Resource Management for IP Telephony Networks

Matthew Chapman Caesar[1], Dipak Ghosal[2], and Randy H. Katz[1]

[1]Department of Computer Science, University of California at Berkeley, Berkeley, CA 94270
E-mail: {mccaesar,randy}@cs.berkeley.edu
[2]Department of Computer Science, University of California at Davis, Davis, CA 95616
E-mail: ghosal@cs.ucdavis.edu

***Abstract*—**

The two key resources in an IP Telephony network are the Internet Telephony Gateways (ITGs) and the IP network. These resources must be effectively managed to simultaneously provide good QoS to calls and maximize network resource utilization. This paper presents two main contributions. First, we design a call admission policy based on congestion sensitive pricing. As the load increases, this policy preferentially admits users who place a higher value on making a call while simultaneously maintaining a high utilization of network resources. We derive the function mapping congestion to price for the admission policy that maximizes revenue. Second, we design a call redirection policy to select the best ITG to serve the call. The policy balances load to improve network efficiency and incorporates QoS sensitivity to improve call quality. Simulation results show the following: (i) Congestion pricing based admission control lowers call blocking probability, increases provider revenue, and improves economic efficiency over a static flat-rate admission control scheme. (ii) Congestion sensitivity in the redirection policy balances load across all the ITGs while QoS sensitivity improves call audio quality. (iii) Incorporating price sensitivity in the redirection policy improves the economic efficiency, i.e., ensures that users who pay more get higher QoS. The techniques studied in this paper can be combined into a single resource management solution that can improve network resource utilization, provide differentiated service, and maximize provider revenue.

***Index Terms*—**

IP Telephony, Call Admission Control, Congestion sensitive pricing, QoS sensitive routing, Blocking probability, Economic efficiency, Revenue.

I. Introduction

While Internet Telephony (IP Telephony) encompasses many different architectures and services, the key idea is the transport of real-time voice traffic over the Internet. The IP Telephony architecture [1] allows the entire end-to-end path to be routed over the Internet. In this case, the endpoints are regular personal computers (PCs) that are equipped with IP Telephony software. The IP Telephony architecture also allows one or both the endpoints to be connected to the PSTN (Public Switched Telephone Network). For these cases, a portion of the end-to-end path is routed over the Internet. This requires interoperability between the Internet and the PSTN which is achieved using gateways that act as application level interfaces between the two networks. These gateways are referred to as Internet Telephony Gateways (ITGs) [2].

This research was supported through NSF grants NCR-9703275 and ANI-9741668. Matthew Chapman Caesar is supported by a National Defense Science and Engineering Graduate (NDSEG) Fellowship.

In the PSTN, voice traffic is carried over dedicated circuits established using the Signaling System Number 7 (SS7) protocol [3]. As a result, the only delay suffered by the voice traffic in the PSTN network is the propagation delay which is fixed once the circuit has been established. On the other hand, the Internet is still inherently a best-effort network and provides no end-to-end bandwidth guarantees. Thus, transporting packetized voice over the Internet can result not only in variable delays but also losses which can cause poor audio quality at the receiver.

One important feature of IP Telephony is that it provides a rich signaling architecture that can extend up to the endpoints when they are PCs enabled with IP telephony software. This not only enables new services [4] but also allows service providers to provide differentiated services by offering tiered service levels each with a different guarantee on QoS [5]. A differentiated service architecture would allow service providers to implement flexible and dynamic pricing policies that can maximize provider efficiency (revenue) and network efficiency (utilization of network resources).

The key resources in an IP Telephony network are the IP network and Internet Telephony Gateways (ITGs). Congestion in the IP network increases delays and loss of audio packets which degrades the quality of the received audio. As a result, paths to different ITGs can have different QoS. ITG resources include the protocol processing capacity and the number of voice ports. Congestion at the ITG can add to overall audio packet delay and increase call blocking due to unavailability of voice ports. In order to implement the differentiated architecture the key issues are (1) a call admission algorithm, (2) a load balancing algorithm, and (3) QoS sensitive call redirection.

In this paper, we investigate the above issues as part of an integrated resource management scheme for IP telephony networks. First, we study a congestion sensitive pricing based call admission scheme in which the price charged for a call increases with congestion at the gateway. This creates an incentive for users to place calls at a later time. We give a price-congestion function for the admission control policy that maximizes provider revenue. This scheme is compared with a Flat-rate based Admission Control (FAC) scheme and a scheme with No Admission Control (NAC) as baseline cases. Next, we design a call redirection policy to select the ITG best suited to serve the call based on the number of voice ports in use at the ITG and the real-time measurement of path quality. This policy (CQR) balances load to improve network efficiency, and incorporates QoS sensitivity to improve call quality. The scheme can be tuned to become more sensitive to gateway load, which decreases call blocking probability, or more sensitive to QoS,

which improves call audio. We improve this scheme by using price to tradeoff between these two factors in order to improve economic efficiency. This scheme (PCQR) is compared with a Random Redirection (RR) scheme, in which the ITG is selected at random.

The evaluation of the call admission and redirection schemes is done under a realistic workload based on standard telephony models. This evaluation is performed with respect to three measures, namely, network efficiency, provider efficiency, and economic efficiency. Network efficiency corresponds to the blocking probability and the utilization of network resources. Economic efficiency reflects the correlation between the price paid by users and the QoS received, and provider efficiency corresponds to the total revenue generated by each of the schemes. Simulation results show the following: (i) Congestion pricing based admission control lowers call blocking probability, increases provider revenue, and improves economic efficiency over a static flat-rate admission control scheme. (ii) Congestion sensitivity in the redirection policy improves the capacity by balancing load, while QoS sensitivity can greatly improve call audio quality. (iii) Price sensitivity in the redirection policy can be used to improve the economic efficiency. Finally, we also investigate the effect of network topology, levels of background traffic, and the relative percentage of RTP and background traffic on the results.

The rest of the paper is organized as follows. In Section II we discuss the IP Telephony architecture, in particular, the various network entities and the protocols that are used in IP Telephony. In Section III, we introduce call admission control in the form of congestion sensitive pricing and conduct a simple queuing analysis to investigate the effect of different methods of congestion sensitive admission control. In Section IV we present a generalized redirection model and our call redirection technique. In Section V, we discuss the experimental setup, the parameters, and the various performance measures that have been used in the comparative analysis. The results are discussed in Section VI. Finally, in Section VII, we conclude with a summary of the results and some future research directions.

II. IP Telephony Architecture

Figure 1 shows the components of an IP Telephony architecture and the manner in which it inter-operates with the PSTN system. A key entity in the architecture is the IP Telephony Gateway (ITG) which provides interoperability between PC-based IP Telephony users and PSTN endpoints. An ITG operates at the application level, with connectivity to the PSTN on one side and the Internet on the other [2]. To connect to a PSTN endpoint, an IP host first connects to an ITG, which terminates the IP portion of the call and initiates a PSTN call to the PSTN endpoint by allocating one of its many circuits (voice ports) to the call. To perform the function as an application level proxy each ITG is capable of initiating and terminating IP Telephony signaling protocols, such as H.323 [6] and/or the Session Initiation Protocol (SIP) [7], and also the Signaling System 7 (SS7) protocol [3].

With large numbers of ITGs, discovery and selection of the suitable ITG to service a call is an important problem. In this study, we assume the TRIP (Telephony Routing over IP) architecture [8] for the gateway location, gateway discovery, and gateway routing problems. The key entities defined in TRIP include the Administrative Domain (AD), the ITG, and the Location Server (LS). The Internet is viewed as a collection of ADs that are connected by multiple backbone networks. Each AD contains one or more ITGs, one or more LSs, and users, which are customers that initiate calls.

Given an ITG and a list of ITG attributes, the LS [1] finds the best ITG for a particular call. Towards this end, each LS maintains a database containing information about all the other ITGs. This database is built using advertisements that are exchanged by the LSs. Each advertisement contains multiple attributes, including (i) phone numbers serviced by the ITG, (ii) the IP address of the ITG, (iii) the AD identification number of the ITG, and (iv) the number of available voice ports at the ITG. Additional information like service features, protocols, and codecs supported by the ITG may also be part of the advertisement. In addition, the LS acquires and records statistics regarding the average call quality achieved from calls originating in its AD to the ITGs in the system.

A typical call setup sequence with reference to Figure 1 consists of the following steps:

1) When a user initiates a call, a client on behalf of the user sends a request to the LS providing the destination phone number and other parameters including the QoS the user requires and the price the user is willing to pay.
2) The LS does a local database lookup to determine which ITG will service the call. This operation can have one of three outcomes: (i) The LS finds that the price requested by the user is less than the admission price. In this case, the call is denied and appropriate information is passed back to the client. (ii) The price offered by the user is higher than the admission price and the LS finds an ITG which can service the call, in which case it returns the IP address of the selected ITG to the client. (iii) The price offered by the user is higher than the admission price, but none of the ITGs have available resources to service the call. In this case also the call is denied and the client is informed of the non-availability of resources.
3) If the previous step succeeds, the client initiates a SIP transaction to begin a session with the selected ITG. The ITG and client then perform initial session agreements and setup an RTP session to transfer audio data [9]. On the PSTN side the ITG uses the SS7 protocol to setup a circuit to the destination.

In IP Telephony, calls may originate and terminate in either the PSTN or IP networks. Furthermore, part or all of the call path may take place over the IP infrastructure. Any routing algorithm used to determine the path must take into account the desired QoS and the cost of the call to both the network provider and the user. We present a call routing architecture with high network efficiency (i.e., low blocking probabilities and high network utilization) and high economic efficiency (i.e., it allocates higher QoS paths to users who place a greater value on good audio quality).

III. Congestion Sensitive Pricing-Based Call Admission Control

Congestion sensitive pricing is implemented in the PSTN in the form of time-of-day pricing. However, unlike in the PSTN, congestion sensitive pricing in IP Telephony can be implemented much more dynamically particularly for net-to-net

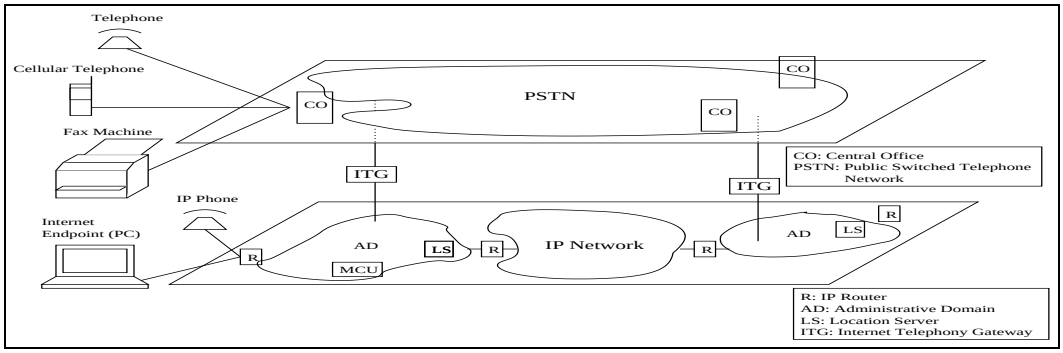

Fig. 1. IP Telephony architecture.

and net-to-phone calls where the network provider can exploit the high quality user interface to indicate the cost of making a call.

In this study, we assume a cooperative environment, in which a single resource management scheme is implemented over all ITGs in the network [10]. This would correspond to an IP Telephony service provided by an ISP with Points-of-Presence (PoPs) in a number of different geographical areas. We assume that the total cost of the call is based on the duration of the call and a per unit time charge which remains the same throughout the call[1].

In our congestion sensitive pricing-based call admission scheme, referred to as CAC, the admission price depends on the total number of voice ports in use at each of the ITGs in the system. Each ITG occasionally sends its current load advertisements to its LS, which is responsible for propagating this information to the other LSs in the system. Each LS calculates the total number of voice ports in use over all the gateways in the system from these advertisements, and uses this value to determine the admission price according to the price-congestion function discussed below. The LS offers a lower price to callers when there are many voice ports free, i.e., when the system is under-utilized, and increases the admission price as the number of ports in use increases. For simplicity, we assume the user offers a bid during call setup uniformly distributed between 5 and 15 cents per minute, or between 15 and 45 cents for an average 3-minute call.

There are two important aspects of this scheme. The first is the manner in which the congestion is measured. In this study, we measure the congestion using an exponential averaging scheme [6]. The average number of voice ports in use at each ITG is recalculated every time a connection is made to or leaves from the ITG. The second important aspect of this scheme is the function that maps the congestion to the price that the system offers, referred to as the price-congestion function[12].

A. Analytical Model

We develop a queuing model to determine the price-congestion function that maximizes revenue generated by the provider[2]. The network consists of n voice ports and is modeled as a n server loss system. Arrival of call requests follow a Poisson process with rate λ. Without loss of generality, we assume the user is unaware of the current system price and that each call request comes with a price p that the user is willing to pay which is uniformly distributed between $p_{min}/3$ and $p_{max}/3$ cents per minute, or between p_{min} and p_{max} for a three minute call.

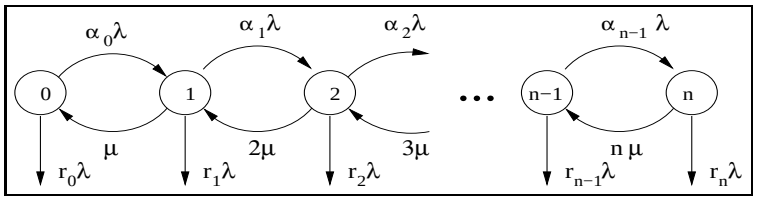

Fig. 2. State transition diagram of the network under CAC.

In CAC, as more voice ports are utilized, fewer users' bids will be greater than the price charged by the system, causing a larger number of calls to be blocked. This corresponds a queuing system with discouraged arrivals. The probability a user will be discouraged and not accepted into the system is state dependent and depends on the price-congestion function. Let r_k be the probability that in state k an arriving call is rejected by the system because the user's bid is lower than the current admission price for that state, denoted by A_k. Hence, $\alpha_k = 1 - r_k$ is the probability that a call is accepted into the system and the effective arrival rate in state k is $\alpha_k \cdot \lambda$.

Using standard techniques [13], it can be shown that p_k, the probability that the system is in state k, is given by

$$p_k = p_0 \cdot \prod_{i=0}^{k-1} \frac{(1 - r_k) \cdot \lambda}{(i+1) \cdot \mu}, \tag{1}$$

where

$$p_0 = \frac{1}{1 + \sum_{k=1}^{n-1} \prod_{i=0}^{k-1} \frac{(1-r_i) \cdot \lambda}{(i+1) \cdot \mu}}. \tag{2}$$

The total revenue is equal to the probability of the system being in a particular state, multiplied by the revenue earned in that state and the duration of time under which the system was operating. Therefore, the total normalized revenue R earned by the system over some duration δ is given by:

[1] When the call is made from a PC, the caller may be provided real-time information on the cumulative cost based on which the caller may limit the duration of the call [11]. In this study, we do not consider such a scenario.

[2] Determining the price-congestion function with the best performance in a real-world IP Telephony system is a very difficult problem due to the complexities in user modeling. However, this analysis allows us to more rigorously compare CAC with the other admission control schemes.

$$R = \delta \cdot \sum_{k=0}^{n-1} p_k \cdot r_k \cdot \alpha_k \cdot \lambda. \quad (3)$$

Note that R is not equal to the actual revenue, as it is calculated based on r_k. The actual revenue may be calculated by substituting A_k for r_k, where A_k is the admission price in state k. Since the user's bid is uniformly distributed between p_{max} and p_{min}, it is easily shown that A_k is equal to $r_k \cdot (p_{max} - p_{min}) + p_{min}$.

We consider the following three types of price-congestion functions:

1) Linear: Here A_k is a linear function of k, i.e., $A_k = m \cdot k + b$. b was fixed at 0, and m was varied to determine the linear function that generates the maximum revenue.
2) Stepwise Linear: We consider a stepwise linear function of the form $A_k = h(\lfloor \frac{k}{c} \rfloor)$, where h is a function relating k to the price and c is the horizontal length of the step. In our study, we choose $c = 5$ and vary h to determine the stepwise function that maximizes revenue.
3) Exponential: Here we consider an exponential price-congestion function of the form $r_k = c \cdot e^{m \cdot k} + b$. Both m and b are varied to determine the function that maximizes the revenue. c was chosen to make the y intercept close to zero.

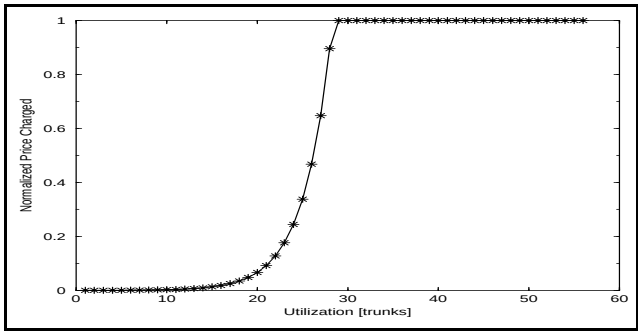

Fig. 3. Exponential price-congestion function that maximizes revenue.

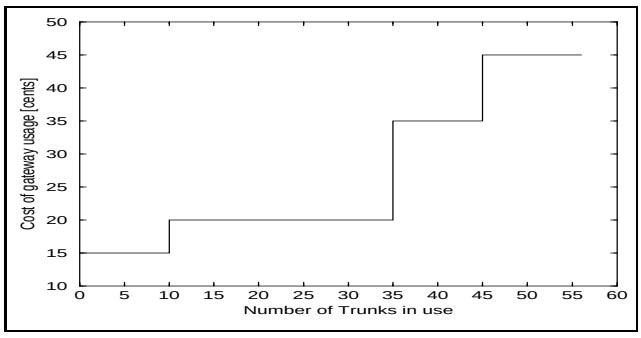

Fig. 4. Price-congestion function used in this study.

From the analysis we observe that when the offered load and the number of voice ports are varied, the exponential function outperforms the other two functions. We also note that that the stepwise linear function performs almost as well as the exponential function. Finally, each of the best performing functions begins charging the maximum system price well before the system reaches peak utilization. This shows the importance of maintaining excess capacity to service high paying callers during times of heavy load in expectation of later bursts of traffic. The exponential function that yielded the maximum revenue is shown in Figure 3. Our CAC implementation utilizes the discretized exponential price-congestion function shown in Figure 4. This approximation is necessary to limit fluctuations in the admission price.

B. Baseline Schemes

We consider two admission control schemes as baseline cases. First, we consider a Flat rate based Admission Control (FAC) scheme. In this scheme the admission price is a fixed per unit time charge. The cost per minute does not vary and is set at 10 cents per minute. Since the average call holding time is assumed to be a negative exponential distribution with a mean of 3 minutes, the average cost of each call is 30 cents. In this scheme, the advertisements are sent out only when the ITG enters or leaves a state in which it is fully utilized. This is opposed to sending out advertisements whenever the number of voice ports in use changes. Second, we consider an alternate baseline case in which all calls are admitted, referred to as No Admission Control (NAC). Under this scheme, the average call price is set to 15 cents per minute.

IV. Redirection Schemes

The LS maintains a table of the average path quality to, and the current number of voice ports in use at each gateway in the system. Admitted calls are assigned to different gateways by selecting an entry from this table. There are several schemes which the LS may use to perform this assignment. In this section, we first introduce two existing simple redirection schemes. We compare these schemes to a simple Random Redirection (RR) technique that selects the servicing ITG at random. We then present a generalized redirection model based on the best features of each of the schemes.

A. Congestion Sensitive Redirection (CR)

Congestion sensitive redirection has been previously used to balance load on a connection level granularity across a set of replica servers [31] [14] [15]. We adapt these techniques for use in our system. We consider a Congestion Sensitive Redirection (CR) scheme in which calls are redirected to the least loaded ITG in the system. Unlike RR, CR can dampen oscillations and can more effectively balance across heterogeneous gateways. However, it relies on propagation of load information which may be out of date and is hence less resilient to sudden overloads.

Each ITG advertises its congestion, measured as the number of voice ports in use, to its local LS. The LS then propagates this information to the other LSs. A LS may have out-of-date state information for a particular gateway due to delayed or dropped advertisements. To reduce control traffic, an ITG advertises its load to its local LS whenever the congestion rises above or falls below a threshold. In this study, we chose to place thresholds at multiples of 5 voice ports. Experimental results showed performance had little sensitivity to the distribution or number of these thresholds.

B. QoS Sensitive Redirection (QR)

In the current best-effort Internet, audio packets can be dropped or delayed depending on the path quality between the source and the destination. This will affect the quality of the received audio. Other factors, such as jitter, protocol errors, and codec delays also affect the received audio quality [16] [17]. We consider a QoS Sensitive Redirection (QR) scheme in which calls are redirected to the ITG which has recently provided the best QoS to callers, in an attempt to maximize call QoS. For simplicity, we assume that the majority of QoS degradation takes place between Administrative Domains (ADs), and not in the network path between the client and its AD. Although the access network sometimes accounts for significant quality degradation in today's Internet due to underprovisioning, no redirection scheme can mitigate this loss in quality.

One important aspect of this scheme is the mechanism used to measure the quality of the received audio as a function of the losses and delays of audio packets in the network [18] [19] [20]. In this study, we use the Mean Opinion Score (MOS) for audio quality [19]. This measure is based on scores given by participants on the quality of the received audio as the type of codec and packet loss rates are varied. We calculate the received audio quality as a function of the number of lost RTP packets based on the study done in [19].

The statistics of packet loss are obtained through the Sender Reports and Receiver Reports in the RTCP protocol, which is part of the RTP session [9]. Each ITG maintains a window of these reports to all other ITGs in the network. These path qualities are then used to estimate the MOS.

C. Congestion and QoS Sensitive Redirection (CQR)

We define a generalized redirection scheme, called Congestion and QoS Sensitive Redirection (CQR), that combines the best features of each of these schemes. To the best of our knowledge, this work is the first to propose such a scheme. The key idea behind CQR is to choose a gateway based on both the number of voice ports in use and the expected quality of the received audio, in a manner that is resilient to sudden overload conditions. We define the Redirection Metric (Rdm) for gateway i as follows:

$$Rdm_i \ = \ \beta * M_i + (1 - \beta) * Q_i, \qquad (4)$$

Where M_i is the latest estimate of the number of voice ports in use at gateway i, and Q_i is the latest estimate of the audio quality at gateway i. Both Q_i and M_i are normalized to a value between 0 and 1 by dividing their measured values by their maximum possible values. The LS calculates the Rdm for each ITG in the system, sorts the ITGs in order by Rdm, and chooses randomly from the k ITGs with lowest Rdm. Note that when k is equal to the number of gateways in the system, CQR becomes RR. When $\beta = 1$, CQR becomes CR, and when $\beta = 0$, CQR becomes QR. We can hence use the variable k to trade off between load balance and protection against sudden overloads, and the variable β to trade off between call quality and load balance. We also consider a version of CQR in which beta is inversely proportional to the user's bid, called Price-based CQR (PCQR). In particular, we set $\beta = (45 - B)/30$, where B is the user's bid price. The aim is that users with high bids should be allocated higher quality paths, and users with low bids should be redirected so as to balance the load.

V. Simulation Model

The experimental setup used in our study adopts the architecture proposed in TRIP [8]. Our simulation architecture, shown in Figure 1, is implemented as a modified version of the ns-2 simulator [21] [22]. Over the base functionality implemented in ns-2, we implemented an ITG module, an LS module, and a user module based on the description given in Section II.

A. User and Server Models

This experimental study is based on a user model in which users expect to get the best quality of service that is available for some maximum price that they are willing to pay. During connection setup, the user offers a particular price, which the LS uses to perform call admission. If the call is admitted, the LS then redirects the call to an ITG that has available voice ports.

B. Network & System Parameters

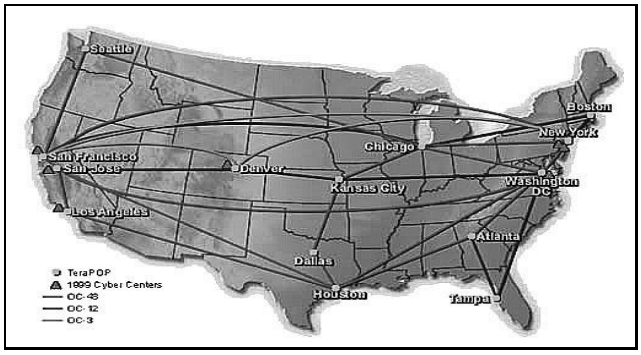

Fig. 5. The Qwest National IP backbone from 1999.

The results presented below are based on the following network and system parameters.

1) Simulation analysis is carried out on two network topologies: (1) The Qwest United States IP backbone from 1999 as shown in Figure 5 with an AD placed at each of the 14 Points-of-Presence (PoPs) [23], and (2) the Excite@Home OC-48 IP backbone with an AD at each of the 29 @Work Super Nodes [24]. The topology characteristics differ in several respects, for example, the average distance between ITGs in the Qwest network is lower. Unless otherwise stated, results are acquired from the Qwest network topology.
2) The call arrival in each AD follows a Poisson process with rate λ calls per second. This parameter is varied to study the sensitivity of the results to the offered load. Call holding time follows a negative exponential distribution with mean 180 seconds. These parameters are based on standard telephony traffic models.
3) Each ITG has 56 voice ports.
4) The background traffic in each link follows a Pareto distribution with a shape parameter of 1.2 and with mean link utilization uniformly distributed from 30% to 60% [25]. Sensitivity to the level of background traffic is considered.
5) The voice traffic is modeled by an exponentially distributed "on-off" Markov modulated process utilizing a

G.723 codec [6]. RTP traffic is tuned to constitute 1% of the total traffic [26]. Sensitivity to this percentage is also studied. Finally, we considered a playback buffer of 300 ms, and a packet is considered lost if it is not received within 300 ms of being sent. The International Telecommunications Union (ITU) defines this value to be the required maximum delay time for voice communication.

6) The ITGs are replicated and calls can be serviced by any ITG. Furthermore, all calls cost the same, i.e., unlike in the PSTN, there is no difference between local calls and toll calls.
7) α is the filter gain used in the exponential averaging scheme to computer the mean number of voice ports at an ITG. In this study we set $\alpha = 0.5$.

C. Performance Metrics

To compare the various management strategies we use the following three performance measures.

1) Provider Efficiency: This is the total revenue generated, and is measured as the total call revenue averaged over all ITGs.
2) Economic Efficiency: A management strategy is economically efficient if it ensures that callers who place a higher value on services are allocated resources before callers that place a lower value on those services. Furthermore, callers with higher bids should be allocated higher quality paths over users with lower bids. A strong correlation between price paid and QoS achieved shows that users who are willing to pay a high price tend to be allocated paths with good service quality over users who offer a lower bid, which causes user benefit to be maximized in our user model [27]. We use simple linear regression techniques to find the correlation coefficients between the price charged to the user and the blocking probability and received audio quality.
3) Network Efficiency: This is the call blocking probability. A call can be blocked due to two reasons: (1) It could be blocked at the LS. This could happen if the LS cannot find an ITG that can service the call. Note that the call could be blocked either because all ITGs are busy, i.e., no free voice ports at any of the ITGs (NFP), or because there is no ITG that is willing to service the call at the price offered by the user (OPR). (2) It could be blocked at the ITG (GRC). This could happen if the LS has outdated/incorrect load information about the ITG. Thus, the LS initiates a call setup to an ITG which actually has no voice ports available. GRC and NFP blocks are less desirable than OPR blocks, as they are not caused by price based admission control and might cause a high paying user to be denied service.
   We additionally measure the average service distance and average audio quality achieved by callers in each of the schemes to investigate how well resources are utilized. The service distance is defined to be the number of intermediate network hops between the user and the ITG used to service the call.
   a) Average Service Distance: This is the average distance in terms of the number of hops between the user and the ITG that services the call.
   b) Received Audio Quality: This is the average audio quality in terms of MOS measured at the receiver as discussed in Section IV-B.

VI. Results and Discussions

We collect performance metrics in a manner that ensures the system reaches a steady state. While experiments are conducted for 90 minutes, the results presented in this paper are based on system state sampled during the last 60 minutes of each experiment. This is done to eliminate the effects of cold start [28].

A. Admission Control

An admission control scheme has three desirable properties. It should maximize network efficiency by allowing as many calls as possible into the system. It should maximize economic efficiency, i.e., as resources become limited it should preferentially admit users who place a higher value on making a call. It should maximize provider efficiency by generating as much revenue as possible. In this section we compare Flat Admission Control (FAC) and Congestion Sensitive Admission Control (CAC). We fix the redirection scheme as QoS Sensitive Redirection (QR). We refer to the resulting schemes as QR+FAC and QR+CAC, respectively.

1) Network Efficiency: Unlike in the PSTN, we can use the rich signaling network to implement complex pricing policies. To achieve good network efficiency and mitigate overload conditions, we use congestion sensitive pricing to perform call admission.

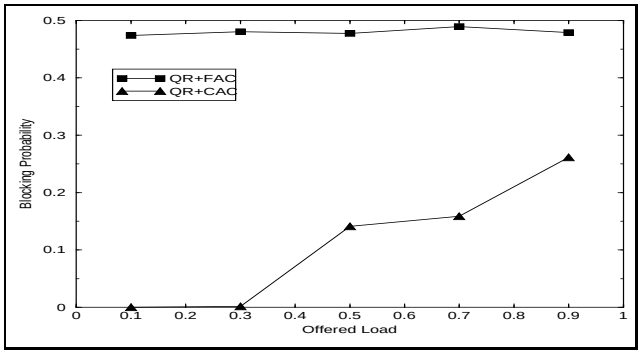

Fig. 6. Average blocking probability as a function of offered load.

Figure 6 compares the call blocking probability as a function of the offered load for the two management schemes. Table I tabulates the various components of the blocking probability.

In QR+FAC, the blocking probability of a call is about 50%, regardless of the offered load. Since the price of a call is fixed at 30 cents and the user's bid is uniformly distributed between 15 and 45, half of the calls are blocked in this scheme. The effective voice port utilization is low and there are always free ports available for users who offer higher bids. Consequently, as seen in Table I, all calls blocked in this scheme are blocked due to overprice, i.e., OPR blocks.

In QR+CAC, the price changes dynamically with load. At low loads, the average price is low, allowing more calls to be serviced than in QR+FAC. With higher offered load, the number of blocked calls increases. However, as can be seen from Figure 6, the blocking probability for QR+CAC never exceeds

TABLE I

COMPARISON OF DIFFERENT TYPES OF BLOCKING PROBABILITIES (P(OPR/NFP/GRC) → BLOCKING PROBABILITY DUE TO OPR/NFP/GRC).

Management	Load = 0.3		Load = 0.5		Load = 0.7		Load = 0.9	
Strategy	P(OPR)	P(NFP)+ P(GRC)	P(OPR)	P(NFP)+ P(GRC)	P(OPR)	P(NFP)+ P(GRC)	P(OPR)	P(NFP)+ P(GRC)
QR+FAC	0.480	0	0.477	0	0.489	0	0.479	0
CQR+CAC	0.002	0	0.144	0	0.154	0	0.261	0

that of QR+FAC. This happens because the price-congestion function of QR+CAC admits a large number of calls, causing a higher system utilization. Although the system has high utilization, the price-congestion function effectively prevents the system from overload, and hence no calls are blocked due to NFP (no free voice ports) or GRC (gateway rejected call setup). At low loads, the price-congestion function generates a low admission price, allowing almost all calls to be admitted.

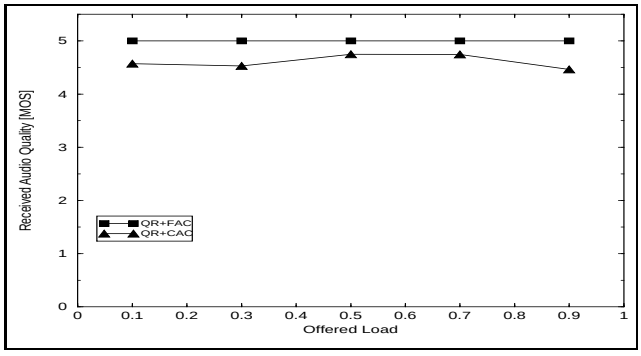

Fig. 7. Average received audio quality of calls as a function of load.

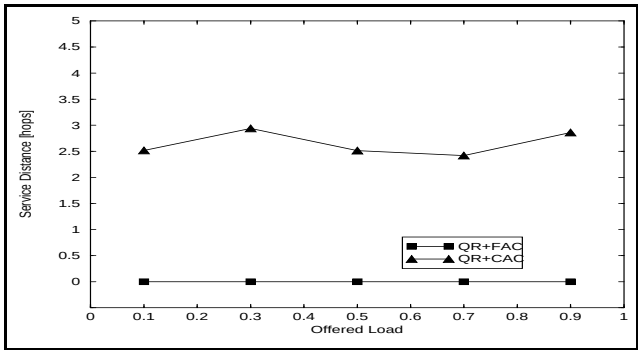

Fig. 8. Average service distance of calls as a function of load.

2) Economic Efficiency: At high loads, congestion arises in terms of high demand for voice ports at the ITGs, which increases the call blocking rate. Our system preferentially allocates voice ports to users with high bids over users with low bids by increasing the call admission price during times of overload. We consider two metrics to compare the economic efficiency of the different management schemes: the relationship between admission probability and price paid, and the average audio quality achieved.

- *Average QoS Achieved:* Figures 7 and 8 show the relationship between the average received audio quality and service distance as a function of the load for each of the resource management schemes with price based admission control. We observe that QR+FAC provides excellent quality for all calls, since all calls are terminated locally due to the large number of OPR blocks under this scheme as shown in Figure 8. QR+CAC also provides an excellent audio quality, as low paying calls tend to be blocked and high paying calls can be satisfied locally. The received audio quality does not vary with load, as our price-congestion function does a good job of blocking a larger number of low paying calls when load increases. This keeps the system from becoming overutilized and hence forces calls to use paths with poor quality. As load increases, the greater the demand for high QoS paths, and hence the smaller the percentage of calls that can be handled locally. However, a large percentage of calls that would have been allocated paths with poor quality are blocked, keeping audio quality constant across load. QR+CAC provides lower audio quality than QR+FAC due to the higher system utilization. All the high QoS paths become allocated forcing calls to travel farther from the local ITG.
- *Relationship between Price and Admission Probability:* We measure the relationship between the user's bid price and the probability that the user is admitted into the system by calculating the correlation coefficients of the two variables. In QR+FAC, all calls with a bid price greater than 30 cents are admitted, and all other calls are blocked by the system. This results in a strong correlation of 0.866 regardless of offered load. However, this correlation is achieved by unnecessarily blocking a large number of calls. We find that QR+CAC also had a very strong relationship between the two variables. However, the correlation coefficients for this scheme are slightly less than than that of QR+FAC under all loads due to the fact that the admission price fluctuates with load. A sudden burst of call traffic causes the admission price to raise, blocking high paying users. If a statistical fluctuation suddenly decreases the amount of call traffic, the admission price will drop, allowing more low paying calls to be admitted into the system. In this manner, callers with a low bid now have a method to "sneak in" and acquire a set of voice ports at a low price. More detailed information is available in our technical report [28].

3) Provider Efficiency: We achieve provider efficiency by making the admission price congestion sensitive. This allows us to run the system at high utilization and allocate voice ports to users who are willing to pay more. In Figure 9 we compare the total revenue generated from each scheme. For QR+FAC, the revenue increases linearly with load. This is not surprising to find since in this model prices do not fluctuate with load.

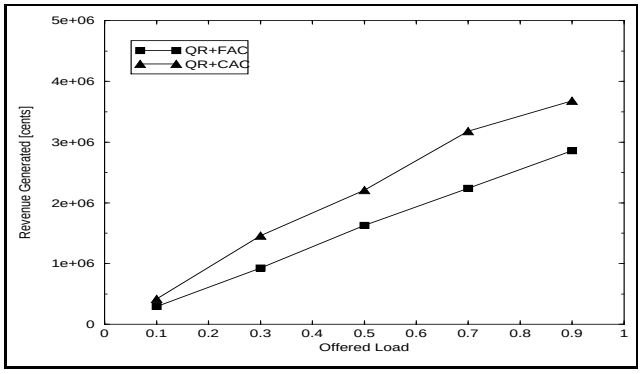

Fig. 9. Total revenue as a function of load.

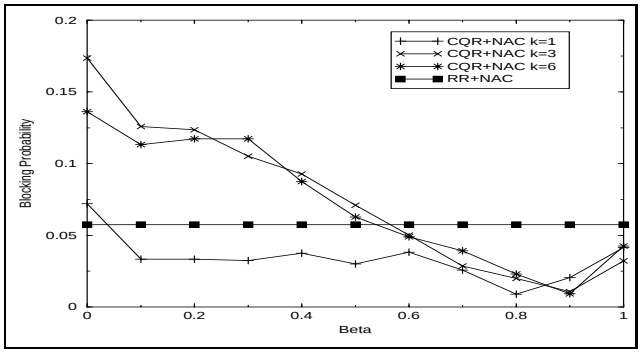

Fig. 10. Average blocking probability as a function of β.

Since a call is equally likely to be blocked regardless of load in this scheme, the total revenue will be a function strictly based on the number of calls coming into the system.

In QR+CAC we find a similar effect: the more traffic coming into the system increases the number of calls that are handled, causing revenue to increase. The total revenue generated by this scheme is higher than that of QR+FAC for high load, since the admission price increases with load. The total revenue is higher than that of QR+FAC for low load for a less intuitive reason. At low loads, we block much fewer calls than in QR+FAC. Although we tend to charge a lower price per call, we service more calls and hence generate more revenue. Our analysis explains this result: the admission price for a particular system load is designed to generate the maximum amount of revenue.

B. Redirection Schemes

Once a call is admitted into the network it should be redirected to the gateway best suited to provide service. Such a redirection scheme has two desirable properties. First, it should maximize network efficiency by achieving a balanced load and dampening oscillations. Second, it should maximize economic efficiency by ensuring that calls with high bids acquire high quality paths ahead of calls with lower bids. In this section we compare the following redirection schemes: Random Redirection (RR), Congestion and QoS Sensitive Redirection (CQR), and Price-based CQR (PCQR). We use No Admission Control (NAC) for all results shown in this section. We refer to the resulting schemes as RR+NAC, CQR+NAC, and PCQR+NAC, respectively.

1) Network Efficiency: To maximize network efficiency, it is necessary to have a load balancing policy to prevent oscillations in call traffic to each of the ITGs. This policy must be distributed in nature, as it must run across multiple LSs with potentially out of date information. We hence use congestion sensitive redirection to balance call traffic over the set of ITGs, thereby increasing the number of calls that can be admitted into the system.

In Figure 10, we measure the blocking probability of the two redirection schemes. We use β to tune the relative weights of congestion and QoS sensitivity (note that CQR with $\beta = 1$ is the same as CR). RR+NAC does not take β as a parameter and hence is shown as a flat line.

As the level of congestion sensitivity is increased in the hybrid scheme, the call blocking probability decreases significantly. The strictly QoS sensitive scheme causes many calls to be routed to a few ITGs in the center of the network topology that provide good QoS to all callers. Sudden increases in call traffic make these ITGs more likely to saturate, resulting in larger numbers of blocked calls due to GRC. Larger amounts of congestion sensitivity cause the load to be more evenly distributed across the set of ITGs, thereby decreasing the blocking probability. Hence, some amount of congestion sensitivity is necessary in the redirection scheme in order to increase system capacity.

Surprisingly, the call blocking rate increases slightly as we increase the relative weight of congestion sensitivity from $\beta = 0.9$ to $\beta = 1.0$. This happens because adding QoS sensitivity causes calls to be routed to closer ITGs. Advertisements sent by these ITGs are less likely to be dropped or delayed by the IP network, and hence calls directed toward these ITGs are less likely to result in a blocked call. Furthermore, although we see smaller values of k result in better performance, smaller values also cause an increased sensitivity to sudden overload conditions. Experimental results showed $k = 3$ to be a reasonable tradeoff.

At high loads, statistical fluctuations can cause all voice ports at certain ITGs to saturate due to sudden influxes of calls. This change in system state is reflected as NFP and GRC blocks. This can be explained by noting that the system can exist in one of the following two states: (1) When there is at least one ITG with no free voice ports. In this case no incoming call will be blocked. (2) When none of the ITGs have free voice ports. In this case all incoming calls will be blocked. NFP blocks can only occur if the system is in State 2. Because of the large amounts of background traffic, advertisements indicating that the ITG is full can get dropped or delayed, and hence state 2 tends to be manifested by GRC blocks. The large relative percentage of GRC type blocks emphasizes the importance of congestion sensitivity in a redirection scheme when the IP network is heavily loaded.

2) Economic Efficiency: The IP network is not homogeneous and hence flows traversing different paths in the network may experience widely varying levels of QoS. Flows passing through congested areas of the IP network sustain high packet loss, thereby decreasing call quality. Our system allocates high quality paths to users with high demand over users with less demand by performing call redirection based on price.

1) *Average Call QoS:* Figure 11 shows the average audio quality achieved in the hybrid scheme without call admission control. We notice that QoS sensitivity can sig-

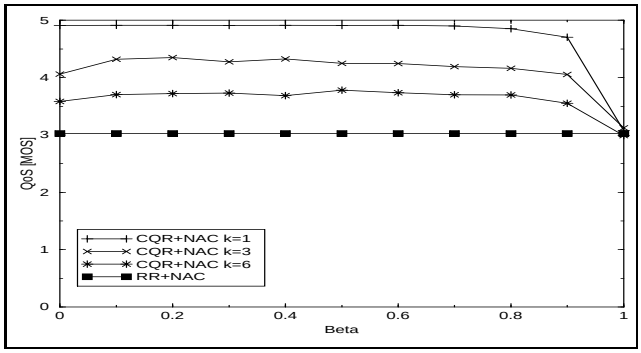

Fig. 11. Average call quality achieved as a function of β.

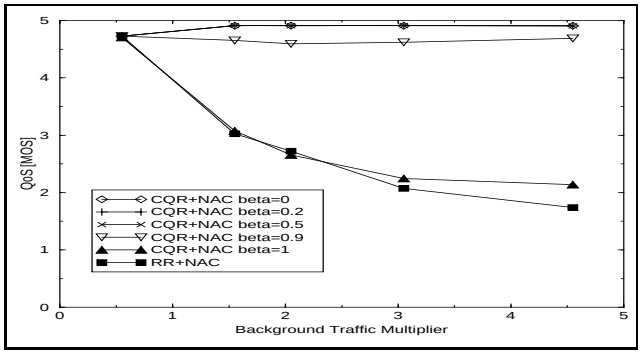

Fig. 12. Average call quality achieved as a function of cross traffic for $k = 1$. The horizontal axis is the multiplier by which we increase link utilization.

nificantly increase the average call quality. For example, quality increases from a poor MOS of 3 to a good MOS of 4.5 for CQR when $k = 3$. This happens because the LS can leverage QoS sensitivity to distinguish between ITGs with nearly equal levels of congestion based on their respective path qualities. Although smaller values of k improve performance, they decrease resilience to sudden overload conditions. Figure 12 shows how QoS sensitivity can improve resilience to IP network congestion. We also observe that congestion sensitivity offered a similar benefit in terms of blocking probability under increased offered load. It is interesting to note that a relative weight β in the range $[0.7..0.8]$ offers both a very high QoS and a very low blocking probability. This shows that we can achieve the benefits of congestion sensitivity with QoS sensitivity in a single hybrid redirection scheme. Additional experiments showed that there existed a good operating point for the system under a variety of different parameters, although the best value of β varied.

2) *Relationship between Price and Audio Quality:* We measure the relationship between the bid price of the user and the audio quality achieved by calculating the correlation coefficients of the two variables. In RR+NAC, all calls are terminated at an ITG independently of the bid price. Hence, all calls receive the same average audio quality and so the correlation coefficients are close to zero under this scheme. The correlation coefficients under CQR+NAC are also close to zero for the same reason. We find that adding price sensitivity to the redirection method in PCQR+NAC results in correlation coefficients close to 1. This happens because users with higher bids are assigned a lower value of β under this scheme, resulting in higher quality calls. Low paying users are redirected to more distant ITGs, leaving higher quality paths available for high paying users and decreasing the blocking probability for all users. Furthermore, we notice that the correlation improves with load under this scheme. At low load, most ITGs had small numbers of voice ports in use. This causes the QoS sensitive component of the scheme to have a larger effect, thereby giving good QoS to all calls and decreasing the correlation. Note that this is a desirable result, as at low loads congestion sensitive redirection becomes unnecessary. Furthermore, these schemes are particularly susceptible to overload due to the lack of call admission control. Since the large number of blocks occurs independently of price, economic efficiency is also reduced in the sense that many high paying calls are blocked. This shows the importance of an effective admission control scheme. More detailed information is available in our technical report [28].

3) Provider Efficiency: RR+NAC and CQR+NAC generate revenue that rises linearly with load. This happens because we charge all callers the same fixed rate. In PCQR+NAC, we achieve provider efficiency by charging users more for access to higher quality paths. This allows the system to generate more revenue from these paths from users with high bids, while still generating revenue from users with low bids on the lower quality paths. This scheme generates more revenue than all the other redirection schemes under high load. This is because the system is always providing a wide range of prices for incoming calls. For example, if a caller with a high bid enters the system it will likely be handled at the home ITG which will allow the system to charge the maximum possible amount for the call. In addition, we block very few calls, and of those which are not blocked we can charge a price very close to the bid.

C. Sensitivity to the Network Parameters

To investigate the sensitivity of our main results to changes in network parameters, we vary the network topology and the percentage of RTP traffic of the total traffic load. The following is a summary of the main observations.

- We carried out simulations based on the Excite@Home topology mentioned earlier. The larger network diameter causes the average service distance to rise for all schemes. However, the average audio quality also increases with load. Due to the larger number of ITGs in the system, each LS has a larger number of entries for ITGs with a particular load, and can choose the best from a wider range of path qualities to serve the user. Furthermore, the larger network diameter causes the number of GRC blocks for CQR+NAC to increase significantly with load, causing revenue to drop. Calls that would have been directed to ITGs via a poor QoS path tend to get blocked first, causing the average received audio quality to improve with system load under this scheme.
- RTP traffic is by nature less bursty than the Pareto distributed background traffic. Hence, as we increase the relative percentage of RTP traffic, the average received audio quality achieved by each of the management schemes increases. However, the average received audio quality

in PCQR+NAC decreases, as each LS now has a smaller number of entries for ITGs under a particular load. This causes the congestion sensitive component of CQR to have a larger effect, thereby forcing a larger number of calls to be pushed to ITGs offering a poor QoS. Furthermore, increasing the offered load causes the average call quality to decrease. Since RTP traffic now constitutes a larger percentage of total network traffic, increasing the number of calls in the system has a significant effect on IP network load. This effect is smallest for the schemes with admission control, due to the large number of OPR blocks.

VII. Conclusion

The underlying client-server architecture in IP Telephony allows service providers to enable new services and efficiently manage network resources. In this paper, we presented a comparative study of a few simple but representative resource management strategies that take into account both network resource congestion, the QoS requirements of the users, and the price that the users are willing to pay for a certain QoS. For the current best-effort Internet, we introduced two congestion sensitive call redirection mechanisms. We introduced an admission control policy that changes price based on network resource utilization to avoid system overload. The results show that adding QoS sensitivity to congestion sensitive redirection maximizes resource utilization while simultaneously maximizing the economic efficiency, i.e., it provides higher QoS to users who are willing to pay for it. We compared a hybrid redirection scheme that takes into account the fact that the quality of received audio depends on the characteristics of the path chosen by the network with a simple flat and congestion sensitive redirection techniques.

There are many items that need further investigation. Firstly, there are different ways to combine CS and QoS sensitivity into a single resource management model; these approaches need to be compared to determine their relative strengths and weaknesses. Secondly, the problem of ITG placement in wide area networks needs to be addressed. Thirdly, the study needs to be carried out using more realistic user models. Finally, it would be of some interest to simulate a competitive network where management schemes can differ at each administrative domain.

VIII. Acknowledgements

The authors would like to thank Lakshminarayanan Subramanian and Zhuoquing Morley Mao from UC Berkeley and Takashi Suzuki from NTT DoCoMo for their helpful comments and suggestions. We are also grateful to the anonymous reviewers for helping us to improve the paper.

References

[1] J. Rosenberg and H. Schulzrinne, "The IETF Internet Telephony Architecture and Protocols," Technical Report, Columbia University, 1999.

[2] J. Rosenberg and H. Schulzrinne, "Internet Telephony Gateway Location," *IEEE INFOCOM*, San Francisco, California, March-April 1998.

[3] T. Russell, *Signaling System Number 7*, McGraw-Hill Series on Computer Communications, New York, 1995.

[4] N. Anerousis, R.Gopalakrishnan, C. R. Kalmanek, A. E. Kaplan, W. T. Marshall, P. P. Mishra, P. Z. Onufryk, K. K. Ramakrishna, and C. J. Sreenan, "TOPS: An Architecture for Telephony over Packet Networks," *IEEE Journal of Selected Areas in Communications*, Vol. 17, No. 1., pp. 91-108, January 1999.

[5] N. Semret, R. Liao, A. Campbell, and A. Lazar, "Pricing, Provisioning and Peering: Dynamic Markets for Differentiated Internet Services and Implications for Network Interconnections", *IEEE Journal on Selected Areas of Communications*, Vol. 18, No. 12, pp. 2499-2513, December 2001.

[6] J. Kurose and K. Ross, *Computer Networking, A Top-Down Approach Featuring the Internet*, Addison-Wesley, 2000.

[7] M. Handley, H. Schulzrinne, E. Schooler, and J. Rosenberg, "SIP: Session Initiation Protocol", IETF, *RFC 2543*, March 1999.

[8] J. Rosenberg and H. Schulzrinne, "A Framework for Telephony Routing over IP", IETF, *RFC 2871*, June 2000.

[9] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", IETF, *RFC 1889*, January 1996.

[10] A. Gupta, D. Stahl, and A. Whinston, *The Economics of Network Management*, Communications of the ACM, Vol. 42, No. 9, pp. 57-63, September 1999.

[11] R. J. Edell and Pravin Varaiya, "Providing Internet Access: What we learn from the INDEX Trial," Index Project Report 99-010W, University of California, Berkeley, April 27, 1999.

[12] A. Ganesh and K. Laevens, "Congestion Pricing and User Adaptation", in *Proceedings IEEE INFOCOM*, Anchorage, USA, April 2001.

[13] L. Kleinrock, *Queuing Systems Volume I: Theory* , John Wiley & Sons, New York, 1975.

[14] Nortel Networks, Ltd., "Alteon Content Director" *http://www.nortelnetworks.com/products/01/pcd/index.html*.

[15] Akamai Technologies, Inc., *http://www.akamai.com*.

[16] A. Rayes and K. Sage, *Integrated Management Architecture for IP-Based Networks*, IEEE Communications Magazine, Vol. 38, No. 4, pp. 48-53, April 2000.

[17] M. Hassan, A. Nayandoro, and M. Atiquzzaman, *Internet Telephony: Services, Technical Challenges, and Products*, IEEE Communications Magazine, Vol. 38, No. 4, pp. 96-103, April 2000.

[18] S. Pracht and D. Hardman, "Voice Quality in Converging Telephony and IP Networks", Technical white paper, *http://literature.agilent.com/litweb/pdf/5980-0989E.pdf* .

[19] A. Watson and M. Sasse, "Multimedia Conferencing via Multicast: Determining the Quality of Service Required by the End User," in *Proceedings of the International Workshop on Audio-Visual Services over Packet Networks (AVSPN)*, pp. 189-194, Aberdeen, Scotland, September 1997.

[20] M. Podolsky, C. Romer, and S. McCanne, "Simulation of FEC-Based Control for Packet Audio on the Internet," in *Proceedings IEEE INFOCOM*, Vol. 2, pp. 505-515, San Francisco, USA, March-April 1998.

[21] S. McCanne, S. Floyd, et. al. "The Network Simulator - ns-2," *http://www.isi.edu/nsnam/ns/*.

[22] M. Caesar and D. Ghosal, "IP Telephony Simulator Based on ns-2," *http://www.cs.berkeley.edu/∼mccaesar/research/iptelsim.tar.gz*.

[23] Qwest, "Qwest National IP Infrastructure," *http://www.upnetworks.com/companyinfo/images/ipnetwork.jpg*.

[24] Excite@Home, "@Work – Excite@Home Dual OC-48 IP Backbone," *http://work.home.net/whitepapers/backbone.html*.

[25] V. Paxson and S. Floyd, "Wide-Area Traffic: The Failure of Poisson Modeling," in *IEEE/ACM Transactions on Networking*, Vol. 3 No. 3, pp. 226-244, June 1995.

[26] S. McCreary and K. Claffy, "Trends in Wide Area IP Traffic Patterns: A View from Ames Internet Exchange," CAIDA Technical Report, *http://www.caida.org/outreach/papers/AIX0005/* .

[27] L. Murphy and J. Murphy, "Feedback and Pricing in ATM networks", *ATM Networks: Performance Modeling and and Evaluation*, Vol. 2, pp. 197-212, Chapman and Hall, 1996.

[28] M. Caesar and D. Ghosal, "Resource Management for IP Telephony Networks," Technical Report CSE-2002-2, Computer Science Department, University of California at Davis.

[29] M. Caesar and D. Ghosal, "IP Telephony Annotated Bibliography," *http://www.cs.berkeley.edu/∼mccaesar/research/iptel_litsurv.html*.

[30] E. Crawley, R. Nair, B. Rajagopalan, and H. Sandick, "A Framework for QoS-based Routing in the Internet", IETF, *RFC 2386*, August 1998.

[31] A. Fox, S. Gribble, Y. Chawathe, E. Brewer, and P. Gauthier, "Cluster-Based Scalable Network Services," *the Proceedings of ACM Symposium on Operating Systems Principles*, Vol. 31, pp. 78-91, October 1997

[32] F. Kelly, A. Maulloo, and D. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," in *Journal of the Operational Research Society* , Vol. 49, pp. 237-252, May 1998.

[33] S. Shenker, "Fundamental Design Issues for the Future Internet," *IEEE Journal on Selected Areas in Communications*, Vol. 13, No. 7, pp. 1176-1188, September 1995.