A Model Analysis of Pricing and Link Bandwidth Allocation in a Multiple Class-of-Service Network

Hao Che and Siyang Zheng
Department of Electrical Engineering, The Pennsylvania State University
University Park, PA 16802, USA
Xinwei Hong
Department of Electronic and Information, Huazhong University of Science and Technology
Wuhan, China 430074

Abstract— **In this paper, an analytic model is developed for the study of the impact of user behaviors on the effectiveness of pricing, service quality, and link bandwidth allocation/sharing in a multiple class-of-service Internet. The model characterizes a variety of user behaviors, such as user sensitivity to service price/quality and user migration effect from one class of service to another. It also incorporates a usage-based pricing scheme and a general trunk reservation policy. In particular, a three-tier service model for a single link is considered. Numerical studies reveals the richness of this research issue and provides insights in how to optimally select service price, quality and trunk reservation values with given user behaviors. We find that trunk reservation is effective in terms of service protection only when the call migration effect is small and call migration effect can be harnessed by a proper design of service pricing-quality structure. We also find that user sensitivities have great impact on service price-quality structure design and care must be taken to avoid sudden network performance degradation at certain price-quality parameter values.**

Keywords— **Differentiated Services, Quality of Services, Assured Service for TCP, Performance measurement, Adaptive algorithm design, End-to-End performance**

I. Introduction

The network design methodology which separates service design from traffic behavior has attracted much attention recently, e.g. the differentiated services (DS) model [1], [2] and DS in multi-protocol label switching (MPLS) networks [3]. To provide this kind of services, several key design issues are identified: (1) what CoSs should be offered and how to set service quality for each CoS; (2) what pricing policy should be used and how to set the price values for individual CoSs; (3) how to allocate network resources and allow resource sharing among different CoSs. The decoupling of service design from traffic behavior results in a loose tie between services and applications. Besides application requirements the service price-quality structure largely determines user behaviors. User behaviors impact the effectiveness of resource allocation policies, which in turn determine the service quality. The strong coupling and intertwining of different design aspects calls for an integrated approach with all different aspects being taken into account.

Unlike the previous work on the optimal resource allocation and pricing based on microeconomic and game-theoretic approaches [5], [6], [7], this paper focuses on the performance analysis of the impact of user behaviors on the effectiveness of a simple pricing scheme and a bandwidth allocation policy.

Whether dynamic pricing or static pricing policies should be used in multiple CoS-enabled Internet has been a widely debated issue [8], [9]. Dynamic pricing can potentially improve network performance by regulating user behaviors in response to network congestion [10]. On the other hand, it may introduce psychologically negative effect on users due to unpredictable service rates. In this paper, we adopt a simple static pricing policy and usage charges are used for reservation based CoSs.

A commonly adopted bandwidth allocation/sharing technique is to allow bandwidth sharing among different CoSs while using *trunk reservation* to protect one CoS from being overloaded by other CoSs. Here trunk reservation means that bandwidth sharing among different CoSs is allowed, provided that a given amount of bandwidth is reserved for the underloaded CoSs. The trunk reservation technique was successfully in a multi-CoS circuit switched link to allocate link bandwidth [16], and in a circuit switched network to guard against network instability [4]. Recently, Gerald Ash [11] pointed out the importance of trunk reservation for quality of service (QoS) based routing in IP networks, traffic engineering in MPLS networks, and bandwidth sharing in multiple CoS-enabled IP networks. In this paper, we examine a simple but general link bandwidth allocation policy. It combines trunk reservation and bandwidth partitioning and it takes complete sharing and complete partitioning policies [17] as its special cases.

With a simple user utility function, we were able to characterize rather complex user behaviors. In particular, we were able to quantitatively characterize an important phenomenon known as *call migration effect*, i.e., a user whose call gets blocked by one CoS may decide to migrate to another CoS. E. Altman, et al. first studied the call migration effect between QoS traffic and the best-effort traffic. [12]. In this paper, we study the call migration effect from a more fundamental point of view by formulating it as a direct consequence of user behaviors.

The remainder of this paper is organized as follows. In Section 2, the problem is formulated. Section 3 develops an analytical model to solve the problem posed in Section 2. Section 4 presents the numerical results and analyses of the results. Finally, conclusions and future work are given

0-7803-6494-5/00/$10.00 © 2000 IEEE

in Section 5.

II. Problem Formulation

A. Service Model

We consider a three-tier model with two reservation-based service classes and a best-effort service class. In this model, traffic is described at *call (session) level* and each reservation-based call requires a fixed bandwidth. Throughout the paper, we shall use the terms "call" and "session" interchangeably. The three CoSs defined in this model bear some resemblance to the three DS classes described in [14]. However, our model is different from the DS model in the sense that ours is a *generic* performance analysis tool without specification on implementation.

B. Admission Control and Bandwidth Allocation/Sharing

A call for the best-effort service is always admitted. The admission control policy for the other two CoSs depends on the actual bandwidth allocation/sharing policy in use. For simplicity, the two reservation-based CoSs and best-effort service are referred to as the 1st, 2nd, and 3rd CoSs, respectively, where the 1st CoS requires more stringent QoS guarantee than the 2nd CoS. Let W_1, W_2, and W_3 be the *nominal bandwidths* allocated to the 1st, 2nd, and 3rd CoSs, respectively, and $W_t = W + W_3$ and $W = W_1 + W_2$, where W_t is the total link bandwidth and W is the total bandwidth allocated to the 1st and 2nd CoSs. We consider the following link bandwidth allocation/sharing policy: Admit an i-th CoS session with requested rate u_i, if and only if

$$u_i \leq W - U_1 - U_2 - \Delta_i, \qquad \text{for } i = 1, 2 \tag{1}$$

where

$$\Delta_i = \begin{cases} 0, \quad \text{if } U_i < W_i - u_i \\ min\{R_j, W_j - U_j\}\theta(W_j - U_j), \\ \qquad \text{if } U_i > W_i - u_i \text{ for } i \neq j = 1, 2 \end{cases} \tag{2}$$

where R_j is the trunk reservation for the j-th CoS, U_i is the aggregate bandwidth in use for the i-th CoS, and $\theta(t)$ is a step function where $\theta(t) = 1$ for $t > 0$, and 0, otherwise. Note that Δ_i is the bandwidth reservation *against* the admission of the i-th CoS session. Also note that by setting $R_i = 0$ for $i = 1, 2$, the above policy reduces to the complete sharing policy for the 1st and 2nd CoSs, and by setting $R_i = W_i$, it degenerates to the complete partitioning policy for the two CoSs.

C. Performance Measures

Three performance measures are used in this paper. One is *network revenue* V, which is the average rate of network monetary income, subtracted by an implicit average cost rate. The implicit average cost rate is generated by user dissatisfaction and the loss of revenue due to call blocking. The other two performance measures we use are the call blocking probabilities for the two reservation-based CoSs. The network revenue is used as a measure of the overall network performance. The blocking probabilities are used to measure the performance of the reservation-based CoSs.

D. Pricing and Service Quality

First, we define our pricing policy as follows. A fixed monthly fee is applied to the service subscription. There is no per call charge for the best-effort service. A user of any of the other two services incurs *usage charge* with a fixed price rate and proportional to the reserved bandwidth multiplied by the call duration. Note that in our model analysis, the monthly fee is irrelevant because we are only concerned with short-term user behaviors (or dynamics) relevant to the bandwidth allocation/sharing. We define the usage prices for the 1st, 2nd, and 3rd CoSs as p_1, p_2, and p_3 respectively with $p_3 = 0$. The prices are measured in the unit of dollars per unit bandwidth per unit time.

In general, the service quality of a given CoS can be characterized by a multi-dimensional vector with components such as delay, call blocking probability, packet loss rate, and so on. It can be a function of time and should be measured based on the current network congestion situation. For simplicity, in this paper, we use a single-valued static parameter q_i to characterize the quality of the $i-th$ CoS ($i = 1, 2, 3$), which is determined by measured service quality of a session in isolation or measured average quality of a session over a long period of time.

E. User Behavior

In order to quantitatively characterize user behavior as a consequence of service quality and price tradeoff, here we make a strong assumption that service qualities q_i can be measured in the same unit as the prices p_i. By following Sairamesh and Kephart [15], a user behavior is characterized by the following user utility function:

$$\begin{aligned} U_i(p, q, \gamma) = \quad & \{\gamma(p - p_i) + (1 - \gamma)(q_i - q)\} \\ & \theta(p - p_i)\theta(q_i - q), \quad \text{for } i = 1, 2, 3 \end{aligned} \tag{3}$$

where $U_i(p, q, \gamma)$ is the user utility with respect to the $i-th$ CoS, and p_i and q_i are the price and quality for the $i - th$ CoS, respectively. p is the maximum price a user is willing to pay. q is the minimum quality the user can tolerate. $\theta(x)$ is a step function which is 0 for $x \leq 0$ and 1 for $x > 0$. γ is a weight that quantifies a user's sensitivity to service price as compared to service quality. A user with $\gamma = 1$ is at the extreme limit of price sensitivity: it will choose a CoS with the lowest price, so long as its received quality is no less that q. On the other hand, a user with $\gamma = 0$ is at the extreme limit of quality sensitivity: it will choose a CoS with the highest quality, so long as what is paid is no more than p. A user's decision-making process is as follows. He will select a CoS for which his utility is maximal among three CoSs. If the maximal utility is zero, he will not use any of the services. If the access gets blocked, he will then try the CoS with the next highest utility, and so on, until he is granted an access, otherwise, he will drop the call. Hence, with our modeling technique, the call migration effect can be quantitatively characterized by the service quality-price tradeoff.

III. Analytic Approach

A. Link Model

To be mathematically tractable, we formulate the overall problem as a two-CoS, multi-rate loss model plus a "sink". The "sink" is a channel with bandwidth W_3 which drains the best-effort traffic as well as the traffic migrated to the best-effort service from the other two CoSs. The link bandwidth allocation/sharing is dictated by (1). For simplicity, we further assume that 1st and 2nd CoS calls will request a fixed bandwidth of integer k units and 1 unit respectively. Then it makes sense to assume that W_1 and R_1 are integer multiples of k. The maximum numbers of calls, N_1, and N_2, that the 1st and 2nd CoSs can carry are given by,

$$N_1 = W_1/k, \qquad N_2 = W_2 \tag{4}$$

B. Call Arrival and Departure Processes

We assume that the call arrival is Poisson process with mean arrival rate λ. Each arrived call *independently* makes its decision as to which CoS it should request an access to. For any given user population, the probability for a user to access any given CoS is fixed, denoted by β_i $(i = 1,2,3)$. It can be shown [18] that the resulting arrival process for each CoS is still Poisson with mean arrival rate given by,

$$\lambda_i = \beta_i \lambda \qquad \text{for } i = 1,2,3 \tag{5}$$

To account for the call migration effect, we assume that a blocked request for the i-th CoS has a probability α_{ij} to migrate to the j-th CoS, for $i \neq j$, and $i,j = 1,2,3$. Note that $\alpha_{31} = \alpha_{32} = 0$, since a request for the best-effort service is always admitted.

We further assume that the call durations are *i.i.d.* and they follow an exponential distribution with departure rate μ_i for the i-th CoS $(i = 1,2)$. Fig. 1 gives two example model scenarios under two different overload/underload situations.

C. Model of User Behavior

Throughout the paper, we assume $p = q$, i.e., when a user has a high quality requirement, he will be willing to pay a high price, and vice versa. We also assume that p (or q) and γ are independent of each other. We further impose the following conditions on the relative values of p_i and q_i:

$$\begin{cases} q_i > p_i & \text{for} \quad i = 1,2,3 \\ q_1 > q_2 > q_3 \\ p_1 > p_2 > p_3 \end{cases} \tag{6}$$

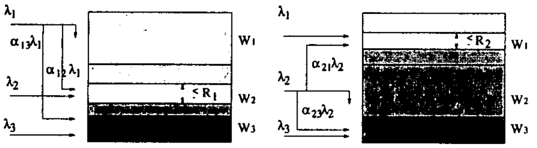

Fig. 1. Two model scenarios: the 1st CoS is overloaded and the 2nd CoS is underloaded (left), and vice versa (right)

$q_i > p_i$ is required because otherwise $U_i = 0$ for any p and no user will make an attempt to access the $i-th$ CoS. Since there is no charge for the best-effortservice users, $p_3 = 0$. Without loss of generality, we set $q_1 = 1$.

With the above assumptions, we can now use *price-quality phase diagrams* to classify user behaviors. In Fig. 2, five possible phase diagrams are plotted, corresponding to five relative values of p_1 and p_2 with respect to q_2 and q_3. Let A_i be the set defined by

$$A_i = \{a | a \in (p_i, q_i)\} \qquad \text{for} \quad i = 1,2,3. \tag{7}$$

Note that $U_i(p, p, \gamma)$ is positive when p falls into A_i. Hence, we call A_i the feasible region for the i-th CoS.

In the first phase diagram, no customer will have more than one non-zero utilities, and thus call migrations cannot occur for this case. If p falls into any of the two shaded areas, called *forbidden zones*, no service will meet his price-quality requirements and he will simply drop the call. In diagram 2, region $A_1 \cap A_2$, the intersection between A_1 and A_2, is most interesting. A request falls into this region may migrate from the 1st CoS to the 2nd CoS if $U_1 > U_2$, and vice versa. Diagram 3 represents a case when call migrations can take place between the 2nd and 3rd CoSs. In diagram 4, call migrations can occur between the 1st and 2nd CoSs, and the 2nd and 3rd CoSs as well. In diagram 5, call migrations can occur among all three CoSs in region $A_1 \cap A_2 \cap A_3$.

Let Ω be the sample space $[0, 1]$ for any given diagram. A probability space $\{\Omega, A, P(A)\}$ can be defined with $A \subseteq \Omega$ and

$$P(A) = \int_A f(p)dp, \qquad P(\Omega) = 1, \qquad P(\emptyset) = 0. \tag{8}$$

We define:

$$\begin{aligned} \beta_i(\gamma) = \quad & P\{p | U_i(p,\gamma) > max\{U_j(p,\gamma), U_k(p,\gamma)\}, \\ & i \neq j \neq k\} \quad i,j,k = 1,2,3 \end{aligned} \tag{9}$$

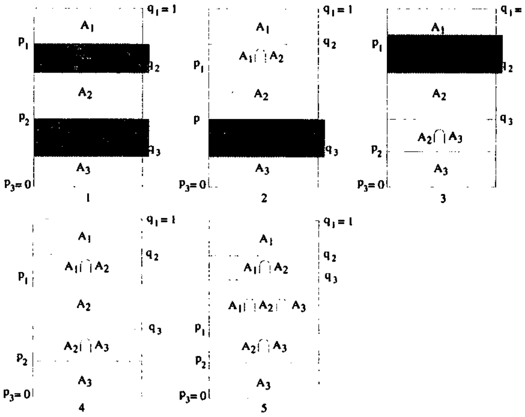

Fig. 2. Price-quality phase diagrams

$$\alpha_{ij}(\gamma) = \begin{cases} \frac{P\{p|U_i(p,\gamma)>U_j(p,\gamma)>0\&U_j(p,\gamma)>U_k(p,\gamma), i\neq j\neq k\}}{\beta_i(\gamma)} \\ \quad \text{i=1,2, and j,k=1,2,3} \\ 0 \\ \quad \text{i=3, j=1,2.} \end{cases} \tag{10}$$

where $U_i(p,\gamma) \equiv U_i(p,p,\gamma)$.

Three user sensitivities groups are considered, i.e., an incoming call has probabilities r_1, r_2 and r_3 ($r_1 + r_2 + r_3 = 1$) to be extremely quality sensitive ($\gamma = 0$), equally sensitive to quality and price ($\gamma = 0.5$), and extremely price sensitive ($\gamma = 1$), respectively. We then have,

$$\begin{aligned} \beta_i &= r_1\beta_i(0) + r_2\beta_i(0.5) + r_3\beta_i(1), \\ \alpha_{ij} &= r_1\alpha_{ij}(0) + r_2\alpha_{ij}(0.5) + r_3\alpha_{ij}(1). \end{aligned} \tag{11}$$

D. Model Analysis

The model is formulated in mathematical terms as a Markov process. Each state in the Markov process is uniquely identified by a tuplet (n_1, n_2) where n_1 (n_2) is the number of active calls in the 1st (2nd) CoS. According to the bandwidth allocation/sharing policy, a feasible state should satisfy the following conditions: The total occupied bandwidth should not exceed the total bandwidth W for CoSs 1 and 2; the occupied bandwidth of the 1st (2nd) CoS should not exceed $W - R_2$ ($W - R_1$). So, we define,

$$\begin{aligned} \mathbf{S} := \{\mathbf{n} = \ & (n_1, n_2)|kn_1 + n_2 \leq W, \\ & 0 \leq n_1 \leq N_1 + (W_2 - R_2)/k, \\ & 0 \leq n_2 \leq N_2 + W_1 - R_1\} \end{aligned} \tag{12}$$

as the set of feasible bandwidth occupancies. Denote by the set $\mathbf{D_1}$ ($\mathbf{D_2}$) the acceptance region for the 1st (2nd) CoS calls. We then have,

$$\begin{aligned} \mathbf{D_1} \ := & \{\mathbf{n} \in \mathbf{S}|(n_2 \geq N_2, kn_1 + n_2 \leq W - k) \cup \\ & (n_2 < N_2, kn_1 \leq W_1 + (W_2 - n_2 - R_2)^+ - k)\}, \end{aligned} \tag{13}$$

$$\begin{aligned} \mathbf{D_2} \ := & \{\mathbf{n} \in \mathbf{S}|(n_1 \geq N_1, kn_1 + n_2 \leq W - 1) \cup \\ & (n_1 < N_1, n_2 \leq W_2 + (W_1 - kn_1 - R_1)^+ - 1)\}, \end{aligned} \tag{14}$$

where $(x)^+$ equals x when x is larger than zero, otherwise it takes value zero. Define $\mathbf{M_1}$ ($\mathbf{M_2}$) to be the region where there are the 1st (2nd) CoS call blocking and call migration. Then we have,

$$\begin{aligned} \mathbf{M_1} &:= \{\mathbf{n} \in \mathbf{S}|\mathbf{n} \notin \mathbf{D_1}\}, \\ \mathbf{M_2} &:= \{\mathbf{n} \in \mathbf{S}|\mathbf{n} \notin \mathbf{D_2}\}. \end{aligned} \tag{15}$$

The bandwidth occupancy constitutes an irreducible Markov process with the state space $\mathbf{S}$. The steady-state probability distribution, $\pi(\mathbf{n})$, for each state $\mathbf{n} \in \mathbf{S}$ is determined by the following global balance equations, which can be solved using any linear equation solution procedure, for example, Gauss-Siedel method,

$$\begin{aligned} & \pi(\mathbf{n}) \textstyle\sum_{\mathbf{i=1,2}}(\lambda_\mathbf{i}\mathbf{I}(\mathbf{n} \in \mathbf{D_i}) + \alpha_\mathbf{ij}\lambda_\mathbf{i}\mathbf{I}(\mathbf{n} \in \mathbf{M_i}) + \mathbf{n_i}\mu_\mathbf{i}) \\ & = \textstyle\sum_{i,j=1,2,i\neq j}[\lambda_i I(\mathbf{n} - \mathbf{e_i} \in \mathbf{D_i}) + \alpha_\mathbf{ji}\lambda_\mathbf{j}\mathbf{I}(\mathbf{n} - \mathbf{e_i} \in \mathbf{M_j})] \\ & \pi(\mathbf{n} - \mathbf{e_i}) + \textstyle\sum_{\mathbf{i=1,2}} \pi(\mathbf{n} + \mathbf{e_i})(\mathbf{n_i} + 1)\mu_\mathbf{i}\mathbf{I}(\mathbf{n} + \mathbf{e_i} \in \mathbf{S}), \\ & \quad \forall\ \mathbf{n} \in \mathbf{S}, \textstyle\sum_{\mathbf{n}\in\mathbf{S}} \pi(\mathbf{n}) = 1. \end{aligned} \tag{16}$$

where $\mathbf{I}(x)$ is the indicator function of event x, and $\mathbf{e_1} = (1, 0), \mathbf{e_2} = (0, 1)$.

Now, the performance measures can be easily calculated. The average network revenue per unit time, V, can be expressed as:

$$\begin{aligned} & V = \textstyle\sum_{i=1}^2 \sum_{\mathbf{n}\in\mathbf{D_i}} k_i p_i n_i \pi(\mathbf{n}) \\ & - \textstyle\sum_{i=1,j\neq i}^2 c_i k_i p_i \frac{(1-\alpha_{ij})\lambda_i}{\mu_i} \sum_{\mathbf{n}\in\mathbf{S}-\mathbf{D_i}} \pi(\mathbf{n}) \\ & - \textstyle\sum_{i=1,j\neq i}^2 c_i k_i p_i \frac{\alpha_{ji}(1-\alpha_{i3})\lambda_j}{\mu_i} \sum_{\mathbf{n}\in\mathbf{S}-(\mathbf{D_1}\cup\mathbf{D_2})} \pi(\mathbf{n}) \end{aligned} \tag{17}$$

where $k_1 = k, k_2 = 1$,and c_i ($i = 1, 2$) is the penalty (cost) coefficient for an i-th CoS call blocking without call migration. A call blocking not only means that there is a loss of network revenue but it also causes user dissatisfaction. So in general, c_i is set at a value larger than 1 [16]. Note that any call migrations to the 3rd CoS will be accepted and hence there is no penalty associated with the call migrations to the 3rd CoS. We also use call blocking probabilities of the 1st and 2nd CoSs, denoted by B_1 and B_2, respectively, as performance measures to reflect the dynamic nature of the service qualities. They are,

$$B_1 = \sum_{\mathbf{n}\in\mathbf{S}-\mathbf{D_1}} \pi(\mathbf{n}), \qquad B_2 = \sum_{\mathbf{n}\in\mathbf{S}-\mathbf{D_2}} \pi(\mathbf{n}). \tag{18}$$

IV. Numerical Analysis

Due to space limitation, only a few typical case studies are presented in this paper. We set $f(p) = 1$, which means we only consider a uniformly distributed user population in terms of p and q ($p = q$ as assumed). For all our numerical studies, the following parameters are fixed: $W_1 = 40$, $W_2 = 20$, $k = 2$, $\mu_1 = 2$, $\mu_2 = 1$, $p_3 = 0$, $p_2 = 0.1$, $q_1 = 1$, $q_3 = 0.1$, $c_1 = c_2 = 4$, and we only consider $q_2 \geq p_1$. Consequently, we focus on the study of a special case of phase diagram 4 in Fig. 2, i.e, $A_2 \cap A_3 = \phi$ without creating a forbidden zone.

A. Offered Load

We study the performance under different overall offered loads. The following parameters are used for the numerical analysis: $r_1 = 0.5, r_2 = 0, r_3 = 0.5$, $p_1 = 0.6$, and $q_2 = 0.8$. The first row in Fig. 3 presents V, B_1, and B_2 against the overall offered load λ at four different trunk reservation values: $R_1/W_1 = R_2/W_2 = 0\%, 20\%, 50\%$, and 100%. To identify the call migration effect, the three performance measures without call migrations are calculated and presented in the second row in Fig. 3.

First, we note, from the first plot for V, that when the link is lightly loaded, trunk reservation is unnecessary and the complete sharing policy offers the highest revenue. However, the complete sharing policy results in a quick drop of the revenue when the offered load λ exceeds 50 calls per second and it also causes a sharp increase of the blocking probability for the 1st CoS (see the first plot for B_1) as the offered load increases. On the other hand, the complete partitioning policy suffers from an overall low V values and a rather high B_2 values due to the lack of resource sharing. One can see that the case with 20% trunk

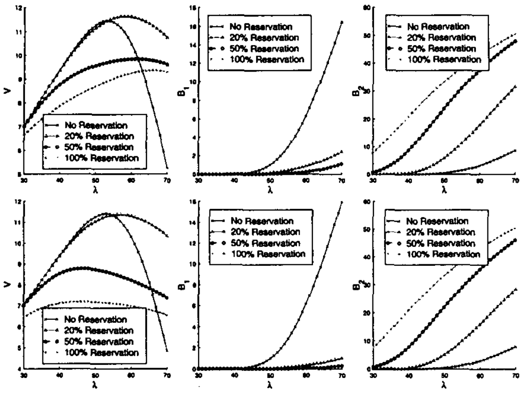

Fig. 3. Performance measures versus traffic load with and without call migration effect

reservations for both CoSs offers the best tradeoffs. It offers the highest overall V values, a good protection for the 1st CoS calls (low B_1 values) and lower overall B_2 values than the case for complete partitioning. In fact, $10 - 20\%$ trunk reservation has been found to be sufficient for all the cases we studied.

Second, by comparing with the case without call migration, we find that B_1 reduces when call migration is not allowed. This suggests that the call migration effect reduces the effectiveness of trunk reservation in protecting the 1st CoS against the 2nd CoS calls. As the overlap region $A_1 \cap A_2$ in diagram 4 of Fig. 2 increases when either p_1 decreases or q_1 increases (not given here), we find that the negative impact of the call migration effect becomes significant. On the other hand, the call migration effect improves network revenue by allowing better bandwidth sharing.

B. Effect of User Behaviors

In what follows, we study the effect of distinct user sensitivities in isolation. We set the trunk reservation values at 20% for both CoSs.

First, we study an extremely price sensitive ($r_1 = 0, r_2 = 0, r_3 = 1$) user population. With the other parameters unchanged from the previous case, we calculated the three performance measures at different p_1 and q_2 values, subject to $p_1 \leq q_2$, as shown in Fig. 4. For any fixed q_1 value, increasing p_1 both improves revenue and reduces the blocking probabilities for both CoSs and in a large quality range (>0.5), reducing q_2 generates higher revenue without significantly changing the blocking probabilities for both CoSs. This observation has a significant implication on the optimal price and quality setting for price sensitive population, i.e., set $p_1 = q_2$, or $A_1 \cap A_2 = \phi$. The key to understand this unique phenomenon is to realize that all the calls falling into the overlap region $A_1 \cap A_2$ will choose the 2nd CoS and reducing the overlap region by raising p_1 will not change the call arrival rate for the 2nd CoS. The only effect of raising p_1 is to reduce the call migrations from the 2nd CoS to the 1st CoS. Hence B_1 decreases. The increase of V is due to increased price p_1 for the 1st CoS calls. However, it seems counter intuitive that B_2 also drops when raising p_1 since the call arrival rate for the 2nd CoS does not change with the increase of p_1. The reason is that the reduced traffic load for the 1st CoS provides better sharing of its own bandwidth with relatively heavily loaded 2nd CoS. Here we see, once again, how remarkable the trunk reservation mechanism is in providing CoS quality protection and bandwidth sharing. Finally, reducing q_2 reduces the call arrival rate for the heavily loaded 2nd CoS and increases the call arrival rate for the lightly loaded 1st CoS. Consequently, reducing q_2 helps to improve V at any given p_1.

Second, let's study the case when user population is extremely quality sensitive ($r_1 = 1, r_2 = 0, r_3 = 0$). The numerical results are presented in Fig. 5. First, we note that all three performance measures are rather insensitive to the variation of the overlap region $A_1 \cap A_2$ caused by the changes of q_2. Notice that all the calls falling into this overlap region belong to the 1st CoS and changing q_2 does not alter the call arrival rate for the 1st CoS. This explains why B_1 does not vary much as q_2 changes. Neither does B_2 because the call migration effect from the 1st CoS to the 2nd CoS is small due to small blocking probability B_1 (as p_1 increases). Consequently V does not vary much as q_2 changes. Second, one observes that V increases with p_1 up to a certain point and then drops. This is because increasing p_1 improves per call revenue for the 1st CoS but at the same time, the call arrival rate for the 1st CoS decreases. Hence there is a p_1 value where V is maximized. As p_1 increases, B_2 also increases but then drops after a certain point. This can be explained as follows: The call arrival rate for the 2nd CoS increases as p_1 increases and consequently B_2 increases initially. However, after a certain p_1 value, the 1st CoS becomes highly underloaded and the 2nd CoS starts to benefit from bandwidth sharing and take over

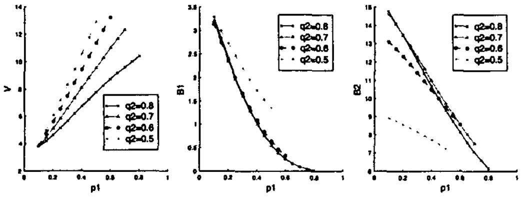

Fig. 4. Performance for extremely price sensitive population

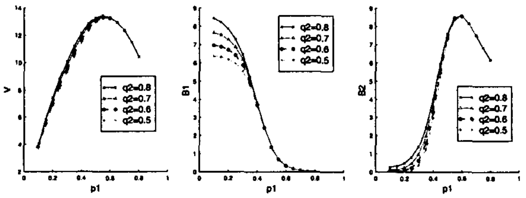

Fig. 5. Performance for extremely quality sensitive population

the available bandwidth (except the 20% reserved bandwidth) from the 1st CoS. Consequently B_2 drops. Again, the trunk reservation mechanism plays a key role here. In summary, for the extremely quality sensitive population, the performance measures are insensitive to the selection of q_2 but sensitive to the selection of p_1. There is a unique p_1 where V is maximized.

Finally, we examine the user population in which users are equally sensitive to the price and quality $(r_1 = 0, r_2 = 1, r_3 = 0)$. The numerical results are presented in Fig. 6. Interesting enough, all the performance measures have sudden changes at some p_1 value. To understand this phenomenon, we note that $U_i = q_i - p_i$ $(i = 1, 2)$, when p_1 falls into $A_1 \cap A_2$. Since $U_1 > U_2$ when p_1 is small, all the calls in $A_1 \cap A_2$ belong to the 1st CoS. As p_1 increases and reaches a critical point, the inequality is reversed and all the calls in $A_1 \cap A_2$ shift to the 2nd CoS, resulting in sudden value changes for the performance measures. Although this phenomenon is a consequence of our using a simplified user behavior model, it reveals that the network may experience severe performance degradation when prices are improperly adjusted. In practice, the operational point should stay away from this critical point. The performance behavior as q_2 changes is similar to the price sensitive case.

We make a few general observations based the above analyses. First, it suffices to set trunk reservation values at 10-20% of the nominal bandwidth to achieve a good service protection. Second, call migration effect undermines the effectiveness of trunk reservation as a service protection measure and should be prevented. The key is to design pricing-quality structure in such a way that the overlap regions $A_i \cap A_j$ $(i, j = 1, 2, 3,\ i \neq j)$ are kept small. Finally, user sensitivities have great impact on service price-quality design. Care must be taken to avoid the possible sudden network performance degradation at some operating points in price-quality parameter space.

V. Conclusions and Future Work

In this paper, an analytical model was developed for the study of the impact of user behavior on the effectiveness of pricing, service quality declaration, and link bandwidth allocation/sharing policies in a multiple-CoS Internet. In particular, a single link was considered. A static usage-charge based pricing policy and a bandwidth allocation/sharing policy with trunk reservation were examined. With the aid of a set of price-quality phase diagrams, we were able to capture the impact of distinct user behaviors on the performance of various design aspects.

In this paper, we restricted ourselves to the analysis of a single link case. An extension of the paper to a multiple-node case was presented in [13]. However, in that report, only fixed routing was studied and most of the observations made in the present paper was found to be still valid. An interesting research direction is to extend the present work to a multiple-node network with CoS-based routing. Among other issues, the performance of trunk reservation in the context of protecting shortest path connections is worth studying. Note that with the bandwidth allocation/sharing policy proposed in this paper, CoS-based routing algorithm design can take advantage of the service segregation and abstraction provided by the allocated nominal link bandwidths for individual CoSs.

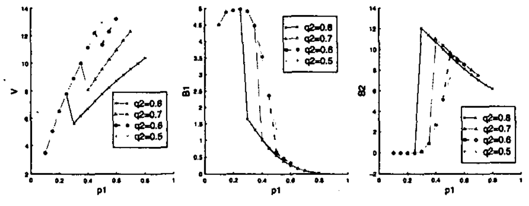

Fig. 6. Performance for equally sensitive population

References

[1] D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Services," RFC2475, Dec., 1998.

[2] Y. Bernet, J. Binder, S. Blake, M. Carlson, S. Keshav, E. Davies, B. Ohlman, D. Verma, Z. Wang, and W. Weiss, "A Framework for Differentiated Services," Internet Draft <draft-ietf-diffserv-framework-01.txt>, Oct., 1998.

[3] T. Li and Y. Rekhter, "Provider Architecture for Differentiated Services and Traffic Engineering," Internet Draft <draft-li-paste-01.txt>, Sept. 1998.

[4] R. J. Gibbens, F. P. Kelly, and P. B. Key, "Dynamic Alternative Routing," in *Routing in Communications Networks*, edited by M. Steenstrup, Prentice Hall, 1995.

[5] E. W. Fulp, M. Ott, D. Reininger, and D. S. Reeves, "Paying for QoS: An Optimal Distributed Algorithm for Pricing Network Resources," *Proceedings of IEEE INFOCOM'98*, p. 75, 1998.

[6] J. Sairamesh, D. F. Ferguson, and Y. Yemini, "An approach to pricing, optimal allocation and quality of service provisioning in high-speed packet networks," *Proceedings of IEEE INFOCOM '95*, Vol. 3, p. 1111, 1995.

[7] R. Cocchi, S. Shenker, D. Estrin, and L. Zhang, "Pricing in computer networks: motivation, formulation, and example," *IEEE/ACM Transactions on Networking*, Vol. 16, p. 614, Dec. 1993.

[8] J. K. KacKie-Mason and H. R. Varian, "Pricing Congestible Network Resources," *IEEE J. Selected Areas in Communications*, Vol. 13, No. 7, Sept. 1995.

[9] L. P. Breker and C. L. Williamson, "A Simulation Study of Usage-Based Pricing Strategies for Packet-Switched Networks," *IEEE INFOCOM'96*, 278, 1996.

[10] J. M. Peha, "Dynamic pricing as congestion control in ATM networks," *IEEE GLOBECOM'97*, Vol. 3, 1367, 1997.

[11] G. R. Ash, "Routing Guidelines for Efficient Routing Methods," Internet Draft <draft-ash-itu-sg2-routing-guidelines-00.txt>, Oct. 1999.

[12] E. Altman, I. France, A. Orda, and N. Shimkin, "Bandwidth Allocation for Guaranteed versus Best Effort Service Categories," *Proceedings of INFOCOM'98*, March, 1998.

[13] N. Shameer, and H. Che, "A Simulation Study of Trunk Reservation in A Multiple Class of Service Network," Technical Report No. 022500.

[14] K. Nichols, V. Jacobson, and L. Zhang, "A Two-Bit Differentiated Services Architecture for the Internet," Internet Draft <draft-nichols-diff-svc-arch-02.txt>, 1998.

[15] J. Sairamesh and J. Kephart, "Price Dynamics of Vertically Differentiated Information Markets," To appear in *Proceedings of First International Conference on Information and Computational Economics*, Charleston, S.C., Oct. 1998.

[16] S. C. Borst and D. Mitra, "Virtual Partitioning for Robust Resource Sharing: Computational Techniques for Heterogeneous Traffic," *IEEE J. Select. Areas Commun.*, vol. 16, No. 5, p. 668, June, 1998.

[17] K. W. Ross, "Multiservice Loss Models for Broadband Telecommunication Networks," Springer-Verlag London Limited, 1995.

[18] D. Bertsekas and R. Gallager, "Data Networks," Prentice-Hall, 1987.