IP Telecommunications QoS (Quality of Service)
Is Service Quality a Sustainable Metric?[1]

Terrence P. McGarty[2]

Abstract

This paper is an attempt to bridge the gap between the issues of service quality as perceived by the Internet community wherein the maxim is "every packet is an adventure" to that of the telecommunications community wherein they frequently have to provide monthly service metric reports to regulatory bodies as regards to the service level that they provide the end user. In the Internet world there is no reasonable measure of end user service quality wherein the more classic telecommunications world the issue of service quality is generally both well understood and well managed.

Contents

---

[1] Presented at MIT ITC Meeting, at Telecom Italia, L'Aquila, Italy, June 8, 1999.

[2] Dr. McGarty is with The Telmarc Group, 24 Woodbine Road, Florham Park, NJ, and a member of the Steering Committee of the MIT ITC, mcgarty@rpcp.mit.edu. He is also CEO of Zephyr Telecommunications.

1. INTRODUCTION

IP telecommunications is a broadly defined set of services that include the standard set we know as the Internet service set as well as the expanding set including the standard telephony set, such as voice. In the telephony world, voice has a well established set of metrics for the determination of service quality. The measurement called the Mean Objective Score, MOS, is a standard process wherein the psychometric measurements are made as to speech quality and then the system parameters such as echo, signal distortion, loss, and other factors can be measured and if they are in a certain range then the QoS can be guaranteed to be within a certain window.

This paper presents a set of tools and methodologies wherein the service quality for IP telecommunications can be developed. It should be pointed out that the IP world view is dramatically different than the telecommunications world view. This difference is a powerful barrier to effective communications between the two communities. This paper will attempt to address this issue and then with that fundamental philosophical difference allow for a reinterpretation of service quality in the IP world.

The issue of service quality is very complex. First, generally the service quality is generally reflected ultimately in the quality of the service offering provided the end user, not necessarily the metric as measured within the network. The metrics measured in the network, albeit well defined measurements that can be both measured and generally controlled are reflected as end user metrics through what is termed a subjective or psychometric tool. Namely, voice quality is in the "ear of the beholder". Thus several naive users may be used as a test ensemble and they are asked what the level of service is as one modifies some of the well understood network metrics such as delay, echo level, channel isolation, and other similar metrics. Then the subjective metric, say the MOS, is determined and the MOS is then correlated to each of the metrics on a statistical basis. The network engineers are then told to keep the network at the measurable network levels wherein the subjective levels can be guaranteed. One never measures subjective values "on the fly" rather they are measured in benchmark levels and then projected to metrics that can be measured "on the fly".

As one evolves into a global IP platform, the issue then becomes one wherein the question asked is what metrics in an IP network are important and in turn what values are acceptable for those IP metrics to ensure that the subjective end user levels are met.

The challenge in an IP environment for service quality is several fold:

1. ***Services:*** What are the service descriptions that will be provided. One know voice, one know web browsing, one is familiar with web video and web audio. The latter two are poor quality now but they may have quality standards applied. There are metrics for broadcast audio and broadcast video. Can similar standards be applied for IP base video and audio, or is it too early. In the case of new services, what are the service metrics, how can they be determined, and who specifies them.

2. ***Metrics of Services:*** The service metrics are generally subjective and psychometric. We know voice, video, and broadcast audio. We know voice in the context of an ITU international environment. We have different video standards, PAL, CECAM and NTSC, for example. Will there be national standards or should they be international. When should these standards evolve and in what manner.

3. ***Correlates of System Metrics with Service Metrics:*** What are the system metrics. The IP and ATM world are generating what they call QoS system metrics which in many cases the engineers doing so believe that they are the end of the process. How do we correlate them with the service metrics and who specifies that. In many telecommunications interconnection agreements service quality is determined by system quality and remedies are available if the provider fails to meet the levels specified. How do we monitor, manage, and in turn incorporate these into the IP interconnection world.

4. ***Management of System Metrics:*** What process is created for the management of the standards and of the measurements. This is a process which is actually a dialectic, one of almost Hegelian dialectic of thesis, antithesis and synthesis. Where is the venue for this process, how do network providers provide overall end to end management, and what transparency is required.

We address these issues from both a top down and bottom up view in this paper.

2. ARCHITECTURAL EVOLUTION

The strength of the Internet has been in specifying as little as possible and specifying that little extremely well. IP and transport interfaces (the thin middle of the Internet hourglass model) must be stable to maximize connectivity. At the same time, there are few or no technological barriers to innovations at the link and content layer. Public policy should continue to promote this balance. This is the key to the Internet and corresponding IP architecture. The essence of this architecture is minimalism.

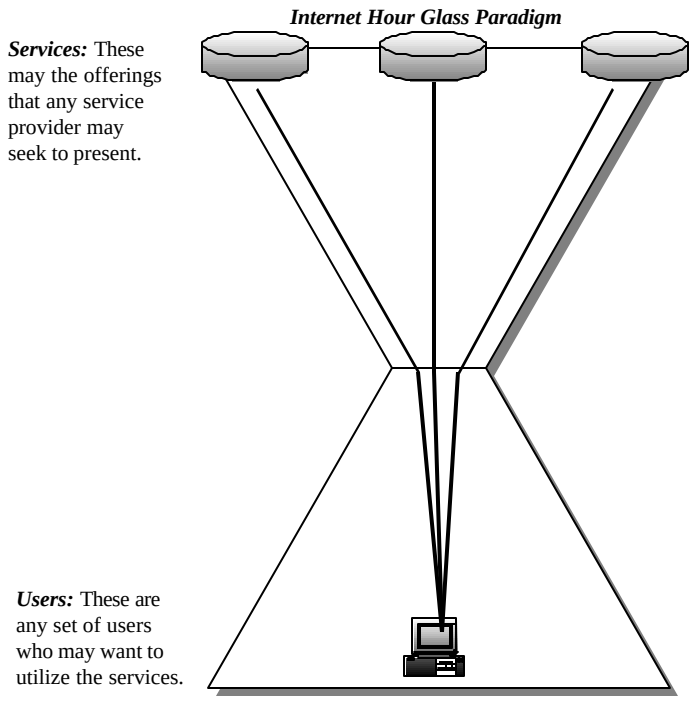
However, there are many areas where this hourglass approach, minimalist as it may be, clearly does not address the need for the broader interconnection and interfaces with the other telecommunications networks throughout the world. There is a set of issues wherein there are many questions that can be posed and that the answers to these questions in many cases will require further development of the Internet and the political landscape. However, these questions in and of themselves provide basis for understanding the directional pushes and pulls upon the Internet.

The technological issues relating to broad based interconnectivity are:

- Common channel signaling (issues in integrating the phone and Internet networks): How does the Internet, wherein IP is the lingua franca, deal with the proliferation of signaling languages that proliferate in normal telecommunications networks.
- Naming and address portability: How does one deal with the issues of naming and addressing across networks and moreover how does one deal with the issues of the naming and addressing as portable elements, that are not geographically fixed, but have virtually in a broad construct.
- Multi-tier QoS (service offerings): The issue of Quality of Service, Level of Service and Grade of Service, will dominate the evolution of Internet II as well as the evolution of new IP based networks such as those proposed by Bell Atlantic and the AT&T and British Telecom joint venture. Will the Internet evolve into the network of last resort if QoS, LoS, GoS are better on private IP based networks. Is this threat to the Internet or will their be a natural tiering of such Service grades.

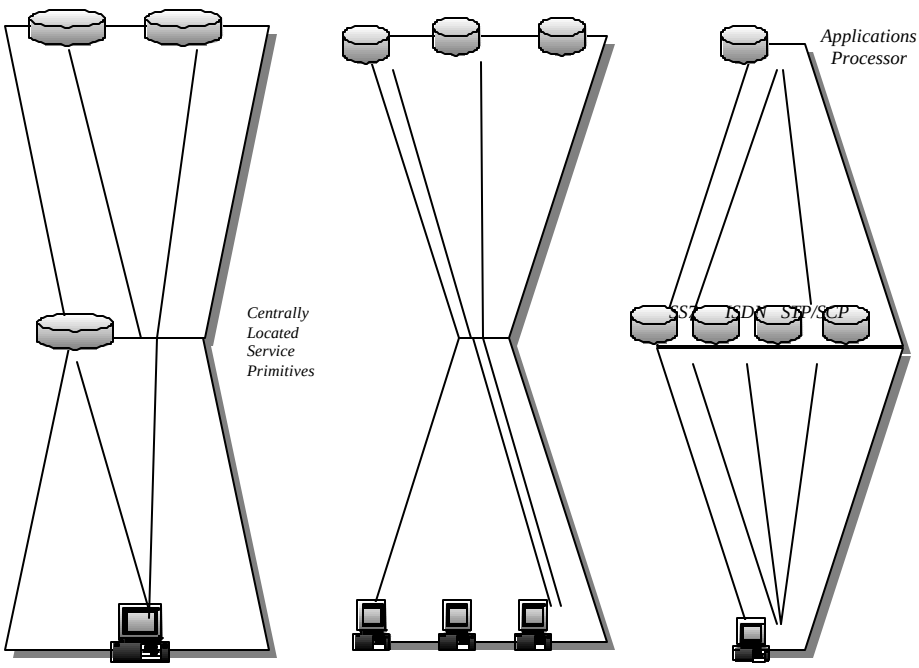
2.1 THE HOURGLASS

The view of the Internet that has been developed is the concept of the Internet as “hourglass”. The hourglass approach has connected Applications and Technology. In this report we present the model of the Internet as connection of Services and Users. The connection is via the IP based center of the hourglass. This is NOT a block or bottleneck to communications but a representation of a minimalist set of interfaces. The following depicts the general view of the Internet. The IP protocol sits at the center and allows for the connection of the user to a set of services that allow the use of any and all possible applications that the user may seek to have access to. The applications may be effected by the interconnection of the user to any sets of services in a connectable fashion.



In contrast the telecommunications architecture is dramatically different than that of the Internet. The telecommunications approach places sets of the architectural elements within the network. The access to these elements may be through a set of primitives which can be combined, managed, controlled, effected by the applications provider and then allowing the user in a similar way to get access to the service provided in the overall application. In the telecommunications environment, the end users are generally "dumb" devices as regards to the communications and the devices may be more sophisticated at higher levels.

A third evolution is one which compares the telecommunications environment to the IP environment. A third is the IP "private" carrier which may, in addition, place service elements in their network that enhance the service. These service elements may be QoS elements, caches, and other enhancements that may make the service provision more useful and efficient. However, this third approach is best understood as one which can be best delivered over a private or semi-private IP network. This is shown below.



In the next section we take this issue of architecture and expand it for a better understanding of the Internet and its corresponding expectations of performance.

2.2 INTERCONNECTION ALTERNATIVES AND ISSUES

There are three general views of interconnection that are valid today; the Telecom, the Computer Scientist, and the User. The Telecom view is based on the assumption of voice based transport with universal service and the assumption of the inseparability of interconnect and control.[3] The Computer Scientist view is based upon the assumption that the network, as transport, is totally unreliable, and that computer hardware and software must be used in extremis to handle each data packet. Furthermore the Computer Scientist's view of the network is one where timeliness is secondary to control.

The Computer Scientists view has been epitomized in the quote, "Every Packet is an Adventure". This is said with glee, in that each data packet is set out across the network and it is through the best of hacking that the Computer Scientist saves the packet from the perils of Scylla and Charybdis. This is why we are so frequently driven by the QoS issue. The third view is that of the user, who is interested in developing an interconnect capability that meets the needs and minimizes cost.

In the current telephone system, the interconnect element of the architecture is provided by the Central Office Switch and the physical interconnection of the wires from the street to that switch. The point a which the many wires from the street meet the switch are at a device called the Main Distribution Frame (MDF).[4] The Frame must be able to connect any incoming wire to any outgoing wire. The MDF, as it is called, has been the same for over fifty years. It is a manually connected system , where the craft person must connect each incoming telephone wire to a corresponding location on the switch, each time a customer moves or changes their phone number. In computer systems, this is all done in an electronic fashion.

[3] See McGarty From High End User to New User, Harvard, 1995.

[4] See Freeman.

In contrast, the central processing unit in computers goes through changes once every two years. The standard processing capacity curves show a doubling of processing capability in the same two year period. Computer users have a more rapid turnover of technology because they generally work in an environment with no regulation, shorter depreciation schedules and a focus on meeting specific business needs.

Consider what was written by a Bell System polemicist in 1977 at the 100th anniversary of the Bell System at MIT. The author was ***John R. Pierce, Executive Director at Bell Labs***, who stated:[5]

" Why shouldn't anyone connect any old thing to the telephone network? Careless interconnection can have several bothersome consequences. Accidental connection of electric power to telephone lines can certainly startle and might conceivable injure and kill telephone maintenance men and can wreak havoc with telephone equipment. Milder problems include electrically imbalanced telephone lines and dialing wrong and false numbers, which ties up telephone equipment. An acute Soviet observer remarked: "In the United States, man is exploited by man. With us it is just the other way around." Exploitation is a universal feature of society, but universals have their particulars. The exploitation of the telephone service and companies is little different from the exploitation of the mineral resources, gullible investors, or slaves." .[6]

This was written nine years after the Carterfone decision and five years before the announced divestiture. Pierce had a world view of an unsegmentable telephone network. This paper has the view of a highly segmentable communications system. The world view of the architecture has taken us from "slavery" of Pierce to the freedom of the distributed computer networks of today. Kuhn has described technologists as Pierce as the "Old Guard", defenders of the status quo. They defend the old paradigms and are generally in controlling positions for long periods of time.

However, since we now have two world colliding, namely the Telecom hierarchical network with well defined and controlled QoS and the Internet with “catch as catch can” QoS we have a major concern that possibly the interconnection and lack of balance of QoS can lead to instabilities. That in fact Pierce was correct, not necessarily in the way he stated, but in the principles that he was building upon.

Consider the following. Recently an Internet Telephony Company, ITC, terminates on local switches in Korea. The local network had problems on the trunk side and the normal International Carriers dropped. Traffic increased on the ITC. The ITC connects to multiple local switches. The local switches see dramatically increased loads and are over loaded. The local network is then put in a failure mode and the national network crashed. The network is traffic engineered at the local and trunk level. If local traffic is now trunk traffic, namely coming from the Internet but really trunk like in characteristics, then the load balancing that was done by the Telco is no longer valid. Network instability is strongly possible.

Consider the following case:

1. *There are two telecommunications networks; Local Telecom Carrier and Internet Telecom Carrier.*

2. *Local Telecom Carrier get traffic from International gateway, send it to its Tandem Network, and connects to its local network.*

3. *Internet Telecom Carrier get traffic from International Internet Gateway, via Tandem network of Local Telecom Carrier, and then sends it back to local Telecom Carrier via local switch.*

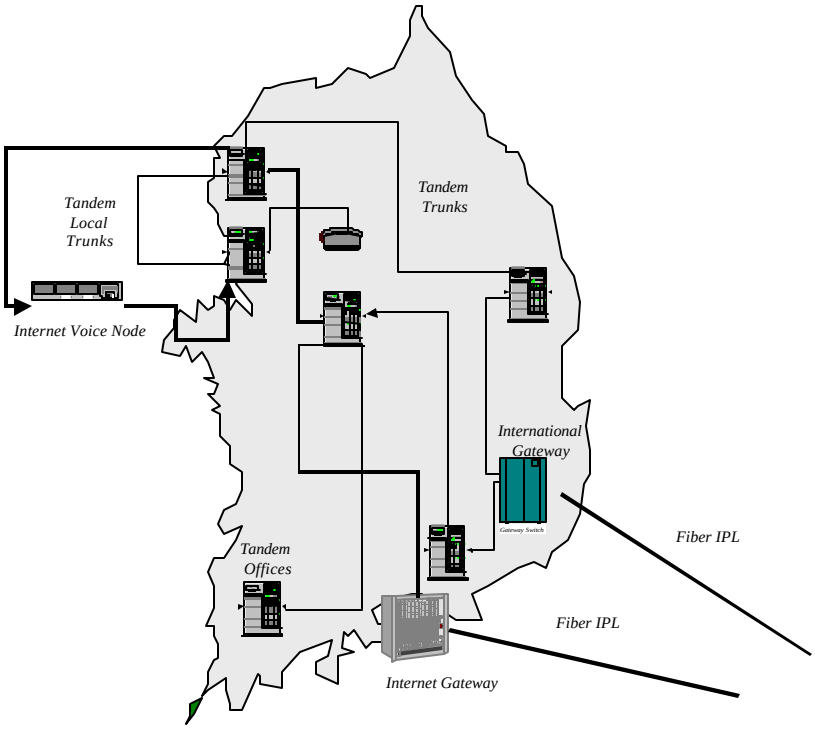
---

[5] See McGarty, Internet Architectural and Policy Implications, Harvard, 1993, and Alternative Networking Architectures, Harvard, 1990.

[6] *de Sola Pool Ed, Pierce, pp 192-194*

4. *This causes significant potential for overloading local switches which were traffic engineered for traffic coming from Tandem Network and not from the Local Network as if it were tandem traffic.*

The following Figure depicts the typical connections.



We argue that this clash of world views, architectures in the QoS space, can and most likely will lead to network instabilities and dramatic losses in traffic and network handling capabilities.[7] Namely, there are "loops" which are positive feed back loops which may result in system wide instabilities. For example, if the maximum load capacity of the Class 5 Local Switch, CL5, is 100,000 call attempts per hour, then if the tandem network, which is the load balancing network, fails, the Internet Telephony players will "dump" large loads on the CL5 network, and as one CL5 fails, the Internet automatically routes to the next CL5 thus disabling it, but increasing the overall load on the remaining CL5s dramatically. This "house of cards" approach leads to overall network instability.

2.3 ARCHITECTURE AND ELEMENTS

Based upon the above analysis we can now compare the stand Telecom network with that of an IP carrier. Specifically, we compare both a shared and dedicated IP network.

[7] The reference by Arthurs and Stuck clearly detail an analysis of the stability problem of networks and queues. We argue here, without proof, that if one looks at the telephone network and sees that the local switch is typically the least stable node, then the Internet Telephony world presents a potentially clear and present danger to global network stability.

Element	*Telco*	*IP Shared (Internet)*	*IP Dedicated*
Control Elements	In the network	At the boundary	Both at boundary as well as in the network. "Value Added" may be placed in network by carrier such as caching.
Control Processes	Service provider and is fully end to end.	At best it is via end to end protocol overlay.	Service provider intervention proposed.
Service Descriptions	Controlled, defined, managed by network architecture.	Defined by the end user.	A combination of network and user defined.
Traffic Management & Engineering	Carefully balanced by trunk and local line engineering.	None	Some by provider.
Grade of Service	Defined, measured, monitored and reported.	None performed.	TBD
Quality of Service	Wee defined and reported metrics.	None in place.	Proposed to be overlaid by protocol development.
QoS Design	Over Capacity and Over Engineered		Over Capacity and Over Engineered or Increased Overhead and Management (ATM and IPv6)
Reconfiguration	Lengthy process, no real time, and generally slow.	Real time	Quasi real time and will depend on end to end control.

3. GRADE, LEVEL, QUALITY OF SERVICE

Quality of Service, Levels of Service, Grades of Service may be characterized by various metrics.[8] There is generally is no general consensus at this time as to what metrics are the most useful for quality differentiation. However, the ITU and the IETF are currently working on such standards. There are also many government entities that have the role and responsibility for obtaining and facilitating the dissemination of the information provided by carriers. There may already be entities that have the authority to provide then end users with this information. The general position is that there should be no mandated values and further that it is generally appropriate to respond to end user concerns and that government should not mandate certain levels of performance but at most should be an information gathering and disseminating entity.

The major problem generally is defining what a QoS is and how it may relate to Levels of Services and Grade of Service. The terms have mixed understandings.

[8] See definitions in Newton's Dictionary.

Grade of Service: This, for example, is the probability that a call presented to a telephone system is actually carried by that system.[9] In more general terms, Grad of Service is the mapping between a system factor and a service factor, between some measure of packet delay and some measure of voice quality.

Quality of Service: This is a widely disputed term. In the telephony world, the QoS is generally stipulated by the Telco along with mandated values given by the Public Utility Commission, "PUC", that details such elements as time to get dial tone, voice quality, subjectively determined, time to respond to customer calls, and similar measures. In the ATM world the issues are such as cell error rations, cell loss rations, and cell delay variability. In fact there are currently 4 classes of ATM service, ranging from that of a private line to that of a connectionless data protocol.

3.1 SYSTEM AND SERVICE LEVEL ELEMENTS

The system is defined as the set of any and all elements which comprise the underlying communications fabric in support of the IP network and its environs. The system elements can be characterized quantitatively via well defined metrics or measurements. For example, the system elements may be comprised of packet delay, lost packets, bit error rate, phase jitter, and similar system elements and parameters.

The service is what the end user perceives and the service elements may be characterized by quantitative objective or subjective measurements. The Mean Objective Score, "MOS", is a subjective or psychometric measurement of service quality.

3.1.1 SYSTEM ELEMENTS

The system elements that may be measured are contained in the following Table. These represent typical ones and are not inclusive.

System Elements	*Measurement*
Packet Delay	Average Delay and Standard Deviation
Packet Loss	Average Packet Loss
Bit Error Rate	Standard BER
Phase Jitter	Phase stability and standard deviation.

3.1.2 SERVICE ELEMENTS

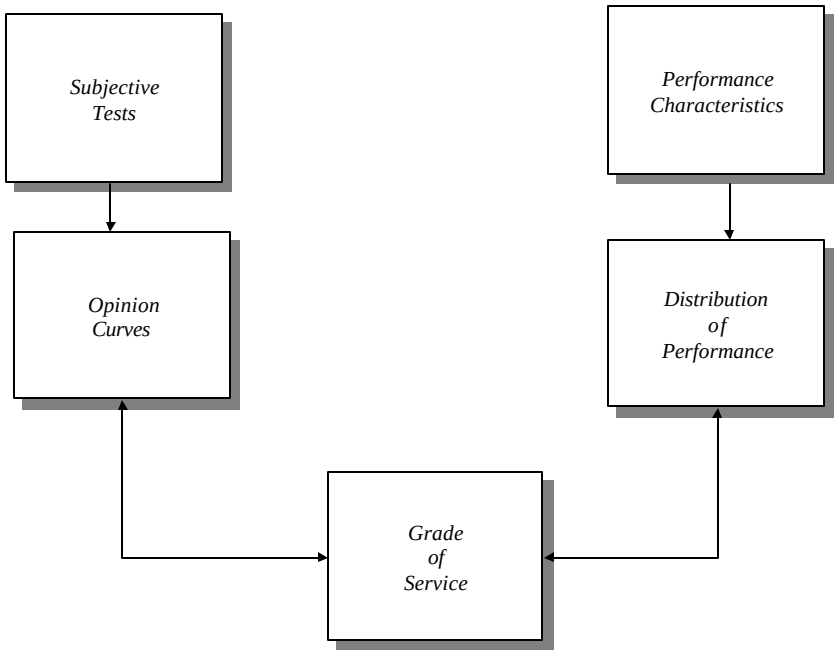
The service elements that may be measured are contained in the following Table. These represent typical ones and are not inclusive.

[9] Newton, Telecom Dictionary, p. 329, or Freeman, Telecommunications Engineering, p. 2.

Service Elements	*Measurement*
Voice Quality	MOS
Video Quality	NTSC, SECAM, PAL standards
Data Quality	BER
Fax Quality	Group 3 Standard
Multimedia Quality	Unknown
Video Conference Quality	Unknown
Completed Calls	Percent Complete, ASRs
Call Set Up Time	Average, distribution analysis
Audio/Video Synch	Standards.

3.2 GRADE OF SERVICE

Grade of Service has been defined by AT&T as follows.[10] Let R be the rating of a call by a customer in category R by a customer and let M be the system performance measurement. Let M be measured from the system. For example, in a voice call, a customer may measure a call quality, MOS score, as "4.1", given packet delay of 20 msec. The following is from the TT report and depicts the process of Grade of Service.



We can now expand on the above concept and define Grade of Service as a relationship between the system variable, such as packet delay or loss, and the service variable such as voice quality. Let x be a variable that is a system variable. For example x may be the packet delay or the noise on the circuit.

Let M be a performance measure parameter, namely a service variable, such as the MOS score.

[10] Rey, p. 674-675.

Let us assume that one can determine via psychometric testing the probability density of:

$$p_{M/x}(M/x)$$

where this is the conditional probability density of M given the system variable x.

Let us also assume that one can determine the density of x, namely:

$$p_X(x)$$

The we can determine:

$$p_M(M) = \int_{-\infty}^{\infty} p_{M/x}(M/x)p_x(x)dx$$

The on can create a Grade of Service metric, say the average M, or MOS score, given by:

$$\overline{M} = \int_{-\infty}^{\infty} Mp_M(M)dM$$

The we can say what the average M is as a function of the psychometric filter of the conditional probability and of the performance of the system by the probability density of the system variable, say the packet delay.

We can now pose the following design problem:

If $p_{M/x}(M/x)$ is known, what is the acceptable set of $p_X(x)$ such that $\overline{M} \geq M^*$.

The Grade of Service, GoS, is thus defined as:

$$\text{GoS} = \int P(R/M)\,p(M)dM$$

Namely, GoS is the expected R averaged over the anticipated M.

3.3 QUALITY OF SERVICE

Quality of Service is a term now used for the actual system level elements. There is a growing issue regarding the Quality of Service on the Internet. For the most part, the Internet was an “as is” facility, namely the user took what they got and liked it. There were and are ways around this issue but for the most part they are patch works of improvement. The issue of Quality of Service, Level of Service and Grade of Service, will dominate the evolution of Internet II as well as the evolution of new IP based networks such as those proposed by Bell Atlantic and the AT&T and British Telecom joint venture. Will the Internet evolve into the network of last resort if QoS, LoS, GoS are better on private IP based networks. Is this threat to the Internet or will their be a natural tiering of such Service grades.

3.3.1 THE DEVELOPMENT OF HIERARCHICAL ENTITIES

The development of extra-Internet entities, as may be envisioned by certain carriers, who desire to ensure a better quality of service at a higher price, may result in the ghettoization of the internet and may result in a

segmentation and fragmentation of the Internet. This may result in the establishment of separate IP networks that have restricted connectivity, which may allow for "improved" service when one agrees to be controlled by larger entities.

3.3.2 NO QUALITY-OF-SERVICE (QOS)

As noted above the Internet is more distributed and adaptive but more difficult to control if a QOS is to be achieved. In the PSTN traffic congestion is managed such that under overloads connections may not be made. In the Internet originating traffic will access the destination endpoint and receive some level of service even though that service level may not be useful – referred to as "best effort". Thus, certain applications, e.g., voice or video, may be restricted in their use unless service management capabilities are introduced to ensure acceptable performance levels. And, although higher-priced high-speed links may be made available, there is no guarantee that the unmanaged core of the network will provide high-speed throughput.

3.4 CAPACITY VERSUS CAPABILITY

QoS can be achieved by two extreme methods; overcapacity and by over capability. The latter is what is some sense is being accomplished with ATM and has been proposed in the implementation of IPv6.[11] This approach places a significant amount of controlling signals and intelligence in an overhead communications system. It tries to use the system intelligence at the periphery to maximum advantage.[12]

The former approach, over capacity, is what telecommunications systems have done for a century. Since intelligence was more difficult than capacity, there was a great deal of effort expended in the deployment of excess capacity.[13] Nothing was "cute", it was simple and it was effective.[14] Unfortunately it was also expensive. However, in today's world where capacity is significantly less expensive, albeit possibly more costly than processing, it may be more effective to over provision.

We depict these alternative below. The top embodiment is excess capacity and over provisioning. The resources used may be significant, but traffic engineering is a fairly mature discipline, and even using upper bounds to provision, one may achieve significant levels of quality improvement over protocol implementation. The second depiction shows the attempt to control the network via a signalling protocol. The intent is to minimize the data transfer pipe width while placing an additional burden on a signalling pipe. The net result, it is argued, is that in many cases the resultant resources used are comparable to those of over-provisioning alone.

4. VOICE NETWORK METRICS

Service Quality in voice networks have various elements of effectiveness.

4.1 TESTS AND METRICS

The following Table presents a set of tests, objectives, procedures, results and levels of performance for service quality in an IP telecommunications world wherein the voice service quality is of concern.

---

[11] See Gai regarding Cisco implementation issues.

[12] See Stallings.

[13] See Arthurs and Stuck for many detailed examples.

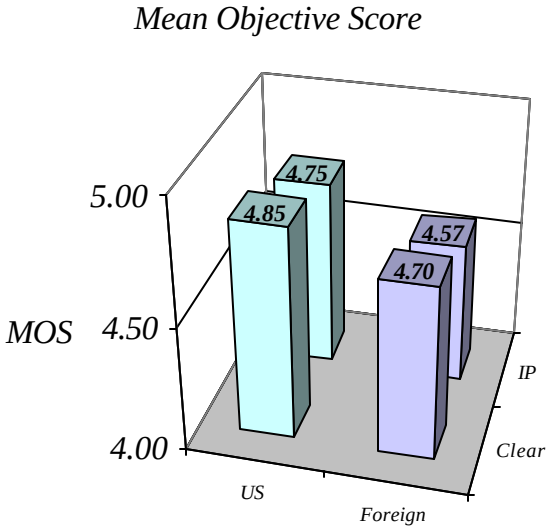
[14] The issue of quality of service and grade of service in a telecommunications network has been detailed in Rey.

Test	*Objective*	*Procedure*	*Results*	*Level of Performance*
Call Set Up Time	To test the time to set up a call measured from the time the last digit is entered until time end party ringing is commenced.	This will be an A/B test procedure using the circuit and a standard reference on a time of day basis. The Standard Reference shall be a US generated AT&T clear channel circuit. The test shall include the measurement of the time between the last digit dialed and the time of the commencement of end number ringing. The test shall include measurements for 50 calls of Type A and B.	The results shall display the histogram of calls.	<25% difference
Link Loading	To determine the maximum loading on each circuit to meet the Blocking Probability requirement.	This will include the loading of the circuit to its maximum handling capability by placing calls onto the circuit and determining the maximum number of simultaneous calls before blocking exceed the 5% level.		Maximum loading with blocking probability < 5%
Call Completion	To determine the fraction of calls that are terminated without problem.	This test shall consist of the placing of 100 calls in a row and determining the number which are terminated successfully. This shall be done on both ends of the circuit.		> 95% call completion
Call Blocking	To determine the call blocking probability on the circuit.	Calls shall be made at three levels of loading, 50%, 100% and 125% of maximum peak busy hour capacity and call completion shall be recorded.	The test shall report the blocking percentage at the three test points	< 5% call blocking at load
Voice Call Quality	To determine the voice quality of the circuit.	This will be an A/B test procedure using the circuit and a standard reference on a time of day basis. The Standard Reference shall be a US generated AT&T clear channel circuit. The procedure will be to place twenty calls on each end of the circuit in a double blind fashion. There will be a 50:50 mix of the Standard Reference and the company circuit. The caller will be asked to determine whether the quality was acceptable or not. Then the two will be compared for statistical significance of difference using a Student t Test.	Student t Test results and measures of difference significance.	<15% determining difference in average
Bit Error Rate	To determine the end to end bit error rate, BER, of the circuit.	Use a BER tester on the loop back circuit.	Standard BER testing.	BER < 10^{-6}
Fax Quality	To determine if the fax transmissions are acceptable.	Transmit fax ten times.		Readable fax

Test	*Objective*	*Procedure*	*Results*	*Level of Performance*
Modem Test	To determine if the modem transmissions are acceptable.	Try data modems up through 56 Kbps		Modem connection via synch.
Failure Reporting Tests	To determine if the failure reporting procedure is followed.	Failures will be generated at each end of the circuit. Calls will be placed to the NOC and the time to determine and report the trouble will be measured.		Time to Rep[ort < 15 min Failure to report rate < 5%
Trouble Tickets	To determine if Trouble Tickets are prepared properly and if the clearing process is commenced.	Trouble Tickets shall be prepared and circulated.		Time to issue shall be < 15 min.
Trouble Clearing	To determine if trouble clearing process is working.	This will entail the end to end clearing of the created trouble.		Time to clear shall be < 75 min.
Billing	To determine if billing system integrity is in operation.	Traffic shall be loaded for 48 hours from both ends of the circuit. Bills and CDRs shall be prepared.		< 1% billing errors.
Customer Care Test	To determine if customer care system integrity is working.	Calls shall be placed at random to customer care.		Time to answer shall be < 45 sec

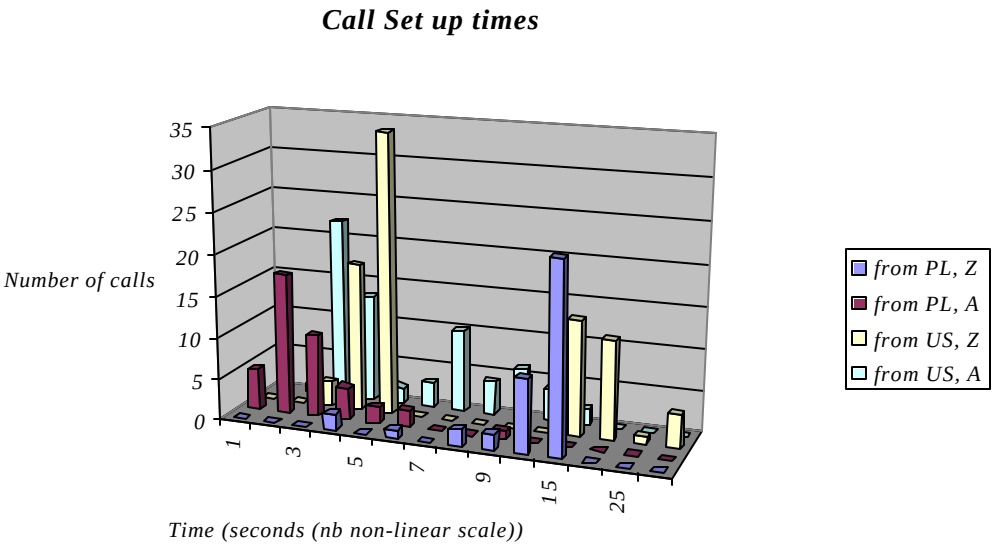
4.2 TYPICAL QOS RESULTS

Results on voice quality are shown below for a link on a clear channel IP telephony system as measured by both ends. Note that on this link we compare IP to the clear channel AT&T links. There is no noticeable difference in voice MOS, mean objective score, QOS. Several carriers have achieved this QOS at this time and we further believe that few IP carrier has even begun to test for this factor no less achieve it.

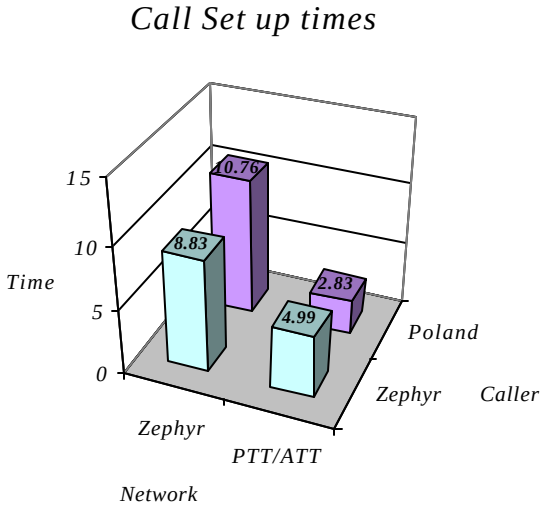


This above test shows that MOS scores for IP telephony clearly match those of the clear channel telephony whether they originate from the US or from a foreign country. There is at most a 10% difference which given the size of the sample makes its statistically insignificant.

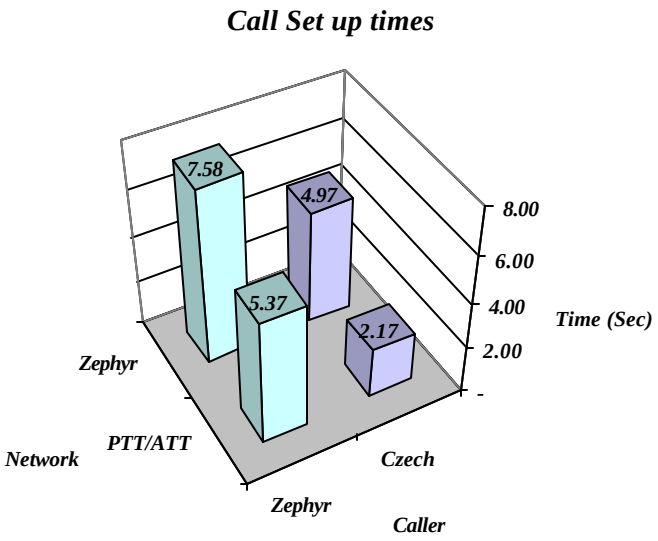
The following show results from call set up time measurements on an IP versus standard call set up procedure. Note the bi-modal characteristics.



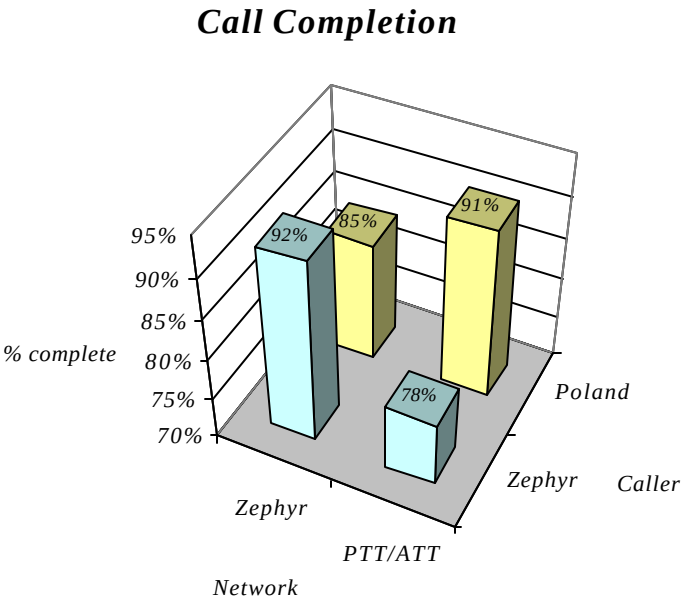
The following is the summary of these call set up tests. Namely that an IP network may have longer call set-up times. In actuality this was due to the process of having a non-PRI interface and in establishing a call by a second dial tone basis.



Calls, however, terminating via a C7 interface are shown below. Note the dramatic difference in call set up time. There are still some differences since the interface is a local switch, albeit with C7 and not an international tandem switch with C7.



The last measurement that can be shown is the call completion rates. The following Table depicts some of these rates. Note that they comparable but generally reflect the network configuration.



5. ARCHITECTURAL ELEMENTS

There are several new IP telecommunications architectures. These will each impact the way one sees QoS and how it will be applied. In this section several of them are discussed.

5.1 THE IP UTILITY

The IP Utility is a concept that private entities will generate an Internet based and connected utility or network that will guarantee a quality, level, or grade of service to the end user and will allow the end user to connect at IP level. It will further allow the end *user to interconnect to the generally accepted Internet and then also possibly to other similar IP Utilities. The IP Utility is derivative from the need by certain Internet based purveyors of services to have a more reliable, more secure, more predictable, network that can be currently afford by the general Internet. This approach as has already been set forth by companies such as Bell Atlantic and the proposed AT&T BT joint ventures.

The IP Utility concept is a compartmentalization and hierarchialzation of the Internet. It is based upon the premise that there is a need and demand for IP connectivity as contrast to just bandwidth or per minute clear channel connectivity. It further assumes that there is a need for a Quality, Level, or Grade of Service, QoS, LoS, or GoS, wherein the purchaser would want to communicate with many others who have access to this better and more reliable network element. This allows the owner, a commercial and singly controlled entity to charge a price for access that reflects the costs of the increase in service. The question is does this evolution create a threat for the natural evolution of the more public network structure. Is this evolution form the better but more costly Internet extension good for the Internet as a whole, is it competition, or furthermore is it a threat?

Internet II allows for certain improvements of Quality of Service ("QoS") and the proposed private networks by the likes as AT&T/British Telecom and Bell Atlantic portent to provide the improvements that many users are seeking.

The businesses market clearly would place a financial premium for the Internet's valuable ubiquity but such a user base cannot obtain adequate performance on that network.[15] It is essential to improve the Internet Protocol (IP) with quality of service (QoS) mechanisms that can support multiple performance-differentiated services.

ATM has developed a methodology for improving this service quality issue and it is a QoS feature now common in ATM networks. However, altering a deployed network protocol such as IP is highly complex. The networking industry's long development of frame relay and ATM technology has provided many of the mechanisms required for implementing QoS in IP networks. IP can now be carried over frame relay and ATM networks, thereby leveraging their QoS mechanisms, but the Internet Engineering Task Force, the Internet's standards body, is also borrowing these QoS concepts and transforming them to operate natively on IP traffic.

In the detailed article by Kaufman, he states, "The Differentiated Services (DiffServ) Working Group has been chartered to define these new mechanisms and encode them in a set of standards for all vendors to implement. The new mechanisms will bring QoS to all IP traffic, whether transmitting over leased lines, frame relay, ATM or even new packet-over-SONET links. They use the existing IPv4 protocol definitions with full backward compatibility for existing routers, end stations and applications. They even enable new highly

[15] Kaufman, D.H., Delivering Quality of Service on the Internet, Telephony Magazine, February, 1999. This discussion parallels Kaufman's comments in detail.

dynamic QoS services that were not possible with pure ATM- or frame relay-based mechanisms and deliver new congestion management tools to cope with the continued growth of IP applications and traffic."

5.2 THE ATM QOS MODEL

ATM implements QoS by defining a set of service classes. For example: constant bit rate (CBR), variable bit rate real time (VBR-rt), VBR non-real time (VBR-nrt), and unspecified bit rate (UBR). The service class defines parameters for acceptable cell loss and delay variation through the network. As each virtual connection (VC) in an ATM network is created, it is given an associated service class. Each VC also has a traffic contract, which specifies how much data may be sent on a given VC. A traffic contract specifies three parameters to accommodate data's inherent burstiness:

- *Burst rate (the maximum permitted throughput);*
- *Sustained rate (the minimum rate guaranteed from the network); and*
- *Maximum burst size.*

Service providers offer QoS-differentiated services based on the different service classes and varying bandwidth guarantees in the traffic contract. Service classes and traffic contracts are enforced using input policers and output schedulers.

ATM input policers measure input traffic rates and identify traffic that is violating the contract. Traffic that exceeds maximum burst size can be dropped by the policer at ingress or be marked for dropping later on congested links by setting ATM's cell loss priority bit.

ATM output schedulers manage the traffic from the different service classes. Traffic from each class is generally queued separately, but traffic from multiple queues must be combined and transmitted from a common output port, while preserving the bandwidth and delay guarantees of the particular service class.

5.3 ADAPTING QOS TO IP

Because IP is "connectionless" and without traffic contract concepts, the DiffServ group had to solve the problem of how to mark IP data packets with both traffic class and traffic contract information. Fortunately the IPv4 packet header had a largely unused type of service (TOS) field that could be reused with minimal impact. It is redefined as the DiffServ field and carries information analogous to QoS class and traffic contract compliance, defining acceptable packet delay and loss, much like the ATM service classes. Because the DiffServ field is 6 bits wide, it allows a set of encodings richer than ATM's single-bit cell loss priority field.

The information encoded in the DiffServ field is called the per-hop behavior (PHB) and refers to the treatment that the packet expects to receive in transit. Each DiffServ-capable router examines the packet's DiffServ field and implements the expected PHB. DiffServ routers implement PHBs using input policers and output schedulers very similar to those in an ATM switch. They also add congestion management mechanisms specific to IP:

IP input policers/classifiers measure input traffic rates and identify packets that are out-of-contract, enforcing both sustained and burst rates. Out-of-contract or burst packets can be marked with a new PHB or dropped. IP policers/classifiers also include traffic classifiers that replace the QoS-class association inherent in an ATM VC. A packet's physical port, IP source address, IP destination address or TCP/UDP ID can be used to assign, modify or verify a particular PHB for the packet. IP output schedulers include priority queues and traffic shapers that schedule packets for output based on traffic contracts. To avoid traffic overloading, IP routers must also implement congestion management algorithms. Congestion management algorithms help routers cope with the traffic congestion on the Internet. Random early detection (RED) and weighted RED (WRED) are two proven algorithms that detect impending congestion and drop randomly

selected packets before the router becomes overloaded. These intentionally dropped packets cause IP's Transmission Control Protocol (TCP) to slow the sender's transmission. (TCP operates end-to-end between applications.) Because IP's QoS classes must be apparent even through congested networks, WRED adjusts the discard parameters so that "better-QoS" packets are far less likely to be dropped when congestion occurs.

5.4 PER-HOP BEHAVIORS

Service providers are already offering performance-differentiated services based on DiffServ. At the same time, the DiffServ Working Group is defining standard PHBs, which IP router vendors will implement. Standardized PHBs will ensure compatibility between service providers, enabling standard services across the Internet. DiffServ is standardizing the following three PHBs. They range from best effort (no guarantees) to an IP equivalent of ATM CBR (very strict guarantees). Best effort standardizes the default QoS available today. Best effort has no traffic contract, so best effort traffic gets whatever bandwidth is left over after the other PHBs have been processed. Assured forwarding delivers a guaranteed sustained rate, with bursts up to a defined maximum. Input policers mark the burst packets and out-of-contract packets with different values of the DiffServ field, so that they can be dropped if congestion occurs. Burst packets allow statistical multiplexing (like ATM's VBR service classes), without rigorous end-to-end traffic engineering to guarantee bandwidth and delay. Expedited forwarding delivers guaranteed bandwidth, delay and packet loss based on ATM CBR-like traffic engineering. Input policers drop all out-of-contract packets at ingress, and packets are shaped on egress. Expedited forwarding enables a virtual leased line service that delivers the benefits of a traditional leased line--guaranteed bandwidth and minimal delay--but at a lower cost, because it is carried over the shared IP network.

IP QoS has additional flexibility that is not possible in ATM's strict circuit-based design. DiffServ supports dynamic QoS based on time of day, application or user identity. For example, businesses might give traffic from a preferred customer (or the CEO's workstation) a higher QoS. Similarly, Web-based e-commerce might be given priority by giving it an assured forwarding QoS, while e-mail and other less-critical traffic could be left as best effort. Service providers will also offer off-peak QoS, at lower rates during off-peak times. These services are possible because IP QoS classifiers use existing information in the IP header, not new connection-time control protocols.

5.5 END-TO-END IP QOS

DiffServ implements QoS for service provider backbones. Complete end-to-end IP QoS requires integrating DiffServ with the LAN-oriented QoS mechanisms used by enterprises. Traffic measurements show that most end-to-end IP connections are very short-lived and that there are several thousand active connections at any time in a backbone router. DiffServ's simplicity allows IP QoS to scale into the backbone, while still leveraging more granular QoS in enterprises. Several different QoS mechanisms are popular in the LAN, including Resource ReSerVation Protocol (RSVP) and link-layer protocols such as Ethernet 802.1p. QoS integration involves translating LAN QoS into DiffServ PHBs. Some translations are simpler to implement than others, depending on how dynamic the LAN QoS is. QoS integration takes place in a business context: Enterprises must negotiate a menu of services with their service providers. Some traffic will be statically steered into a DiffServ PHB, while IP QoS-aware applications will signal QoS requests to the service provider, using RSVP or other protocols. This configuration allows corporations to assume control over the QoS they request from the service provider for each individual application.

Service providers already deploy border routers to connect to customer sites. In DiffServ, the border router performs QoS integration and receives QoS signals from enterprise applications. Static configuration implements permanent QoS assignments based on IP traffic classifiers. Each LAN QoS is translated into a DiffServ PHB marking, which is carried through the backbone.

Translating LAN QoS into DiffServ PHBs also simplifies mapping IP traffic onto ATM VCs. Most service providers use ATM in their backbones; by provisioning multiple VCs (with different service classes) between their routers, the service provider can steer IP packets into the appropriate QoS. The DiffServ PHB is translated into an ATM service class. In this design, DiffServ acts as a QoS interworking layer rather than an actual QoS implementation.

Service providers must also maintain QoS guarantees when IP packets cross more than one service provider's backbone. Service providers will agree on service menus including the standard PHBs. The DiffServ standards allow large corporations to use multiple service providers and still receive QoS guarantees, rather than being locked into a single provider.

The DiffServ group defines PHBs that allow IP QoS to scale in service provider backbones and to integrate with pre-existing LAN QoS at border routers. IP QoS uses the mechanisms developed for ATM QoS, while adding some new mechanisms specific to IP. It allows service providers to offer new services to their customers, while its ATM heritage simplifies QoS interworking in hybrid IP-ATM networks.

5.6 ATM QOS (QUALITY OF SERVICE) LEVELS

ATM, with its multiplexing architecture, is designed to support traffic with various bandwidth, jitter, and delay requirements. This design feature allows ATM networks to support voice, video, and data multiplexed on the same links. Quality of service is established at the time that the connection is made. Implementing quality of service is dependent upon ATM being a connection-oriented protocol. The ATM Forum has defined four quality-of-service types that are architected to handle the different types of traffic.

Constant Bit Rate (CBR) and Variable Bit Rate (VBR) are particularly well-suited for supporting applications with stringent requirements for quality of service, such as multimedia transmission or high-quality videoconferencing.

5.6.1 CONSTANT BIT RATE

CBR is a reserved bandwidth service. A contract is established between the network and the end station. The end station provides the network with parameters describing the traffic for that specific connection at call setup time. The network, in turn, allocates resources that match the parameters or, if the resources are not available, rejects the call. This is called call admission control. Once the call is accepted, it is the end station's responsibility to send only traffic that is compliant with the contract. The network checks the traffic against the contract, and noncompliant cells are discarded.

5.6.2 VARIABLE BIT RATE

Like CBR, VBR is a reserved bandwidth service. The network allocates resources to the end station at call setup in response to the traffic parameters requested by the end station. However, in the case of VBR, in addition to a peak rate, a sustainable rate and a maximum burst size are established. The sustainable rate is the upper limit of the average rate, and the maximum burst rate limits the duration of cell transmission at peak rate. These additional parameters allow the network to achieve statistical multiplexing by allocating fewer resources for the connection than would be required by the peak cell rate.

In most campus environments today, the majority of traffic is data transfer that, for the foreseeable future, will operate over ATM using either LAN Emulation or Classical IP mode. These legacy applications are not able to specify the quality of service that they will require. The ATM Forum is proposing that this traffic employ either Unspecified Bit Rate (UBR) or Available Bit Rate (ABR).

5.6.3 UNSPECIFIED BIT RATE

UBR is a non-reserved bandwidth service. The cell loss ratio is unspecified, which means that the network is not required to provide resources for a proposed UBR connection. No flow control parameters are specified in the ATM Forum for UBR service. Consequently, when UBR service is employed, cell discard seriously impacts the overall performance of the system. For example, a single cell discarded in a 192-cell packet (the default size for an IP packet when using Classic IP over ATM) triggers retransmission of the whole packet. The network has transmitted 191 cells needlessly. To avoid wasting network resources in this way, early packet discard and partial packet discard can be implemented in any intermediate node (switch) of the network. If a switch recognizes that a cell has been lost, it discards the rest of the packet. If a sending station fails to acknowledge a congested condition, the incoming switch in the network will reject packets until the congestion disappears. When early packet discard and partial packet discard are implemented in conjunction with virtual circuits, fairness and hop-by-hop backpressure mechanisms ensure loss-free UBR operation.

5.6.4 AVAILABLE BIT RATE

ABR service can be seen as a mix of reserved and non-reserved bandwidth service. Periodically, a connection polls the network and, based upon the feedback it receives, adjusts its transmission rate. Polling is done by Resource Management cells sent by the source and looped back at the destination so that the network elements and the destination can provide feedback information. In addition, network elements can create and insert RM cells in the backward direction to provide feedback to the source more quickly.

6. CONCLUSIONS

This paper is a first attempt to establish a methodology for the presentation of service quality in the IP world. The challenge will be several fold. First, with the opening of global telecommunications to competition, service quality will generally be relegated to a third or fourth place as regards to price and market share. This may be acceptable for the initial phase of deregulation but as is well known the consumers will soon rebel against low cost and poor quality.

Second, quality can be regulated by one of two extremes; regulation and market forces. Generally regulation is the poorer form of service quality management since regulators generally do not understand the rapidly changing technology environment and further and more significantly they are controlled by the larger and established players who are seeking to maintain the status quo. Thus market regulation is the better form and this can be accomplished via a well understood correlation between service quality and system quality.

Third, standards must evolve and this is the challenge between the worlds of the ITU and the IETF. The IETF has established the "libertarian" view to the network, namely stipulates as little as possible and let whatever happen occur. In contrast the ITU is the classic standards process of the embedded players who look for slow evolution and for whom time is a strategic competitive tool. Somehow there must be a rapid facilitation of the two extremes into a single integrated body the reflects the timeliness of the IETF and the concern for quality of the ITU.

Fourth, there must be a dialog and a research process for the development of metrics that allow for the establishments of the correlates of the system and service metrics and this must be accomplished in as neutral an environment as possible.

7. REFERENCES

Arthurs, E., B.W. Stuck Network Performance Analysis, Prentice Hal (New York) 1985

De Prycker, M. Asynchronous Transfer Mode, Prentice Hall (New York) 1995.

de Sola Pool, I., Technologies Without Barriers, Harvard University Press (Cambridge, MA), 1990.

Ferguson P., G. Huston Quality of Service, Wiley (New York) 1998.

Freeman, H.F. Telecommunications Engineering, Wiley (New York) 1994.

Gai, S. Internetworking IPv6 with Cisco Routers, McGraw Hill (New York), 1998.

Isenberg, David Is Quality of Service Necessary, America's Network, *www.americasnetwork.com*, 990301.htm.

Kaufman, D.H., Delivering Quality of Service on the Internet, Telephony Magazine, February, 1999.

Kleinrock, L. Queuing Systems, Wiley (New York) 1975.

Martin, J. Systems Analysis for Data Transmission, Prentice Hall (New York) 1972.

McDysan, D., D. Spohn ATM Theory and Applications, McGraw Hill (New York), 1999.

McGarty, T.P Internet Architectural and Policy Implications, Kennedy School of Government, Harvard University, Public Access to the Internet, May 26, 1993.

McGarty, T.P From High End User to New User: A New Internet Paradigm, McGraw Hill (New York), 1995

McGarty, T.P International IP Telephony, to be Published, MIT Press, 1999.

McGarty, T.P. Alternative Networking Architectures; Pricing, Policy, and Competition, Information Infrastructures for the 1990s, John F. Kennedy School of Government, Harvard University, November, 1990.

Newton, H. Telecom Dictionary, 1999.

Rey, R.F. Engineering and Operations in the Bell System, ATT Bell Labs (Murray Hill, NJ) 1983.

Stallings, W. *High Speed Networks, Prentice Hall (New York) 1998.*