Price-Based Quality-of-Service Control Framework for Two-Class Network Services

Whan-Seon Kim

Abstract: **This paper presents a price-based quality-of-service (QoS) control framework for two-class network services, in which *circuit-switched* and *packet-switched services* are defined as "premium service class" and "best-effort service class," respectively. Given the service model, a customer may decide to use the other class as a perfect or an imperfect substitute when he or she perceives the higher utility of the class. Given the framework, fixed-point problems are solved numerically to investigate how static pricing can be used to control the demand and the QoS of each class. The rationale behind this is as follows: For a network service provider to determine the optimal prices that maximize its total revenue, the interactions between the QoS-dependent demand and the demand-dependent QoS should be thoroughly analyzed. To test the robustness of the proposed model, simulations were performed with gradually increasing customer demands or network workloads. The simulation results show that even with substantial demands or workloads, self-adjustment mechanism of the model works and it is feasible to obtain fixed points in equilibrium. This paper also presents a numerical example of guaranteeing the QoS statistically in the short term—that is, through the implementation of pricing strategies.**

Index Terms: **Fixed-point problem, pricing, quality-of-service (QoS), two-class network services.**

I. INTRODUCTION

Since recently, the telecommunications engineering community has been paying much attention to quality-of-service (QoS) issues. This is due mainly to the wide dissatisfaction with the perceived performance of telecommunications networks, including the Internet.

With regard to the Internet, the flat-rate pricing scheme, based on the best-effort service model, has contributed to the explosive growth of the Internet user population and of their Internet usage. With this pricing scheme, however, Internet users or customers do not have any incentive to cut down their network usage, thereby leading to the congestion of the network.

The customers' demand for high-quality and premium services in the face of this congestion has stimulated the development of various protocols to achieve a QoS guarantee, which are supposed to provide differentiated QoS levels. These protocols, however, may be hard to implement due to their technical complexity, and may require major changes in the network.

This paper presents a simple approach to implementing differentiated QoS services by adopting an integrated packet-switched and circuit-switched services model, in which circuit-switched service is the premium service while packet-switched service is the current best-effort service. The circuit-switched or connection-oriented service is a premium service in that transmission speed is guaranteed by reserving the required bandwidth and buffers in the network. There is no information loss during the transmission.

Given the engineering model, usage-sensitive pricing schemes are applied to the service model to induce the customers to optimize their network resource usage. More specifically, a volume-sensitive pricing scheme is used in the packet-switched service,[1] and a time-sensitive pricing scheme is used in the circuit-switched service.[2] Therefore, the charge for packet-switched service is determined based on the volume of data to be served in a job, while the charge for circuit-switched service is determined based on the amount of network connection time used to carry out a job.

The prices that are considered in this paper are static prices, or those that do not change as the traffic load in the network changes. The static-pricing scheme has the following drawbacks: Prices should be chosen *a priori*—that is, without yet knowing the actual demands; and prices do not react to the changes in the network's traffic load. One obvious advantage of the static-pricing scheme, however, is that it is much simpler to implement than the dynamic-pricing scheme. Therefore, no real-time mechanism is needed to communicate prices to the customers. Moreover, as the charges do not change frequently, the customers can predict them. The purpose of this paper is to look into how static pricing can be used to control the demand and the QoS of each class in the two-class network services. Ultimately, the objective of the proposed model is to find the prices that maximize the network service providers' total revenues.

Pricing in communication networks has received much attention, as shown in the literature. Regarding packet-switched network services, [1] and [2] propose packet-based pricing as an incentive for a more efficient flow control. Especially, usage-based schemes, which charge based on the actual resources used, have been proposed for multiple-service classes in [3] and [4]. Regarding circuit-switched network services, [5] and [6] adopt a conventional practice of pricing per connection, and show that pricing schemes can offer efficient resource allocations in connection-oriented networks offering QoS guarantees. Static pricing has been studied in [7]–[9] for the cases of a single-service class, two priority classes, and a continuum of priorities, respectively. Particularly, [10] quantifies how much is lost when

[1] Volume-sensitive pricing scheme has been adopted for data services in which the charge is proportional to the job size. For example, countries such as Australia, Austria, Belgium, Germany, New Zealand, and Singapore have adopted various forms of usage-based pricing scheme for Internet services where the total charge depends on the total job size measured in bytes.

[2] Time-sensitive pricing scheme has been adopted for voice (telephone) services where the total charge depends on the total service time.

Manuscript received January 23, 2006; approved for publication by Min Young Chung, Division III Editor, April 23, 2007.

W.-S. Kim is with the Business School at Myongji University, Seoul, Korea, email: whankim@mju.ac.kr.

1229-2370/07/$10.00 © 2007 KICS

using static pricing instead of dynamic pricing with a finite number of priority classes.

This paper proposes a means of estimating the demands of two-class network services, based on a network service provider's service quality as well as the consumers' willingness to pay. Among the goals of this paper are to combine the engineering and economic aspects of the two-class network services model, and to obtain a better understanding of the interactions between the demand-dependent QoS and the QoS-dependent demand.

This paper is organized as follows: Section II presents an overview of the proposed model and the basic assumptions made in the modeling process. Section III presents the basic modeling framework and develops mathematical models. Also, systematic ways of checking the stability of an equilibrium point are derived in solving fixed-point problems. In Section IV, a numerical example is presented to illustrate the models that were developed, and the simulation results are shown and discussed. Finally, Section V presents the conclusions and the suggestions for further research.

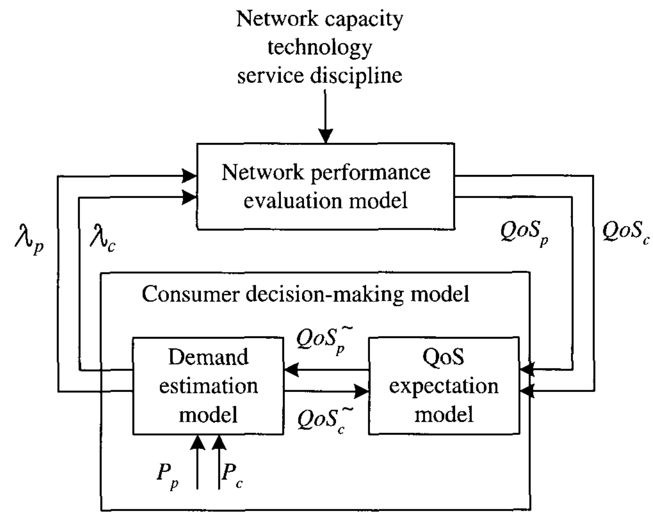


Fig. 1. Price-based QoS control framework in two-class network services. (P: price, λ: demand, QoS: quality-of-service, $QoS^{\sim}$: expected QoS, and the subscripts p and c: packet-switching services and circuit-switching services, respectively.)

II. OVERVIEW OF THE MODEL

Service class is defined according to the type of service it is associated with. Therefore, there are two class options in this model, namely:

- The premium service class, in which an entire job is processed using the circuit-switched technology. (In this class, the so-called *connection establishment stage* is a prerequisite, in which an end-to-end route reservation takes place. Therefore, this class, based on a successful connection establishment, provides a fixed bandwidth end-to-end. This service class may be implemented using the RSVP-like route reservation technologies.); and
- The best-effort service class, in which an entire job is processed using the packet-switched technology. (In this class, the entire information stream is chopped into multiple packets. These packets are then queued and processed. Therefore, a packet transmission is subject to queueing delay.)
- Given the engineering model, demand substitution effects are incorporated into the economic model as follows: When a customer requests a job process or information transmission to a network service provider, a pop-up menu listing the class options and their prices is shown on the monitor screen of his/her local machine. Then, based on the price and the expected QoS of each class, the customer chooses a service class with which the job will be processed. This means that the customer may decide to use the other class as a perfect or an imperfect substitute[3] when he/she perceives the class as having higher utility. Therefore, cross-elasticity of demand exists.

The assumptions used in the modeling process are as follows:

- For simplicity, this model assumes that a predetermined bandwidth is guaranteed to all the customers in the circuit-switched service. Therefore, it does not consider a customer-specific or application-specific QoS request.
- The size and characteristics (e.g., income, taste or preference, willingness to pay) of the customer pool that a network service provider considers do not vary in the short term.
- The revenue and QoS of a network service provider are affected only by the pricing strategy it uses, given that the network capacity is fixed. Therefore, for analytical convenience, this paper excludes network-capacity-decision problems and focuses mainly on price-decision problems.
- The customers or their local machines know the total cost of the job in each class before they make their decision, given the job length and the estimated service time based on the bandwidth allocated.
- In the packet-switched service, the customers are charged for the packets submitted to the network rather than for the packets actually delivered. This simplifies the implementation since there is no need to track the dropped packets for accounting and billing purposes [11].
- In the circuit-switched service, in order to implement end-to-end route reservations, peering agreements among network service providers are necessary. The implementation of peering agreements, however, and the corresponding charging schemes are beyond the scope of this paper, and will therefore not be dealt with.

III. MODELING FRAMEWORK

A. Price-Based QoS Control Framework in Two-Class Network Services

The price-based QoS control framework in two-class network services is shown in Fig. 1. In Fig. 1, the entire model consists of two submodels, namely: The network performance evaluation model and the consumer decision-making model. The latter is further subdivided into the QoS expectation model and the

[3]This depends on the types of applications and on the network conditions. For example, real-time applications may work reasonably well only in the premium-service class. They may also work well in the best-effort service class, but only when its QoS is good.

demand estimation model.

The network performance evaluation model derives an actual QoS value for each class based on customer demands, taking the network capacity, adopted technologies, and service disciplines as given. The QoS expectation model derives a QoS value for each class that consumers predict for the immediate future, based on the actual QoS values collected in the present and in the past. Then the expected QoS value of each class is used as an input to the demand estimation model, in which the actual demand on each class is derived. The mathematical representations of these submodels are shown in Section III-B.

Fig. 1 shows that the demand vector (λ_p and λ_c) and the QoS vector (QoS_p and QoS_c) are controlled by the price vector (P_p and P_c). Since all the submodels shown in Fig. 1 form a closed loop, the demand vector and the QoS vector may converge[4] at different fixed points in equilibrium, respectively, depending on the price vector. The ultimate goal of this price-decision model is to find an optimal price vector that will provide maximum revenue to the service provider.

B. Mathematical Models

As mentioned earlier, the network performance evaluation model derives an actual QoS value for each class based on customer demands. Taking the network capacity, adopted technologies, and service disciplines as given, the QoS_p and QoS_c in Fig. 1 can be written in the following simple forms:

$$QoS_p = f_p(\lambda_p, \lambda_c) \tag{1}$$

$$QoS_c = f_c(\lambda_p, \lambda_c) \tag{2}$$

where λ_p and λ_c are unit-time demands of packet-switched service and circuit-switched service, respectively, and $f_p(\cdot)$ and $f_c(\cdot)$ represent some nonlinear functions.

In this paper, the QoS parameters QoS_p and QoS_c are limited to the average delay per packet for the packet-switched service class, and the blocking probability for the circuit-switched service class. A specific model is presented and analyzed in Section IV as a numerical example.

Customers may not believe that in the future, network performance will be exactly the same as that in the present or in the past. There are several factors that affect customers' expectations of QoS, among them knowledge by experience, word-of-mouth referrals, and advertising.

As mentioned earlier, the QoS expectation model derives the QoS value of each class that the customers predict at present for the immediate future, based on the actual QoS values collected in the present and in the past. They may use convex combinations of the past values to estimate the next period's value. This prediction constitutes an exponential smoothing forecast based on past information obtained [16]. This expectation model is called *adaptive expectation model* [2].

For example, customers may use convex combinations of the past two periods' QoS values for the next period's estimated QoS values. Assume that time is divided into an infinite number of non-overlapping periods of fixed length and a succession of discrete time periods are labeled $m = 1, 2, 3, \cdots$. Mathematically, for $w \in (0, 1)$ and $i = p$ or c,

$$QoS_i^{\sim}(m) = (1 - w)QoS_i(m) + wQoS_i(m - 1) \tag{3}$$

where $QoS_i^{\sim}(m)$ is the "predicted" QoS value estimated at period m; and $QoS_i(m)$ and $QoS_i(m-1)$ are the "actual" QoS values measured at period m and $m-1$, respectively. Let this model be called *adaptive expectation model with window size 2* since the customers rely on the past two periods' (m and $m-1$) values to estimate the next period's value.

In (3), as m becomes large, it is possible for $QoS_i^{\sim}(m)$ to reach a steady state in equilibrium. That is, by controlling w in (3), given the specific price vector in Fig. 1, it is possible to make $QoS_i^{\sim}(m)$ converge to a fixed point in equilibrium. If the QoS vector converges, the demand vector converges as well, as shown in Fig. 1.

In general, fixed points remain identical for any $w \in (0, 1)$. In the expectation model, releasing or advertising the past network performance data will induce the customers to choose a larger smoothing factor w, which will in turn lead to a more stable network behavior [12]. This does not mean, however, that a large w is always the preferred choice. This is because a larger w requires more aggressive advertising of the past performance data, thereby incurring a higher advertising cost. Given this, as an effort to minimize the cost, the service provider should try to find the smallest w that will stabilize the system.

Whatever the value of w, however, there may be cases where the system will not be stabilized at all. This means that in this model, the QoS vector and the demand vector will continue to fluctuate as $m$goes to infinity. In this case, the service provider has two options, namely:

- To increase the capacity of the network elements so as to eliminate congestion as a source of instability [12]; or
- to consider expanding the window size.

The first option is good in that it is a more direct approach to stabilizing the system. Its downside, however, is that it cannot be achieved in the short term. On the other hand, the second option is achievable in the short term, through advertising. It is an indirect approach, though, and the service provider needs to explicitly reveal past performance information by advertising. In any case, either option will incur a fixed cost, namely: The cost of purchasing new network elements and the cost of advertising, respectively. Therefore, the service provider should conduct cost/benefit analyses of the two options before deciding which option to choose.

Now, let the variables associated with the customers' willingness to pay be defined as follows:

- WTP_p: Maximum willingness to pay for a job in the packet-switched service, given the best QoS value ($QoS_p = 0$); and
- WTP_c: Maximum willingness to pay for a job in the circuit-switched service, given the best QoS value ($QoS_c = 0$).

Since the customers' willingness to pay is heterogeneous, this variable has some probabilistic distributions. Let the corresponding probability density functions be denoted as follows:

- f_{WTP_p}: pdf of random variable WTP_p;
- f_{WTP_c}: pdf of random variable WTP_c;
- f_{WTP_p,WTP_c}: Joint pdf of WTP_p and WTP_c; and
- f_{WTP_c/WTP_p}: Conditional pdf of WTP_c given WTP_p.

[4]To be exact, it is possible to make them converge. More details will be discussed in Section III-*B*.

Given that the circuit-switched service is a premium service compared to the packet-switched service, the customers are assumed to have a higher willingness to pay for the former than for the latter. Therefore, WTP_p and WTP_c are dependent on the relationship $0 \leq WTP_p \leq WTP_c$, and the joint pdf f_{WTP_p,WTP_c} can be written as follows:

$$f_{WTP_p,WTP_c} = f_{WTP_p} f_{WTP_c/WTP_p},\ 0 \leq WTP_p \leq WTPc. \quad (4)$$

Now, for each customer j, the utility function for service class i ($i = p$ or c) is defined as follows:

$$U_i^j = WTP_i^j - C_i^j(QoS_i^{\sim}) - TP_i. \quad (5)$$

Here $C_i^j(\cdot)$ is a cost function reflecting the customer's valuation for the expected service quality. Therefore, its value decreases as $QoS_i^{\sim}$ improves. Furthermore, $C_i^j(\cdot)$ converts $QoS_i^{\sim}$ into a monetary value so that the units of all the terms in (5) are standardized. TP_i is the total price for a job in service class i. Therefore, $C_i^j(QoS_i^{\sim}) + TP_i$ is the total cost occurring to customer j for using service class i.

Let S denote a random variable of job size and f_S the corresponding probability density function of S. As for the packet-switched service, let V be a unit size (measured in bytes) for charging and P_p the charge per Vin the volume-sensitive pricing scheme. Likewise, as for the circuit-switched service, let T be a unit time[5] (measured in seconds) for charging and P_c the charge per T in the time-sensitive pricing scheme. Then the total price, $TP_i(i = p$ or $c)$, can be determined as follows:

$$TP_p = Pp\lceil\frac{S}{V}\rceil \quad (6)$$

$$TP_c = Pc\lceil\frac{S/b_c}{T}\rceil \quad (7)$$

where b_c is the bandwidth allocated to a job in the circuit-switched service. Therefore, S/b_c in (7) represents the total time required to process a job.

Let Xdenote the total number of subscribers or customers of a network service provider. Then each customer j makes the following decisions:

- If $U_p^j \geq 0$ or $U_c^j \geq 0$, choose the service class with a higher utility value.
- If $U_p^j < 0$ and $U_c^j < 0$, do not choose any service class.[6]

Given that each customer decides based on the above rules whether to join and which service class to join if he/she decides to join, the overall demand on each class can be estimated from Fig. 2.[7] In Fig. 2, the region I indicates that the utility value for the packet-switched service is positive and that for the circuit-switched service is negative, while the other way around is true for the region IV. The region II indicates that the utility value for the packet-switched service is higher than that for the circuit-switched service, while the other way around is true for the region III. Therefore, the unit-time average demands for the packet-switched service and the circuit-switched service can be estimated by integrating the joint pdf, f_{WTP_p,WTP_c}, over the regions I+II and III+IV in Fig. 2, respectively, as follows:

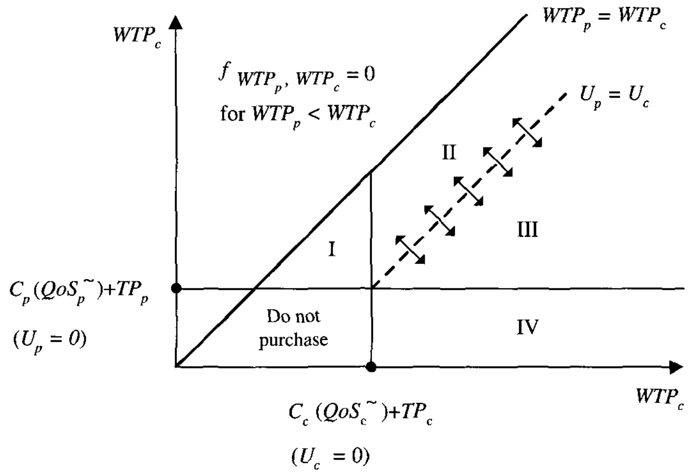


Fig. 2. Graphical representation of customers' choices.

$$\lambda_p = Xr_{avg} \int\int\int_{\text{I+II}} f_{WTP_p} f_{WTP_c/WTP_p} f_S dWTP_p dWTP_c dS \quad (8)$$

$$\lambda_c = Xr_{avg} \int\int\int_{\text{III+IV}} f_{WTP_p} f_{WTP_c/WTP_p} f_S dWTP_p dWTP_c dS \quad (9)$$

Here, r_{avg} is the average number of jobs sent by a customer per unit time.

C. Optimization Formulations

When the size of a job is divided by the unit size V to estimate a charge in the volume-sensitive pricing scheme, the remainder may be less than V. Likewise, when the time required to process a job is divided by the unit time T to estimate a charge in the time-sensitive pricing scheme, the remainder may be less than T. Therefore, to estimate the average charge for a job, it is necessary to round up the remaining part of the job, which is less than Vor T.

Mathematically, this rounding up process can be accomplished by converting a continuous distribution of job size into a discrete distribution of that. Let m_p denote the number of unit size of a job and m_c the number of unit time of a job. The average numbers of unit size and unit time of a job can then be estimated from (10) and (11), respectively, as follows:

$$E[m_p] = \sum_{m_p=1}^{\infty} m_p \int_{V(m_p-1)}^{Vm_p} f_S dS \quad (10)$$

$$E[m_c] = \sum_{m_c=1}^{\infty} m_p \int_{b_c(m_c-1)}^{b_c m_c} f_S dS \quad (11)$$

Ultimately, the optimal price-decision model can be represented mathematically as follows:

[5]This unit time T is different from the unit time mentioned in (1) and (2).

[6]In this case, the customer may wait for a random time and may keep trying until a service class provides a positive utility value.

[7]Fig. 2 shows a two-dimensional graph viewed from the top of a three-dimensional graph.

$$\max_{P_p, P_c} \lambda_p P_p \sum_{m_p=1}^{\infty} m_p \int_{V(m_p-1)}^{V m_p} f_S dS + \lambda_c (1 - QoS_c) P_c \sum_{m_c=1}^{\infty} m_c \int_{b_c(m_c-1)}^{b_c m_c} f_S dS \quad (12)$$

$$\text{subject to } QoS_p = f_p(\lambda_p, \lambda_c) \quad (13)$$
$$QoS_c = f_c(\lambda_p, \lambda_c) \quad (14)$$
$$\lambda_p = g_p(P_p, P_c, QoS_p, QoS_c) \quad (15)$$
$$\lambda_c = g_c(P_p, P_c, QoS_p, QoS_c) \quad (16)$$
$$P_p, P_c > 0 \quad (17)$$

The objective function (12) represents the unit-time total revenues. More specifically, the first term of (12) is revenues from the packet-switched service and the second term of (12) is revenues from the circuit-switched service. The circuit-switched service involves admission control and therefore its unit-time total demand is controlled by $\lambda_c(1 - QoS_c)$, while the unit-time total demand for the packet-switched service is just λ_p. Given the assumptions of finite customer pool as well as finite willingness-to-pay ranges,[8] this objective function is bounded. The constraints (13)–(16) are fixed-point problems, and $f_p(\cdot)$, $f_c(\cdot)$, $g_p(\cdot)$, and $g_c(\cdot)$ represent nonlinear functions of (1), (2), (8), and (9), respectively. Regarding the fixed-point problems, systematic ways of checking the stability of an equilibrium point are derived in the appendices. More specifically, the convergence conditions for the case of window size 2 in the adaptive expectation model are derived in Appendix A.1, and they are generalized into the case of window size n in Appendix A.2.

The above optimization problem and the fixed-point problems can be solved numerically through simulations.[9] The time it takes to solve the optimization problems of (12)–(17) depends mainly on two factors. First, it depends on how long it takes to find the QoS values in Fig. 1, which corresponds to solving (13) and (14), given the specific network performance evaluation model. Secondly, it depends on how soon the QoS and the demand values converge at equilibrium points, respectively, in Fig. 1, which corresponds to solving fixed-point problems of (13)–(16).

IV. A NUMERICAL EXAMPLE

This section provides a numerical example to illustrate the price-based QoS control framework shown in Fig. 1. Given that the main focus of this paper is to present an optimal price-decision algorithm given the proposed framework of Fig. 1, rather than presenting a QoS estimation algorithm given the specific network performance evaluation model in Fig. 1, this paper presents a simple example of a single-node case in the network performance evaluation model. The network performance evaluation model presented here is similar to that developed in [6], but the original model has been modified to incorporate batch arrivals so it can be more appropriately applied to the packet-switched service class. A schematic diagram of the two-class model is shown in Fig. 3, and the assumptions about technology choice, network capacity allocations, and service disciplines are as follows:

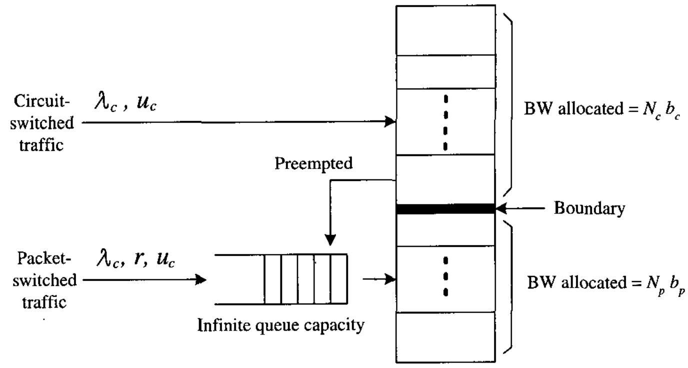


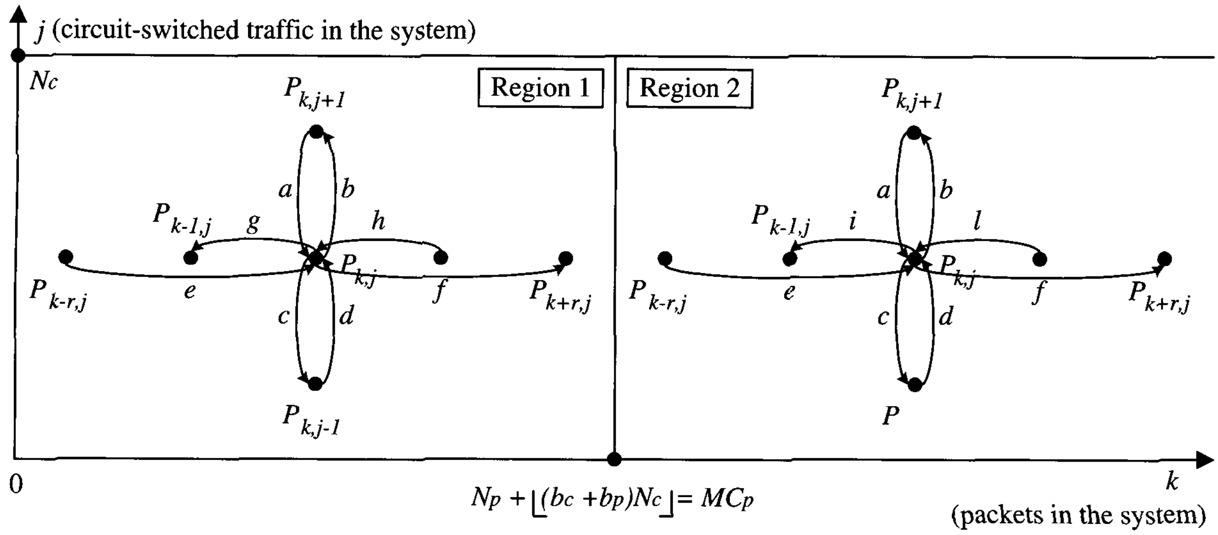
Fig. 3. Schematic diagram of the two-class model with *batch (r)* arrivals and with *movable boundary & preemptive priority* service disciplines.

- In the circuit-switched service class, jobs arrive according to Poisson, at a rate of λ_c, and each job is processed in a channel of bandwidth b_c. The job size S is exponentially distributed. Therefore, the corresponding service time is also exponentially distributed, with a mean of $1/\mu_c$.
- In the packet-switched service class, jobs arrive according to Poisson, at a rate of λ_p. A job is chopped into r-batch packets, and each packet is processed in a channel of bandwidth b_p. The packet size is exponentially distributed. Therefore, the corresponding service time is also exponentially distributed, with a mean of $1/\mu_p$.
- Given b_p, b_c, and the total available capacity F, N_p channels are initially assigned to the packet-switched service class, and N_c channels are initially assigned to the circuit-switched service class. The infinite queue capacity is assumed to derive analytical-form solutions for QoS_p.
- *Movable boundary* and *preemptive priority* service disciplines: A job that arrives at the circuit-switched service class is blocked if, on arrival, all N_c channels are busy carrying out other jobs. A packet that arrives at the packet-switched service class is queued in a FIFO (first-in first-out) manner, and then processed when any N_p channel is available. If all the N_p channels are busy, the packets are allowed to use any unused bandwidth assigned to the circuit-switched service class (Movable Boundary). A job that arrives at the circuit-switched service class, finding all the N_c channels busy, preempts the packets currently using the channels (preemptive priority). Those preempted packets then go back to the head of the queue, waiting for any channel to become available.

Given the service disciplines, circuit-switched traffic experiences no interference from packet-switched traffic. Therefore, the Erlang loss model of $M/M/N_c/N_c$ [1] can be applied, and the blocking probability QoS_c can be represented by the Erlang

[8] Finite willingness-to-pay ranges imply finite price ranges.

[9] If the Jacobian matrix shown in the Appendix cannot be solved analytically to get the corresponding eigenvalues, the trial-and-error method for choosing the value of w should be applied in the simulations.



$a = (1 - \delta_{jN_c})\lambda_c$, where $\delta_{mn} = 1$ if $m = n$ and 0 otherwise
$b = (1 - \delta_{jN_c})(j + 1)\mu_c$
$c = \lambda_c,\ d = j\mu_c$
$e = \lambda_p,\ f = \lambda_p$
$g = \min[k, N_p + \lfloor (b_c/b_p)(N_c - j)\rfloor \mu_p$
$h = \min[k+1, N_p + \lfloor (b_c/b_p)(N_c - j)\rfloor \mu_p$
$i = [N_p + \lfloor (b_c/b_p)(N_c - j)\rfloor \mu_p$
$l = [N_p + \lfloor (b_c/b_p)(N_c - j)\rfloor \mu_p$

Fig. 4. State diagram of the two-class model.

loss formula, as follows:

$$QoS_c = \frac{\left(\frac{\lambda_c}{\mu_c}\right)^{N_c} / N_c!}{\sum_{i=1}^{N_c} \left(\frac{\lambda_c}{\mu_c}\right)^{i} / i!}. \tag{18}$$

Finding the average delay per packet, QoS_p, is more challenging. Let MC_p denote the maximum number of channels available for packets when there is no circuit-switched traffic. Then MC_p can be represented as follows:

$$MC_p = N_p + \lfloor (b_c/b_p) N_c \rfloor. \tag{19}$$

A two-dimensional state diagram of the two-class model is shown in Fig. 4. The balance equations for Region 1 in the state diagram shown in Fig. 4 can be represented as follows: For $0 \le k \le MC_p - 1$ and $0 \le j \le N_c$,

$$\begin{aligned}&[\lambda_p + (1 - \delta_{jN_c})\lambda_c + j\mu_c + \\ &\quad \min(k, N_p + \lfloor (b_c/b_p)(N_c - j)\rfloor)\mu_p] P_{k,j} \\ &= \lambda_p P_{k-r,j} + \lambda_c P_{k,j-1} (1 - \delta_{jN_c})(j+1)\mu_c P_{k,j+1} + \\ &\quad \min(k+1, N_p + \lfloor (b_c/b_p)(N_c - j)\rfloor P_{k+1,j}.\end{aligned} \tag{20}$$

These are $MC_p(N_c + 1)$ equations with $MC_p(N_c + 1) + (N_c + 1)$ unknowns, among which $N_c + 1$ extra unknowns are boundary probabilities and $P_{MCp,j}$'s $(j = 0, 1, 2, \cdots, N_c)$ in Region 1. Therefore, additional $N_c + 1$ equations are needed to solve (20). These equations can be obtained from Region 2. The balance equations for Region 2 shown in Fig. 4 can be represented as follows: For $MC_p \le k$ and $0 \le j \le N_c$,

$$\begin{aligned}&[\lambda_p + (1 - \delta_{jN_c})\lambda_c + j\mu_c + \\ &\quad (N_p + \lfloor (b_c/b_p)(N_c - j)\rfloor)\mu_p] P_{k,j} \\ &= \lambda_p P_{k-r,j} + \lambda_c P_{k,j-1} (1 - \delta_{jN_c})(j+1)\mu_c P_{k,j+1} + \\ &\quad (N_p + \lfloor (b_c/b_p)(N_c - j)\rfloor \mu_p P_{k+1,j}.\end{aligned} \tag{21}$$

To solve (21), the following z-transform or generating function should be applied:

$$G_j(z) = \sum_{k=MC_p}^{\infty} P_{k,j} Z^{k-MC_p},\ j = 0, 1, \cdots, N_c \tag{22}$$

and the same approach used in [6] and [14] is followed with some modifications. Let $E(L)$ denote the average number of packets in the system (queue + server). Then $E(L)$ can be determined in the following way:

$$E(L) = \sum_{j=0}^{N_c} \sum_{k=0}^{MC_p - 1} k P_{k,j} + \sum_{j=0}^{N_c} [MC_p G_j(1) + G_j(1)] \tag{23}$$

where $G_j(1) = \frac{G_j(z)}{dz}|_{z=1}$.

From Little's result [19], the average delay per packet, QoS_p, can be represented as follows:

$$QoS_p = \frac{E(L)}{\lambda_p}. \tag{24}$$

Regarding the consumer decision-making model shown in Fig. 1, uniform distributions are used[10] for the customers' will-

[10]Given that real-world data on the actual distributions do not exist for the two-class services, simple uniform distributions are chosen among distributions with finite ranges. The uniform distributions can be replaced by any other distribution later, if necessary, to further refine this model.

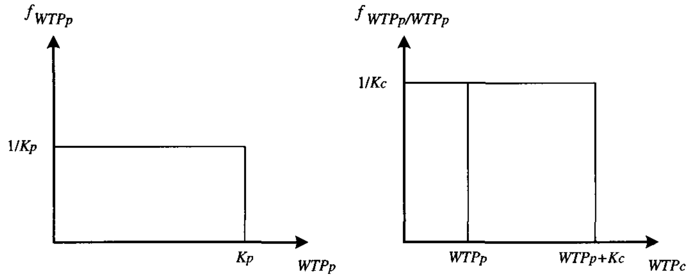


Fig. 5. Willingness-to-pay distributions used in the simulations.

ingness to pay. As shown in Fig. 5, the willingness-to-pay ranges for the packet-switched and circuit-switched services are $0{\sim}K_p$ and $0 \sim (K_p + K_c)$, respectively. For the customers' cost function, a simple linear form of $C_i^j(QoS_i^{\sim}) = v_i QoS_i^{\sim}$ is used, assuming the homogeneity[11] of the customers' valuation of the expected QoS value.

The algorithm that should be used in solving the fixed-point problems is programmed in Mathematica. Using the program, simulations to find an optimal price vector have been undertaken in Excel. The program is designed in such a way that the iterations stop when the difference between the successive values is less than 1/1000.[12] In searching for the optimal price vector, the exhaustive search algorithm is applied. This means that given the finite willingness-to-pay ranges, all possible price ranges are searched with a fixed step size. When the exhaustive search algorithm is used, the whole process is quite tedious and inefficient,[13] but it is necessary to view the overall shape of the iso-revenue regions, from which useful implications can be gathered.

A simulation output[14] of a numerical example (with $b_p = 64$ K; $b_c = 256$ K; $N_p = 3$; $N_c = 5$; $\mu_p = 4$; $\mu_c = 4$; $X = 50$; $r_{avg} = 1$; $K_p = 3$; $K_c = 3$; $v_p = 1$; $v_c = 1$; $V = 50$ K; $T = 1$; and $\Delta P_i = 0.1$) is shown in Fig. 6. By appropriately rounding up the revenue values found for all the price vectors, the graph shows iso-revenue regions or bands centered at the maximum revenue. Even within an iso-revenue band, however, each point or cell has different implications in terms of price, QoS, and demand. This implies that the network service provider can strategically control the QoS and the demand of each class while maintaining the same revenue. Moreover, the oval shapes of the iso-revenue bands in Fig. 6 imply that revenues are less sensitive to the changes in P_c than to the changes in P_p. Therefore, the network service provider has more flexibility in changing P_c than in changing P_p, without affecting the total revenue.

The center point of the iso-revenue bands in Fig. 6 represents the unit-time maximum revenue(R^*), and the corresponding optimal price vector is as follows:

$$(P_p^*, P_c^*) = (1.1, 9.4). \tag{25}$$

Table 1. Simulation results with gradually increasing demands or workloads.

r_{avg}	P_p^*	P_c^*	R^*	windowsize
1	1.1	9.4	32.61	2
2	1.2	10.8	43.92	2
3	1.3	11.5	48.84	2
4	1.4	11.9	51.59	2
5	1.4	12.2	53.38	2
6	1.4	12.3	54.66	3
7	1.4	12.5	55.64	3
8	1.5	12.6	56.41	3
9	1.5	12.7	57.05	4
10	1.5	12.8	57.59	4
11	1.5	12.8	58.06	5
12 ~	not converging (no fixed points)			

Now, to test the robustness of the model, the service request rate, r_{avg}, which is an exogenous variable in this model, is increased from the original value of 1, and the simulation results are shown in Table 1. Here increased service request rate implies increased customer demand as well as increased network workload. Table 1 shows that given the values of r_{avg} ranging from 1 to 11, it is feasible to obtain fixed points by controlling the window size. This implies that with substantial demands or workloads, self-adjustment mechanism of the model works and the demand vector as well as the QoS vector shown in Fig.1 converge at fixed points in equilibrium, although the model may require a larger window size as shown in Table 1. However, given the extremely large demands or workloads, which correspond to the values of r_{avg} greater than or equal to 12 in the simulations, it is not feasible to obtain fixed points with any values of the window size. In these cases, the only way to make the network stable is to expand its capacity, as explained in Section III-B.

Given the bandwidth allocation policy, consider the case where the QoS is guaranteed[15] by using the pricing strategy. Since the circuit-switched service is a premium service, this paper considers a case where only QoS_c is guaranteed.

The optimal point in Fig. 6 has the following QoS vector:

$$(QoS_p, QoS_c) = (1.07, 0.17). \tag{26}$$

Now, consider a specific case where $QoS_c < K$ $(K = 0.10)$, is guaranteed. By appropriately rounding up the QoS_c values found for all the price vectors, an iso-QoS_c graph is constructed, as shown in Fig. 7. Even though circuit-switched traffic experiences no interference from packet-switched traffic in Fig. 3, the shapes of the iso-QoS_c bands show that the QoS_c values are affected not only by P_c, but also by P_p. This result can be attributed to the demand substitution effects. Since the packet-switched service and the circuit-switched service are perfect or

[11]The homogeneity assumption is made for computational simplicity. This model can be refined further by relaxing this assumption and incorporating a probabilistic distribution, at the cost of increased computational complexity.

[12]This convergence tolerance is quite an acceptable small value.

[13]More efficient search algorithms, such as the steepest-ascent algorithm, can be used, but they may not provide information as useful as those that are provided by the exhaustive search algorithm.

[14]In this simulation with the window size 2, the same form of (3) is used for each service class in the simulation algorithm. In the simulation, with $w = 0.5$ in (3), (13)–(16) all converged at fixed points.

[15]Given that QoS is an average value, it is a statistical guarantee of QoS [8].

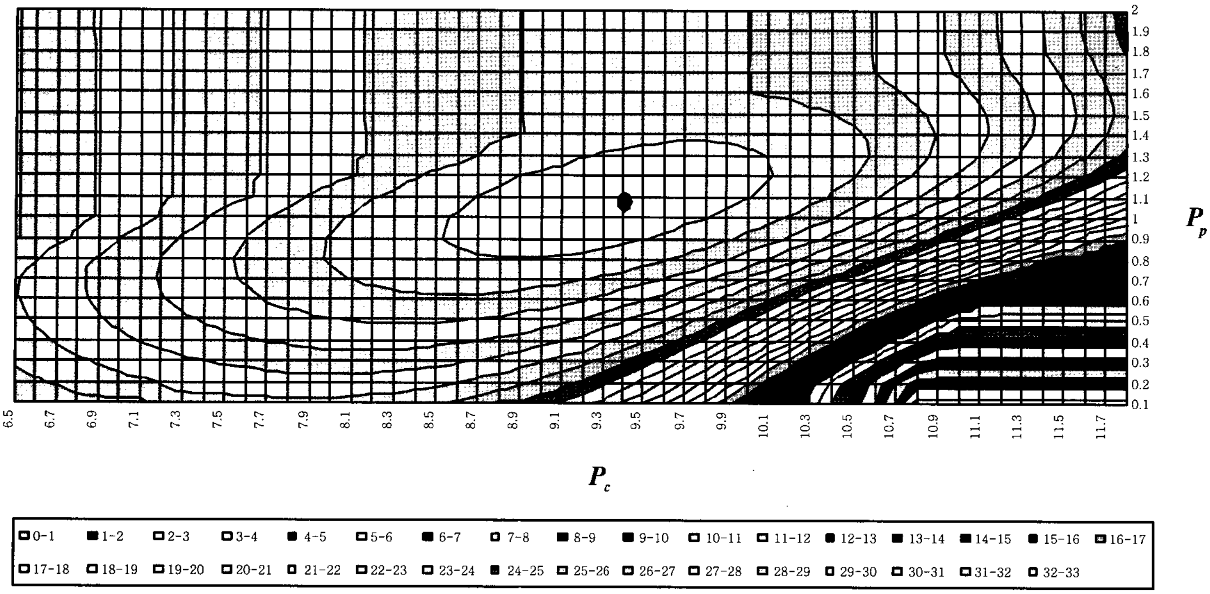


Fig. 6. Iso-revenue graph.

imperfect substitutes, depending on the types of applications and the network conditions, the demand of the circuit-switched service (λ_c) depends on P_p as well as on P_c. The same is true for the demand of the packet-switched service (λ_p). Therefore, the QoS of each class, QoS_p or QoS_c, is affected by both P_p and P_c.

To find an optimal price vector under the condition of guaranteeing QoS_c, the optimization problem, (12)–(17), should be resolved with the following modification of the constraint (14):

$$QoS_c = f_c(\lambda_p, \lambda_c). \tag{27}$$

Now, Fig. 7 shows that the optimal point has changed and that the new optimal point, indicated by an arrow, is as follows:

$$(P_p^*, P_c^*) = (1.2, 10.2). \tag{28}$$

This result means that when the price increases, the demand decreases, thus improving the QoS but decreasing the revenue.

V. CONCLUSION

This paper presents a price-based QoS control framework for two-class network services, in which *circuit-switched* and *packet-switched services* are defined as a "premium service class" and a "best-effort service class," respectively. Given the framework, fixed-point problems are solved numerically to investigate how static pricing can be used to control the demand and the QoS of each class. The rationale behind this is as follows: For a network service provider to determine the optimal prices that maximize the total revenue, close analyses of the interactions between the QoS-dependent demand and the demand-dependent QoS should be conducted. That is, the network service provider should consider the consumers' willingness to pay and their sensitivities regarding QoS when estimating their demands. The network service provider must then understand the possible impact of these demands on the QoS.

To test the robustness of the model, simulations were performed with the increased customer demands or network workloads. The simulation results show that up to a certain value of demand or workload, self-adjustment mechanism of the model works and it is feasible to obtain fixed points by controlling the window size. However, beyond that value, expanding the capacity of the network is the only solution to make it stable.

This paper proposes a means to guarantee the QoS statistically in the short term—that is, through the use of pricing strategies. The simulation results show that given the network capacity, when the price increases, the demand decreases, thus improving the QoS but decreasing the revenue. Beyond the short-term measure that relies on the pricing strategy, expanding the network capacity as a mid- and long-term strategy can be one way of increasing the revenue and of improving the QoS at the same time. Even though this is beyond the scope of this paper, it is necessary to point out that since the size and the characteristics of the customer pool ultimately vary in the mid- and long term, decision-making over this kind of investment should be made carefully based on accurate forecasts of the changes in the consumers' willingness-to-pay distributions and of their valuations of the service quality.

Regarding the implementation issues, it is very inconvenient for the customers to make a decision every time they request for a job process. Given this, the role of decision-making can be transferred to their local machines, so that the machines can make frequent decisions automatically. This can be implemented by pre-specifying in their machines the reservation prices for both classes, as well as the maximum budget to spend during the billing cycle. It is also possible to implement the system in which, for the non-emergency and elastic jobs, the customers can indicate in their machines the minimum utility level they are requiring so that the job can wait and be carried out when the minimum utility value has been met.

In these schemes, however, the QoS expectation model shown in Fig. 1 may no longer work since it is the machine, not the human customer, that makes the decisions. Therefore, either the machines should have some heuristic algorithm to estimate the expected QoS values, or the current and past information about the QoS should be explicitly provided to the machines. Further

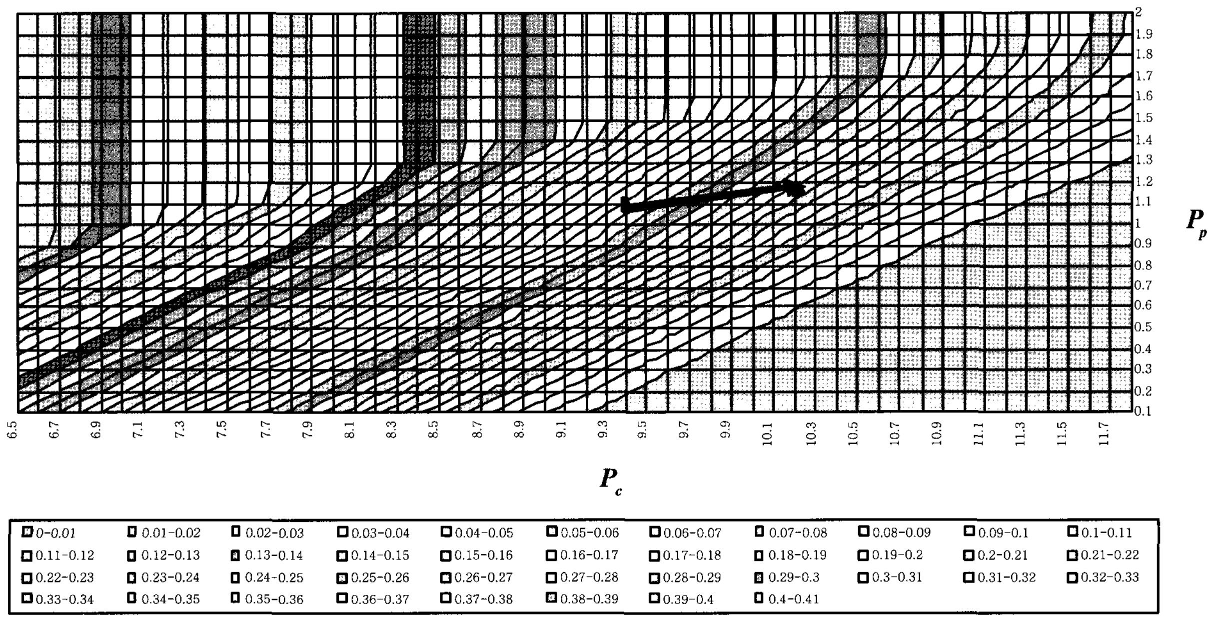


Fig. 7. Iso-QoS_c graph.

research is needed in this area.

Depending on the types of applications the customers are using, they may have different levels of willingness to pay and different valuations of the service quality. The current models may be refined to incorporate this consideration by defining application-specific customer pools and utility functions. Also, by allocating variable bandwidths instead of fixed bandwidths to job requests in the circuit-switched service, the current models can be extended to accommodate customer-specific QoS requests, instead of guaranteeing QoS statistically. Further research is needed as well in all these areas.

APPENDIX: SYSTEMATIC WAYS OF CHECKING THE STABILITY OF AN EQUILIBRIUM POINT

A.1 Derivation of the Conditions for Convergence in the Adaptive Expectation Model with Widow Size 2

For the nonlinear functions $g_p[\cdot]$ and $g_c[\cdot]$,

$$\lambda_p(m+1) = g_p[QoS_p^{\sim}(m), QoS_c^{\sim}(m)] \tag{29}$$

$$\lambda_c(m+1) = g_c[QoS_p^{\sim}(m), QoS_c^{\sim}(m)] \tag{30}$$

where

$$QoS_p^{\sim}(m) = (1-w)QoS_p(m) + wQoS_p(m-1),\ w \in (0,1) \tag{31}$$

$$QoS_c^{\sim}(m) = (1-w)QoS_c(m) + wQoS_c(m-1),\ w \in (0,1) \tag{32}$$

and for the nonlinear functions $f_p[\cdot]$ and $f_c[\cdot]$,

$$QoS_p(m) = f_p[\lambda_p(m), \lambda_c(m)] \tag{33}$$

$$QoS_c(m) = f_c[\lambda_p(m), \lambda_c(m)]. \tag{34}$$

Integrating all of the above, the followings can be obtained:

$$\begin{aligned}\lambda_p(m+1) &= g_p[QoS_p(m), QoS_c(m),\\ &\qquad QoS_p(m-1), QoS_c(m-1)]\\ &= g_p[f_p(\lambda_p(m), \lambda_c(m)], f_c(\lambda_p(m), \lambda_c(m)],\\ &\qquad f_p(\lambda_p(m-1), \lambda_c(m-1)],\\ &\qquad f_c(\lambda_p(m-1), \lambda_c(m-1)]]\\ &= h_p[\lambda_p(m), \lambda_c(m), \lambda_p(m-1), \lambda_c(m-1)]\end{aligned} \tag{35}$$

$$\begin{aligned}\lambda_c(m+1) &= g_c[QoS_p(m), QoS_c(m),\\ &\qquad QoS_p(m-1), QoS_c(m-1)]\\ &= g_c[f_p(\lambda_p(m), \lambda_c(m)], f_c(\lambda_p(m), \lambda_c(m)],\\ &\qquad f_p(\lambda_p(m-1), \lambda_c(m-1)],\\ &\qquad f_c(\lambda_p(m-1), \lambda_c(m-1)]]\\ &= h_c[\lambda_p(m), \lambda_c(m), \lambda_p(m-1), \lambda_c(m-1)]\end{aligned} \tag{36}$$

Here, $h_p[\cdot]$ and $h_c[\cdot]$ are some nonlinear functions as well.

Now, conditions for convergence to a fixed point in equilibrium are derived as follows. Let λ denote a 2-dimensional state vector, $\lambda = (\lambda_p, \lambda_c)$, and λ^* denote a 2-dimensional equilibrium point, namely, $\lambda^* = (\lambda_p^*, \lambda_c^*)$. And let H denote a 2-dimensional vector function whose components are the $h_i[\cdot]$'s. Then, with the vector notations,

$$\lambda(m+1) = H[\lambda(m), \lambda(m-1)]. \tag{37}$$

Then, in equilibrium, this becomes

$$\lambda^* = H[\lambda^*, \lambda^*]. \tag{38}$$

Let a vector $\Delta(m)$ denote the deviation from the fixed point as follows:

$$\Delta(m) = \lambda(m) - \lambda^*. \tag{39}$$

Then, $\lambda(m+1) = H[\lambda(m), \lambda(m-1)]$ can be expressed as

$$\lambda^* + \Delta(m+1) = H[\lambda^* + \Delta(m), \lambda^* + \Delta(m-1)]. \quad (40)$$

Now, using Taylor expansion, the above can be approximated as follows:

$$\lambda^* + \Delta(m+1) \approx H[\lambda^*, \lambda^*] + \nabla_{\lambda(m)} H(\lambda^*, \lambda^*)(\lambda(m) - \lambda^*) \\ \nabla_{\lambda(m-1)} H(\lambda^*, \lambda^*)(\lambda(m-1) - \lambda^*). \quad (41)$$

Since $\lambda^* = H[\lambda^*, \lambda^*]$ in equilibrium, this becomes

$$\Delta(m+1) \approx \nabla_{\lambda(m)} H(\lambda^*, \lambda^*)(\lambda(m) - \lambda^*) \\ \nabla_{\lambda(m-1)} H(\lambda^*, \lambda^*)(\lambda(m-1) - \lambda^*). \quad (42)$$

Let's rewrite the above as follows:

$$\Delta(m+1) \approx \begin{bmatrix} \frac{\partial \lambda_p(m+1)}{\partial \lambda_p(m)} & \frac{\partial \lambda_p(m+1)}{\partial \lambda_c(m)} \\ \frac{\partial \lambda_c(m+1)}{\partial \lambda_p(m)} & \frac{\partial \lambda_c(m+1)}{\partial \lambda_c(m)} \end{bmatrix}_{\text{at } \lambda^*} \Delta(m) + \\ \begin{bmatrix} \frac{\partial \lambda_p(m+1)}{\partial \lambda_p(m-1)} & \frac{\partial \lambda_p(m+1)}{\partial \lambda_c(m-1)} \\ \frac{\partial \lambda_c(m+1)}{\partial \lambda_p(m-1)} & \frac{\partial \lambda_c(m+1)}{\partial \lambda_c(m-1)} \end{bmatrix}_{\text{at } \lambda^*} \Delta(m-1) \\ = J_p(\lambda^*)\Delta(m) + J_c(\lambda^*)\Delta(m-1). \quad (43)$$

Here $J_p(\lambda^*)$ and $J_c(\lambda^*)$ represent Jacobian matrices.

Since the above equation is a 2nd-order difference equation, I will convert it into the equivalent system of 1st-order difference equation, so that Liapunov's first method [9] can be applied in checking the stability. Let's define a vector,

$$Z(m) = \begin{bmatrix} Z_1(m) \\ Z_2(m) \end{bmatrix} \quad (44)$$

where $Z_1(m) = \Delta(m-1)$ and $Z_2(m) = \Delta(m)$.

With these definitions, it follows immediately that

$$Z_1(m+1) = Z_2(m) \quad (45)$$

$$Z_2(m+1) = J_p(\lambda^*)Z_2(m) + J_c(\lambda^*)Z_1(m) \quad (46)$$

and in matrix form,

$$\begin{bmatrix} Z_1(m+1) \\ Z_2(m+1) \end{bmatrix} = \begin{bmatrix} 0 & I \\ J_c(\lambda^*) & J_p(\lambda^*) \end{bmatrix} \begin{bmatrix} Z_1(m) \\ Z_2(m) \end{bmatrix}. \quad (47)$$

That is,

$$Z(m+1) = J_T(\lambda^*)Z(m) \quad (48)$$

where

$$J_T(\lambda^*) = \begin{bmatrix} 0 & I \\ J_c(\lambda^*) & J_p(\lambda^*) \end{bmatrix} \\ = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \frac{\partial \lambda_p(m+1)}{\partial \lambda_p(m-1)} & \frac{\partial \lambda_p(m+1)}{\partial \lambda_c(m-1)} & \frac{\partial \lambda_p(m+1)}{\partial \lambda_p(m)} & \frac{\partial \lambda_p(m+1)}{\partial \lambda_c(m)} \\ \frac{\partial \lambda_c(m+1)}{\partial \lambda_p(m-1)} & \frac{\partial \lambda_c(m+1)}{\partial \lambda_c(m-1)} & \frac{\partial \lambda_c(m+1)}{\partial \lambda_p(m)} & \frac{\partial \lambda_c(m+1)}{\partial \lambda_c(m)} \end{bmatrix}_{\text{at } \lambda^*} \quad (49)$$

As long as all eigenvalues of the combined Jacobian matrix, $J_T(\lambda^*)$, lie inside the unit circle of the complex plane, the deviation from the fixed point, $Z(m+1)$, will converge to zero, thereby implying the convergence to a fixed point in equilibrium. Therefore, a necessary and sufficient condition for asymptotic stability is

$$\forall |\text{eigenvalue of } J_T(\lambda^*)| < 1. \quad (50)$$

All elements of the combined Jacobian matrix, $J_T(\lambda^*)$, can be numerically estimated using Chain rules and therefore, 4 eigenvalues of $J_T(\lambda^*)$ can be found.

A.2 Generalization into Window Size n in the Adaptive Expectation Model

By generalizing the window size into n, the linear approximation of the nonlinear system results in the following:

- An nth order difference equation;
- n Jacobian matrices;
- $2n \times n$ combined Jacobian matrix; and
- $2n$ eigenvalues.

A larger n makes the stability conditions more difficult to analyze. In the extreme case of the "entire-history" adaptive expectation model, in which the window size corresponds to the total time it takes to converge, it is very difficult to check the local stability analytically. Therefore, it is important to find the appropriate window size from the analytical point of view.

REFERENCES

[1] R. J. Gibbens and F. P. Kelly, "Resource pricing and the evolution of congestion control," *Automatica*, vol. 35, pp. 1969–1985, Dec. 1999.

[2] S. Kunniyur and R. Srikant, "End-to-end congestion control: Utility functions, random losses and ECN marks," in *Proc. IEEE INFOCOM 2000*, Tel Aviv, Israel, 2000, pp. 1323–1332.

[3] R. J. Edell, N. Mckeown, and P. P. Varaiya, "Billing users and pricing for TCP," *IEEE J. Sel. Areas Commun.*, vol. 13, pp. 1162–1175, Sept. 1995.

[4] O. Rose, "Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems," in *Proc. 20th Annu. Conf. Local Computer Networks*, Minneapolis, MN, 1995, pp. 397–406.

[5] P. Thomas, D. Teneketzis, and J. K. Macki-Mason, "A market-based approach to optimal resource allocation in integrated-services connetion-oriented networks," *Operations Research*, vol. 50, pp. 603–616, 2002.

[6] G. de Veciana and R. Baldick, "Resource allocation in multiservice networks via pricing: Statistical multiplexing," *Comput. Networks ISDN Syst.* vol. 30, pp. 951–962, 1998.

[7] Y. Jin and G. Kesidis, "Nash equilibria of a generic networking game with applications to circuit-switched networks," in *Proc. IEEE INFOCOM 2003*, San Francisco, CA, 2003, pp. 1242–1249.

[8] M. Mandjes, "Pricing strategies under heterogeneous service requirements," in *Proc. IEEE INFOCOM 2003*, San Francisco, CA, 2003, pp. 1210–1220.

[9] P. Marbach, "Priority service and max-min fairness," in *Proc. IEEE INFOCOM 2002*, New York, 2002, pp. 266–275.

[10] P. Marbach, "Analysis of a static pricing scheme for priority services," *IEEE/ACM Trans. Net.*, vol. 12, pp. 312–325, Apr. 2004.

[11] D. Bertsekas and R. Gallager, *Data Networks*. Prentice Hall, 1992.

[12] M. Carter and R. Maddock, *Rational Expectations*, Macmillan, London, 1984.

[13] B. Kraimeche and M. Schwartz, "Analysis of traffic access control strategies in integrated service networks," *IEEE Trans. Commun.*, vol. pp. 1085–1093, Oct. 1985.

[14] W. Lee, "QoS provisioning technologies for media streaming transmission," *Telecommunications Review*, vol. 13, pp. 188–197, 2003.

[15] D. G. Luenberger, *Introduction to Dynamic Systems: Theory, Models & Applications.* John Wiley & Sons, Inc., 1979.

[16] Y. Masuda and S. Whang, "Dynamic pricing for network services: Equilibrium and stability," *Management Science*, vol. 45, pp. 857–869, 1999.

[17] I. Mitrani and P. J. B. King, "Multiprocessor systems with preemptive priorities," *Performance Evaluation*, pp. 118–125, 1981.

[18] C. M. Rump and S. Stidham, "Stability and chaos in input pricing for a service facility with adaptive customer response to congestion," *Management Science*, vol. 44, pp. 246–261, 1998.

[19] J. Walrand and P. Varaiya, *High-Performance Communication Networks*, Morgan Kaufmann, 2000.

Whan-Seon Kim was born in Seoul, Korea, on July 29, 1966. He received his B.S. and M.S. degrees in Electrical Engineering from Yonsei University, Korea in 1989 and University of Wisconsin at Madison, U.S.A. in 1992, respectively. He also received his M.B.A. degree from University of California at Berkeley, U.S.A. in 1994 and Ph.D. degree in Management Science from Stanford University, U.S.A. in 2002. From 1994 to 1996, he worked at Korea Information Strategy Development Institute (KISDI) in Seoul, Korea. He is currently an associate professor in the Business School at Myongji University, Seoul Korea. His major research interests are in information and communications technology, information and telecommunications policy and economics, digital convergence, and digital ecosystem.