INFORMS Journal on Computing

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

# Pricing Communication Services with Delay Guarantee

Zhongju Zhang, Debabrata Dey, Yong Tan,

To cite this article:
Zhongju Zhang, Debabrata Dey, Yong Tan, (2007) Pricing Communication Services with Delay Guarantee. INFORMS Journal on Computing 19(2):248-260. https://doi.org/10.1287/ijoc.1050.0159

*INFORMS Journal on Computing*
Vol. 19, No. 2, Spring 2007, pp. 248–260
ISSN 1091-9856 | EISSN 1526-5528 | 07 | 1902 | 0248

DOI 10.1287/ijoc.1050.0159


# Pricing Communication Services with Delay Guarantee

Zhongju Zhang
School of Business, University of Connecticut, One University Place, Stamford, Connecticut 06901–2315, USA,
john.zhang@business.uconn.edu

Debabrata Dey, Yong Tan
University of Washington Business School, Box 353200, Seattle, Washington 98195–3200, USA
{ddey@u.washington.edu, ytan@u.washington.edu}

Although pricing communication services has received attention, there has been little work where quality-of-service (QoS) guarantees are incorporated into ex ante pricing of communication links. In recent years, however, data networks providing services with QoS guarantees have become quite popular. We study optimal pricing for communication links provided over packet-switching networks with explicit QoS guarantees in terms of an expected per-packet delay. We first develop a pricing scheme for multiple classes of service when the delay guarantees are fixed and exogenously determined and analyze this scheme with one and two classes. Subsequently, we extend this analysis to the case where a provider can choose her own QoS parameters. We also examine long-run pricing and capacity planning for one and two classes of service. Finally, we consider other QoS guarantees such as jitter, packet-loss ratio and maximum delay.

*Key words*: data network pricing; quality of service; queueing effects; network externality
*History*: Accepted by Ramayya Krishnan, former Area Editor for Telecommunications and Electronic Commerce; received August 2003; revised September 2004, February 2005, April 2005; accepted July 2005.

## 1. Introduction

Over the last two decades, distributed computerized information systems have become popular. These systems, along with the widespread use of e-commerce systems over the Internet, have led to proliferation of data-communication networks. Most large organizations operate their distributed information systems over private networks, often built around leased point-to-point links from different network carriers. A point-to-point data-communication link can be leased as either a dedicated or a switched service. Among different types of data-communication links, packet-switching services exhibit high efficiency and reliability, at a fraction of the cost of a dedicated link and have naturally become popular. In a packet-switching network, a message is broken into small blocks or segments called *packets* that are routed via a set of intermediate nodes to the destination. To manage the task of routing packets, a packet-switching network requires a set of devices (routers/switches) that control delivery of packets—a router queues packets and forwards them to the destination according to specific criteria, such as minimum number of hops or minimum delay (McDysan 2000).

To the subscriber, the routing details are transparent. The end-to-end connection appears similar to a dedicated link, except for the variable delay. Because of the queueing effects on a packet-switching network, the transmission time of an individual packet is unpredictable, so traditional packet-switching networks typically operate on a *best-effort* basis within available capacity. Since packet transfer delays and loss characteristics depend on the instantaneous load and state of the network, there can be no quantitative assurance about the delivered quality of service (QoS). QoS on data-communication links, on the other hand, has become increasingly important with sophisticated applications and services. Along with the growth in bandwidth, there has been growth in applications (such as videoconferencing, medical imaging, and real-time financial transactions) that are more susceptible to delay and require a certain level of QoS. In addition, not all data from various applications are equally important within an organization, and it is necessary to prioritize communication services for mission-critical business units (Kesh et al. 2002). To deal with these issues, many service providers have employed the fast packet-switching technologies (such as frame relay and cell relay) to offer multiple classes of priority-based point-to-point links on their networks. In a priority-based network, a user chooses a priority class for sending a message—a higher priority class is expected to deliver the message faster.

Most fast packet-switching networks have established end-user service-level agreements along several QoS dimensions, such as availability, average per-packet delay or *latency*, maximum per-packet delay, per-packet delay variance or *jitter*, and packet-loss ratio. Among the delay-related QoS dimensions, latency is the most common in service-level agreements, so is our focus. Latency includes the propagation delay, the transmission time, as well as the average queueing delay experienced by a packet, and is often measured as the average round-trip time (RTT) for a packet on a point-to-point link, excluding access and delay in devices in customer premises (Peterson and Davie 2003).

To maintain the service-level agreement, it is necessary to design and implement a proper pricing scheme (DaSilva 2000). Cocchi et al. (1993) and Parameswaran et al. (2001) argue that pricing is an integral part of the operations of a fast data-communication network and that appropriate pricing schemes are necessary as incentives to make users choose appropriate QoS levels. MacKie-Mason and Varian (1994) note that one of the most important elements of network cost is the social cost of congestion. They argue that nonpricing approaches to reduce congestion cost are either flawed or have undesirable side effects. Pricing, on the other hand, has an advantage because it permits users to express the value they place on obtaining a service. Users can trade off their benefits from the service against its price (plus the delay cost) to make informed decisions. The number of low-value and frivolous network users is thereby likely to decrease, which could provide a better quality of service for high-value users (MacKie-Mason and Varian 1994).

We develop an incentive-compatible pricing scheme for point-to-point data-communication links offered over a provider's private packet-switching network with one or two classes of priority-based services. Within the network, a user chooses point-to-point communication service from one of the classes and pays a fee for the delivery of each packet with an advance assurance that the expected end-to-end per-packet delay he will experience is at most a prespecified level. After discussing prior research and the basic modeling framework in Sections 2 and 3, we analyze, in Section 4, the pricing scheme when the expected-delay guarantee is fixed and exogenously determined. In Section 5, we extend the basic pricing model to the situation where a provider can choose her own delay parameter as well. Long-run pricing along with network capacity planning problems are then studied in Section 6. Section 7 extends the basic model to incorporate additional QoS dimensions such as jitter, maximum per-packet delay, and packet-loss ratio. Section 8 concludes.

## 2. Literature Review

Pricing communication services has received a lot of attention in the literature. An important stream of research is use of price and current congestion level of the network to control user traffic patterns. For example, MacKie-Mason and Varian (1995) analyze congestion-based pricing of data networks and explore the implications of a two-part tariff structure in three different settings: centrally-planned, competitive, and monopolistic. MacKie-Mason et al. (1997) propose a dynamic iterative pricing scheme that attempts to increase network and economic efficiencies simultaneously.

Another stream of research attempts to achieve network efficiency based on efficient resource allocation through dynamic pricing. Low and Varaiya (1993), for instance, maximize network welfare by allocating resources among competing service requests differentiated by traffic burstiness and maximum end-to-end delays. Gupta et al. (1997, 1999) present an optimal priority pricing scheme for a network with multiple nodes and multiple service classes, and examine the role of pricing as a resource-allocation mechanism to facilitate QoS. Wang et al. (1997) analyze pricing and capacity investment for a monopolistic provider with guaranteed QoS. Keon and Anandalingam (2003) and Thomas et al. (2002) allocate resources in connection-oriented networks to offer multiple services to users with QoS guarantees.

Other work compares the efficiency of different pricing schemes. For example, Cocchi et al. (1991) compare user benefits under flat-rate and priority schemes, finding multiple service classes to be advantageous. Wang and Peha (1999) compare the benefits under state-dependent pricing and long-term average pricing schemes, showing that dynamic state-dependent pricing improves social welfare as well as profit to the service provider, but may sometimes hurt consumers.

Pricing models can also include queueing externalities (Marchand 1974, Mendelson 1985, Mendelson and Whang 1990, Rao and Petersen 1998). This has been extended to pricing communication services, where queueing externalities are common. For example, Masuda and Whang (1999) study a situation where the network administrator has incomplete information about the arrival rate of messages, and suggest a way to adjust the price dynamically, but do not guarantee QoS.

Our main innovation is considering "ex ante" pricing of communication links, while earlier research considers "dynamic" prices calculated from instantaneous congestion levels and resource availability. Though theoretically efficient, dynamic pricing is difficult to implement because the actual dynamic price is known only "ex post," after the service has been

provided. For such pricing to work well, users must monitor the prices and the QoS levels continuously and use this information dynamically to make every transmission decision, which is not how communication services operate. For these services, providers usually publicize their "ex ante" prices and QoS guarantees, and users enter into service-level agreements that remain in effect for months or years. Our results could help a provider understand better how to draw these service-level agreements, price the communication links, and plan for network capacity.

## 3. The Modeling Framework

We consider a general situation where a provider offers multiple classes of priority-based links over a packet-switching network, though we analyze only the cases of one or two classes. Each class has an expected per-packet delay. Initially, we assume that the expected delay is fixed for each class (say, as an industry standard); we relax this in Section 5. The capacity of the network is also assumed to be fixed (in the short run), because it is expensive to expand network capacity; this assumption is relaxed in Section 6. A user chooses a class and the message is broken into fixed-size packets; for example, in an ATM network, a message is broken into 53-byte cells. Packets from a message are queued with the priority chosen by the user for the entire message. In the queue, packets from a high-priority class take precedence; within a class, the queue is first-come first-served (McDysan 2000). Packets are reassembled at the destination to recreate the message.

At the time of subscribing to the packet-switching service, the user knows the class characteristics in terms of the price per packet (set by the provider) in each class and the per-packet guaranteed delay in that class. The user then decides if he wants to obtain the link and, if so, on which priority class. Given the pattern of user behavior, the provider can estimate the demand for each class (in terms of an arrival rate), and hence the expected delay per packet in each class. The provider's problem then becomes one of setting prices to maximize her profit while maintaining the expected-delay assurances. This is a classic sequential game (Tirole 1989), where the provider is the leader and the user is the follower.

There are $N$ priority classes, with $d_i$ being the expected per-packet delay guarantee in class $i$, $i = 1, 2, \ldots, N$. Packets from class $i$ have priority over those from class $i+1$. The provider sets the price vector $\mathbf{p} = (p_i)$, where $p_i$ is the price per packet transmitted in class $i$. The user then chooses an appropriate class for his message. The arrival rate of messages in class $i$ is $\lambda_i(\mathbf{p})$. The actual expected delay for a packet in class $i$, $\delta_i(\mathbf{p})$, can be estimated from these arrival rates using an appropriate queueing model. The provider needs to guarantee that the actual expected delay $\delta_i(\mathbf{p})$ will be at most $d_i$. The provider's problem is

$$\max_{\mathbf{p}} \sum_{i=1}^{N} \lambda_i g_i p_i$$
$$\text{s.t. } \delta_i(\mathbf{p}) \le d_i, \quad \forall\, i \in \{1, 2, \ldots, N\},$$

where $g_i$ is the average number of packets in a message sent using priority class $i$. We maximize the total revenue generated by a pricing strategy. In a packet-switching network, the marginal cost of a packet-transmission service is negligible (in the short run when the capacity is fixed), so this is the profit-maximizing formulation as well.

### 3.1. Characterization of User Behavior

Since the arrival rate of messages in each class depends on user self-selection, we need a model of user behavior. We characterize a user by two parameters: his value parameter $v$ and his delay-sensitivity parameter $h$. A user's value parameter $v$ indicates the per-packet value realized from the transmission of a message, and we assume it is uniformly distributed on $[0, 1]$. A user's delay-sensitivity parameter $h$ is the disutility to the user if the transmission of a packet is delayed by one time unit. The delay cost is assumed linear in the delay-sensitivity: $h\gamma$ per time unit of delay, where $\gamma$ is constant. We assume that $h$ is also uniformly distributed on $[0, 1]$.

If the user with $(v, h)$ chooses to send a message in priority class $i$, then he receives a value of $v$ per packet, pays a per-packet price of $p_i$, and incurs a per-packet delay cost of $h\gamma d_i$. One may argue that the per-packet delay cost to the user is actually $h\gamma\delta_i$, and not $h\gamma d_i$, but at the time of choosing the service, the user does not know $\delta_i$, and would thus base his decision on $d_i$ (the upper bound on the expected delay). Thus, the delay-adjusted net value (to the user) of sending the packet in class $i$ is $v - p_i - h\gamma d_i$, so for a given message to be sent, the user's problem is $\max_i(v - p_i - h\gamma d_i)$.

### 3.2. Estimation of Arrival Rate

Let $i = N + 1$ be the no-service class; i.e., $p_{N+1} = 0$ and $\gamma d_{N+1} = 1$. Let $H_i$ be the $h$-value so that a user is indifferent between priority classes $i$ and $i+1$. This implies that $v - p_i - H_i\gamma d_i = v - p_{i+1} - H_i\gamma d_{i+1}$, or

$$H_i = \frac{p_i - p_{i+1}}{\gamma(d_{i+1} - d_i)}, \quad i = 1, 2, \ldots, N.$$

A user with $H_i \le h < H_{i-1}$ should prefer class $i$, where $H_0 = 1$; this is the *incentive-compatibility constraint* (ICC). Though a class may dominate another, the former would be chosen by the user only if it

results in a nonnegative net value since the user has the option of not sending the message at all (i.e., of not choosing any of the classes), getting a net value of zero. This implies $v - p_i - h\gamma d_i \geq 0$, or $v \geq p_i + h\gamma d_i$; this is the usual *individual rationality constraint* (IRC).

Before we can develop the IRC further, we need to understand the inter-dependence between $v$ and $h$. For example, one could reasonably claim that a high value of $v$ is usually associated with a high value of $h$. However, it is difficult to ascertain the exact nature of the relationship between $v$ and $h$. We assume that $v$ and $h$ are linearly related. The other extreme case where $v$ and $h$ are independent is similar, both in terms of the analysis and results, and is summarized in the Online Supplement to this paper on the journal's website. Since $v$ and $h$ are both uniformly distributed on $[0, 1]$, either $v = h$ or $v = 1 - h$; we simply assume that $v = h$. Hence, the IRC for class $i$ can be written as $h \geq p_i/(1 - \gamma d_i)$, where $0 \leq \gamma d_i < 1$. Let $V_i = p_i/(1 - \gamma d_i)$. Then combining ICC and IRC, the arrival rate of messages in class $i$ is $\lambda_i = \lambda_0[\max\{H_{i-1}, V_i\} - \max\{H_i, V_i\}]$, where $\lambda_0$ is the total traffic intensity; i.e., $\lambda_0$ is the total arrival rate of all messages in all classes when there is no delay and the communication service is free.

### 3.3. Estimation of Expected Delay

To estimate the expected delay per packet in each priority class, we assume a homogeneous priority batch arrival queueing model $M^x/G/1/Pr$, where each batch represents a message consisting of packets. Although the stationary Poisson process has been commonly used to describe the arrival process of messages in networks, doubts have been raised about the validity of this assumption (Leland et al. 1994, Paxson and Floyd 1995). It is, for example, well-known that the arrival rate of messages tends to change with the time of the day. However, such diurnal changes have been observed to be very slow (Nuzman et al. 2002), and it is possible to find time intervals during which the arrival rate remains steady (Roberts 2004). For each such time period, we model the arrival of messages as a stationary Poisson process. Further empirical justification for this assumption, especially in a high-speed network with a high degree of multiplexing, is provided by Karagiannis et al. (2004). They analyze three different traffic traces from high-speed backbone networks and observe that network traffic is well characterized by a stationary Poisson process at a sub-second level and, beyond that, by a sequence of stationary Poisson processes.

The batch-arrival process ensures that all packets in a message are of the same priority. We use nonpreemptive priority at the packet level, which is reasonable. For the sake of analytical tractability, we treat the entire network as a single server with a capacity limit; this may appear to be an oversimplification, but a private network of a single provider can often be viewed as a single-node network. Since we study pricing issues for a single provider, this assumption is reasonable as well.

For nonpreemptive priority queues, the mean waiting time of an *arbitrary* packet of class $i$ is (Takagi 1991, Takagi and Takahashi 1991)

$$E[W_i] = \frac{\sum_{k=1}^{i} \lambda_k g_k^{(2)} b_k^2 + \sum_{k=1}^{N} \lambda_k g_k b_k^{(2)}}{2\left(1 - \sum_{k=1}^{i-1} \rho_k\right)\left(1 - \sum_{k=1}^{i} \rho_k\right)} + \frac{g_i^{(2)} b_i}{2g_i\left(1 - \sum_{k=1}^{i-1} \rho_k\right)}.$$

Here $g_i$ and $g_i^{(2)}$ are the mean and the second factorial moment of the number of packets included in each batch of class $i$, $b_i$ and $b_i^{(2)}$ are the mean and the second moment of the service time for each packet of class $i$, and $\rho_i = \lambda_i g_i b_i$.

Since packets are of the same size, service time in each class is deterministic, so denote it as $b$; then $b^{(2)} = b^2$. Further, let $m$ be the number of packets in a message. Clearly, $m$ does not depend on the user's valuation of the message and hence the priority class. Therefore, the batch size for all priority classes follows the same distribution, so denote $g_i = g$ and $g_i^{(2)} = g^{(2)}$. In data communication, shorter messages are more likely than the longer ones, without a pre-specified upper bound on the message size. The geometric distribution fits these characteristics nicely. Hence, we assume that $m - 1$ follows a geometric distribution, so $g^{(2)} = 2g(g - 1)$. The expected delay an arbitrary packet is then

$$\begin{aligned}\delta_i &= b + E[W_i] \\ &= b + \frac{(\rho b/2)\sum_{k=1}^{N}(\lambda_k/\lambda_0) + (g-1)b}{\left(1 - \rho\sum_{k=1}^{i-1}(\lambda_k/\lambda_0)\right)\left(1 - \rho\sum_{k=1}^{i}(\lambda_k/\lambda_0)\right)},\end{aligned}$$

where $\rho = \lambda_0 g b > 0$ is the normalized total traffic intensity.

We need to clarify our strategy of pricing the expected delay of an arbitrary packet. One could argue that a user may not care about the delay experienced by an arbitrary packet in a message, and is more likely to care about the delay of the last packet. In a large network, however, data traffic would be aggregated over a large population, and the expected delay of an arbitrary packet is a good indicator of overall performance, so the expected delay of the last packet is closely related to the expected delay of an arbitrary packet. In fact, they are identical in our model. Since the number of packets in a message is assumed to follow a shifted geometric distribution, the waiting-time distribution of the last packet coincides with that of an arbitrary packet (Halfin 1983, Takagi 1991). Since the service time for a packet is fixed, we get

PROPOSITION 3.1. *The expected delay of the last packet in a message is the same as the expected delay of an arbitrary packet.*

## 4. Pricing Scheme with Fixed Delay

We now illustrate the optimal pricing for the special cases of a single class and two classes of service.

### 4.1. Case P1: Single Class of Service

Since there is only one class, we drop the subscript denoting the class of service. Thus, the effective arrival rate of messages is simply $\lambda = \lambda_0(1 - p/(1-\gamma d))$ and the service provider's optimization problem becomes

$$\textbf{(P4.1)} \quad \max_p \; z = p\left(1 - \frac{p}{1-\gamma d}\right)$$
$$\text{s.t.} \quad \frac{(\rho b/2)\left(1-p/(1-\gamma d)\right)+(g-1)b}{1-\rho\left(1-p/(1-\gamma d)\right)} \le d-b,$$
$$0 \le \frac{p}{1-\gamma d} \le 1.$$

Proposition 4.1. *The optimal price $p^*$ as a solution to* **(P4.1)** *is given by*

$$p^* = \begin{cases} \dfrac{1}{2}(1-\gamma d), & \text{if } \rho < u = \dfrac{4(d-gb)}{2d-b}, \\ (1-\gamma d)\left[1 - \dfrac{2(d-gb)}{\rho(2d-b)}\right], & \text{otherwise.} \end{cases}$$

For proof, see the Online Supplement. It is easy to verify that this solution satisfies the stability condition that $\lambda g b < 1$ (i.e., $p/(1-\gamma d) > 1 - 1/\rho$), and that the optimal price $p^*$ decreases with $\gamma$ and $d$. Clearly, an increase in $\gamma$ or $d$ implies an increase in users' disutility from delay, so to provide a net positive value to the user, one must decrease $p^*$. On the other hand, $p^*$ does not change with $b$ when the traffic intensity is low, but, beyond a threshold ($\rho = u$), $p^*$ increases with $b$. This is because, at low traffic intensity, the delay constraint is not binding and the optimal price does not depend on capacity. However, as traffic intensity increases beyond a threshold, the delay constraint becomes binding—any decrease in capacity (i.e., an increase in $b$) has to be accompanied by a corresponding increase in price to maintain the delay guarantee.

Figure 1 shows how the optimal price changes with the normalized total traffic intensity, $\rho$, when $g = 100$, $b = 0.000275$ second, and $d = 0.06$ second, for three different values of $\gamma$. The optimal price is flat under low values of $\rho$ since there is no externality effect when the traffic intensity is low. However, the price starts increasing after the traffic intensity reaches a threshold characterized by $\rho = u$; beyond this point, the provider must keep increasing the price to deliver guaranteed service. Variation of price with the traffic intensity has clear implication for the service provider. Since traffic intensity may be different at different time periods, the provider can design and post an efficient price schedule for each time period, based on her prior knowledge of diurnal traffic patterns.

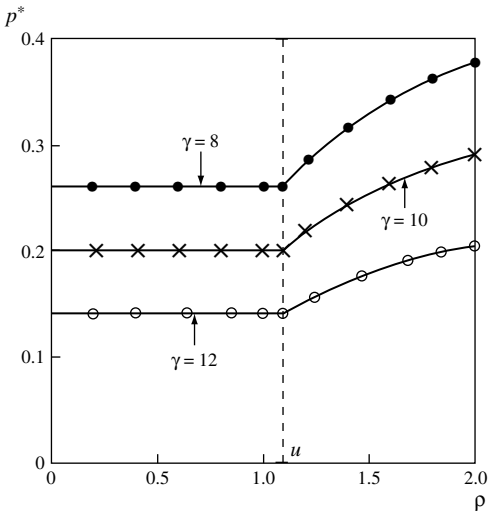


**Figure 1 Optimal Price as a Function of $\rho$ in Case P1**

By substituting the optimal price into the expressions for the consumer surplus ($\sigma_c$) and the provider surplus ($\sigma_p$), we get

$$\sigma_c = \lambda_0 g \int_{p^*/(1-\gamma d)}^{1} (h - p^* - h\gamma d)\, dh$$
$$= \begin{cases} \dfrac{\rho}{8b}(1-\gamma d), & \text{if } \rho < u, \\ \dfrac{\rho}{2b}(1-\gamma d)\left(\dfrac{2(d-gb)}{\rho(2d-b)}\right)^2, & \text{otherwise.} \end{cases}$$

and

$$\sigma_p = \lambda_0 g p^* \left(1 - \frac{p^*}{1-\gamma d}\right)$$
$$= \begin{cases} \dfrac{\rho}{4b}(1-\gamma d), & \text{if } \rho < u, \\ \dfrac{\rho}{b}(1-\gamma d)\dfrac{2(d-gb)}{\rho(2d-b)}\left[1 - \dfrac{2(d-gb)}{\rho(2d-b)}\right], & \text{otherwise.} \end{cases}$$

Both $\sigma_c$ and $\sigma_p$ decrease with $\gamma$ or $d$—as the disutility from delay increases, both the consumers and provider are taxed. The capacity parameter $b$ does not affect the surplus when $\rho < u$, but as the traffic intensity increases beyond this threshold, both $\sigma_c$ and $\sigma_p$ start decreasing with $b$. Figure 2 shows how $\sigma_c$, $\sigma_p$, and the overall social welfare, $\sigma_t = \sigma_c + \sigma_p$, change with the normalized total traffic intensity $\rho$ when $g = 100$, $b = 0.000275$ second, $\gamma = 10$, and $d = 0.06$ second. The consumer surplus increases with $\rho$ until the delay constraint becomes binding because more consumers are served at a flat price as $\rho$ increases. However, as $\rho$ increases beyond $4(d-gb)/(2d-b)$, the consumer surplus decreases because the price increases while the effective demand does not change. On the other hand, the provider surplus and the total social welfare consistently increase with $\rho$, although at lower rates (due

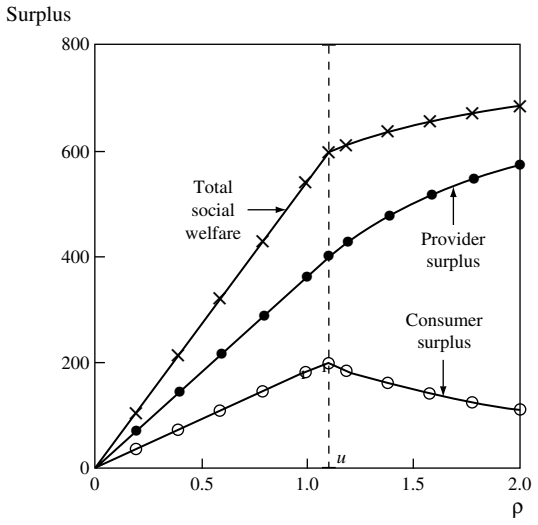


**Figure 2** **Consumer Surplus, Provider Surplus, and Total Social Welfare in Case P1**

to negative network externality) after the delay constraint becomes binding.

### 4.2. Case P2: Two Classes of Services

Here, $H_1 = (p_1 - p_2)/(\gamma(d_2 - d_1))$, $V_1 = p_1/(1 - \gamma d_1)$, and $H_2 = V_2 = p_2/(1 - \gamma d_2)$. Thus the arrival rates in the two classes are $\lambda_1 = \lambda_0[1 - \max\{H_1, V_1\}]$ and $\lambda_2 = \lambda_0[\max\{H_1, V_2\} - V_2]$. The following result is necessary (see the Online Supplement for proof).

Proposition 4.2. *Both $V_1$ and $V_2$ are on the same side of $H_1$; i.e., they are either both less than $H_1$ or they are both greater than $H_1$.*

Therefore, we need to examine only two simple cases: (i) $V_1, V_2 \le H_1$, or (ii) $V_1, V_2 > H_1$. If $V_2 \le H_1$, $\lambda_1 = \lambda_0(1 - H_1)$ and $\lambda_2 = \lambda_0(H_1 - V_2)$, where $\lambda_0$ is the total traffic intensity, as defined earlier. Also, $p_2 = V_2(1 - \gamma d_2)$ and $p_1 - p_2 = H_1\gamma(d_2 - d_1)$. Substituting these into the objective function and dropping the multiplicative constant $\lambda_0 g > 0$, we get the provider's optimization problem

$$\textbf{(P4.2)} \quad \max_{H_1, V_2} \; z = \gamma(d_2 - d_1)H_1(1 - H_1) + (1 - \gamma d_2)V_2(1 - V_2)$$
$$\text{s.t.} \quad \frac{(\rho b/2)(1 - V_2) + (g - 1)b}{1 - \rho(1 - H_1)} \le d_1 - b,$$
$$\frac{(\rho b/2)(1 - V_2) + (g - 1)b}{(1 - \rho(1 - H_1))(1 - \rho(1 - V_2))} \le d_2 - b,$$
$$0 \le V_2 \le H_1 \le 1.$$

If, on the other hand, $V_1 > H_1$, then $\lambda_1 = \lambda_0(1 - V_1)$ and $\lambda_2 = 0$. Hence, we can substitute $p_1 = V_1(1 - \gamma d_1)$ to obtain

$$\max_{V_1} \; (1 - \gamma d_1)V_1(1 - V_1)$$
$$\text{s.t.} \quad \frac{(\rho b/2)(1 - V_1) + (g - 1)b}{1 - \rho(1 - V_1)} \le d_1 - b,$$
$$\frac{(\rho b/2)(1 - V_1) + (g - 1)b}{(1 - \rho(1 - V_1))^2} \le d_2 - b,$$
$$0 \le V_1 \le 1.$$

However, the second optimization problem is subsumed by **(P4.2)**. To see this, assume that $V_1^*$ is the solution to the second problem. By setting $H_1 = V_2 = V_1^*$, we get a feasible solution to **(P4.2)** with the same objective-function value.

**(P4.2)** is a nonlinear constrained optimization problem with multiple inequality constraints. To solve it, we need to examine the delay constraints for the two priority classes more closely. Temporarily assume that the delay constraint of the low-priority class would be binding only after that of the high-priority class becomes binding. This would be the case when $d_1$ and $d_2$ are well separated. More specifically, as long as $d_2/d_1 \ge d_1/(gb)$, the delay constraint of the high-priority class would become binding before that of the low-priority class. A provider would like to have a significant level of differentiation in the quality levels of the two classes, so this condition is reasonable.

To simplify this case further, we note that, in reality, the expected number of packets in a message would be large, i.e., $g \gg 1$. This implies that $d_1, d_2 > gb \gg b$. Under these assumptions, we have (see the Online Supplement for proof).

Proposition 4.3. *The solution to* **(P4.2)** *is given by $H_1^* = 1 - \min\{1/2, u_1/(2\rho)\}$ and $V_2^* = 1 - \min\{1/2, u_2/(2\rho)\}$, where $u_1 = 2 - 2gb/d_1$ and $u_2 = 2 - 2d_1/d_2$. The resulting optimal prices are $p_1^* = V_2^*(1 - \gamma d_2) + H_1^*\gamma \cdot (d_2 - d_1)$ and $p_2^* = V_2^*(1 - \gamma d_2)$.*

Suppose $g = 100$, $b = 0.000275$ second, $\gamma = 10$, $d_1 = 0.03$ second, and $d_2 = 0.06$ second. Then $d_2/d_1 = 2 > 1.09 = d_1/(gb)$, so the results in Proposition 4.3 hold. Figure 3 shows how the optimal prices change with $\rho$ in this case. There are several interesting observations. First, the optimal prices for both the classes are flat under low traffic intensity ($\rho \le u_1$). This is expected. For low values of $\rho$, a user message imposes little queueing delay (i.e., no negative externality) on the other messages. Since the marginal cost is zero, the flat prices are simply the provider's monopoly premium. Further, for low values of $\rho$, the effective demand to the low-priority class is zero (as $H_1^* = V_2^*$), i.e., the provider would offer only high priority. Second, under moderate traffic intensity ($u_1 < \rho \le u_2$), the price of the high-priority class increases while the price of the low-priority class remains flat. When the normalized traffic intensity increases beyond $u_2$, both the delay constraints become binding. As a result, both prices start to increase, but the price of high priority increases at a higher rate. This phenomenon of

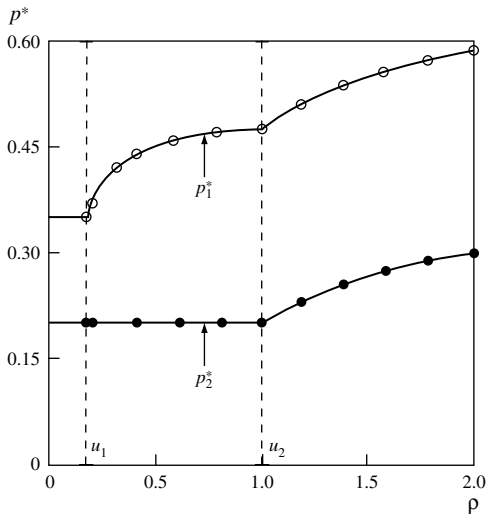


**Figure 3** **Optimal Prices as a Function of $\rho$ in Case P2**

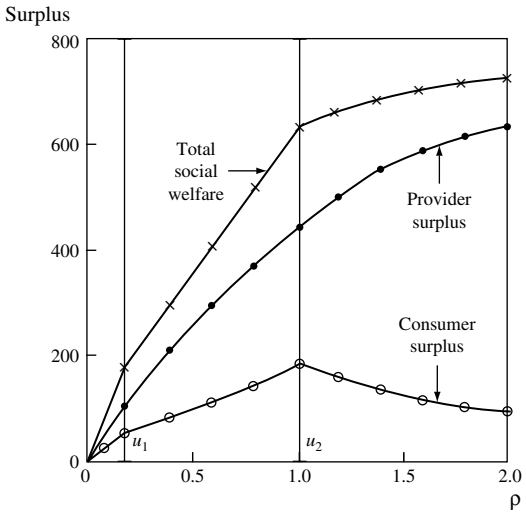


**Figure 4** **Consumer Surplus, Provider Surplus, and Total Social Welfare in Case P2**

different rates of increase can be easily explained via negative externality. A message from a low-priority class can only delay messages in its own class, thereby imposing a lower delay cost and a smaller negative externality effect. A message from a high-priority class, on the other hand, imposes a delay cost on messages in its own class as well as on messages in the low class. As traffic intensity increases, in order to maintain the delay guarantees, it is necessary to tax the high-priority users at a higher rate for this negative externality. This effect, of course, is negligible for the low-priority users.

We now estimate the consumer surplus, the provider surplus, and the total social welfare arising from this pricing scheme. The consumer surplus ($\sigma_c$) at the optimal prices is

$$\begin{aligned}\sigma_c &= \lambda_0 g\left[\int_{H_1^*}^{1}(h-p_1^*-h\gamma d_1)\,dh+\int_{V_2^*}^{H_1^*}(h-p_2^*-h\gamma d_2)\,dh\right]\\ &= \frac{\rho}{b}\left[\frac{1-\gamma d_1}{2}(1-H_1^{*2})+p_1^*(H_1^*-1)\right.\\ &\qquad\left.+\frac{1-\gamma d_2}{2}(H_1^{*2}-V_2^{*2})+p_2^*(V_2^*-H_1^*)\right].\end{aligned}$$

The provider surplus ($\sigma_p$) is simply the provider's profit

$$\sigma_p=\frac{\rho}{b}[\gamma(d_2-d_1)H_1^*(1-H_1^*)+(1-\gamma d_2)V_2^*(1-V_2^*)].$$

Figure 4 shows how the consumer surplus, the provider surplus, and the total social welfare ($\sigma_t=\sigma_c+\sigma_p$) change with $\rho$. The consumer surplus increases with $\rho$ until both the delay constraints become binding ($\rho \le u_2$), although at a lower rate after the delay constraint of the high-priority class becomes binding ($\rho \ge u_1$). This is because the effective demand for both classes increase with $\rho$ while the prices do not change ($\rho < u_1$), but the price of the high-priority class increases after $\rho \ge u_1$. As $\rho$ increases beyond $u_2$, the consumer surplus starts decreasing because both prices increase while the total effective demand $\lambda_1+\lambda_2$ remains constant. On the other hand, both the provider surplus and the total social welfare consistently increase with $\rho$, although at lower rates (due to negative network externality) after delay constraints become binding.

Now consider the unlikely case where the delay parameters for the two classes are not well separated, i.e., $d_2/d_1 < d_1/(gb)$. The delay constraint of the high-priority class would always be slack. At low traffic intensities, both the delay constraints would be slack, leading to a solution of $H_1^*=V_2^*=\frac{1}{2}$. When the traffic intensity increases beyond a threshold, the delay constraint of only the low-priority class becomes binding. Substituting this constraint into the objective function and maximizing it, we get a quartic (fourth order) equation that can be solved. We exclude the details here because, as mentioned earlier, this situation arises only when $d_1$ and $d_2$ are close to each other, which is unrealistic.

## 5. Pricing Scheme with Variable Delay

In real situations, there could be lack of consensus regarding the standard QoS level, and different service providers may offer different per-packet delay guarantees. While the difference in the QoS guarantee might be attributed to technology disparity among providers, it is usually more of a provider's strategy because of the significant polarization of the data-networking market. This section considers a situation where a provider can choose her own delay as well as price parameters. In this case, the provider first decides the expected per-packet delay and sets

a per-packet price for each class. As before, the users then self-select an appropriate class of service according to their delay and price sensitivities. Hence, the provider's problem is now choosing the delay as well as the price parameters to maximize her expected profit while maintaining the delay guarantees.

### 5.1. Case Q1: Single Class of Service

In this case, the provider's problem is similar to **(P4.1)** except that both $p$ and $d$ are decision variables, so the optimal solution is obtained only when the delay constraint is binding. To see this, temporarily assume that $(p^*, d^*)$ is the optimal solution and the delay constraint is not binding. Then it is possible to increase the objective function by lowering $d$ (because $\lambda$ increases) until the delay constraint becomes binding. The original solution, therefore, could not have been optimal. Thus, the delay constraint must be binding. Furthermore, as before, $g \gg 1$ implies that $d > gb \gg b$. The binding delay constraint can then be simplified as $p = (1-\gamma d)[1-(d-gb)/(\rho d)]$. Substituting $p$ into the objective function, we get

$$\textbf{(P5.1)} \quad \max_d z = (1-\gamma d)\frac{d-gb}{\rho d}\left[1-\frac{d-gb}{\rho d}\right].$$

Proposition 5.1. *The solution $d^*$ to* **(P5.1)** *satisfies $d^{*3} + Qd^* + R = 0$, where $Q = (\rho gb - 2gb - g^2b^2\gamma)/(\gamma(1-\rho))$ and $R = 2g^2b^2/(\gamma(1-\rho))$.*

See the Online Supplement for proof. The above cubic equation can be solved analytically. Once $d^*$ is obtained, $p^*$ can be found from the above binding delay constraint. Numerical results are in Figure 5 where $g = 100$, $b = 0.000275$ second, and $\gamma = 10$. The optimal $d^*$ strictly increases with $\rho$, so the expected delay guaranteed by a provider increases as more consumers join the network. In a market without externality effects, this would often imply lower prices (to encourage consumers to buy inferior products or services). However, this situation is not always true in a data-communication network where strong negative externality effects might occur. It is clear from Figure 5 that the optimal price $p^*$ decreases slightly (commensurate with the low service quality) under low traffic. However, as $\rho$ reaches a threshold (around $\rho = 0.6$), the negative externality becomes significant—the provider must keep increasing $p^*$ to deliver the guaranteed expected delay. In other words, when the traffic intensity is low, the provider uses the delay parameter as a tool for admission control. However, for high traffic intensity, admission control happens using price as well. At still higher traffic intensity, the optimal delay becomes approximately fixed and admission control primarily happens through price.

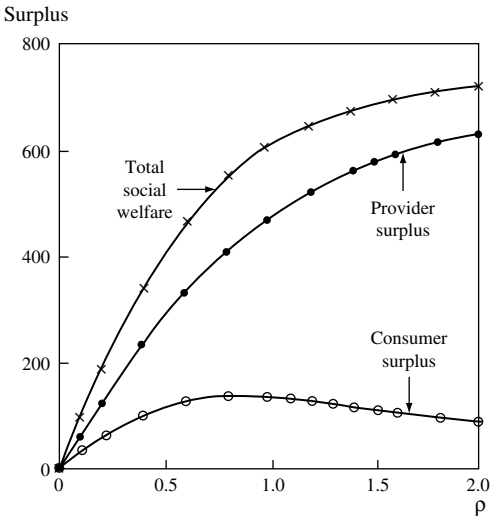


**Figure 6** **Consumer Surplus, Provider Surplus, and Total Social Welfare in Case Q1**

Figure 6 shows how $\sigma_c$, $\sigma_p$, and $\sigma_t$ change with $\rho$ when $g = 100$, $b = 0.000275$ second, and $\gamma = 10$. As expected, the provider surplus increases with $\rho$. The consumer surplus increases with $\rho$ only under low traffic; when the traffic is above a threshold (around $\rho = 0.6$), the consumer surplus decreases with $\rho$ because of the increasing price and high delay. Total social welfare increases steadily with $\rho$, but at a lower rate once $\rho$ is above around 0.6.

### 5.2. Case Q2: Two Classes of Services

This problem is similar to **(P4.2)** except that $d_1$, $d_2$, $p_1$, and $p_2$ are all decision variables. Both the delay

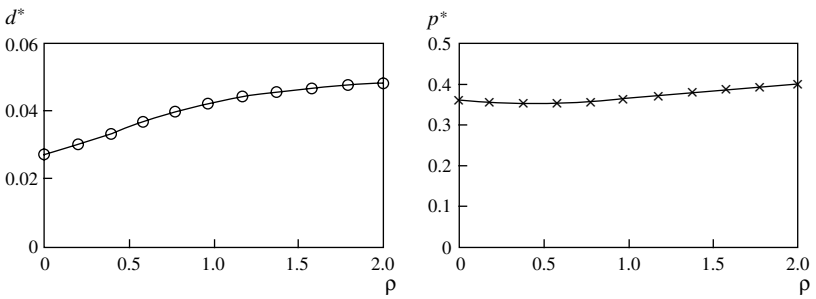


**Figure 5** **Optimal Delay and Price as Functions of $\rho$ in Case Q1**

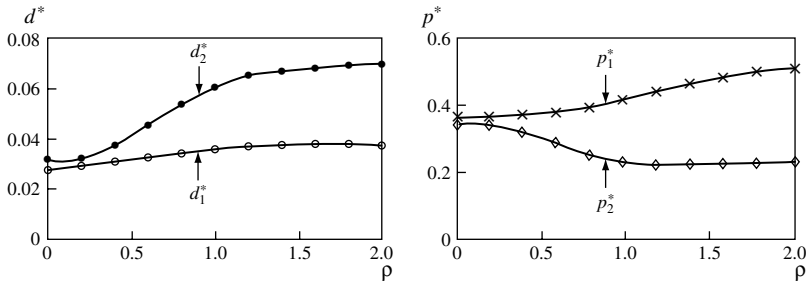


**Figure 7 Optimal Delays and Prices as Functions of $\rho$ in Case Q2**

constraints must be binding since otherwise it is possible to increase the objective function by lowering $d_1$ or $d_2$ until both the delay constraints become binding. From Proposition 4.3, we get $H_1 = 1-(d_1-gb)/(\rho d_1)$ and $V_2 = 1-(d_2-d_1)/(\rho d_2)$. Substituting these into the objective function of **(P4.2)**, we get

$$\textbf{(P5.2)} \quad \max_{d_1,\,d_2} z = \gamma(d_2-d_1)\left(\rho-1+\frac{gb}{d_1}\right)\left(1-\frac{gb}{d_1}\right) + (1-\gamma d_2)\left(\rho-1+\frac{d_1}{d_2}\right)\left(1-\frac{d_1}{d_2}\right).$$

Proposition 5.2. *The solution to* **(P5.2)** *is* $d_1^* = gbf$ *and* $d_2^* = gbf^3$, *where* $f$ *satisfies* $(2-\rho)\gamma gbf^6 - \gamma gbf^5 - \gamma gbf^3 - (2-\rho)f^2 + 2 = 0$.

See the Online Supplement for proof. Once we get $d_1^*$ and $d_2^*$, we can substitute these values to get $H_1^*$ and $V_2^*$, and then $p_1^*$ and $p_2^*$. Numerical results (in terms of optimal prices and delays) are in Figure 7 where $g = 100$, $b = 0.000275$ second, and $\gamma = 10$. As expected, both the prices and the delays are relatively flat at low traffic. As traffic increases, optimal delays steadily increase before leveling off beyond a threshold (around $\rho = 1.4$). Beyond this, the provider increases prices to deliver guaranteed service. Overall, the price of the low-priority class first decreases with $\rho$ and then increases under heavier traffic because of negative externality effects. The price of the high-priority class, on the other hand, monotonically increases with $\rho$. Finally, Figure 8 shows how the consumer surplus, the provider surplus, and the social welfare change with $\rho$. These trends are similar to the ones observed in the single-class case.

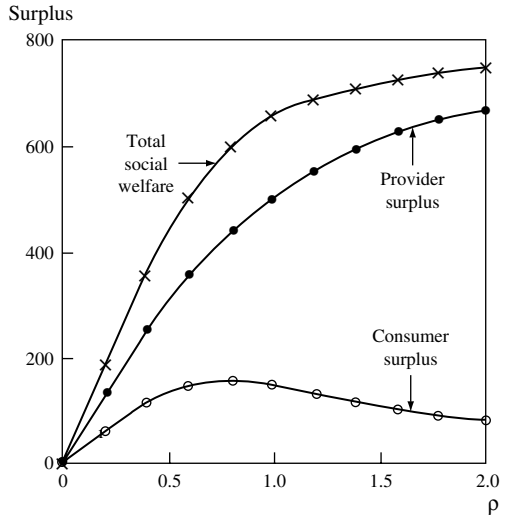


**Figure 8 Consumer Surplus, Provider Surplus, and Total Social Welfare in Case Q2**

## 6. Capacity Planning and the Long-Run Problem

We now consider long-run pricing where the provider can also plan for network capacity. The capacity cost function is assumed linear (Mendelson 1985), i.e., the cost for capacity $1/b$ is simply $c/b$, where the constant $c$ is the cost of a unit capacity per unit time (amortized over the life of the capacity). For simplicity, we assume that the delay parameters are fixed. The long-run pricing problem is then

$$\max_{\mathbf{p},\,b} \sum_{i=1}^{N} \lambda_i g_i p_i - \frac{c}{b}$$
$$\text{s.t.} \quad \delta_i(\mathbf{p}) \le d_i, \quad \forall\, i \in \{1, 2, \ldots, N\}.$$

### 6.1. Case L1: Single Class of Service

The provider's problem simplifies to

$$\max_{p,\,b} \lambda g p - \frac{c}{b}, \quad \text{s.t.} \quad \delta = d.$$

The delay constraint is strictly binding in the above formulation. To see this, temporarily assume that the optimal solution to the long-run problem is obtained at $(p^*, b^*)$, but the delay constraint is not binding at the optimum. Then the provider could increase her profit simply by increasing $b$ until the delay constraint becomes binding, without changing $p$ from its original value of $p^*$. This is because the average response time $\delta$ increases if the provider increases $b$; the effective demand $\lambda$, on the other hand, remains unchanged because it depends only on price $p$. Clearly, the original solution could not have been optimal. This fact,

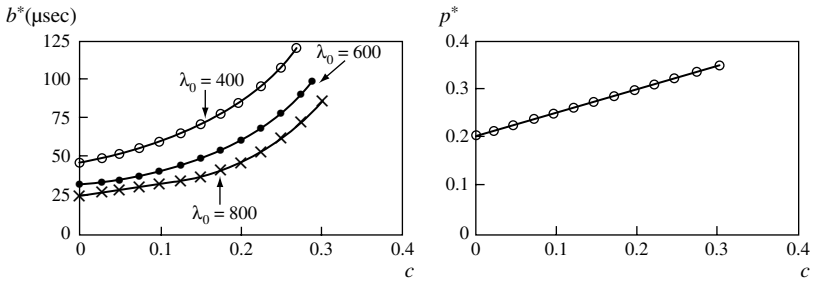


**Figure 9 Optimal Service Time and Price as Functions of $c$ in Case L1**

along with $d \gg b$, allows us to simplify the delay constraint to $p = (1-\gamma d)[1-(d-gb)/(\rho d)]$, which can be substituted into the objective function to get

$$\textbf{(P6.1)} \quad \max_b z = (1-\gamma d)\frac{d-gb}{\rho d}\left[1-\frac{d-gb}{\rho d}\right]-\frac{c}{\rho}.$$

Proposition 6.1. *The optimal solution to* **(P6.1)** *is* $b^* = 2d(1-\gamma d)/(g[(1-\gamma d)(2+\lambda_0 d)-c\lambda_0 d])$ *if* $c < \bar{c}$, *where* $\bar{c} = ((1-\gamma d)/\lambda_0 d)[1-\sqrt{1+\lambda_0 d}]^2$. *The provider will neither invest in capacity nor provide any service if* $c \geq \bar{c}$.

See the Online Supplement for proof. Substituting $b^*$, we get $p^* = (1-\gamma d+c)/2$. While $b^*$ depends on the traffic intensity $\lambda_0$, $p^*$ does not depend on $\lambda_0$, i.e., the provider tries to accommodate overall demand by capacity planning. Of course, if the demand is later different from the estimate used in capacity planning, the provider may charge a different price if changing capacity is not feasible. This is consistent with our basic intuition on how a provider should plan her capacity and react to changing demand.

Figure 9 shows how the optimal capacity and price changes with the cost of capacity $c$, for $g = 100$, $\gamma = 10$, and $d = 0.06$ second. As expected, capacity decreases (i.e., per-packet service time $b$ increases), and optimal price increases, as the cost of capacity increases. Also, for a given $c$, as the traffic intensity $\lambda_0$ increases, the provider attempts to accommodate the extra demand by increasing capacity (i.e., decreasing $b$).

### 6.2. Case L2: Two Classes of Services

The provider's optimization problem becomes

$$\begin{aligned}\max_{H_1,V_2,b}\ & \lambda_0 g[\gamma(d_2-d_1)H_1(1-H_1)\\ &\qquad +(1-\gamma d_2)V_2(1-V_2)]-\frac{c}{b}\\ \text{s.t.}\ & \frac{(\rho b/2)(1-V_2)+(g-1)b}{1-\rho(1-H_1)} \leq d_1-b,\\ & \frac{(\rho b/2)(1-V_2)+(g-1)b}{(1-\rho(1-H_1))(1-\rho(1-V_2))} \leq d_2-b,\\ & 0 \leq V_2 \leq H_1 \leq 1.\end{aligned}$$

Since $b$ is a decision variable, at least one of the delay constraints must be binding. Assuming that $d_1$ and $d_2$ are well separated (i.e., $d_2/d_1 \geq d_1/(gb)$), we claim that the constraint for class 1 is binding while that for class 2 may be slack. Then, using the result from Proposition 4.3, we rewrite the optimization problem as

$$\textbf{(P6.2)} \quad \max_b z = \gamma(d_2-d_1)H_1(1-H_1) + (1-\gamma d_2)V_2(1-V_2)-\frac{c}{\rho},$$

where $H_1 = 1-(d_1-gb)/(\rho d_1)$ and $V_2 = 1-\min\{1/2, (d_2-d_1)/(\rho d_2)\}$.

Proposition 6.2. *If* $\lambda_0 \geq \bar{\lambda}_0$ *or* $c < \bar{c}$, *the optimal solution to* **(P6.2)** *is*

$$b^* = \frac{2d_1\big(\gamma+((d_2-d_1)/d_2)B\big)}{g\big[\gamma(2+\lambda_0 d_1)+B\lambda_0 d_1 - c\lambda_0 d_1/(d_2-d_1)\big]}, \quad \textit{where}$$

$$B = \begin{cases}\dfrac{1-\gamma d_2}{d_2}, & \textit{if } c > \dfrac{2\gamma(d_2-d_1)}{\lambda_0 d_1}-\gamma d_1,\\ 0, & \textit{otherwise,}\end{cases}$$

*and* $\bar{\lambda}_0$, $\bar{c}$ *are*

$$\bar{\lambda}_0 = \frac{2\big[\gamma(d_2-d_1)+\sqrt{\gamma(d_2-d_1)(1-\gamma d_1)}\big]}{d_1(1-\gamma d_2)}, \quad \textit{and}$$

$$\bar{c} = \frac{\gamma(d_2-d_1)}{\lambda_0 d_1}\left[2+\lambda_0 d_1+\frac{B\lambda_0 d_1}{\gamma}-2\left(\left(1+\frac{B(d_2-d_1)}{\gamma d_2}\right)\cdot(1+\lambda_0 d_1)+\frac{(Bd_2+\gamma d_2-1)\lambda_0^2 d_1^2}{4\gamma(d_2-d_1)}\right)^{1/2}\right].$$

*The provider will not provide any service if* $\lambda_0 < \bar{\lambda}_0$ *and* $c \geq \bar{c}$.

See the Online Supplement for proof. Substituting $b^*$, we obtain $H_1^*$ and $V_2^*$, and then $p_1^*$ and $p_2^*$. Figure 10 shows how the optimal capacity and prices change with the cost of capacity $c$, for $g = 100$, $\gamma = 10$, $d_1 = 0.03$ second, and $d_2 = 0.06$ second. As before, capacity decreases and optimal prices increase as the cost of capacity increases. For a given $c$, as the traffic intensity $\lambda_0$ increases, the provider attempts to accommodate the extra demand by increasing capacity (i.e.,

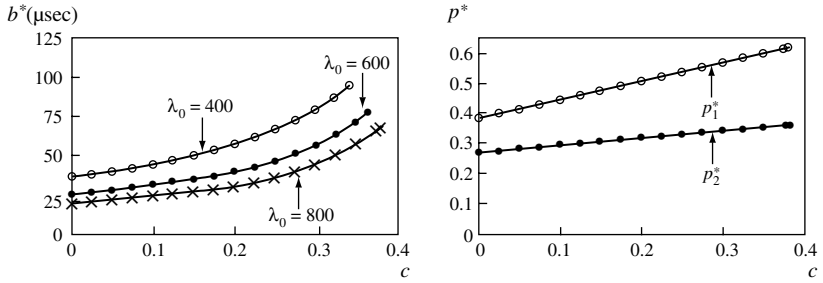


**Figure 10 Optimal Service Time and Prices as Functions of $c$ in Case L2**

decreasing $b$), but, as before, $\lambda_0$ has little impact on prices.

## 7. Other QoS Guarantees

We now extend our framework to other QoS dimensions such as jitter, packet-loss ratio, and maximum delay. For simplicity, we consider only the single-class case with a fixed delay guarantee (Section 4.1). Extending to multiple classes or to a variable delay guarantee poses no conceptual difficulty, but the details may be cumbersome.

When an additional QoS dimension is considered as a part of the pricing scheme, a new constraint must be introduced into the model. Except for a degenerate case, the QoS constraints in the optimization problem cannot both be binding. If the expected-delay constraint is binding, then the result of the pricing model would be exactly the same as in Section 4.1. Therefore, we illustrate the more interesting case where the expected-delay constraint is slack, and the new QoS constraint is binding.

### 7.1. Jitter

The variation in per-packet delay is called *jitter* (Peterson and Davie 2003). It can be measured as the variance of the per-packet delay and, for our queueing model, is (Takagi 1991)

$$\begin{aligned}\mathrm{Var}[\delta] &= \mathrm{Var}[W] = \mathrm{E}[W^2] - (\mathrm{E}[W])^2 \\ &= \frac{b^2(4\rho' - \rho'^2 + 12g(g-1))}{12(1-\rho')^2},\end{aligned}$$

where $\rho' = \rho(1 - p/(1-\gamma d))$. Let $\psi$ be the guaranteed maximum variance of the per-packet delay. The service provider's problem is then

$$\begin{aligned}\textbf{(P7.1)}\quad \max_p\ & z = p\left(1 - \frac{p}{1-\gamma d}\right) \\ \text{s.t.}\ & \frac{(\rho b/2)\big(1 - p/(1-\gamma d)\big) + (g-1)b}{1 - \rho\big(1 - p/(1-\gamma d)\big)} \le d - b, \\ & \frac{b^2(4\rho' - \rho'^2 + 12g(g-1))}{12(1-\rho')^2} \le \psi, \\ & \rho' = \rho\left(1 - \frac{p}{1-\gamma d}\right), \quad 0 \le \frac{p}{1-\gamma d} \le 1.\end{aligned}$$

Proposition 7.1. *When the jitter constraint is binding, the optimal price $p^*$ as a solution to* **(P7.1)** *is*

$$p^* = \begin{cases} \dfrac{1}{2}(1-\gamma d), & \text{if } \rho < 2\left(1 - \sqrt{\dfrac{g(g-1)b^2}{\psi}}\right), \\ (1-\gamma d)\left[1 - \dfrac{1}{\rho}\left(1 - \sqrt{\dfrac{g(g-1)b^2}{\psi}}\right)\right], & \text{otherwise.}\end{cases}$$

*When the jitter constraint is slack and the expected-delay constraint is binding, the solution to* **(P7.1)** *is obtained from Proposition* 4.1.

See the Online Supplement for proof. Figure 11 shows how the optimal price changes with $\rho$, when $g = 100$, $b = 0.000275$ second, $d = 0.06$ second, and $\psi = 0.003$ second$^2$, for three different values of $\gamma$. Comparing this with Figure 1 (obtained using a guarantee on only the expected per-packet delay), the results are similar. At low traffic, the jitter constraint remains slack, and the optimal price is flat. However, as traffic increases, the jitter constraint becomes binding at $\rho = 2(1 - \sqrt{g(g-1)b^2/\psi})$, beyond which the provider must keep increasing the price to deliver the promised jitter guarantee.

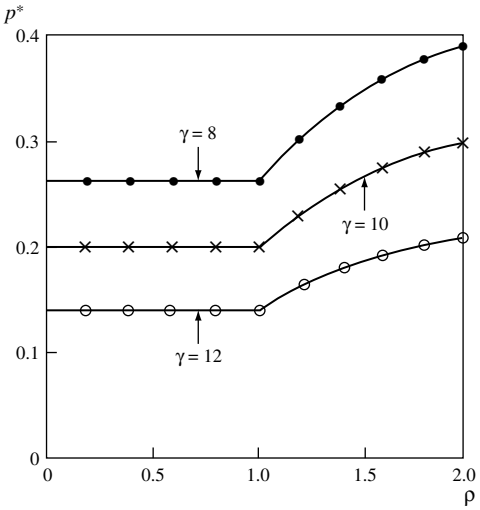


**Figure 11 Optimal Price as a Function of $\rho$ with Jitter Guarantee**

### 7.2. Maximal Delay and Packet-Loss Ratio

Now consider two other QoS dimensions, maximum per-packet delay and packet-loss ratio. The maximum per-packet delay is usually used as the time-out period in connection setup—a packet not delivered within this period is considered lost or dropped. Maximum per-packet delay is thus related to the packet-loss ratio: the larger the maximum delay, the lower the packet-loss ratio. Consider a service-level agreement that guarantees the packet-loss ratio to be 3% or less and the maximum per-packet delay 0.2 second or less. We demonstrate how such a constraint on the delay distribution can be incorporated into the pricing model.

For our queueing model, the distribution of the delay for an arbitrary packet is known (in the form of its Laplace transform). For a single-class case, the Laplace transform of the distribution for the waiting time (the total delay minus the service time) of a packet is given by Takagi (1991). Substituting $\lambda = \lambda_0(1-p/(1-\gamma d))$ in that expression, we get

$$\mathcal{W}^*(s) = \frac{s[(1-\rho)(1-\gamma d)+\rho p]}{g(1-e^{-sb})[(s-\lambda_0)(1-\gamma d)+p\lambda_0]+se^{-sb}(1-\gamma d)}.$$

The cumulative distribution for the waiting time of a packet can then be expressed in terms of an inverse Laplace transform

$$F_W(x) = \mathcal{L}^{-1}\left[\frac{(1-\rho)(1-\gamma d)+\rho p}{g(1-e^{-sb})\left[(s-\lambda_0)(1-\gamma d)+p\lambda_0\right]+se^{-sb}(1-\gamma d)}; x\right].$$

Let $d_m(f)$ be the maximum per-packet delay guarantee for $f$ fraction of packets; for example, if $f = 0.97$, $d_m$ would be the maximum delay allowable for 97% of the packets. The optimization problem becomes

$$\textbf{(P7.2)} \quad \max_p \; z = p\left(1-\frac{p}{1-\gamma d}\right)$$
$$\text{s.t.} \quad \frac{(\rho b/2)\left(1-p/(1-\gamma d)\right)+(g-1)b}{1-\rho\left(1-p/(1-\gamma d)\right)} \le d-b,$$
$$F_W^{-1}(f) \le d_m - b,$$
$$0 \le \frac{p}{1-\gamma d} \le 1.$$

Of course, analytical tractability is lost and a closed-form solution can no longer be found. However, we can obtain a numerical solution via Stehfest's (1970) approximation

$$F_W(x) \approx \frac{\ln 2}{x}\sum_{i=1}^{J}\phi_i \mathcal{W}^*\left(\frac{i\ln 2}{x}\right),$$

where the coefficients $\phi_i$ are

$$\phi_i = (-1)^{(J/2)+i}\sum_{j=\lfloor (i+1)/2\rfloor}^{\min(i,J/2)}\frac{j^{J/2}(2j)!}{\left(\frac{J}{2}-j\right)!j!(j-1)!(i-j)!(2i-j)!}$$

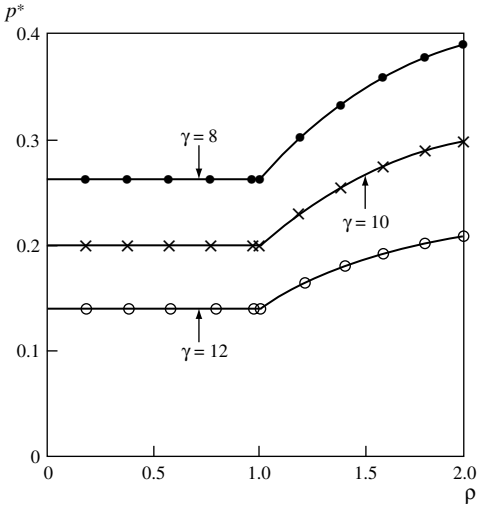


**Figure 12** **Optimal Price as a Function of $\rho$ with Maximum Delay Guarantee**

and $J$ is an even number, typically 14, 16, or at most 18. If we enforce the constraint $F_W^{-1}(f) \le d_m - b$ as $F_W(d_m - b) \ge f$, we get a nonlinear optimization model that can be solved. The results are presented in Figure 12 for $f = 0.97$ and $d_m = 0.2$ second, while all the other parameters are the same as in Figures 1 and 11. Comparing Figures 1, 11, and 12, a maximum-delay guarantee has a similar effect on the per-packet price as does an expected-delay or a jitter guarantee: At low levels of traffic intensity, the maximum-delay constraint is slack, resulting in a flat price. As traffic increases beyond a threshold, the price must be increased to meet the promised maximum-delay guarantee.

## 8. Conclusions

QoS and priority-based services on communication links have become increasingly important, so pricing communication services with QoS guarantees in multiple service classes is relevant for managing networking resources. We have proposed and analyzed an incentive-compatible pricing scheme for point-to-point data-communication links offered over a provider's private packet-switching network. We considered only a single monopolistic profit-maximizing service provider, who offers multiple classes of priority-based services with corresponding prices.

Our results would be of interest to service providers for better managing their resources in terms of designing service-level agreements, pricing communication links, and planning for capacity. Optimal prices and QoS guarantees change with traffic intensity, which helps the provider design and post an efficient price and QoS schedule. When traffic is low, it may be optimal for the provider to offer only a single class of service.

There are several possible directions for future research. We have considered pricing schemes within a single-node private network. Clearly, these results cannot be directly applied to a large public network, such as the Internet, where a large number of providers interconnect. Investigating pricing schemes in such networks, as well as transfer prices between providers, would be an exciting and a relevant exercise. In general, pricing a multi-node queueing network, with little or no central control, is a topic we are exploring. Analytical results are difficult to obtain, but one could resort to simulation experiments to obtain a more accurate representation and insights into the problem.

## Acknowledgments

The authors gratefully acknowledge the helpful comments and suggestions received from the Editor-in-Chief, the Area Editor, the Associate Editor, and the Reviewers of *INFORMS Journal on Computing*. The second and third authors were partially supported by a grant from the Ford Motor Company Fund.

## References

Cocchi, R., D. Estrin, S. Shenker, L. Zhang. 1991. A study of priority pricing in multiple service class networks. *ACM SIGCOMM Comput. Comm. Rev.* **21** 123–130.

Cocchi, R., S. Shenker, D. Estrin, L. Zhang. 1993. Pricing in computer networks: Motivation, formulation, and example. *ACM Trans. Networking* **1** 614–627.

DaSilva, L. A. 2000. Pricing for QoS-enabled networks. *IEEE Comm. Surveys* **3** 2–8.

Gupta, A., D. O. Stahl, A. B. Whinston. 1997. A stochastic equilibrium model of Internet pricing. *J. Econom. Dynam. Control* **21** 697–722.

Gupta, A., D. O. Stahl, A. B. Whinston. 1999. The economics of network management. *Comm. ACM* **42** 57–63.

Halfin, S. 1983. Batch delays versus customer delays. *The Bell System Tech. J.* **62** 2011–2015.

Karagiannis, T., M. Molle, M. Faloutsos. 2004. A nonstationary Poisson view of Internet traffic. *Proc. IEEE Infocom* **3** 1558–1569.

Keon, N. J., G. Anandalingam. 2003. Optimal pricing for multiple services in telecommunications networks offering quality of service guarantees. *IEEE/ACM Trans. Networking* **11** 66–80.

Kesh, S., S. Nerur, S. Ramanujan. 2002. Quality of service—Technology and implementation. *Inform. Management Comput. Security* **10** 85–91.

Leland, W. E., M. S. Taqqu, W. Willinger, D. V. Wilson. 1994. On the self-similar nature of Ethernet traffic. *IEEE/ACM Trans. Networking* **2** 1–15.

Low, S. H., P. P. Varaiya. 1993. A new approach to service provisioning in ATM networks. *IEEE/ACM Trans. Networking* **1** 547–553.

MacKie-Mason, J. K., H. R. Varian. 1994. Economic FAQs about the Internet. *J. Econom. Perspectives* **8** 75–96.

MacKie-Mason, J. K., H. R. Varian. 1995. Pricing congestible network resources. *IEEE J. Selected Areas Comm.* **13** 1141–1149.

MacKie-Mason, J. K., L. Murphy, J. Murphy. 1997. Responsive pricing in the Internet. L. McKnight, J. Bailey, eds. *Internet Economics*. MIT Press, Cambridge, MA, 279–303.

Marchand, M. G. 1974. Priority pricing. *Management Sci.* **20** 1131–1140.

Masuda, Y., S. Whang. 1999. Dynamic pricing for network service: Equilibrium and stability. *Management Sci.* **45** 857–869.

McDysan, D. 2000. *QoS and Traffic Management in IP & ATM Networks*. McGraw-Hill, New York.

Mendelson, H. 1985. Pricing computer services: Queueing effects. *Comm. ACM* **28** 312–321.

Mendelson, H., S. Whang. 1990. Optimal incentive-compatible priority pricing for the M/M/1 queue. *Oper. Res.* **38** 870–883.

Nuzman, C. J., I. Saniee, W. Sweldens, A. Weiss. 2002. A compound model for TCP connection arrivals for LAN and WAN applications. *Comput. Networks: Internat. J. Comput. Telecomm. Networking* **40** 319–337.

Parameswaran, M., J. Stallaert, A. B. Whinston. 2001. A market-based allocation mechanism for the DiffServ framework. *Decision Support Systems* **31** 351–361.

Paxson, V., S. Floyd. 1995. Wide area traffic: The failure of Poisson modeling. *IEEE/ACM Trans. Networking* **3** 226–244.

Peterson, L. L., B. S. Davie. 2003. *Computer Networks: A Systems Approach*. Morgan Kaufmann, San Francisco, CA.

Rao, S., E. R. Petersen. 1998. Optimal pricing of priority services. *Oper. Res.* **46** 46–56.

Roberts, J. W. 2004. Internet traffic, QoS, and pricing. *Proc. IEEE* **92** 1389–1399.

Stehfest, H. 1970. Numerical inversion of Laplace transforms. *Comm. ACM* **13** 47–49.

Takagi, H. 1991. *Queueing Analysis: A Foundation of Performance Evaluation, Volume 1: Vacation and Priority Systems, Part 1*. North-Holland, Amsterdam, The Netherlands.

Takagi, H., Y. Takahashi. 1991. Priority queues with batch Poisson arrivals. *Oper. Res. Lett.* **10** 225–232.

Thomas, P., D. Teneketzis, J. K. MacKie-Mason. 2002. A market-based approach to optimal resource allocation in integrated-services connection-oriented networks. *Oper. Res.* **50** 603–616.

Tirole, J. 1989. *The Theory of Industrial Organization*. MIT Press, Cambridge, MA.

Wang, Q., J. Peha. 1999. State-dependent pricing and its economic implications. *Proc. Seventh Internat. Conf. Telecomm. Systems Model. and Anal., Nashville, TN, March* 1999, IEEE Press, New York, 61–71.

Wang, Q., J. M. Peha, M. A. Sirbu. 1997. Optimal pricing for integrated services networks. L. McKnight, J. Bailey, eds. *Internet Economics*. MIT Press, Cambridge, MA, 353–376.