USAGE-BASED PRICING OF PACKET DATA GENERATED BY A HETEROGENEOUS USER POPULATION

Michael L. Honig
Dept. of EECS
Northwestern University
Evanston, IL 60208

Kenneth Steiglitz
Dept. of Computer Science
Princeton University
Princeton, New Jersey 08544

ABSTRACT

Usage-based pricing of offered traffic to a data network can be an effective technique for congestion control. To gain insight into the benefits usage-based pricing offers, we propose and study a simple model in which many users wish to transmit packets to a single-server queue. Based on the announced price per packet and the available Quality of Service (QoS) (e.g., mean delay), each user independently decides whether or not to transmit. Given statistical assumptions about the incoming traffic streams and the QoS as a function of offered load, the equilibrium relationship between price and QoS is determined by a fixed-point equation. The relationships among price, QoS, revenue, and server capacity are illustrated numerically, assuming a particular type of random user population. These examples indicate that adjusting the price to maximize revenue results in an efficient use of service capacity with an associated small mean delay.

1. Introduction

As the demand for telecommunications services accelerates, the likelihood of user dissatisfaction due to network congestion increases. Although in the near term broadband networks will be designed to accommodate far more than the anticipated offered traffic, future demands for high-speed telecommunications services may eventually strain available resources. In such an environment efficient allocation of network resources, such as available bandwidth and switch capacity, becomes an important problem. Usage-based pricing is an attractive approach to solving this problem.

In the context of high-speed networks serving heterogeneous traffic, usage-based pricing can offer many benefits. First, pricing can be used to provide closed-loop congestion control. As the price (per packet or service) increases, users may decide to withhold transmission, or transmit at a reduced rate, depending on the degree of urgency of the applications. This use of pricing assumes that the network can adjust and communicate the price to the users within a short time period (i.e., within the period of congestion). In addition, users must be willing and able to use the pricing information to determine an appropriate response which maximizes the value they obtain from the service.

A second benefit of pricing, which is related to congestion control, is traffic smoothing. By varying the price of a service with time, users can be encouraged to alter usage patterns so that the load on the network is shifted from normal peak hours to times when the network is usually lightly loaded. In contrast to the congestion control application, this use of pricing does not require a tight feedback loop between the users and the congested elements. Rather, extensive traffic measurements are needed to determine a price schedule, which applies to some period of time (i.e., 24 hour period), and is announced to the users in advance. This type of pricing scheme has been studied in [4] in the context of voice telephony.

A third benefit of pricing is the ability to discriminate among different applications that require different qualities of service. One way to do this is to assign priorities to packets, and attach different prices to different priorities. In this way the network may be able to accommodate more traffic, and make more efficient use of its resources than possible without price discrimination. Revenues derived from different classes of service can then be used to guide bandwidth allocations. The user has more options, which can be selected to maximize the overall utility of the network.

In addition to the preceding potential benefits of usage-based pricing, standard economic and marketing considerations are clearly important. Namely, pricing plays a critical role in winning market share from competitors, and revenues, which depend on pricing policies, can be used to decide whether or not to add more capacity to the network. An overview of economic issues in the pricing of broadband services is given in [1] and [12]. The focus of this paper; however, is on the relation between pricing and network performance, assuming a single statistically multiplexed data service.

The study of pricing in the context of network engineering presents a difficulty which is generally not present in more conventional performance studies. Namely, it is necessary to model user behavior in addition to the behavior of the network. This is illustrated in Figure 1 which shows the relation between the two models. The network model takes as input the offered traffic generated by the user model, and produces as output a quality of service (QoS) seen by each of the users. The user model takes as input the QoS output by the network model, and the price, which is a control parameter. The output of the user model is the traffic offered to the network model.

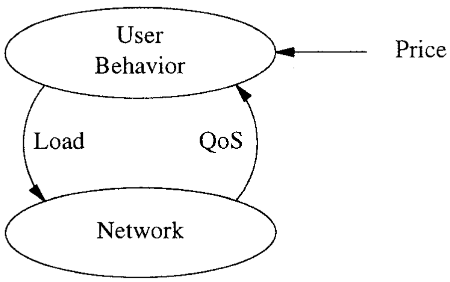

Figure 1. System model.

Given the complexity and diversity of user behavior, constructing an accurate model of user behavior that can be used in the context of Figure 1 can be quite challenging. This is in addition to the network modeling problem, which poses a different set of complications. In this paper we propose, analyze, and simulate a simple model based on the approach in Figure 1. The network is modeled as a single queue, and users generate traffic in a simple way that depends on price and QoS. For this model we characterize stable operating points (i.e., equilibria) and specify the relation between price and performance. Although the model is too complicated to derive analytic relationships among system parameters such as price, performance, and revenue, we present many numerical results which illustrate various tradeoffs.

In the next section we briefly mention related work on usage-based pricing and explain how our model differs from previous models considered. Section 3 specifies our model and states the equilibrium, or fixed-point equation, that relates price to QoS. This fixed-point equation can be used to plot revenue vs. price, so that the set of prices that maximizes revenue can be found numerically. Illustrative examples are given in Section 4, along with an example of transient behavior associated with a sudden price change. In Section 5 we indicate briefly how the approach presented here can be used to study pricing of different types of traffic with multiple priorities, and usage-based price schedules (time of day pricing).

2. Background

References [1]-[12] discuss recent efforts pertaining to usage-based pricing in voice and data networks. A fundamental difference between the approach taken here and previous approaches that model user behavior is that we assume a heterogeneous finite-user population. That is, the value perceived by each user from a price/QoS combination may be different from that perceived by other users. In addition, the model considered here differs in one or more of the following ways from previous models studied:

- The network is assumed to be a private good, so that the objective of any pricing scheme is to maximize total revenue. (In contrast, much of the work previously mentioned assumes that the network is a public good. In that case the objective typically is to maximize the total utility of the network, defined as the sum over all users of each user's utility.)
- Each user is assigned a utility function, which is the maximum price per packet s/he would be willing to pay for a specified QoS. The utility functions are assumed to vary randomly among the user population, and are known only to the assigned users.
- The network measures and announces the QoS (e.g., mean delay or cell loss probability). Each user transmits data only if his/her value for the announced QoS exceeds the price.
- The users wish to transmit at different, fixed data rates (characterized by mean packet arrival rates), which are determined according to some probability distribution.

The preceding assumptions were chosen to capture important elements of an actual data network which implements usage-based pricing. Although the model we consider is quite simple, it has the disadvantage of being difficult to analyze, if not intractable. Nevertheless, numerical results are easily generated, and give insight into the price/performance tradeoff for different user populations.

3. Model Description

3.1. Assumptions

We assume that there are K users who each have data to transmit through a gateway to a high-speed network. This gateway might be a switch somewhere in the network, and the users can be viewed as the incoming links. The gateway announces a price P per packet. The charge to the data source would presumably be the sum of charges imposed by each switch encountered in the end to end path. Here we focus on a single gateway in isolation ignoring the effect each gateway has upon output traffic and associated congestion elsewhere in the network. We also ignore any additional traffic and cost due to the computation and transmission of billing information (i.e., see [5] and [11]).

The gateway is modeled as a queue, and user k, $1 \le k \le K$, is assumed to transmit data with mean rate λ_k. We assume that all users are transmitting a single traffic type, i.e., data (as opposed to voice or video), so that the QoS for all users is measured in the same way (i.e., mean delay or cell loss probability). We later consider the generalization of this discussion to multiple traffic types with different QoS's. The QoS is assumed to depend only on the offered load to the gateway. That is, we define a QoS function $D(\Lambda/C)$, which applies to every user, where the offered load is

$$\Lambda = \sum_{\text{active users } k} \lambda_k \tag{1}$$

and C is the capacity (service rate). Since we will often assume that C is fixed, C will sometimes be omitted as a function argument. Throughout this paper the QoS is assumed to be some form of delay (i.e., mean cell delay), denoted by δ. This is simply to make the discussion more concrete. Other forms of QoS, such as cell loss probability or delay jitter, could just as easily be considered.

In order to examine the effect of an announced price on the offered load to the network, it is necessary to make some assumptions about user behavior. Here we assume that associated with each user k is a willingness to pay, or utility function $u_k(\delta)$. This function is the maximum amount user k would be willing to pay for a QoS δ. We will assume that u_k is a monotonically decreasing function of δ. Although we do not explicitly model competing services or networks, this might be incorporated into the choice of utility functions. We will present a specific form for the functions $u_k(\delta)$ which depends on a few (random) parameters. This class of utility functions will be used to generate the numerical results in Section 4.

Finally, we assume that the gateway measures and announces the QoS δ. If $u_k(\delta) \ge P$, then by assumption, user k is willing to pay price P/packet to transmit his/her data. However, if $u_k(\delta) < P$, then user k is not willing to pay price P/packet for the announced QoS, and does not transmit. (We assume that this user either uses a competitor to send the message, or stores the data and waits for the price to drop to an acceptable level.)

3.2. Fixed-Point Equation

Given the preceding modeling assumptions we can compute the QoS offered by the gateway as a function of price. Namely, in equilibrium the QoS announced by the network must be what each user observes. Stated precisely,

$$D\left(\sum_{\{k:u_k(\delta) \ge P\}} \lambda_k\right) = \delta \tag{2}$$

This fixed-point equation (FPE) clearly relates the announced price P to the resulting QoS seen by the users. In general, the FPE may not have a solution, or may have multiple solutions. However, by restricting the class of allowable utility functions and delay function $D(\cdot)$, the FPE describes a unique operating point in the following sense.

Theorem 1: Let $u_k(\delta)$ for each $k = 1, \ldots, K$ be strictly positive, continuous, monotonically decreasing, and $\lim_{\delta \to \infty} u_k(\delta) = 0$. Let $D(\Lambda)$ be strictly positive, finite, continuous, and monotonically increasing. Then given $P > 0$, there is a unique δ_0 such that

$$D\left(\sum_{\{k:\, u_k(\delta) \ge P\}} \lambda_k\right) > \delta \quad \text{for } \delta < \delta_0$$

$$\text{and } D\left(\sum_{\{k:\, u_k(\delta) \ge P\}} \lambda_k\right) < \delta \quad \text{for } \delta > \delta_0$$

Proof: Since each $u_k(\delta)$ continuously decreases to zero, the load $\Lambda(\delta) = \sum_{\{k\,:\,u_k(\delta) \ge P\}} \lambda_k$ is a decreasing step-function and $\lim_{\delta \to \infty} \Lambda(\delta) = 0$. Consequently, $D[\Lambda(\delta)]$ is also a decreasing step-function. The result then follows from the fact that $D[\Lambda(0)] > 0$.□

Figure 2 illustrates this FPE for a particular set of utility functions $u_1, \ldots, u_K$, arrival rates $\lambda_1, \ldots, \lambda_K$, and delay function $D(\Lambda)$. As δ increases, fewer users wish to transmit data, and so the load Λ decreases in a stepwise manner causing the announced delay $D[\Lambda(\delta)]$ to decrease accordingly. In this case there is no solution to the FPE. Specifically, there is one user k for which the decision to transmit data causes the network delay to increase to the point where k is unwilling to pay the announced price P. However, if k does not transmit, then the delay falls to a value for which this user is willing to transmit at the announced price. To avoid ambiguity, we define the equilibrium in this situation as the point below the line $D(\Lambda) = \delta$. That is, we assume that k does not transmit, so that the equilibrium delay is defined by

$$\max \{ D(\Lambda) \mid D(\Lambda) \le \delta \}$$

With this modification to (2), it is clear that given the assumptions for Theorem 1, there is a unique equilibrium delay for each price P.

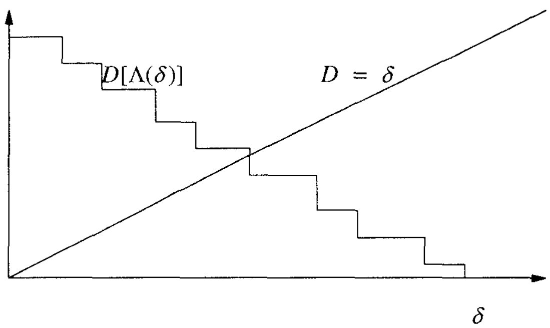

Figure 2. Illustration of fixed-point equation.

Given $D(\Lambda)$, one can plot QoS as a function of announced price for a fixed user population. We would like to choose the price to maximize revenue

$$R(P) = P\Lambda = P \sum_{\{ k \,:\, u_k(\delta) \ge P \}} \lambda_k . \quad (3)$$

3.3. Utility Functions

Our objective is to propose a class of utility functions which is characterized by only a few (random) parameters, and which captures generally accepted notions for user behavior in response to price variations. For each user k we assume that $u_k(\delta)$ is a continuous monotonically decreasing function of delay, and that $\lim_{\delta \to \infty} u_k(\delta) = 0$. Consequently, there is a delay $\delta_{0,k}$ for which $u_k(\delta_{0,k}) = \frac{1}{2} u_k(0)$. We would like to characterize the "abruptness" with which each u_k tends to zero. That is, we assume that there are some users for which $u_k(\delta) \approx u_k(0)$ for $\delta < \delta_{0,k}$ and $u_k(\delta) \approx 0$ for $\delta > \delta_{0,k}$. For other users $u_k(\delta)$ may decay more gradually. We therefore assume that

$$u_k(\delta) = \frac{U_{0,k}}{1 + (\delta / \delta_{0,k})^n} \quad (4)$$

where $U_{0,k} = u_k(0)$, $\delta_{0,k}$, and n are parameters to be determined. As $n \to \infty$, $u_k(\delta)$ converges to a step-function.

4. Numerical Results

4.2. Equilibrium

In this section we present numerical results, based on the preceding model, which relate announced price, network performance, and revenue. The following selection of model parameters simply illustrates our approach to the pricing problem, and does not reflect actual marketing data for any specific data service. We assume that the network gateway is an M/D/1 queue and that the QoS function is mean delay, so that

$$D(\rho) = \frac{\rho}{2(1-\rho)} \quad (5)$$

where $\rho = \Lambda / C$. The arrival rate λ_k is assumed to be a normally distributed random variable where the mean $\bar{\lambda}$ determines the average load $(K\bar{\lambda})$ for an announced price $P = 0$. Given that the service rate is C, the mean normalized load, or requested bandwidth, is

$$B = K\bar{\lambda}/C \quad (6)$$

In what follows we assume that $B > 1$, meaning that if $P = 0$, then the queuing delay at the network gateway tends to infinity. Denoting the standard deviation of the distribution for λ_k as $\sigma_k^{(\lambda)}$, the following results assume that $\sigma_k^{(\lambda)}/\bar{\lambda} = 1/2$.

The parameters $U_{0,k}$ and $\delta_{0,k}$ in the utility function (4) are assumed to be normally distributed random variables. Since each user transmits the same traffic type, we assume that the mean and variance are the same for each user. Specifically,

$$U_{0,k} \sim N(1, \tfrac{1}{2}), \qquad \delta_{0,k} \sim N(20/C, 5/C) \quad (7)$$

Finally, the following results assume that $K = 100$, and that $n = 2$ for each user.

Figures 3a and 3b show, respectively, mean revenue and delay vs. price for a fixed user population with capacity as a parameter. Random model parameters were sampled from the associated distributions. Each point was obtained by solving the fixed-point equation (2). Note that to maximize revenue, the price should be set at the minimum value for which the mean delay is nearly zero. Also notice that for each C, the jagged decay of the revenue as P increases is identical to the decay corresponding to higher values of C. This is because for large enough P, $\delta \approx 0$, so that the revenue is independent of capacity in this range.

Although the curve of revenue vs. price has a prominent global maximum, there exist numerous local maxima. In general, revenue as a function of price can be quite "choppy". The reason for this is that as the price rises, a user may leave the system which locally decreases revenue. However, as the price continues to rise, there is likely to be a price interval for which no users leave, so that the revenue increases in this interval. When the price is low, the increase in revenue as a function of price is nearly linear, since nearly all potential users are transmitting.

Figure 4 shows probability densities for mean revenue with price as a parameter. The densities are with respect to random user populations. These curves are actually histograms of averaged revenue for a sampling of 1000 different randomly selected user populations. The densities drift to the right as the price increases, until the price is raised to the point where the variance becomes quite large. This reflects the uncertainty in revenue when trying to price for a small population of users with large utilities.

Figure 5 shows a plot of maximized revenue $R_{max}(C)$ vs. capacity C for a fixed user population. The maximization is with respect to the price. For the finite-user model assumed here, $R_{max}(C)$ is piecewise continuous due to the discrete nature of user decisions as to whether or not to transmit for given P and δ. Consider the situation in which capacity C is to be allocated to two types of services, i.e., constant bit rate video and statistically multiplexed data. To maximize total revenue, the capacities allocated to these services, (C_1, C_2) where $C_1 + C_2 = C$, should be chosen so that the marginal increase in revenue due to allocating $C_1 + \varepsilon$ units of capacity to the video service is the same as the marginal decrease in revenue due to allocating $C_2 - \varepsilon$ units of capacity to the data service, where ε is small. Depending on how video revenue depends on capacity, there may be multiple allocations for which this is true.

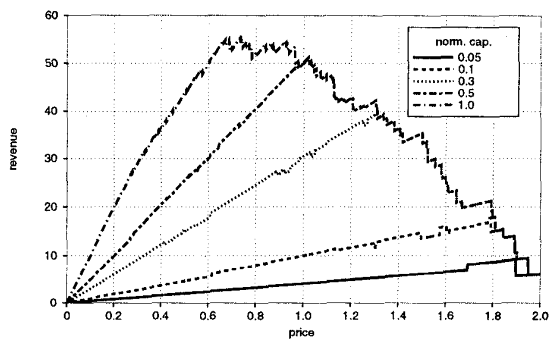

Figure 3a. Revenue vs. price with normalized capacity as a parameter.

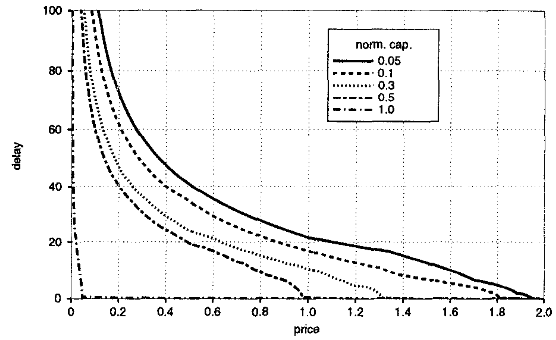

Figure 3b. Delay vs. price with normalized capacity as a parameter.

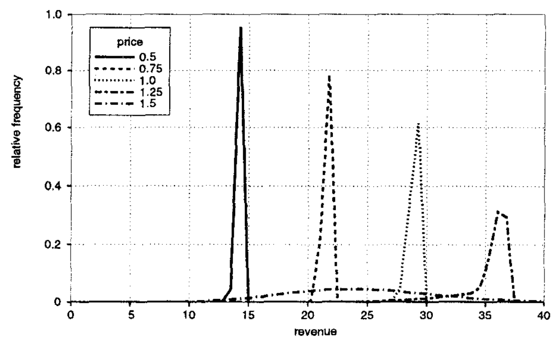

Figure 4. Histogram of revenue for various prices; normalized capacity is 0.3.

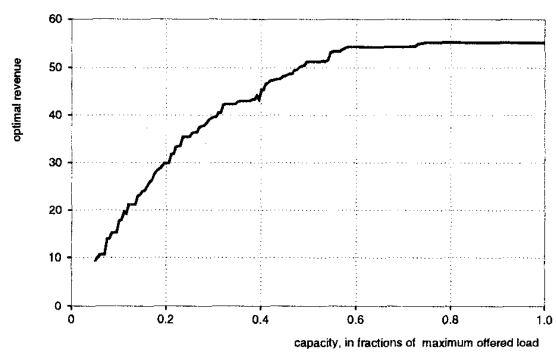

Figure 5. Maximized revenue vs. capacity assuming a fixed user population.

4.3. Simulated Transients

The preceding results assume equilibrium, meaning that the QoS announced by the network is the same as that seen by the users. However, in a dynamic environment where the price can change, transients occur during which the system is not in equilibrium. Consequently, it is of interest to determine if equilibrium can be obtained, given that the price stays fixed at its new value, and if so, how long it takes to achieve equilibrium.

To study the effect of changing price on the preceding model, it is necessary to specify how the network estimates QoS (mean delay). Because we assume that each arrival to the queue is a fixed-length cell, the delay is the same as the queue length. We could assume, then, that the network simply announces the instantaneous delay. However, because delay is a random process, which can vary rapidly, one might expect that announcing a smoothed delay will lead to improved system behavior. The results that follow assume that the mean delay is estimated as

$$\hat{\delta}_{n+1} = \begin{cases} (1-\alpha)\delta + \alpha\hat{\delta}_n & \text{packet exits} \\ \hat{\delta}_n & \text{no packet exits} \end{cases} \quad (8)$$

where δ is the delay experienced by the most recent packet, and $\hat{\delta}_n$ is the estimated mean delay computed from n packets. Of course, $\alpha = 0$ corresponds to taking the current delay as the estimate. The rest of the modeling assumptions remain the same. That is, even though the QoS seen by each user in a dynamic environment is generally different from the QoS announced by the network, the user's decision as to whether or not to transmit is based on the *announced* QoS.[1]

Because the resulting model is complicated, it appears to be difficult to perform a transient analysis without making a number of simplifying assumptions [3]. Transient behavior can, however, be studied by simulation. Figure 6 shows queue length vs. time assuming that the price changes from 1.0 to 0.5, and that the system starts empty. Two cases are shown in Figure 6 corresponding to $\alpha = 0.998$ and $\alpha = 0$ in (8). It is interesting that the queue exhibits oscillatory behavior with smoothing, but without smoothing the queue reacts quickly to the price change and does not oscillate. This is due to the assumption that the users react *instantaneously* to the announced delay and price. Consequently, without smoothing fewer (more) users transmit packets as the delay increases (decreases). This effectively regulates the queue length so that it remains nearly constant. Smoothing actually models the case in which the users do not react instantaneously, but use past announcements to influence their decisions. Consequently, when the price suddenly changes, the users act according to the preexisting steady-state delay longer than they would have without smoothing. This causes the overshoot and oscillations shown in Figure 6. This behavior is consistent with the fluid-flow analysis given in [3].

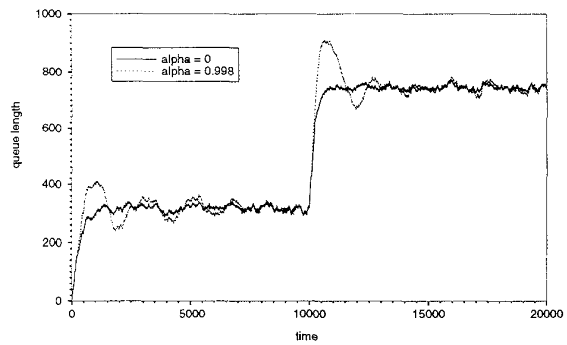

Figure 6. A simulated transient in which the price is suddenly decreased from 1.0 to 0.5 at $t = 10,000$. Shown is the queue length for the cases $\alpha = 0$ (no smoothing in measured delay), and $\alpha = 0.998$. The normalized capacity is 0.3.

5. Extensions

Clearly, the model presented here can be extended in many ways to account for more complicated user behavior as well as more sophisticated network models. Here we briefly indicate how the model presented in Section 3 can be extended to handle different traffic types with different priority status, and how static pricing schemes which vary only with the time of day might be studied. We leave a more detailed examination of these issues for future work.

5.1. Multiple Priorities

Different network applications will require different QoSs, and will thereby pose different burdens on the network. To maximize network efficiency, applications which require strict guarantees on QoS should be charged more than applications which do not require these guarantees. A simple way to achieve different grades of service is to assign different priorities to cells associated with different applications/users. Higher priority cells are served before lower priority cells, and usage-based charges depend on the assigned priority class of the transmitted cells.[2] (The relationship between

[1] We assume that it is in the best interest of the network provider to estimate QoS as accurately as possible. Intentionally inaccurate estimates are likely to frustrate users, causing them to leave the system.

[2] In practice, QoS should be tied to application requirements, such as bounds on absolute cell delay, cell delay jitter, and cell

priority pricing and adaptive, or "spot", pricing is studied in [13] in the context of the electric power industry.)

Consider the model in Section 3 where each user must choose either a high or low priority class. High priority traffic is treated differently from low priority traffic. Namely, we could assume that low priority cells are served only after high priority cells have been served. Of course, other queuing disciplines which preferentially treat high priority cells are possible. We assume that the network measures and announces the QoS (i.e., mean delay) for each priority class. Let P_H (P_L) be the price per packet for high (low) priority cells, and let δ_H (δ_L) be the mean delay for high (low) priority cells. User k chooses to transmit packets via high priority provided that

$$u_k(\delta_H) \geq P_H \quad \text{and} \quad u_k(\delta_H) - P_H > u_k(\delta_L) - P_L,$$

and chooses to transmit packets via low priority provided that

$$u_k(\delta_L) \geq P_L \quad \text{and} \quad u_k(\delta_L) - P_L > u_k(\delta_H) - P_H$$

Denote the set of users who transmit high (low) priority cells as U_H (U_L). If the user population is fixed and the system is in equilibrium, then the delays announced by the network must be the same as the delays seen by the users. This implies the two-dimensional FPE

$$D_H\left(\sum_{k \in U_L} \lambda_k, \sum_{k \in U_H} \lambda_k\right) = \delta_H \qquad (9)$$

$$\text{and} \quad D_L\left(\sum_{k \in U_L} \lambda_k, \sum_{k \in U_H} \lambda_k\right) = \delta_L$$

where $D_H(\rho_L, \rho_H)$ and $D_L(\rho_L, \rho_H)$ relate offered low and high priority loads to mean delays. Of course, these functions depend on the queuing discipline. As in the single-priority, finite-user case, this FPE does not always have a solution, although an operating point can always be defined. This FPE then relates the price vector (P_H, P_L) to the performance vector (δ_H, δ_L). The objective is then to find a price vector that maximizes revenue $C(P_H\rho_H + P_L\rho_L)$. Of course, adding more priority classes results in a higher dimensional FPE.

loss probability. Assigning priorities has the advantage of being relatively simple to implement; however, the relationship between a given priority class and this broader application-oriented notion of QoS may not be easy to characterize in many situations of interest.

5.2. Time of Day Pricing

In practice, an adaptive pricing scheme which responds quickly to variations in offered load is likely to be impractical to implement in the near term. However, it is feasible to announce a static usage-based price schedule in which the price changes every few minutes. This schedule could, for example, cover a 24-hour period. Periods over which the price is constant will depend on the periods during which traffic measurements are collected. (For voice telephony traffic measurements over 15 minute intervals are common [5].)

To illustrate how the approach presented here can be used to study the effect of a static price schedule, suppose that the price schedule contains two entries: a day price P_D and a night price P_N. For simplicity, we assume that each user generates traffic only during the day, and has the choice of sending his/her traffic as soon as it is generated, or storing it and transmitting it during the night. We also assume that "day" and "night" are separated by a fixed time T. The network measures the delay during the day (night) as δ_D (δ_N), and we assume that $\delta_D \ll T$ and $\delta_N \ll T$. User k chooses to transmit traffic during the day provided that

$$u_k(\delta_D) \geq P_D \quad \text{and} \quad u_k(\delta_D) - P_D > u_k(\delta_N + T) - P_N$$

and chooses to store the message and transmit at night provided that

$$u_k(\delta_N + T) \geq P_N \quad \text{and} \quad u_k(\delta_N + T) - P_N > u_k(\delta_D) - P_D$$

Denote the set of users that transmit during the day (night) as U_D (U_N). In equilibrium the delays announced by the network must be the same as those seen by the users, which implies that

$$D\left(\sum_{k \in U_D} \lambda_k\right) = \delta_D \quad \text{and} \quad D\left(\sum_{k \in U_N} \lambda_k\right) = \delta_N \qquad (10)$$

For a fixed user population, these equations specify the relation between day vs. night price differentials and network performance. Depending on the nature of the user utility functions, price differentials can be used to smooth traffic. That is, by choosing P_D sufficiently greater than P_N, the load during the day might be adjusted to match the load during the night. Of interest is the relation between day and night loads when P_D and P_N are chosen to maximize revenue.

5.3. Other Extensions

In practice, pricing of heterogeneous services sharing the same network facilities must take into account additional issues, which have not been mentioned here. For example, different traffic types require different measures of QoS and may incur connection charges in addition to usage-based charges. The general problem of allocating capacity to different services efficiently then becomes much more complicated. The interaction between this problem and pricing clearly deserves further study. Also, the effect of competing services has not been explicitly modeled, and is extremely important in practice. Finally, we mention that the changing regulatory constraints associated with the telecommunications industry are likely to play a major role in determining pricing policies.

References

[1] B. Chakravorty, W. Sharkey, and P. Srinagesh, "Issues in the Pricing of Broadband Telecommunications Services", to appear in *Quality and Reliability of the Telecommunications Infrastructure,* edited by William Lehr, published by Erlbaum Press.

[2] R. Cocchi, S. Shenker, D. Estrin, L. Zhang, "Pricing in Computer Networks: Motivation, Formulation, and Example", *IEEE/ACM Trans. on Networking,* Vol. 1, No. 6, pp. 614-627.

[3] M. L. Honig and K. Steiglitz, "Usage-Based Pricing and Quality of Service in Data Networks", in preparation.

[4] M. A. Koschat, P. Srinagesh, and L. J. Uhler, "Efficient Price and Capacity Choices under Uncertain Demand – An Empirical Analysis", to appear in the *Journal of Regulatory Economics.*

[5] S. Loeb, S. Stanislaw, and J. Alfieri, "Usage Information Collection and Management Paradigm for Broadband and Multimedia Services: Architectural Evolution", submitted to the *1995 International Switching Symposium,* Berlin.

[6] J. K. MacKie-Mason, and H. R. Varian, "Pricing the Internet", to appear in *Proceedings of "Public Access to the Internet",* Harvard University, 1993.

[7] McLean, R.P. and W.W. Sharkey, "An Approach to the Pricing of Broadband Telecommunications Services," Telecommunication Systems, 2, 159-84, 1994.

[8] H. Mendelson and S. Whang, "Optimal Incentive-Compatible Priority Pricing for the M/M/1 Queue, *Operations Research,* Vol. 38, No. 5, pp. 870-883, Sept.-Oct. 1990.

[9] J. Murphy, L. Murphy, and E. C. Posner, "Distributed Pricing for Embedded ATM Networks", preprint.

[10] B. Sanders, "An Incentive Compatible Flow Control Algorithm for Rate Allocation in Computer Networks", *IEEE Trans. on Computers,* vol. 37, pp. 1067-72.

[11] S. Shenker, "Service Models and Pricing Policies for an Integrated Services Internet", Technical Report, Palo Alto Research Center, Xerox Corporation.

[12] P. Srinagesh, "Economics of Broadband Services: Product Line Pricing", unpublished manuscript, 1992.

[13] R. Wilson, "Efficient and Competitive Rationing", *Econometrica,* Vol. 57, No. 1, pp. 1-40, Jan. 1989.