alpha_\lambda^* \theta_R^*), \quad (6.6)$$

where $\theta_A = -(A/\lambda - G) + (R/\lambda - M)$ and $\theta_R = (A/\lambda - G) - (R/\lambda - M) = -\theta_A$.

Proof. The result is immediate from (6.4), and the definition of θ_A and θ_R . \square

Economic interpretation of (6.6) is rather subtle. We imagine an arriving customer at demand rate λ . This customer is accepted with probability α . As in Mendelson [12], this customer experiences the congestion externality G_λ which is the incremental increase of delay cost. In addition, because of submission of marginal jobs at demand λ , this customer foresees increase of rejection probability β_λ . Should the customer be shifted from the primary server to the alternative server because of such marginal jobs, the customer observes the average loss θ_A . The fact that θ_A represents this average loss can be seen in the following manner. A/λ is the average value of accepted customers and therefore $A/\lambda - G$ is the corresponding average profit per job. Similarly, $R/\lambda - M$ is the average profit among rejected customers. Consequently, θ_A represents the average loss of a customer who is shifted from the primary server to the alternative server at demand rate λ . Hence $G_\lambda - \beta_\lambda \theta_A$ is the average total externality per accepted customer at demand rate λ , and $\lambda \alpha (G_\lambda - \beta_\lambda \theta_A)$ is the total externality among all accepted customers at demand rate λ . Similarly, $\lambda \beta (M_\lambda - \alpha_\lambda \theta_R)$ is the total externality among all rejected customers. We note that $\alpha_\lambda = -\beta_\lambda$, and $\alpha_\lambda \theta_R = \beta_\lambda \theta_A$.

Finally, we show in the next corollary that the congestion externality pricing models discussed in Mendelson [12] and Dewan and Mendelson [3] can be derived directly from Corollary 2 by letting $K \rightarrow \infty$.

Corollary 3 *Suppose that $\lim_{K \rightarrow \infty} \alpha = 1$, $\lim_{K \rightarrow 0} \beta_\lambda = 0$ and G , M and M_λ are bounded in K . Then $\lim_{K \rightarrow \infty} (\alpha^* p_A^* + \beta p_R^*) = \lambda^* M C_\lambda^{i*}$.*

Proof. Since $\lim_{K \rightarrow \infty} \alpha = 1$ implies $\lim_{K \rightarrow \infty} \beta = \lim_{K \rightarrow \infty} \beta_\lambda = 0$, the result is immediate from (6.6) and (4.3). \square

We note that the conditions in Corollary 3 are natural ones and hold true for many queueing models.

7 Long Run Problem

In a long-run problem, both the system processing capacity μ and the buffer capacity parameter K become a part of decision variables. The problem facing the management then would be how to jointly determine two prices p_A and p_R and the target demand rate λ along with μ and K so that the resulting resources would be optimally utilized. We now formally formulate this problem.

Long-Run Problem

$$\begin{aligned} & \underset{\lambda, \mu, K, p_A, p_R}{\text{maximize}} && NV - C \\ & \text{subject to} && MV^i = MC^i + \alpha p_A + \beta p_R. \end{aligned}$$

Here $C = C(\mu, K)$ is the cost function representing the investment needed to achieve system capacities μ and K . We note that the cost is spread over a certain planning period so that its unit is dollar per unit of time.

In general, the optimal price is given in terms of externalities and capacity costs as shown in the next theorem, which is an extension of the corresponding result for Short Run Problem.

Theorem 6 *Let $(\lambda^*, \mu^*, K^*, p_A^*, p_R^*)$ be an optimal solution to Long Run Problem. Then*

$$\alpha^* p_A^* + \beta^* p_R^* = - \left((\Delta^{(x,y)} \alpha) A + (\Delta^{(x,y)} \beta) R \right) + \lambda (\Delta^{(x,y)} MC^i) + \Delta^{(x,y)} C, \quad (7.1)$$

where $\Delta^{(x,y)} f \equiv f_\lambda + x f_\mu + y f_K$, and x and y are any real numbers.

Proof. From the first order condition for optimality of Long Run Problem, one has

$$\alpha_\lambda A + \beta_\lambda R + MV^i = MC^i + \lambda MC_\lambda^i, \quad (7.2)$$

$$\alpha_\mu A + \beta_\mu R = \lambda MC_\mu^i + C_\mu, \quad (7.3)$$

$$\alpha_K A + \beta_K R = \lambda MC_K^i + C_K, \quad (7.4)$$

Following the line of the argument given in Theorem 5, one has

$$\alpha^* p_A^* + \beta^* p_R^* = -(\alpha_\lambda A + \beta_\lambda R) + \lambda MC_\lambda^i. \quad (7.5)$$

Multiplying (7.3) by x and (7.4) by y , and then adding the resulting equations to (7.5), the theorem follows. \square

We note that $\Delta^{(x,y)} f$ is the total derivative of f along the direction (x, y) . Setting $x = y = 0$ leads to f_λ . By changing the direction (x, y) , it is possible to give several interpretations of the long run pricing. For instance, let $x = y = 0$. Then, the long run pricing is immediately reduced to the short run pricing, which only contains the loss externality and the congestion externality. If the administrator wishes to incorporate the processing capacity cost but not the buffer capacity cost, then it is possible to set $x = 1$ and $y = 0$. In this case, (7.1) becomes

$$\alpha^* p_A^* + \beta^* p_R^* = E^* + C_\mu^*, \quad (7.6)$$

where $E = -(\alpha_\lambda + \alpha_\mu)A - (\beta_\lambda + \beta_\mu R) + \lambda MC_\lambda^i + \lambda MC_\mu^i$. Thus, the price is seen as the cost for the processor and some externality E . We note that $E \leq LE + \lambda MC_\lambda^i$.

We also note that if $\lim_{K \rightarrow \infty} \alpha = 1$, and if $MC_\lambda^i + MC_\mu^i = 0$, then Eq. (7.5) becomes

$$p_A^* = C_\mu^* \text{ for } K = \infty. \quad (7.7)$$

This is exactly the same marginal capacity cost pricing discussed in Mendelson [12]. The formula $MC_\lambda^i + MC_\mu^i = 0$ is the condition shown in Dewan and Mendelson [3] for guaranteeing the optimality of the marginal capacity cost pricing principle for the infinite buffer case.

The marginal capacity cost pricing is widely used in computing procurement for its simplicity and intuitiveness, see Mendelson [12] and papers referenced therein. However, in systems with non-linearity including many queueing systems, this principle is not necessarily optimal. Dewan and Mendelson [3], and recently Dewan [4] addressed themselves to this issue and compared the marginal capacity cost pricing with the optimal pricing. Such studies are considered to be fundamental for development of appropriate accounting scheme for queueing models describing information systems.

The importance of the marginal capacity cost pricing for accounting purposes has not been fully reflected in the past research because of the lack of consideration of the loss externality. In what follows, we derive a necessary and sufficient condition for the optimality of the marginal capacity cost pricing for the finite buffer case. The following notion associated with the optimal price and marginal capacity costs is employed.

Definition 2 *At optimality, we say that (p_A^*, p_R^*) partially covers the capacity costs C_μ^* and C_K^* with $(x^*, y^*) \in Q_+$ if*

$$\alpha^* p_A^* + \beta^* p_R^* = x^* C_\mu^* + y^* C_K^* = \Delta^{(x^*, y^*)} C^*, \quad (7.8)$$

where Q_+ is the two-dimensional unit cube, i.e.,

$$Q_+ = \{ (x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1 \}.$$

Similarly, we say that (p_A^*, p_R^*) fully covers the capacity costs C_μ^* and C_K^* if they are partially covered by (p_A^*, p_R^*) with $(x, y) = (1, 1)$.

For negative x and/or y , Equation (7.8) loses the economic meaning for cost allocation. For $x > 1$ and/or $y > 1$, users are overcharged for capacities that they do not utilize. Therefore, it is natural that weights of x and y be limited in Q_+ . Of course, the marginal capacity cost pricing for the processor capacity and buffer capacity is optimal if and only if (p_A^*, p_R^*) fully covers the capacity costs C_μ^* and C_K^* .

We now derive a necessary and sufficient condition under which the marginal capacity costs C_μ^* and C_K^* are partially (or fully) covered by (p_A^*, p_R^*) . A preliminary lemma is needed.

Lemma 2 *Suppose that $u \geq 0$, $v \geq 0$, and $w > 0$, and suppose that $uv \neq 0$. Then,*

$$\{ (x, y) : ux + vy = w \} \cap Q_+ \neq \emptyset, \quad (7.9)$$

if and only if

$$u + v \leq w. \quad (7.10)$$

Proof. The proof is easy and omitted here. \square

Theorem 7 *Let $(\lambda^*, \mu^*, K^*, p_A^*, p_R^*)$ be an optimal solution to Long Run Problem. Suppose that $C_\mu C_K \neq 0$ and that $\alpha^* p_A^* + \beta^* p_R^* > 0$. Then, the marginal capacity costs C_μ^* and C_K^* are partially covered by (p_A^*, p_R^*) if and only if*

$$\alpha(\Delta^{(1,1)}G) + \beta(\Delta^{(1,1)}M) \leq (\Delta^{(1,1)}\alpha)(A - G) + (\Delta^{(1,1)}\beta)(R - M). \quad (7.11)$$

Furthermore, the capacity costs C_μ^ and C_K^* are fully covered by (p_A^*, p_R^*) if and only if equality holds in (7.11) at optimality.*

Proof. From (7.5) in Theorem 6, the marginal capacity costs C_μ^* and C_K^* are partially covered if and only if

$$-(\Delta^{(x,y)}\alpha)A - (\Delta^{(x,y)}\beta)R + \lambda(\Delta^{(x,y)}MC^i) = 0 \quad (7.12)$$

has a solution in Q_+ . Now, let

$$u = -\{\lambda MC_\mu^i - (\alpha_\mu A + \beta_\mu R)\}, \quad (7.13)$$

$$v = -\{\lambda MC_K^i - (\alpha_K A + \beta_K R)\}, \quad (7.14)$$

$$w = \lambda MC_\lambda^i - (\alpha_\lambda A + \beta_\lambda R). \quad (7.15)$$

Then, (7.12) becomes $ux + vy = w$. Since $C_\mu \geq 0, C_K \geq 0$ and $C_\mu C_K > 0$, one can easily see that $u \geq 0, v \geq 0$ and $uv > 0$ from (7.2) and (7.3). From Lemma 2, the costs are partially allocatable if and only if $u + v \leq w$, and the result follows. \square

8 Concluding Remarks and Future Research Direction

In this paper, a general microeconomic model has been developed for exploring optimal internal pricing and capacity planning for a service facility with finite buffer capacity. While this model includes the original model of Mendelson [12], the two models differ from each

other in an essential way. The finiteness of system buffer capacity possibly generates jobs that are rejected upon their arrival because of no waiting room available in system. The loss possibility, in turn, creates a gap between the cumulative perception of values observed by individual jobs and the actual achievement of the organizational value. This gap, called the loss externality, plays an important role in providing microeconomic interpretations of optimal pricing schemes.

In queueing systems with finite buffers, there can exist multiple equilibria for a given price. The multiplicity of the equilibrium causes a difficulty in controlling the computing demand through pricing schemes. Even if the optimal price is correctly obtained, the system can settle down in an equilibrium which was not targeted by the system administrator. Several conditions are derived under which the system is controllable through internal prices, i.e., for any targeting demand $\hat{\lambda} > 0$ the system administrator can find a price vector (\hat{p}_A, \hat{p}_R) which enforces the equilibrium of the system to be unique at $\hat{\lambda} > 0$. In other words, by changing the price vector, the system administrator can correctly predict and control the equilibrium point.

It has been shown that the loss externality of the finite buffer system is always non-negative under certain assumptions. Hence the individual users collectively perceive more value in the system than the system administrator. This externality vanishes only when $a(\lambda) = r(\lambda)$ for all $\lambda \leq \hat{\lambda}$ where $\hat{\lambda}$ is the equilibrium. Therefore, every loss system with $a(\lambda) > r(\lambda)$ has strictly positive loss externality.

For the short-run problem, the optimal pricing is expressed as the sum of the congestion externality and the loss externality. We have seen that the loss externality is always non-negative. Therefore, users are penalized by the loss externality as well as the congestion externality. Since the infinite capacity queue is a special case of the finite buffer system with $K \rightarrow \infty$, the optimal price coincides with Mendelson's congestion externality price when $K \rightarrow \infty$.

For the long-run problem, the optimal price is expressed as a linear combination of the externalities and the capacity costs. By changing weights of the combination, one can derive the short run price from the long run price. Different characterizations of the long

run pricing are obtained by changing weights in the linear combination. In particular, a necessary and sufficient condition is obtained under which the long run prices are explained totally in terms of the capacity costs without unmeasurable externalities.

The main concern of the system administrator is that the value functions are not always known. Masuda and Whang [10] examined a dynamics of the economy in the queueing system and showed that the economy would autonomously converge to the socially optimal equilibrium when the initial status of the economy was within a neighborhood of the socially optimal one. An interesting result of the paper was that the price dynamics managed by the system administrator would not require the value functions explicitly and only depend on the realized demand profile in the previous period. Our model may be extended to incorporate such a dynamic setting and similar sufficient conditions may be derived.

References

- [1] P. Afeche and H. Mendelson. Market segmentation for data communication services. In *Proceedings of WISE 96*, 1996.
- [2] R. Cocchi, S. Shenker, D. Estin, and L. Zhang. Pricing in computer networks: Motivation, formulation, and example. *IEEE/ACM Transactions on Networking*, 1:614–627, 1993.
- [3] H. Dewan and H. Mendelson. User delay costs and internal pricing for a service facility. *Management Science*, 36:1502–1517, 1990.
- [4] Sanjeev Dewan. Pricing computer services under alternative control structures: Trade-offs and trends. *Information Systems Research*, 7:301–307, 1996.
- [5] N.M. Edelson and D.K. Hildebrand. Congestion tolls for Poisson queueing processes. *Econometrica*, 43:81–92, 1975.
- [6] Internet Society. *Proceedings of INET 97*. 1997. Available at <http://www.isoc.org>.
- [7] B. Kahin and J. Keller, editors. *Public Access to the Internet*. MIT Press, 1995.

- [8] N.C. Knudsen. Individual and social optimization in a multi-server queue with a general cost-benefit structure. *Econometrica*, 40:515–528, 1972.
- [9] S.A. Lippman and S. Stidham Jr. Individual versus social optimization in exponential congestion systems. *Operations Research*, 25:233–247, 1977.
- [10] Y. Masuda and S. Whang. Dynamic pricing for network service : Equilibrium and stability. Technical report, A.G. Anderson School, University of California, Riverside, 1995.
- [11] L.W. McKnight and J.P. Bailey, editors. *Internet Economics*. MIT Press.
- [12] H. Mendelson. Pricing computer services: Queueing effects. *Communication of the ACM*, 28:312–321, 1985.
- [13] H. Mendelson and S. Whang. Optimal incentive-compatible priority pricing for the $M/M/1$ queue. *Operations Research*, 38:870–883, 1990.
- [14] B.L. Miller and A.G. Buckman. Cost allocation and opportunity costs. *Management Science*, 5:626–639, 1987.
- [15] P. Naor. On the regulation of queue size by levying tolls. *Econometrica*, 38:13–24, 1969.
- [16] OECD. Information infrastructure convergence and pricing : The internet. Technical report, OECD/GD(96)73. Available at <http://www.oecd.org/>.
- [17] S. Stidham. Pricing and capacity decisions for a service facility: Stability and multiple local optima. *Management Science*, 38:1121–1139, 1992.
- [18] S. Whang. Cost allocation revisited: An optimality result. *Management Science*, 35:1264–1273, 1989.
- [19] S. Yamakawa. Optimal pricing and cost recoverability in distributed database environment with value and cost externalities (cis 9605). Technical report, Simon School, University of Rochester, 1996.

- [20] U. Yechiali. On optimal balking rules and toll charges in the $GI/M/1$ queue process. *Operations Research*, 19:348–370, 1971.