Cost/Benefit Tradeoff of Quality of Service Mechanisms in Integrated Services Networks

Junseok Hwang
(hwang@tele.pitt.edu)
Telecommunications Program
University of Pittsburgh
Pittsburgh PA 15260

Martin B. H. Weiss
(mbw@tele.pitt.edu)
Telecommunications Program
University of Pittsburgh
Pittsburgh PA 15260

August 24, 1999

Abstract

Internet Protocol (IP) and Asynchronous Transfer Mode (ATM) are two switching technologies that are being used (or proposed) for large-scale integrated services networks. Integrated services networks are designed to support real time services, such as telephony as well as data services. Both technologies have the potential to radically change the economics and operations of telephone services. In addition, engineers have developed more efficient approaches than simple over-engineering to supporting the quality requirements of real time services (such as voice).

In this paper, we will perform cost-benefit (tradeoff) study of guaranteeing the voice quality of packet telephony of various possible QoS-support approaches (IP/over-engineering, IP/prioritization, IP/RSVP, ATM-CBR, ATM-VBR). We will focus this application on integrated services networks. As with our previous work [19], we will use a simulation model to engineer a large scale network and will then analyze the switching and trunking costs. We will compare the costs of the different approaches needed to meet a prespecified QoS requirement.

Keywords: IP Telephony, QoS, VoATM, IntServ, DiffServ, RSVP, Engineering Economic Model.

1 Introduction

Many telecommunications carriers (RBOCS, CLECS, IXCs, and ISPs) are currently engineering their networks to support integrated services. To do so, they must consider the various quality requirements of different applications, especially voice, which many consider to be an essential application, and which has traditionally been carried over circuit switched networks. Recent developments have supported the commercial emergence of packet voice (packet telephony) with the advantages of statistical multiplexing gain and lower costs. The technologies for supporting packet voice include Internet Protocol (IP), Frame Relay (FR), and ATM (Asynchronous Transfer Mode).

IP telephony is a family of real-time voice applications over IP networks. Voice over IP (VoIP) is implemented using ITU Recommendation H.323 which defines the RTP/UDP/IP protocol stack for packetizing

voice into IP packets. Even though today's best-effort IP networks do not provide guaranteed service quality, telephony applications are rapidly launching into the Internet in various forms in all aspects of the telecommunications market [13], [16].

One of the challenging issues for the successful implementation of IP telephony system is the delivery of PSTN comparable QoS (Quality of Service). In this paper, QoS refers to a set of performance measures associated with basic telephony services. We will focus on delay and delay variation; other possible factors include packet loss and reliability.

Today, many consumers report poorer service quality for Internet telephony due to network delay as well as limitations surrounding the PC [4]. A recent study on the measurement of network delay for various type of the Internet connections by Maxemchuk *et. al.* [12] suggests that the current Internet is better suited for local bypass than its current role as a replacement for long distance or international connections.

IP is not the only packet switching approach to support telephony. Voice over Frame Relay (VoFR) is already in place especially for VPN enterprise applications and ATM was originally designed to meet the diverse QoS requirements of the integrated service environment. Unlike current the IP network, Frame Relay and ATM are connection-oriented protocols, so that the migration to support PSTN comparable telephony is more straight-forward (in terms of QoS).

There are various QoS-Support options in these switching technologies, especially in IP and ATM. Recent developments in IP supporting QoS (IntServ, and DiffServ), and the newly standardized AAL2 (rt-VBR) for telephony applications makes the comparison among these technologies more interesting.

A previous study by Weiss and Hwang [18] found that a pure IP telephony network provides lower switching and trunking cost than the PSTN using an over-engineering approach for voice-only applications. More recent findings on the extended study on this by Hwang [8] using RSVP (Resource ReServation Protocol) suggests that the guaranteed service approach through the reservation may not be as efficient as circuit switching (PCM) if only guaranteed traffic (IP telephony) is carried by the network. This is generally consistent with Baldi *et. al.*[2]. For integrated service networks (not pure Itel), however, QoS approaches seem to be advantageous according to the initial findings from [8, 19].

In this paper, we will extend our previous work to include general QoS-mechanisms of various transfer technologies, especially IP and ATM. The introductory part of this paper provides some background of QoS problems for telephony applications, review some of the related and previous studies and discuss the relevance of relating the cost analysis to the various QoS-mechanisms for the integrated service networks. The second part will:

- provide a brief technical overview of the two packet data networks for telephony applications and the QoS-mechanisms associated with them,
- discuss the motivations of QoS-support for various packet telephony systems, and
- look at the quality and performance requirements for PSTN-comparable quality based packet telephony services.

Various performance parameters for packet telephony applications will be reviewed. The third part will describe the simulation model we implemented for the IP and ATM networks with various QoS-support mechanisms. The fourth part will compare the IP with and ATM (Asynchronous Transfer Mode) based on the simulation results found. Finally, we will develop cost models associated with the QoS mechanisms.

2 Technical Overview

In this section, we will review some of the relevant technical details for the comparison of two different switching architecture for the telephony application.

2.1 IP Telephony System Architecture and Operation

Figure 1 shows the protocol stack typically used for voice communication and signaling in Internet telephony. When a G.729A codec is used, all voice traffic are vocoded from 64 Kbps PCM voice into an 8 Kbps compressed signal. Like most of the low bit rate vocoders, G.729A utilizes silence suppression to further decrease the required data rate. The compressed voice signal is then packetized using an RTP/UDP/IP protocol stack. RTP typically runs over UDP to utilize its multiplexing and checksum services. RTP also specifies the payload type to support multiple data and compression types.

Using a G.729A codec, voice is packetized into 10-byte voice packets every 10 msec and is buffered to make a 10 to 30-byte payload. The payload is then encapsulated with a 40-byte RTP/UDP/IP header. Combined with the RTP/UDP/IP overhead, the suppressed codec output drives the actual average throughput to around 14 Kbps. The RTP sequence numbers and time stamps are used to reassemble digitized voice packets into a real time voice channel. This approach, however, does not provide any QoS guarantees. A packetized voice packet is delivered to its destination through various layer-2 protocols to the access server and multiplexers, and layer-3 switches to finally route the packet to its destination.

A signaling protocol for Internet telephony is essential to achieve a PSTN-comparable QoS performance. There are several possible approaches to providing PSTN compatible signaling capability over the Internet telephony network. The implementation (and, hence cost) varies with the different signaling requirements for Internet telephony calls. In addition, the locus of the signaling function impacts the choice of the call controlling protocol to be used. Three main call control signaling protocols for Internet telephony have been suggested: H.323 Based Signaling [9], SIP (Session Initiation Protocol) [7], and MGCP (Media Gateway Control Protocol) [1]. These protocols differ significantly in terms of control, design, and functionality. Different classes of internet telephony, connectivity options, WAN transport and main applications result in biases toward different signalling protocols. Although quite interesting and important, a detailed discussion of these is outside the scope of this paper.

2.2 VoATM System Architecture and Operation

The system architecture of Voice over ATM (VoATM) resembles VoIP except for the switching method. Unlike IP, ATM is connection oriented cell switching and uses signalling defined in UNI 4.0.

To illustrate the operation of VoATM, let us first assume that:

- a native ATM terminal exists,
- all the COs in the service area are mesh-connected with pre-configured PVCs (Permanent Virtual Circuits), and
- each subscriber establishes individual end-to-end SVCs (Switched Virtual Circuits) through those PVCs.

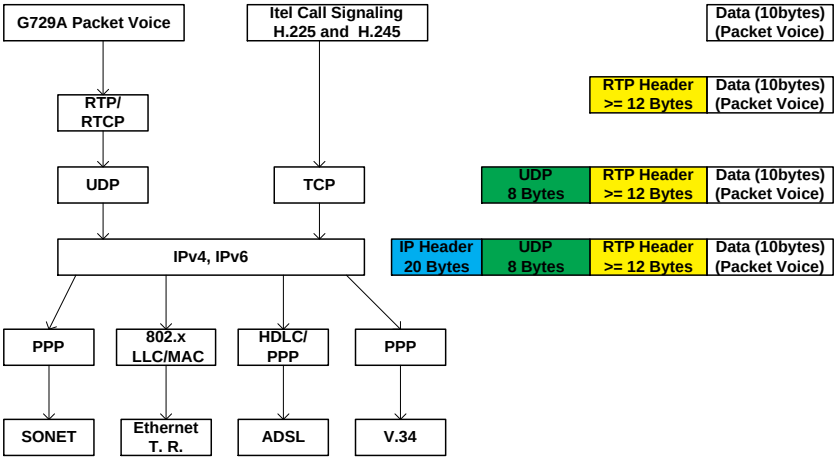

Figure 1: Example of RTP Protocol Stack (10 Byte Packetization) in the IP Telephony System

When a call is originated on the ATM network using a particular virtual circuit, the functions for AAL1 or AAL2 and associated processing are performed at the network edge. QoS negotiation and Call Admission Control (CAC) are performed at the edge ATM switches of each COs and all ATM switches in the network performs traffic shaping based on the negotiated QoS parameters both in the individual SVC level and in the aggregated PVC levels among CO switches. Figures 2 and 3 show the CBR and rt-VBR implementation of VOATM system, respectively.

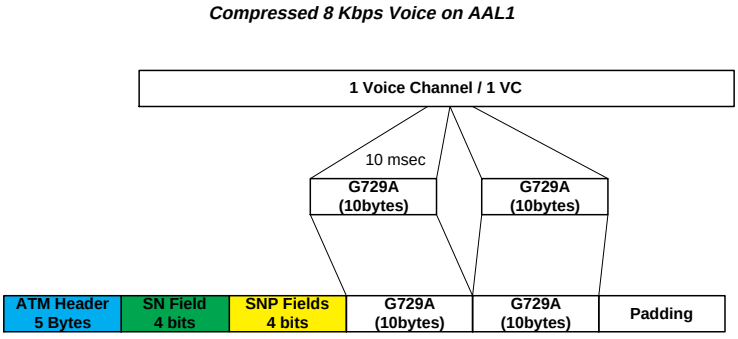

Figure 2: CBR Implementation of VoATM

2.3 Performance Requirements for Packet Telephony

The performance needs of various applications differ greatly. In particular, telephony applications require relatively strict performance limits; in this section, we will discuss those requirements especially for packet telephony network.

The problem of controlling delay and its variation on packet networks is one of the most important

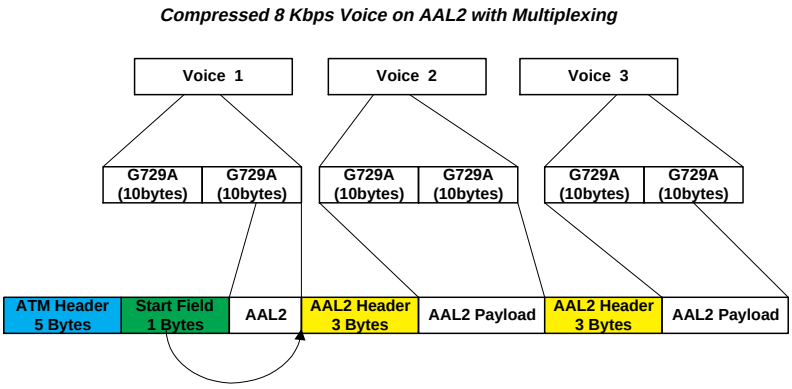

Figure 3: rt-VBR Implementation of VoATM

performance issues of packet-based telephony. To achieve toll-quality service, the end-to-end delay should be below a certain threshold. For PSTN-like quality, the ITU recommends an end-to-end delay of less than 150 milliseconds. Since delay is stochastic in the packet switched networks, this requirement may be interpreted as the 99th percentile delay. In addition, the delay variation should be short enough to enable the dejitter buffer of the receiving phone to equalize the delay and disassemble the voice packet. The jitter requirement may vary depending on the selection of the CODEC and voice packet encapsulation approach.

While packet loss is also an important issue, the human perception on the speech is more sensitive to delay than loss. A carefully-sized receiving buffer can recover the usable late voice packets on an Internet telephony terminal, and a good buffer management protocol is essential to avoid excessive unfair packet loss due to the different sizes of packets. According to [6], the overall packet loss ratio of less than 1% would be able to achieve the toll speech quality of MOS (Mean Opinion Score) around 4.0. We chose not to study this problem in the present paper.

Call set-up delay, also called access delay or post dialing delay, is another important performance parameter for packet telephony. This delay is the value of elapsed time between a access request and successful access and comprises the sum of call request time, selection time, and post selection time [6]. For the PSTN, this call setup performance has been defined to be about 1-3 seconds and assumes SS7 signaling [5], [6]. Currently many Internet telephony applications use H.323 signaling on TCP, and the call-setup delay time is not guaranteed.

The call blocking rate is one of the primary performance measures of circuit switched PSTN networks. The main reasons for call blocking are the network resource congestion (switch blocking and trunk blocking), and network failure. In general traffic engineering practice, the average blocking probability of telephony networks is less than 1%. Adequate performance estimate procedures and methods for QoS-support CAC (Call Admission Control) should be implemented to support a blocking rate of less than 1%.

Dial-tone delay is another important performance measure which is closely dependent on the resource allocation on the access switches or routers in the packet telephony network. This is the time interval between the receipt of a request for dial tone at the local switching office, or packet telephony access servers as a result of off-hook, and the beginning of dial tone transmission to the user [6]. The performance level

required for dial-tone delay in ITU-T are less than 0.1 percent and less than 0.5 percent of call experiencing more than 3 seconds dial tone delay user normal and high loads, respectively [10].

2.4 The QoS-Support Approaches of IP Telephony

QoS (Quality of Service) in Internet telephony can include voice delay and jitter, packet voice loss, voice clipping, call-setup delay, echo, and loudness distortion. Network delay, which is the primary focus of this paper, is dependent on the stochastic behavior of the traffic with different sizes of the packet length and different application requirements. Thus, there should be some approaches which allocate the network resources to compensate for this stochastic behavior. In this section, we will deal with various approaches which would mitigate or bound the such stochastic behavior of the network delay.

The Internet is a "best-effort" network, so QoS cannot be guaranteed. As real time applications, such as Internet telephony, are introduced to the public Internet, QoS issues become more critical because of their real time requirements. To control quality in IP telephony, properly designed IP switches, required QoS signaling protocol support, classification of services, and appropriate dimensioning and traffic engineering of the network are required. Generally, the strategies available for achieving this involve over-engineering the networks, using IP over ATM, using QoS signaling (such as RTCP and RSVP), or class of service differentiation. We classified the possible approaches for providing adequate QoS in the IP telephony architecture as follows;

Over-engineering The simple way of achieving required QoS is providing enough network resources to avoid the high network loads that result in delays and delay jitter. This approach is widely used through the LAN and private network environment to provide IP voice service over packet networks. For a carrier level Internet telephony network, other QoS techniques which will be discussed below would provide better network resource usage.

Resource Reservation: RSVP One way of QoS signaling is reservation of bandwidth on the Internet for time critical services; this is implemented in the RSVP protocol by IETF. RSVP is a reservation based signaling protocol designed enable the allocation of resources to support the QoS requirements applications, such as bandwidth and delay. Since RSVP does not provide QoS-dependent routing, other approaches, such as differentiation of service, must be implemented independently so that the variable delay component produced by router processing can be minimized.

Class of Services(Priority) Differentiation or Classification of Services is another approach to provide better quality of service and better utilization of the network resources. An Itel call can be assigned to the high priority using the ToS (Type of service) field of current IPv4 packet. Some of IP switch vendors (*eg.*, Cisco) use this approach to enable their switch to differentiate the VOIP packet. We also investigate the effectiveness of this approach in the following section. This approach is embedded in IPv6, where four priority bits are introduced to support real time traffic requirement through prioritization. Unlike RSVP signaling, this differentiation approach provides a mechanism for network nodes to use different routing, packet scheduling, and queueing for different of type of services. Fast and intelligent IP switch routers have the functionality of having different routing, scheduling, and queuing techniques such as WFQ (Weighted Fair Queuing).

2.5 The QoS-Support Approaches in VoATM

ATM is a connection-oriented cell (fixed size packet) switching and transfer technologies to support various application services with different QoS-requirements. The fixed short size cell allows network cell delay to

be predictable and controlled. ATM provides stronger QoS-support mechanism than IP and Frame Relay, and supports various options for implementing telephony applications.

CBR (constant bit rate) and rt-VBR (real-time variable bit rate) classes are the ATM service classes to be used for VoATM. CBR is currently the most common approach, using AAL1 to provide CES (Circuit Emulation Service). ATM Permanent Virtual Circuits (PVCs) act like trunk lines in the PSTN network. Switched Virtual Circuits (SVCs) are also possible if the network supports end-to-end ATM services. In CBR, Peak Cell Rate (PCR) is used as a traffic parameter and max CTD (Cell Transfer Delay), CDV (Cell Delay Variation), and CLR (Cell Loss Ratio) are specified as the associated QoS parameters. Accepted CBR calls which transmit the cells at or below the negotiated PCR will achieve the committed QoS performance from the network. Since a CBR connection needs to exchange timing information between source and destination, only a single user of the AAL can be supported on a single ATM connection.

rt-VBR traffic is another ATM option to send voice applications when the source rate is expected to be variable and bursty (using AAL2). In addition to the QoS parameters and PCR parameters specified in the CBR, the additional traffic parameters such as SCR (Sustainable Cell Rate) and MBS (Maximum Block Size) are specified. More efficient bandwidth allocation can be achieved by AAL2 due to variable rates and the support of silence suppression. rt-VBR also enables multiple user channels on a single ATM VC connection. Variable payload size is allowed within cells and across cells which improves the protocol efficiency compared with other "Voice over X" protocols.

3 Simulation Model Description

We have developed the simulation model to compare the performance and cost of those various QoS-mechanisms for the integrated service applications. In this section, we describe the parameters of the simulation models.

3.1 Topological and General Technology Assumptions

To develop a simulation model, we assumed the service area with 1 million population which is equivalent to the U.S. State of Rhode Island. To implement the realistic model close to the practical real network, we have made following assumption:

- Five large COs with core edge switches (ATM or layer 3 switches) are forming part of the SONET Ring in the Service Area.
- Additional edge switches are attached to the core switches with DS3 or Sonet interfaces, as illustrated in Figure 4.
- An average population density of 2.2 person per household uniformly distributed over the COs, and each household has one telephone lines with 0.1 Erlang call density in the busy hours.
- All the switches in the service areas are equipped its associated QoS-support functionalities.
- We assumed the local loop interface with ADSL for both of the network topology.
- All voice traffic would be compressed from 64 Kbps PCM voice to 8 Kbps compressed voice using G729A vocoder at the access network.

3.2 IP Network Model

The Itel simulation model assumed in our analysis is based on RTP/UDP/IP standardized on ITU H.323. In this protocol, sequence numbers and time stamps are used to re-assemble the real time voice traffic. The simulation is concentrated on the IP layer 3 switching and trunking. The assumptions made for Itel simulation model are specified below and in Table A-2.

- Silence suppression will be enabled in each codec, with 60% of a session being silent each way.
- On the suppressed codec output, RTP, UDP and IP overhead will make actual average throughput around 9.6 Kbps (assuming 20 byte packetization).
- Voice is packetized into a 10 byte voice packet every 10 msec and buffered to make a 20 byte payload from compression codec and encapsulated with the 40 byte RTP/UDP/IP header.
- Voice is modeled as an on-off process, with an average 350 msec active state (exponentially distributed) and a 650 msec exponentially distributed silence state.
- Itel calls are modeled as connectionless UDP/IP sessions with exponentially distributed session lengths of 240 sec.
- For Prioritization QoS-support, the highest priority is given to the telephony application.
- For RSVP, we calculated the required reserved bandwidth for each trunk for the telephony only scenario, then calculated the worst case delay. Then we add the integrated service traffic and tuned the simulation to achieve the same quality of voice with RSVP.

  Note that the parameters we chose for RSVP are very conservative (summarized in Table A-3. The bandwidth required by RSVP is very sensitive to the choice of parameters, so that less conservative parameters would result in reduced bandwidth requirements.
- Using the recent traffic data measured by [3], [15], and [11] on the Internet backbone OC-3 trunks, we modeled the integrated service traffic as a cross section of the Internet backbone traffic, and computed the intensity of this traffic relative to Itel call demand.

3.3 ATM Network Model

In our VoATM simulation, we considered CBR (AAL1) and rt-VBR (AAL2) service categories in support of telephony applications. For these services we assumed following conditions.

- We assumed 20 bytes (20 msec of speech) of payload of AALs for both type of the services.
- CBR does not support voice suppression and multiplexing of multiple AAL1s into a ATM cell.
- rt-VBR supports voice suppression and multiplexing of multiple AAL2s into a ATM cell.
- CBR and rt-VBR are given to higher priority than nrt-VBR, ABR and UBR.
- The equivalent load of integrated service traffic used in the IP model is translated into ATM traffic.

Table A-4 summarizes the simulation parameters used in ATM simulation model.

4 QoS Mechanism Simulation Analysis

In this section, we present the main results of the simulation for the various types of QoS-mechanisms for the ISN scenarios. The objectives of this simulation were:

1. to compute the required switching and trunking capacities while maintaining adequate QoS, and
2. to compute the 99th percentile delay for voice packets when integrated services traffic is added or when the number hops is increased.

All simulations were performed using Comnet III.

We assumed the base ISN traffic scenario to be one where the the traffic load of voice and data are identical in terms of byte throughput (9091 erlangs of voice traffic and 128 Mbps of data traffic). The networks were engineered to carry this load based on maintaining acceptable voice delays. To "stress" this model, we added additional data (*i.e.*, integrated service) traffic until the QoS for voice exceed the delay tolerance (without making any additional capacity "investments"). Figure 4 represents the simulation model for the QoS sensitivity of various QoS-mechanisms over various ISN traffic load.

4.1 QoS Sensitivity on Incremental Data Traffic

Figure 5 illustrates the results of the simulations for each QoS-mechanism. The abscissa represents relative throughput of the integrated data traffic. The ISN baseline scenario represents the case when the throughput of IP telephony and data traffic are equivalent (at value of 1 of the abscissa). The ordinate represents 99th percentile variable delay of packets or cells, which includes queuing delay and serialization delay on the core edge switches. At the relative integrated service load level of "0", the networks are pure voice networks and the performance of the approaches are indistinguishable, largely due to the low network utilizations at this level of traffic intensity.

The results show that ATM/VBR and ATM/CBR are statistically indistinguishable from each other and provide the best performance of all the QoS technologies we considered. With ATM/CBR scenario, the network would not be able to utilize more traffic after relative ISN load scale of 2 due to the absence of silence suppression and multiplexing implemented in AAL1 (for VBR). The best-effort and prioritization schemes of IP worked until the point of ISN scale 1 and 1.5 respectively. For the prioritization, simple non-preemptive prioritization is implemented.

Since ATM was designed with integrated services networks in mind, it is not surprising that they performed the best. That ATM/VBR was able to carry such large integrated services traffic loads implies that carriers can add a lot of non-voice traffic without having to make significant network upgrades, which would result in a cost advantage to those carriers. For IP-based networks to carry similar loads at the same performance levels would require additional investments in switching and trunking capacity.

4.2 QoS Sensitivity on Multi-Hop-Networks

Delay in packet networks is known to be dependent on the number of hops that a packet traverses. Thus, we performed more simulation experiments to add a bit more realism to this study, as most packets traverse more than one hop.

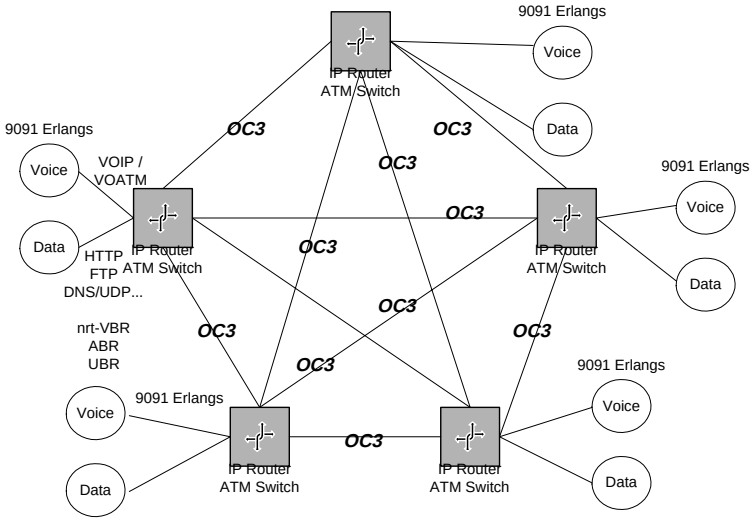

Figure 4: Simulation Model for QoS Sensitivity

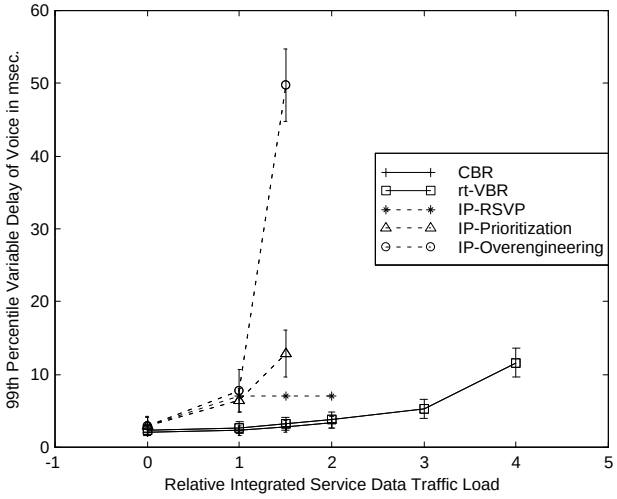

Figure 5: 99th Percentile Variable Delay Trends for Various Traffic Load

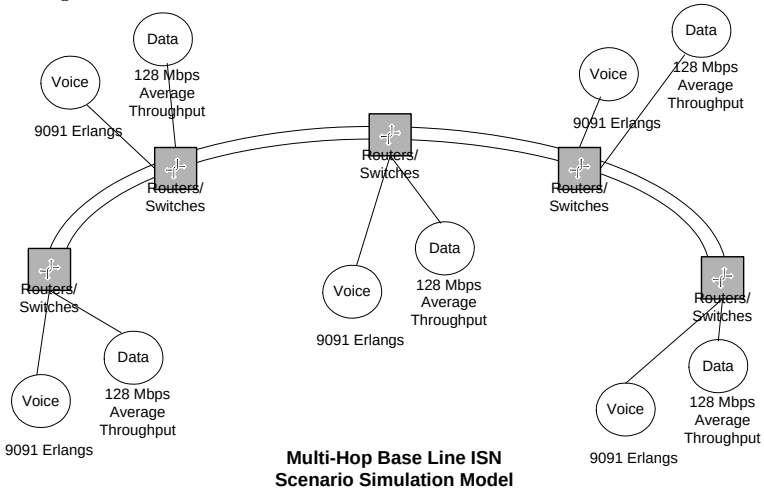

Figure 6: Simulation for Multihop Sensitivity

As it is shown in Figure 6, we consider 5 backbone core switch nodes which are spaced at 100 km evenly. We assumed ISN baseline traffic load to be offered to each of the core switches and to be traversed various number of hops in the model. By monitoring the QoS measures of voice traffic which traversed various hops, we can find that the equivalent utilization of the trunks over various QoS mechanisms provides acceptable QoS. The only exception is the IP/over-engineering case where utilization is around 40%.

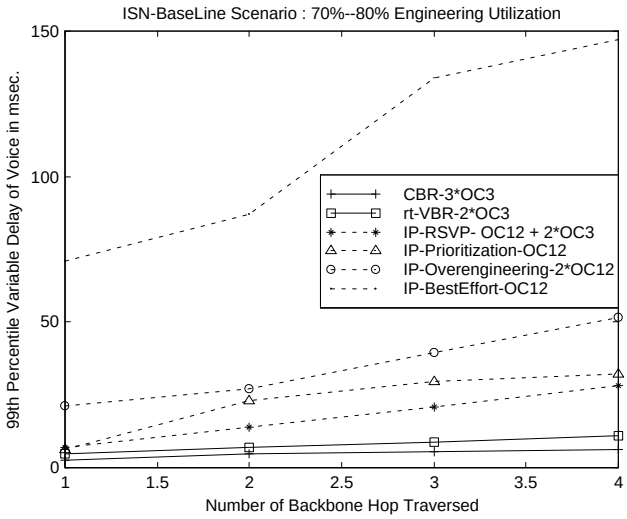

Figure 7: multi-hop sensitivity of ISN-Base line scenario

From the results of this simulation summarized in Figure 7, the CBR is found to be the least sensitive to the number of hops traversed (as expected). IP prioritization was roughly equivalent to rt-VBR after multiple hops are traversed. One of the major findings from this simulation is that each QoS-mechanism shows different bandwidth efficiency, which affects the cost of the network (discussed below).

5 Cost Analysis

Validating the technical comparison of the various QoS technologies was not the goal of this research – understanding the implications on switching and trunking cost was[1]. With designs in hand, we consulted with vendors to review the "reasonableness" of the design and to estimate the cost of the switches or routers. The cost of the transmission links is based on leased line costs from AT&T with bulk price discount rate of 50% and 90% (to assess the sensitivity of the costs on trunking).

We considered five backbone switch nodes which are spaced in 100 km evenly. The cost of the switching technology is composed of two parts; initial investment costs and yearly recurring costs. Most of the switching equipment in a CO will be considered the initial capital investment costs and transmission links (OC-3, *etc.*) will be considered the recurring costs. We assumed the product life time of 3 years for switching equipment and 10 % of MARR (Minimum Attractive Rate of Return) to calculate the NPV (Net Present Value) of trunking costs. The calculated cost of each QoS mechanism is illustrated in Figure 8.

Figure 8 illustrates the costs of the various implementations discussed in this paper. These costs are clearly influenced very heavily by trunking costs, which were calculated to be a 50% discount off of AT&T's

[1] We assume that the access networks are identical for all technologies, so we do not include them in our analysis.

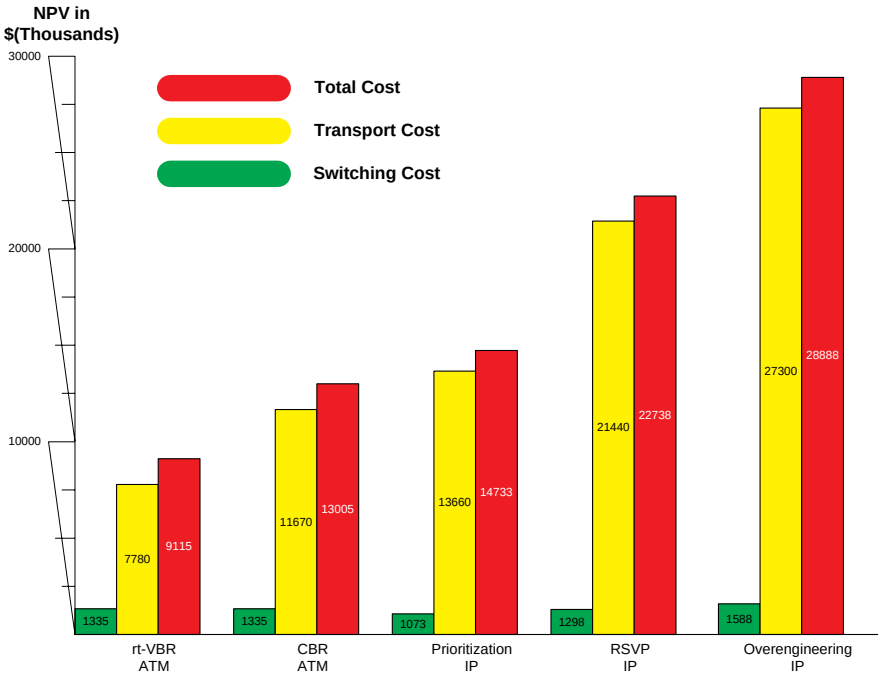

Figure 8: NPV Comparison of Switching and Trunking Costs among Various QoS-Mechanisms

retail rates. Table 1 compares the 50% and 90% discount rates. Even at the lower transmission costs, the same end results hold:

- ATM/VBR is the lowest cost. This is even stronger given the ability of this technology to carry much more integrated service traffic at this level of capacity without significantly degrading the packet delay.
- Especially at the lower trunking cost, the costs of both ATM technologies and IP/Priority appear to be roughly equivalent, as do the costs of IP/RSVP and IP/Over-engineering.

Technology	**Cost**	**50% Discount**	**90% Discount**
ATM/VBR	Transport Cost	\$7.78M	\$1.56M
	Total Cost	\$9.12M	\$2.91M
ATM/CBR	Transport Cost	\$11.67M	\$2.33M
	Total Cost	\$13M	\$3.67M
IP/Priority	Transport Cost	\$13.66M	\$2.73M
	Total Cost	\$14.73	\$3.80M
IP/RSVP	Transport Cost	\$21.44M	\$4.29M
	Total Cost	\$22.74M	\$5.59M
IP/Over-engin.	Transport Cost	\$27.3M	\$5.46M
	Total Cost	\$28.89	\$7.05M

Table 1: Sensitivity of NPV costs to Trunking Costs

6 Summary: Cost and Benefits

The purpose of this paper has been to explore the implications on transmission and switching cost of various QoS support technologies for a relatively large network. We have shown that the transmission and switching costs of even simple QoS technologies (such as prioritization in IP) can be substantial as compared with over-engineering. We previously reported that, using the same basic simulation topology, that IP/Over-engineering resulted in trunking and switching costs that were about 50% lower than an equivalent circuit switched network [18]. This paper claims that this cost advantage can be improved further by adding QoS technologies.

The benefit side of this scenario is more compelling. QoS support technologies allow carriers to support more integrated service (*i.e.*, data) traffic when they deploy QoS support technologies in their networks *without* upgrading the transmission and switching capacity. Circuit switched carriers would have to build a separate infrastructure, incurring more capital and trunking costs, to provide equivalent integrated services traffic (see [19] for a further analysis).

From an engineering perspective, the technical results are not surprising, as that is precisely what QoS support technologies are supposed to do. What was surprising to these researchers was that cost of ATM switching technologies was comparable to the IP technologies. This was not the case a few years ago[2].

From a policy perspective, it illustrates that, depending on the technology, the incremental costs of data traffic can be quite low. This may have substantial bearing on interconnection negotiations, as those have been framed based on these incremental costs. Many of these costs would be "forward-looking," as they represent are next generation technologies for telecommunications service providers.

There were a number of significant costs that we did *not* consider explicitly in this paper. We assert that these omissions are significant only to the extent that the *differences* in costs between these technologies are significant; in many cases, we claim that they are roughly similar, as we are comparing packet technologies with each other and are not comparing them to circuit switching

- One was the cost of gateways and converters. This is typically a large percentage of the capital cost of a "Voice over X" carrier; since we were comparing only these types of carriers, and were not comparing these costs to circuit switching, the significance of this cost is confined only to difference in gateways/converters[3].

- We also did not explicitly account for call processing systems and support. We assert that the only significant bearing of this omission applies when the cost *differences* are significant between the various packet technologies. Signalling for "Voice over X" is an area that is currently evolving very rapidly. It is difficult to get reliable cost estimates for signalling and call processing components.

- Finally, and perhaps most significantly, we did not include operations costs. We believe that this is an area where the differences between the technologies are potentially large. It is often the case that multi-QoS capability requires more engineering and management effort than single-QoS capability does.

[2] This does not imply that operating costs of ATM are comparable to IP operating costs. We have no information on the question of operating costs.

[3] For devices using vocoders, this cost difference is likely to be small, since they would all use Digital Signal Processors (DSPs). This difference could be significant when comparing ATM/CBR to either ATM/rtVBR or IP, since CBR normally packs PCM samples into ATM cells, avoiding the need for vocoding.

We did *not* analyze the implication of interconnection among carriers. Maintaining QoS across different administrative domains in packet networks is quite difficult. Various solutions to this, such as assigning reservations costs (see, generally, Peha *et.al.* [14, 17] seem to be appropriate to circumvent opportunistic behavior on the part of competitors[4].

As current telecommunications carriers including begin deploying packet-based voice technologies in their networks, they are likely to be increasingly interested in various QoS mechanisms. The set that we have studied here are all "pure" implementations. Current engineering discussion is on "hybrid" approaches that are able to adapt to large scale networks. Such hybrids include carrying IP traffic over ATM channels to improve quality, and using different technologies in the edge networks than in the backbone networks to improve scalability. Examples of the latter include RSVP-DiffServ-RSVP, RSVP-IP/ATM-RSVP, IP/ATM-DiffServ-IP/ATM, and DiffServ-IP/ATM-DiffServ, where DiffServ is essentially the IP/Priority approach discussed in this paper, and IP/ATM is IP over ATM.

These hybrid techniques are likely to occur in various forms when networks are interconnected but under differing administrative domains. Technically, engineers have to solve several problems for such interconnected networks, including:

- How to deal with potential QoS losses as packets traverse networks using differing QoS technologies;
- How to develop techniques that scale to large networks, high speed networks, networks with large numbers of users, networks with high connection setup/teardown rates, *etc*; and
- How diverse networks can perform basic (*eg.*, call setup/teardown) functions and more sophisticated functions (*eg.*, QoS signalling).

These are difficult problems that have potentially significant cost implications, none of which we explored in this paper.

References

[1] Arango, M., Dugan, A., Elliott, I., Huitema, C., and Pickett, S. Media gateway control protocol (MGCP). Internet Draft, Internet Engineering Task Force, February 1999. Work in progress.

[2] Baldi, M., Bergamasco, D., and Risso, F. On the efficiency of packet telephony. In *In Proceedings of the 7th International Conference on Telecommunications Systems* (Nashville, TN., March 1999).

[3] Claffy, K. C., Miller, G., and Thompson, K. The nature of the beast: recent traffic measurements from an internet backbone. http://www.cacida.org/Papers/Inet98/index98, April 1998. Proceedings of INET'98.

[4] Clark, D. D. A taxaonomy of internet telephony applications. *Internet Telephony Consortium* (1997), 51.

[5] Duffy, F. P., and Mercer, R. A. A study of network performance and customer behavior during direct-distance-dialing call attemps in the U.S.A. *Bell System Technical Journal 57* (January 1978), 1–33.

[4] For example, a carrier could reserve large pieces of capacity on their competitors network that they have no intention of using. This would effectively raise the cost of the competitor if adequate performance is to be maintained.

[6] Gruber, J. G., and Le, N. H. Performance requirements for integrated Voice/Data networks. *IEEE Journal on Selected Areas in Communications SAC-1*, 6 (December 1983), 981–1005.

[7] Handely, M., Schuzrinne, H., and Schooler, E. SIP: Session initiation protocol. Internet Draft, July 1997.

[8] Hwang, J. Guaranteeing the Delay-Bound of Internet Telephony in the Best-Effort IP Networks: A Tutorial. Technical Report, University of Pittsburgh, March 1999. Work in progress.

[9] International Telecommunication Union. Visual telephone systems and equipment for local area networks which provide a non-guaranteed quality of service. Recommendation H.323, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, May 1996.

[10] International Telecommunication Union (ITU)-T. Specifications of requirements of a switching system. Recommendation GAS 6, ch. VI, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, 1981.

[11] Internet statistics and metrics analysis: Engineering data and analysis. http://www.caida.org/ISMA/isma9809/report.html, September 1998. Workshop Report of ISMA'98.

[12] Maxemchuk, N. F., and Lo, S. Measurement and interpretation of voice traffic on the Internet. In *Conference Record of the International Conference on Communications (ICC)* (Montreal, Canada, June 1997).

[13] Mier, E. Voice-over-IP: Getting started. *Business Communications Review 28*, 5 (May 1998), 18.

[14] Tewari, S., and Peha, J. Competition among telecommunications carriers that offer multiple services. Presented at the Telecommunications Policy Resarch Conference, October 1995.

[15] Thompson, K., Miller, G. J., and Wilder, R. Wide-area internet traffic patterns and characteristics. http://www.vbns.net/presentations/papers/MCItraffic.ps, December 1997. An abridged version appears in IEEE Networks, November/December 1997.

[16] Tracey, L. Voice over ip: Turning up the volume. *Telecommunications 32*, 3 (March 1998), 28.

[17] Wang, Q., Peha, J., and Sirbu, M. Dynamic pricing of integrated services networks. Presented at the Telecommunications Policy Resarch Conference, October 1995.

[18] Weiss, M. B., and Hwang, J. Internet telephony or circuit switched telephony: Which is cheaper? http://www2.sis.pitt.edu/ mweiss/papers/itel.pdf, September 1998. Presented at the 26th Telecommunications Policy Research Conference, Alexandria VA.

[19] Weiss, M. B. H., and Hwang, J. Internet vs. Circuit Switched Telephony: Cost and QoS of Large Scale Integrated Service Networks. Technical Report, University of Pittsburgh, November 1998. Work in progress.

Appendix A: Detailed Modelling Parameters

Attributes	CBR	rt-VBR
Traffic Parameters:		
PCR, CDVT	Specified	Specified
SCT, CDVT, MBS	NA	Specified
QoS Parameters:		
Peak-to-Peak CDV	Specified	Specified
Max CTD	Specified	Specified
CLR	Specified	Specified
QoS Requirements:		
	Low CDV Moderate Loss	Moderate CDV Moderate Loss

Table A-1: ATM Layer Service Category Attributes for Telephony Application

Itel Network Parameters	Values
Packet Switch	IP Switch (>20 Gbps and 2 MPPS)
Fraction of outgoing call	0.1
Originated Traffic per line	0.1 Erlangs
Packet Payload Size	20 bytes
Protocol Overhead	40 bytes (RTP/UDP/IP)
Packet Voice Burst Distribution	burst 350 msec, silence 650 msec

Table A-2: Assumptions for Itel Simulation Model

RSVP Parameters	Values
T_{spec}	
Peak IP datagram rate P	24 kbps
Token Bucket Rate r	24 kbps
Token Bucket Size b	60 bytes
Maximum Voice Packet Size M	60 bytes
Minimum Policed Unit m	60 bytes
Ad_{spec}	
Average Number of Hops H	5
Total C	300 bytes
Total D	20 msec
Delay Budgets	
$D_{Terminal}$	45 msec
$D_{Reassembly}$	50 msec
$D_{propagation}$	20 msec
$V_{queuing}$	35 msec
R_{spec}	
Required Reserved Service Rates R	82.3 Kbps

Table A-3: RSVP Parameters: 20 Byte Packetization with DS0 Line

VOATM Network Parameters	Values
General	
Cell Switch	ATM Switches (>20 Gbps)
ALL Payload Size	20 bytes (G.729a)
Rate Control Parameters for Voice at the Core Switches	
PCR (CBR)	910 Kcells/sec
PCR (VBR)	453 Kcells/sec
SCR (rt-VBR)	184 Kcells/sec
Rate Control Parameters for non-Voice at the Core Switches	
PCR (nrt-VBR)	354 Kcells/sec
SCR (nrt-VBR)	236 Kcells/sec
PCR (ABR)	59 Kcells/sec
MCR (ABR)	10 Kcells/sec
PCR (UBR)	120 Kcells/sec

Table A-4: VOATM Network and Traffic Control Parameters for Baseline ISN Network

Integrated Traffic Parameters	Values
HTTP/TCP: From Server to Client	
Percentage of Packets	38
Percentage of Bytes	70
Percentage of Flow	35
Average Packet Length	791 bytes
Major Packet Length	1500, 40, 552
Packets per Flow	14-18 packets
Average Flow Duration	10-15 seconds
Average Size per Flow	11 KBytes
HTTP/TCP: From Client to Server	
Percentage of Packets	38
Percentage of Bytes	8
Percentage of Flow	35
Average Packet Length	83 bytes
Major Packet Length	40
Packets per Flow	14-16 packets
Average Flow Duration	10-15 seconds
Average Size per Flow	1 KBytes
FTP and Other TCP	
Percentage of Packets	9
Percentage of Bytes	17
Percentage of Flow	5
Average Packet Length	600 bytes
Major Packet Length	40, 1500
DNS/UDP	
Percentage of Packets	5
Percentage of Bytes	2
Percentage of Flow	15
Average Packet Length	165 bytes
Major Packet Length	40
Packets per Flow	2-3 packets
Average Flow Duration	15 seconds
Average Size per Flow	500 Bytes
RTP/UDP	
Percentage of Packets	10
Percentage of Bytes	3
Percentage of Flow	10
Average Packet Length	401 bytes
Major Packet Length	40, 1500
Packets per Flow	50 packets
Average Flow Duration	20-30 seconds
Average Size per Flow	21 KBytes

Table A-5: Summary of Integrated Traffic Parameters