Document Cover Sheet for Technical Memorandum

Title: Report on the Workshop on Quality of Service Issues in High Speed Networks

Author (Computer Address)	Location	Phone Number	Company (if other than AT&T-BL)
S. Keshav (research!keshav)	MH 2C-552	(908)582-3384	

Document No.	Filing Case No.	Project No.
11272-920826-23TM	39199-11	311407-7214

Keywords:

Quality of Service

MERCURY Announcement Bulletin Sections

CMM - Communications CMP - Computing

Abstract

High speed networks of the near future are expected to carry a wide range of traffic types, including data, still pictures, voice, broadcast video and interactive video. Each of these traffic types requires different service from the network. The design of networks that provide a good quality of service to the large variety of expected users is an open and interesting research area. In order to initiate a dialogue between researchers who work in the general area of network design and those who work at the application or user level of integrated networks, AT&T Bell Laboratories sponsored a workshop on ‘Quality of Service Issues in High Speed Networks’ held at Murray Hill, NJ on April 23-24, 1992. This TM is a report on this workshop.

Total Pages (including document cover sheet): 12

Mailing Label

Initial Distribution Specifications **11272-920826-23TM** *(page ii of ii)*

Complete Copy

Executive Director 112, 113
Directors 112, 113
Department Heads 1125, 1126, 1127, 1135
G. E. Nelson
R. Cox
B.T. Doshi
S. Dravida
A.E. Eckberg
A. Elwalid
J. Flanagan
R. Gitlin
E.L. Hahne
B.G. Haskell
P.S. Henry
C.R. Kalmanek
H. Kanakia
S.H. Low
N. Maxemchuk
D. Mitra
G. Murakami
S.P. Morgan
A. Reibman
K.K. Sabnani
K. Sriram

Cover Sheet Only

A. A. Penzias
1125, 1127, 1135 MTS
W. Ryan

Future AT&T Distribution by ITDS

RELEASE to any AT&T employee (excluding contract employees).

Author Signature

____
S. Keshav

Organizational Approval (Optional)

____

For Use by Recipient of Cover Sheet:

Computing network users may order copies via the *library –k* command;
for information, type *man library* after the UNIX prompt.

Otherwise:
Enter PAN if AT&T-BL (or SS# if non-AT&T-BL). ____________________________
Return this sheet to any ITDS location.

Internal Technical Document Service

( ) AK 2H-28
( ) ALC 1B-102
( ) CB 30-2011
( ) HO 4F-112
( ) IH 7M-103
( ) MV 1L-19
( ) WH 3E-204
( ) DR 2F-19
( ) INH 1C-114
( ) IW 2Z-156
( ) MT 3B-117
( ) NW-ITDS
( ) PR 5-2120

AT&T Bell Laboratories

Subject: Report on the Workshop on Quality of Service Issues in High Speed Networks
Work Project No. 311407-7214 -- File Case 39199-11

date: August 26, 1992

from: S. Keshav
Org. 11272
MH 2C-552
(908)582-3384
research!keshav

TM: 11272-920826-23TM

TECHNICAL MEMORANDUM

1. *Introduction*

High speed networks of the near future are expected to carry a wide range of traffic types, including data, still pictures, voice, broadcast video and interactive video. Each of these traffic types requires different service from the network. The design of networks that provide a good quality of service to the large variety of expected users is an open and interesting research area. In order to initiate a dialogue between researchers who work in the general area of network design and those who work at the application or user level of integrated networks, AT&T Bell Laboratories sponsored a workshop on ‘Quality of Service Issues in High Speed Networks’ held at Murray Hill, NJ on April 23-24, 1992. The workshop was attended by about 50 participants from universities and research laboratories who represented a wide variety of research interests. Specifically, some of the areas represented were coding, compression, supercomputing, multimedia systems, B-ISDN, protocol design, queueing theory, network design, hardware design and operating systems. Participants were selected on the basis of research abstracts that were evaluated by the program committee†.

This report describes the main themes of the workshop and some interesting issues that were raised by the participants. While only the speakers’ names are mentioned, quite a lot of the work presented was joint work. Since it would be impossible to mention all the collaborators, interested readers are encouraged to get the collected abstracts of the workshop participants using anonymous FTP from research.att.com:dist/qos/qos.abs.ps.

2. *Overall Themes*

Listening to the talks at the workshop and ensuing discussions, in my opinion, certain themes seemed apparent. First, the speakers at the workshop came from a variety of backgrounds, and each person had a different conception of what QOS meant. The diversity of the notion of QOS was remarkable. For researchers working in video coding, it was a subjective measure of channel quality, whereas others saw it as a need for networks to provide performance bounds, and still others saw it in terms of network availability in the presence of failures. Since the notion of QOS seemed quite fuzzy, hopefully, one outcome of this dialogue was to open researchers to other points of view, perhaps resulting in a single acceptable definition in the future. For the moment, one non-controversial definition of QOS would be ‘network quality sufficient to satisfy user needs, however the needs may be expressed’.

The second theme was that QOS is important in thinking about network design and control. As we move towards integrated networks, we need to ensure that the new facilities are compatible with the existing infrastructure (such as telephony), which means that the integrated network must provide QOS bounds to existing applications (such as voice). Designing networks so as to provide QOS bounds is the critical next step in moving towards the goal of building integrated networks.

Why is providing QOS hard? Most speakers had the same answer: the diversity of user needs and the need for efficient operations. These are, in some sense, lower and upper bounds on the problem. A good solution should be general enough to satisfy most users, yet cheap enough to

†The program committee consisted of A.G. Fraser (AT&T BL, Chair), R. Braden (USC ISI), B.T. Doshi (AT&T BL), D. Ferrari (UC Berkeley), J. Flanagan (Rutgers), I. Gopal (IBM TJ Watson), A. Hopper (Cambridge/Olivetti Research), N.F. Maxemchuk (AT&T BL), P. Messina (CalTech), and S. Weinstein (Bellcore).

implement on a large scale. These constraints are not as simple as they look. Since integrated service networks would have users who need interactive video connections, the network has to deal with large bandwidths and realtime response. Providing large bandwidths requires (usually experimental) leading-edge hardware technology, making the problem harder. The need for realtime response means that resource scheduling and management need to be given careful consideration. The efficiency constraint is also problematic since it rules out the easy solution based on circuit switching. So, the network somehow has to exploit statistical multiplexing of bursty sources to efficiently utilize bandwidth. This is hard.

An interesting observation is that, viewed from a distance, most of the current crop of scheduling disciplines look quite the same. While the schemes differ in details, it is becoming more and more clear that an ability for a scheduling discipline to provide throughput bounds is necessary, to provide delay bounds desirable, and delay jitter bounds debatable. Given a set of QOS specifications, choosing one of the existing schemes, or coming up with a new scheme does not seem particularly hard. That is, this sub-field seems to have reached a level of maturity where the basic principles in designing new schemes are now clear. The pressing problems in network design lie elsewhere.

One point on which speakers agree in general, yet differ widely in the specifics, is in the expected workload of an integrated network. Most speakers agreed that the workload would be diverse, and would probably include video, audio, and data sources. But beyond this, not much is known, and each participant had a different model of the network workload (such as the number of traffic classes, expected requirement of each class, and the number of users in each class). Since the design of the network is directly affected by the workload it assumes, this is a problem. Even comparing the relative merits of two different schemes is hard, since in many cases, the workload assumptions differ radically. The resolution of this problem will probably depend on the relative success of experimental networks that are currently being built.

3. *Summary of Talks*

The workshop was divided into four sessions, each with a different broad focus. The first two sessions discussed network requirements and the second set of sessions discussed the efforts of current researchers to address these requirements as shown by network and end-system design. The first session was a tutorial on network performance requirements as dictated by human perception research and current coding schemes. The second session discussed network requirements of high-end services such as supercomputer interconnects and satellite telemetry. The third session dealt with network support for QOS requirements. The final session discussed end-system technology to support QOS. Each session had a keynote speaker, who presented a survey of the field. These were followed by brief presentations by each participant, moderated by a chair.

3.1. *Coding and Perception of Information Signals and Impacts on Network Specification*

The first session, chaired by Jim Flanagan from Rutgers University, dealt with human factors and coding issues in audio and video communication and QOS specification. In his introduction, Dr. Flanagan made some remarks about the QOS requirements for transmitting speech and video. For good quality transmission, QOS must be above a certain threshold. Research has also shown that variability in QOS is very undesirable: users often remember only the worst portion of a session, or, as he put it, ‘A little vinegar poisons the wine’. Looking into the future, rapid increases in DSP speeds make new compression schemes possible. For example, the human ear has difficulty detecting a weak high-frequency tone sounded in the presence of a strong low-frequency tone. So, if one cleverly shapes the distortion that is inevitable in the coding process, this will not be noticed (this was expanded by Nikil Jayant in his keynote speech). Another interesting idea in his speech was the notion of using force-feedback as part of a virtual-reality based multimedia system. Work in this area is being done under his guidance at Rutgers University.

3.1.1. *Coding Technologies for Video and Audio*

The keynote speech in this session was delivered by Nikil Jayant of AT&T Bell Laboratories. In an excellent talk interspersed with audio and video clips, Dr. Jayant described the state of the art in audio and video compression and coding. He made a few important points. First, network designers need a semi-quantitative feel for QOS in order to design networks. A reasonable measure of quality is the MOS (mean opinion score). This is the mean subjective score, on a scale of 1 to 5, that a fairly large set of viewers give a particular audio or video sequence under controlled conditions. The MOS is the standard way telecommunications engineers determine the end-user perception of transmission quality. Second, the basic technique used to compress audiovisual information is to remove redundant information in the signal (such as by using linear speech prediction or motion compensation). In addition, by matching the quantization level to the perceptual capabilities of the human ear and eye, it is possible to eliminate substantial parts of the signal with no perceived loss of information. This concept is modeled using the notion of Just Noticeable Distortion (JND) level, which is the level at which humans perceive a loss of information. By coding slightly beyond this level, one can compress audio and video signals by orders of magnitude with no noticeable loss of performance. This was demonstrated by playing CD quality audio using only 64kbps (the ISDN rate). Third, traditionally, source and channel coding have been decoupled. Source coding tries to minimize the bits-per-sample, and channel coding tries to maximize the bits-per-second-per-Hertz. If the source coder can be made aware of the channel characteristics (such as loss rate) then the overall coding scheme is much better. For example, the source coder may interleave bits or use smoothing reconstruction filters to cope with packet losses. Fourth, one has to balance several factors in designing integrated networks. These include the signal quality, bit rate, network efficiency, delay, complexity, communication delay and accessibility. The input of coding and perception research is to provide rough estimates of the required capacity in order to engineer the network bandwidth. Further, the channel characteristics can be fed back to coding design to optimize the design. Finally, some numbers: The state of the art compression techniques achieve 'excellent' quality using bit rates shown in the table.

Minimum bit rate for excellent quality transmission

Telephone speech	16kbps
Audioconferencing speech	32kbps
CD quality audio	128kbps
CD-like quality	64kbps
Still images (500x500 color image)	˜256kbits/image
Digital video	1.5Mbps
Medium quality digital video	384 kbps
HDTV	˜20Mbps

In the next talk, Steve Wolf from the Department of Commerce presented a new quantitative measure for video QOS. Using a large number of video samples and users, Mr. Wolf and associates computed the subjective Mean Opinion Score for each clip, then tried to fit an objective performance measure that matched the subjective MOS. They found that a weighted sum of the proportional change in edge energy and proportional motion energy predicted the subjective QOS quite well.

In his talk, Barry Haskell of AT&T Bell Laboratories presented some layered coding schemes and pointed out the pros and cons of using layered coding. The idea is that a signal is coded into essential and enhancement parts. The essential part of the signal requires less bandwidth than the full signal but there can be severe signal degradation if any part of it is lost. The enhancement portion may be lost without much signal degradation. He also talked about the constraints on variable bit rate (VBR) signal coding. Basically, we must ensure that neither the sender nor the receiver decoder buffer overflow. This can be done using leaky bucket input regulation where the leaky bucket parameters are reflected in the encoder control.

Paul Haskell of UC Berkeley then presented a model for composited video called 'Structured Video'. This model imposes structure on the presentation of multiple video streams to a user. Each video stream is represented as an object that can be moved, resized, composited and displayed on multiple displays. The model is being implemented in a prototype 'Videostation' at UC Berkeley under the guidance of Prof. Messerschmitt.

In the next talk, Rich Cox of AT&T Bell Laboratories talked about subjective methods for testing network QOS for speech. Three types of tests are typically used: intelligibility tests, diagnostic tests and category rating tests (MOS tests). The tests are used in network planning, for example, if a statistical multiplexor introduces one unit of degradation, and three units are tolerable, then network planners must ensure that all source destination pairs go through no more than three multiplexors. The most popular tests of the three are the MOS tests. Here, an anchor condition, that is, a condition that is kept invariant, is chosen, and various distortions introduced. The results are then evaluated subjectively on a scale of 1 to 5 by users. A good test would have a range of good as well a poor quality samples so that the score is not biased. While MOS has worked well for standard PCM based coders, more work needs to be done to test its applicability for perceptual coders, video etc. The talk concluded with examples of distorted speech and the corresponding MOS.

3.1.2. *Network Specification and Quality Objectives*

The second part of the first session dealt with the problem of how an application might specify QOS parameters to a network. Craig Partridge from BBN presented his views on defining 'flows'. In his design users specify their desired QOS using a simple set of parameters, since the components of a network can only take simple actions (such as delaying or dropping packets) anyway. Senders are expected to shape their traffic, and, more importantly, receivers should intelligently process this information to recover from distortions introduced by the network. So, the network does not need to provide very strict bounds on quality. One minimal set of flow specifications that he proposed (given that the flow already has been established) is the token bucket size and rate, the loss and corruption rate, the minimum and maximum transit delay, and how strongly these quantities are guaranteed.

The second talk in this part of the session was by Roch Guerin of IBM. He laid out some general problems that need to be solved in order to provide QOS in networks. First, one needs to know how easy or hard it is to build the network. This places fundamental restrictions on the complexity of the actions that we can undertake. Second, we need to know how much resource a particular connection with a particular QOS needs. This depends not only on the traffic on that connection, but may also depend on the other connections currently in service. The translation from the specification to an allocation is constrained by the tractability, accuracy and efficiency of the algorithm. Third, we need to ensure that the accepted connections obey their input traffic specification. This can be done using some sort of leaky bucket, but one must note that leaky buckets are not 'foolproof'. Finally, we need to know how the network's service guarantee maps to the user's point of view. This is a subtle point. For example, consider a network with a loss rate of 1 in a million. Two connections that have the same loss rate but have different bandwidth requirements would see different QOS. A slow speed connection, that sees essentially uncorrelated queue lengths would see random losses, whereas a high speed connection would see correlated queue lengths and probably bursty losses.

3.2. *High speed computing and overall network design*

The second session was broadly on the topic of high speed computing and network design strategies and was chaired by Craig Partridge of BBN. The keynote speech by Jeff Dozier of NASA described the network requirements of the proposed Earth Observation System. Better data management is crucial to the success of scientific investigations of global change. New modes of research about the Earth, especially the synergistic interactions between observations and models, will require massive amounts of diverse data to be stored, organized, accessed, distributed, visualized, and analyzed. Not only is the data voluminous, every bit of data is important, so the techniques of data compression used for video signals cannot be employed. Even using the best coding techniques, one can at the most obtain about a 50% reduction.

Later in this decade (1993-1999 time frame), NASA's Earth Observing System (EOS) will create a new need for a comprehensive data system to handle the large amount of remotely sensed data anticipated from the EOS instruments, related in situ observations, measurements from other satellites, and scientific data products. The estimated data volume at the end of the century exceeds 1 TB/day. Indeed, a single researcher may consume a terabyte during the course of a few days for visualization and modeling. The information system to store, manage, and provide access to these data is as critical to the success of the mission as the measurements from the satellites. The system should be able to move large amounts of data over a wide area and with small enough access times to allow interactive slide show types of applications. To retain integrity, the number of copies of data should be minimized, which means that large capacity network links are necessary. To address some of these technical issues, Sequoia 2000, a collaborative effort between computer scientists and global change scientists at several campuses of the University of California and Digital Equipment Corporation, will apply refinements in computing--involving storage, networking, distributed file systems, extensible data base management, and visualization--to specific global change applications.

3.2.1. *Network design*

The challenges raised by the EOS data system and other such large distributed systems were tackled by the rest of the speakers in this session. Deborah Estrin from USC talked about designing for large networks. She considered networks where the numbers of switches, subnetworks, autonomous domains and bandwidth required for some flows could all be large. It is unlikely that such large networks will be uniformly over-engineered so that it is necessary to intelligently manage resources to maximize utilization. Users should be given flexibility in specifying different performance levels, each at its own price. Routing in such networks is a problem since we need to determine the path that can provide the QOS required by the user at an acceptable price. The important point is that providing QOS to a user in such networks is not just a matter of scheduling discipline and policing - one has also to consider routing, multicasting, pricing and other problems. Her proposed solution is to have adaptive source routing for 'special' traffic and generic routes for other data traffic. For multicast, sources would set up calls to an explicitly known group, but the receivers would initiate reservations. The call setup state should allow switches to aggregate users into classes and decouple setup and routing. The open question is how conservative reservations have to be in order to satisfy the QOS requirements of the users.

While Prof. Estrin concentrated on the issues of large scale, Joseph Hui from Rutgers presented a layered technique for network design. In his opinion, the key problem is to allocate resources to each connection so that it gets the best possible QOS. To do so, the network needs to deal with traffic and manage resources on a number of different time scales. Recognizing the existence of time scales and planning control actions at each time scale is the basis of layered design. Using this technique, Prof. Hui described techniques for path configuration, route planning, dynamic routing, flow control, and cell switching. The notion of layered equivalent bandwidth allows us to simultaneously meet several QOS bounds on the probability of call, burst, and cell blocking. Using this notion of equivalent bandwidth, he considered path dimensioning and configuration for handling heterogeneous and time-varying traffic.

The general area of efficient resource management is complex yet interesting, since this is at the heart of network management. In her talk, Sally Floyd of Lawrence Berkeley Laboratory made a case for separating low level resource management algorithms from high level policy. She presented a hierarchical resource allocation and scheduling mechanism that allows for aggregation of users into classes, reservation, priorities of service and link sharing in networks such as the Internet.The resource management mechanism is composed of four lower-level mechanisms: the classifier, which examines the header of each packet arriving at the gateway and assigns that packet to a class; the selector, which selects the order in which classes send packets on the link; the estimator, which estimates the recent bandwidth used by a class; and the delayer, which delays the packets from classes that have exceeded their throughput assignments. The intention is that these lower-level mechanisms can support a range of higher-level resource-management policies. The specific higher layer policies, such as admission control, and the exact form of specifying QOS are still under investigation.

One aspect of QOS that is often overlooked is network availability. Nick Maxemchuk talked about how to survive failures in MANs. There are two types of possible failures - node failures and link failures. In his talk, Dr. Maxemchuk described failure recovery mechanisms in three MANs. In the Manhattan Street Network, node failures are recovered from using bypass relays and link failures by routing around the failed link. In DQDB, a failure causes movement of the node and frame generators to the endpoints closest to the failed site. However, a higher level protocol is needed to determine the correct bus to transmit on. In FDDI, if a link fails, one of the rings stops functioning and all the traffic is routed on the other ring. Some computation shows that if each link has an expected outage of about 2 hours a year, a 4096 node DQDB network would be down nearly 100 days, whereas a Manhattan Street Network would be down for only about 1 minute! (However a 64 node DQDB network would be down only about 1hr/year.) Thus, evaluating the expected availability of a network should be an important part of the design of a network that provides good QOS.

Telephone companies around the world view ATM based B-ISDN as an integral part of the future. The role of QOS in B-ISDN was explained by Bharat Doshi from AT&T Bell Laboratories. In the original B-ISDN proposals, the technological constraints were thought to be very stringent: high speed switches needed to be simple, and the expense of high speed memory made large buffers infeasible. This led naturally to a design where sources made reservations at the peak rate, and were monitored and policed at the edge by some network selected algorithm. However, this has been shown to be inappropriate for bursty sources. So, in a series of extensions, the basic structure is being altered to allow

- per-class loss QOS
- per class delay QOS
- scheduling disciplines sensitive to differential loss and delay QOS
- fairness per class via service class based cell scheduling
- violation tagging
- source congestion control based on forward or backward congestion indication
- within call modification of negotiated traffic parameters.

Not all of these proposal have been fully accepted, but most of them have been declared optional. The current status is that the peak rate definition is standard and so is the need for some monitoring algorithm at the User-Network Interface (UNI). Work is going on to define additional traffic parameters and QOS parameters to be used during call (and sub-call) set up so that additional monitoring and scheduling mechanisms can be designed (the actual designs are likely to be left to the network providers' discretion). Per service-class scheduling can be done using the virtual circuit id and the QOS parameters negotiated at call set-up and thus does not require standardization. Code points in the ATM cell headers are defined for loss priority and forward congestion notification. Reserved code points may be standardized late for other control functions (backward notification, within call negotiation, etc.)

The next speaker, Nachum Shacham of SRI, raised yet another aspect of diversity of QOS. While most research into QOS has considered the heterogeneity in traffic requirements, in practice if the same data is multicast to different destinations, the heterogeneity in receivers means that each receiver may need to specify a different QOS to the sender. This point becomes clear if one considers that an audio multicast may reach some users over a T1 link, and others through a 19.2 dialup link. Users may prefer to get broadband data with compression ratios that reflect their interest in the data, ability to receive it, or willingness to pay for it. In the proposed solution to this problem, a switching node that is part of the multicast tree is able to determine the part of the signal that should be transferred along each branch of the tree. To do this efficiently, sources would use some form of layered encoding, with each packet containing information from a single layer. A dynamic program can then be set up to find paths from a source to a destination that maximizes the bandwidth from the source to each destination subject to minimizing the transfer path and efficient link utilization.

3.2.2. *Architectures*

The next three talks presented some network architectures for implementing QOS in experimental networks. The first talk, delivered by S. Keshav of AT&T Bell Laboratories described the architecture of the Xunet II network, a prototype wide area high speed network based on ATM. Xunet assumes that sources are broadly divided into constant bit rate (CBR) and variable bit rate (VBR) sources that require service guarantees and best effort traffic that does not need service guarantees. Xunet provides each class with a menu of QOS in the form of delay, throughput and loss bounds. This menu is based on traditional resource reservation, admission control and call set up as well as three novel schemes. First, the Hierarchical Round Robin service discipline allows the network to provide deterministic and statistical delay, throughput and loss bounds to guaranteed service traffic. Best effort traffic is regulated using the Dynamic Adaptive Window mechanism at the switches and the Packet-Pair flow control scheme at the endpoints. By manipulating the parameters of these mechanisms, users can obtain a wide variety of QOS guarantees.

An alternative approach was presented by Aurel A. Lazar of Columbia University, who described QOS control and management on TeraNet. TeraNet is a gigabit lightwave network consisting of 3x3 switches with 1 Gb/s ports. The switching hardware consists of a non-blocking fabric with class oriented output queueing. Four classes of traffic are supported. Network control for guaranteed quality of service is based on the concept of Asynchronous Time Sharing. User traffic is assumed to belong to one of three classes based on time delay and loss requirements on the cell level and blocking constraints on the call level. Cell level QOS requirements for each traffic class are guaranteed by a real-time scheduling algorithm called MARS. An admission control algorithm guarantees QOS requirements on both the cell and the call level. Prof. Lazar made the point that it is important to lay a theoretical foundation for discussing QOS in networks. The approach taken by his group is to evaluate the networking bandwidth of a multiplexer by quantifying the joint performance of the scheduling and admission control algorithms. Given the cell level QOS requirements of the various traffic classes, the cell level characteristics of the traffic and the scheduling algorithm implemented, a stability (schedulable) region can be defined in the space of calls for which cell level QOS guarantees are met. (Interestingly, the schedulable region is the exact analogue of the stability region of the M/M/1 queue.) This region, along with the call level traffic characteristics, call level QOS requirements and the admission control algorithm define a stability region in the space of call intensities called the admissible load region. The schedulable region and the admissible load region together describe the efficiency of the multiplexer and its ability to provide QOS to users.

The last talk of the session, which was about the plaNET/Orbit network, was delivered by Roch Guerin of IBM. Dr. Guerin emphasized that at this time not all the issues in designing networks to provide QOS are clear. The approach of the group at IBM is to build an exploratory system and gain some experience with these issues. plaNET is a wide area high speed network based on packet switching (of either ATM or variable sized packets) and Orbit is the local distribution facility. As in other networks, it is assumed that traffic belongs to one of three classes: long lived connections, bursty connections that need bandwidth on demand, and best effort connections. The plaNET network provides long lived connections with a QOS guarantee by reserving resources based on the equivalent bandwidth concept. Bursty connections are monitored by hardware that can rapidly set up a burst transfer request. Network status is dynamically monitored and distributed to a replicated database. Rapid and efficient flooding of information is achieved through hardware support for multicast at each switch. This allows efficient control of the network at a relatively modest cost.

3.3. *Specific scheduling and traffic management algorithms*

The third session dealt with specific scheduling and traffic management algorithms for supporting QOS. It was chaired by Robert Braden of USC-ISI. The keynote speech by Sandy Fraser of AT&T Bell Laboratories described transmission facilities for computer communications. Currently data communication is only a small fraction of total telecommunication traffic. Given the massive infrastructure in place for telephony, it is important to understand the potential and limitations of this infrastructure.

Today's phone network is divided into local, wide area and signalling components. Local area transmission is usually analog, the wide area is all-digital, and the signalling network is a packet switched overlay on the circuit switched base.

Wide area transmission uses the so-called digital hierarchy of speeds. Existing transmission systems do not have a single master clock so extra overhead must be added at each multiplexing level to allow for clock slippage. The SONET proposal is synchronous and will allow extraction of a single channel without demultiplexing the entire payload. The wide area network is shared and so it is economical to use new technologies for higher speeds: currently it uses optical transmission at 3.4Gbps, and in the future soliton transmission and optical amplification will allow still higher speeds.

The local access component must reach nearly a hundred million endpoints so the technological choices here are highly constrained. Existing technology will allow unshielded copper pairs to carry 1Mbps over short distances and 19.2 kbps over longer distances. In the future fiber to the home, or to the curb, may be feasible (though it requires nearly $50 billion of investment, just for the USA). Another alternative would be to use the cable television infrastructure for data transmission.

Wireless cellular radio communication is a rapidly growing area. The bit rates achieved are rather low (around 16kbps), the cost of bandwidth is high, and the loss rate can be as much as 1%. So it is not a likely candidate for universal high speed local access. Indoor wireless communication, on the other hand, may be a good alternative, especially in buildings where adequate wiring does not exist and installation costs are high.

One interesting point that was raised is the cost of a network breakdown. The economic cost of one national telephone network outage for one day is approximately the same as 1 Terabyte of memory. So, it is worthwhile for a national network to invest in large amounts of hardware in an effort to make network outages very unlikely. In other words, the motivation for providing good quality of service, in terms of network availability, is very high.

3.3.1. *Specific Mechanisms*

The next set of speakers described work in designing specific mechanisms for providing QOS in high speed networks. The first speaker, Domenico Ferrari from UC Berkeley, talked about the requirements a scheduling discipline in a real-time packet-switched internetwork must satisfy. His claim is that the discipline must ensure that given a finite number of packets present in the node at any time, the discipline must not starve (i.e., ignore forever) any of them. However, in order to obtain bounded delays for all packets, even a starvation-free discipline requires that another, external condition be satisfied: that the aggregate real-time packet population be at all times finite in the node. On this basis, a small number of conditions on the source packet generation rate, admission control scheme and bandwidth reservation scheme are imposed. Networks where the conditions hold true will all be capable of real time service. It is interesting that with the right admission control scheme, even the FCFS discipline can provide bandwidth and delay bounds (i.e., real time service).

The next speaker, Hemant Kanakia of AT&T Bell Laboratories, presented a minimalist viewpoint on scheduling algorithms at packet switches. His claim is that one must determine the network service needed by ‘real’ applications and the simplest mechanism that can provide such a service before deciding on any particular mechanism. A minimal scheduling discipline would provide fairness to users, protection, efficient use of resources, guaranteed quality of service and require a minimum characterization of the source traffic. Specific non-goals are shaping of traffic, absolute bounds on delays, collective delay minimization or bandwidth reservation in advance. The Hierarchical Round Robin service discipline used in conjunction with a hop-by-hop flow control scheme would satisfy all these constraints. Simpler scheduling disciplines that satisfy these minimal constraints may also exist, and this is an area for future work.

Taking a cue from the previous speaker, Lixia Zhang from Xerox PARC talked about supporting real-time applications in packet-switched networks. She started her talk by noticing the fact that there are two types of telecommunication networks running today, circuit-switched (CS) telephone networks and packet-switched (PS) data networks, each provides a different service interface. Our goal is to build one single telecommunication infrastructure that can provide integrated services. The fact that almost all realtime applications today are running on CS networks does not necessarily imply that, in order to support realtime applications, this new integrated network must imitate a CS service interface. Rather, Dr. Zhang conjectures that most realtime applications can be characterized by a need to determine a playback point, and according to how they choose this playback point, applications can be sorted into two categories: rigid and adaptive. Rigid applications can be supported by using the Weighted Fair Queueing (WFQ) scheduling algorithm which provides each user with strict delay and throughput bounds. For adaptive applications, an appropriate service commitment from the network is predicted service: the network makes a service guarantee based on post facto measured load, assuming that clients do not change their behavior abruptly. Such applications can be supported by a combined FIFO+ and priority algorithm (details of FIFO+ algorithm are described in a paper by Clark, Shenker, and Zhang in Proceedings of SIGCOMM’92).

The queueing discipline used to multiplex packets onto an internodal link is an important element in the overall bandwidth management of an integrated fast packet network. The next set of talks presented some new service disciplines that allow networks to provide QOS guarantees to users. The first talk, by Abhay Parekh of MIT, presented the Generalized Processor Sharing discipline. This is an extension of Round Robin where each connection may get a different share of the bandwidth (and is identical to the Weighted Fair Queueing discipline invented independently). The discipline allows a token bucket regulated connection to get a worst case end-to-end delay bound independent of the topology of the network. Since the discipline is work conserving, it works well with bursty sources. Thus, it seems suitable for VBR sources.

One aspect of QOS is the ability of the network to ensure that a malicious user does not jeopardize the performance of a well behaved user. This is often called protection or fairness. In her talk, Ellen Hahne of AT&T Bell

Laboratories discussed the fairness properties of Round-Robin servers. Assuming that each user is subject to sliding-window flow control and has ample packet buffers in all its nodes, through analysis and simulation it has been shown that, in contrast to FIFO disciplines, round-robin disciplines protect light users from heavy users. This protection takes a variety of forms. In an uncongested network, the delay of users with short or rare messages is relatively insensitive to the presence of long or frequent messages from other users. As congestion develops at a single link, the local round-robin scheduler maintains throughput fairness among users of that link. As congestion spreads throughout the network, the round-robin link schedulers collectively maintain global fairness in a max-min sense. That is, the smallest user throughput in the network is as large as possible and, subject to that constraint, the second-smallest user throughput is as large as possible, etc. The conclusion from this is that if the service discipline is round-robin like, then the endpoint flow control need only worry about keeping the bottleneck queue fed, and not about fairness.

Michael Hluchyj of Motorola Codex presented a service discipline, a hybrid combination of weighted round robin and head-of-line priority, which provides an appropriate allocation of the internodal link bandwidth among different traffic classes, while allowing delay differentiation within a traffic class. The discipline is based on the observation that the queueing behaviors of CBR traffic, voice traffic and data traffic differ significantly. CBR traffic exhibits a periodic high frequency queueing pattern, whereas low frequency patterns dominate voice and data traffic. Further, voice streams are insensitive to losses, whereas data sources will retransmit lost packets leading to possible congestion. By using weighted round robin for bandwidth allocation between traffic classes and head-of-line priority for delay differentiation within a traffic class, it is possible to provide each class with the class of service it requires. Each traffic class is allocated enough bandwidth to achieve QOS objectives of that class. Within the CBR class, HOLP assigns small packetization delay traffic to the high priority queue, and within the data class, sources generating small bursts are assigned to the high priority queue to give a low end-to-end delay. Using appropriate coding, congestion control and admission control techniques it should be possible to design the network to satisfy QOS goals.

Instead of classifying sources by the type of queueing behavior they produce, the next speaker, K. Sriram of AT&T Bell Laboratories, classified traffic on broadband ATM networks as real-time high-bandwidth (isochronous and statistical), delay-insensitive high-bandwidth, or low-bandwidth. A new scheduling discipline, Dynamic Time-Slice (DTS), allocates and guarantees a required bandwidth for each traffic class and/or virtual circuit (VC). The basic service mechanism is a framed non-work conserving discipline where service slots are dynamically partitioned between traffic classes (similar in spirit to Stop-and-Go and HRR scheduling). Any bandwidth momentarily unused by a class or a VC is made available to the other traffic present in the multiplexer. The scheme guarantees a desired bandwidth to connections which require a fixed large bandwidth. Thus, it facilitates setting up circuit-like connections in a network using the ATM for transport. The DTS scheme is an efficient way of combining constant bit-rate (CBR) services with statistically multiplexed services. Methodologies to schedule delivery of delay- tolerant data traffic within the framework of the DTS scheme were also described. A paper based on this presentation is to appear in a special issue of Computer Networks and ISDN Systems Journal on ‘Traffic Issues in ATM Networks’.

The last speaker in this set of talks, Hui Zhang of UC Berkeley was concerned with the problems associated with currently proposed scheduling disciplines. He classified existing solutions into two categories: one based on a time-framing strategy (e.g., Stop-and-Go, Hierarchical Round Robin) and the other based on a sorted priority queue mechanism (e.g., Virtual Clock, Delay-Earliest-Due-Date). He pointed out that time-framing schemes suffer from the dependencies that they introduce between the queueing delay and the granularity of bandwidth allocation; sorted priority queue is more complex, and may be difficult to implement. To overcome these problems, he presented a new scheduling discipline called Rate Controlled Static Priority (RCSP) Scheduling. The basic idea is that when designing a scheduling discipline, one should separate rate control from delay allocation. A RCSP switch has two parts: a rate allocator, and a set of static priority queues. Packets are held in the rate allocator till they are eligible for transmission (which is determined according to their allocated bandwidth and arrival time) and then placed in some level of the multilevel priority queue. This clean separation of rate control and delay allocation allows decoupling of rate and delay allocation as well as simple design of servers. In practice the rate control is based on a calendar queueing mechanism. The RCSP discipline allows simple calculation of end-to-end delay bounds and, if required, delay jitter bounds.

3.3.2. *Analysis and Call Establishment*

The next set of talks dealt with analysis of QOS in networks. The first speaker, Anwar Elwalid of AT&T Bell Laboratories, described his recent work on fluid analysis of a leaky bucket access regulation mechanism and statistical multiplexing with loss priorities. He has obtained explicit expressions for designing the parameters of the access regulator and has shown that the device provides a 3-way tradeoff between packet marking, packet delay and smoothness of unmarked traffic. A coupled Markov-modulated fluid model which captures the correlation between marked and unmarked streams is used to characterize the output of the regulator. Such accurate characterization of output processes enables him to analyze the statistical multiplexers, further downstream in the network.

In addition to cell marking by the regulator, in real-time applications, cells are assigned priorities which correspond to their importance with respect to service quality so that during congestion low priority cells can be dropped. A number of buffer threshold levels are associated with the priority classes. When the buffer content exceeds a particular threshold, packets of the corresponding priority class are dropped. Analysis shows that threshold levels can be effectively set to improve overall system performance. This work and other related work provide incentive for building the multiple-priority concept into ATM standards (currently only two priority levels are considered.)

The next speaker, Jim Kurose of University of Massachusetts, Amherst, described his work in analyzing multiplexors in tandem. He made the point that most current analysis is valid at the access point of the network. As soon as traffic is multiplexed, it loses its input characterization, and so any analysis that assumes a specific input characterization ceases to hold deep inside the network. In his work, he considers the complex interactions among sessions as they interfere with each other as they pass through various network nodes. His claim is that one needs to identify both intra-session and inter-session packet (cell) interactions and further consider not only external source inputs to the network but also the session-level departure ‘‘processes’’ at the various queues in order to evaluate performance. He presented a technique for computing upper bounds on the distribution of individual per-session performance measures such as delay and buffer occupancy for networks in which sessions may be routed over several ‘‘hops.’’ The approach is based on first stochastically bounding the distribution of the number of packets (or cells) which can be generated by each traffic source over various lengths of time and then ‘‘pushing’’ these bounds (which are then shown to hold over new time interval lengths at various network queues) through the network on a per-session basis. Session performance bounds can then be computed once the stochastic bounds on the arrival process have been characterized.

The last talk of the session was delivered by Anindo Banerjea of UC Berkeley on the Realtime Channel Establishment Protocol (RCAP). The RCAP signalling protocol allows internetworks to establish realtime channels (i.e. those with guaranteed performance bounds). The key characteristic of the protocol is that data transfer and channel control are cleanly separated. A realtime user is assumed to be able to describe his input traffic and expected performance bounds. This information is carried on a RCAP signalling channel to each switch along the path. On the forward path, the switch determines the best possible QOS available and allocates it to the call. It also attaches a record to the setup request describing its allocation. The destination thus receives a list of best possible allocations and uses this to relax the allocations so that the call receives no more resources than it really needs (this is computed from the traffic specification and the desired QOS using a deterministic analysis of worst case delay bounds). On the trip back, each switch carries out the relaxation. Thus, a realtime call can be established in one round trip time. The protocol also supports other functionality such as teardown, status reports etc.

3.4. *Workstation Support and Multimedia*

The last session of the workshop dealt with workstation and operating system support for QOS and multimedia. In his opening remarks, the session chair, Andrew Hopper of Olivetti Research Labs and Cambridge University, made a few observations based on his experience with the Pandora project at Cambridge. First, the system designers found that users tend to use the system in unpredictable ways. They do not use multimedia documents (text + image), since that seems to require a lot of work. Further, they are satisfied with a far lower quality image than what was thought necessary. And, they seem to prefer video mail to real-time interactive video phone. Second, in the future, real time traffic will not dominate network traffic, since ‘Real time traffic needs real time people’, and most people simply do not have enough time to handle large amounts of real time interaction.

The keynote speaker for the session was Forest Baskett from SGI. He started his talk with an interesting video demonstrating the latest multimedia efforts at SGI. The video showed several applications based on video conferencing and multimedia databases. Dr. Baskett’s talk addressed two problems: first, when can we expect the computing power to deal with multimedia in workstations, and second, the relationship between telecommunication and computing. If we look at the trends in silicon performance over the past 25 years, it can be safely concluded that performance increases at the rate of 60% a year. This holds true for DRAM capacity, CPU speeds as well as floating point speeds. So, we can expect to see a 500 Mips workstation around 1996, and a 5000 Mips workstation in 2001. This seems almost inevitable. The computing power to do multimedia work on a workstation will soon be available, including complex coding and compression schemes.

There is no doubt, again, that information is gradually becoming more and more digital, particularly audio and video information. Thus, telecommunication and computing industries, both of which deal with digital information, will come closer together. However, both industries have a significant installed base, and different concerns, so the opportunity to leverage off each other’s technology is limited. For example, data (computer) applications are latency intolerant, error intolerant and compression intolerant. In contrast, traditional telecommunication networks are insensitive to all three. With the advent of ATM and integrated networks, some of these concerns need to be addressed.

In the next talk, Ricky Palmer of DEC talked about the DECspin videoconferencing product that he and his brother, Larry Palmer, developed. In his opinion, about two years ago, it became technologically feasible to build a useful videoconferencing program, in terms of memory speed, LAN speed (FDDI, etc). DECspin is an X11/Motif based application that can support up to 8 simultaneous participants in a videoconference on a RISC workstation. It requires as little as 1.0 - 1.5Mbps of bandwidth, and currently uses software compression to save bandwidth. The application is purely in user space, with no special support from the operating system or network. It uses TCP/IP over FDDI, Ethernet or T1/T3 components. Testing over SMDS is in progress. Audio and video tracks are synchronized and the frame rate and size are variable. In the speaker's opinion, what struck him about the project, initially, was that it was not thought to be feasible; indeed, he himself originally doubted its success. Yet, with not much help from the network or operating system, the system is still able to achieve a rather impressive result. This raises the question of how much support applications really need from a network.

The next speaker, Riccardo Gusella of HP Labs, talked about system issues in building integrated networks, specifically, network design assumptions and the interface supporting motion picture compression and decompression being built at HP. Dr. Gusella claimed that the QOS parameter of interest to a network designer is simply the end-to-end delay since (a) this is the likely to be the only common QOS denominator in an internetwork (b) it reduces the complexity of the design and (c) it is the appropriate metric for a large class of applications. The group at HP makes the additional assumptions that there will be many video and audio channels which will take up between 0.1 and 5Mbps. They plan to pick some classes of service and an appropriate admission control and scheduling discipline to provide end-to-end delay bounds. The final consideration is graceful degradation of service.

One component of the network would be a compression/decompression front-end processor. Currently, compression is done on the CPU, so that uncompressed data flows twice across the system bus. In their design, this functionality is closely coupled to the frame buffer, so that data from the network can be decompressed and displayed on the frame buffer with only compressed data moving across the system bus, and that also only once. The compression engine is programmable, and will explicitly provide hardware support for X windows to allow movement and reshaping of video windows.

It is likely that the first generation of multimedia workstations will work in a high speed LAN environment. There has been a move recently to bring high speed ATM technology to LANs. This work was described by Allyn Romanow of Sun. Dr. Romanow described the design goals of her project as being able to build an ATM LAN at 155Mbps quickly, cheaply and conforming to standards. The proposed design is an S bus based host interface that does rate control in hardware, interacting with third party ATM switches. A signalling protocol reserves bandwidth at the peak rate, though reservations can be dynamically adapted in response to changing traffic conditions. Peak rate allocation is inefficient, but prevents packet losses, is stable at high loads, and requires a minimal traffic description from the user. Further, it does not require any intelligence on the part of switches. So, for the near term, it is the correct solution, though things may change in the future. Bandwidth allocations are managed by a central resource manager that solves a multi-commodity flow problem using linear programming techniques to come up with an optimal solution. This approach is not scalable, but again, is intended for the short term.

It is often claimed that building QOS into networks is an attempt to be responsive to user needs. The subjective elements in QOS were discussed by Steve Weinstein from Bellcore. Dr Weinstein presented four key ideas. First, in a multimedia session, there are several interacting components, such as audio and video streams, that interact with each other. It is necessary to understand the tradeoffs between the QOS demanded by each component. For example, it seems to be better to have a large fuzzy picture with sharp audio than a large crisp picture and poor audio. Thus, video bandwidth can be gainfully traded for audio clarity. Second, session level events can be used to dynamically alter resource requirements. For example, if a user shrinks a video window, the network can reduce the bandwidth allocated to that stream. Of course, this adaptation comes at the expense of simplicity. Third, the network must differentiate between customers and users. Each customer is potentially a company or group who would support several users. So, within a customer connection, some users may need to be given higher priority. Finally, the network may be able to do service substitution without affecting perceived user quality. For example, if a user is sending CBR traffic, the network may divert this to a circuit switched network without notifying the user. It may pay to give the user what he really needs, rather than what he asked for!

In the multimedia domain, a key question is whether to transfer audio, text and video streams together or separately. Sending them together is a clean solution, and avoids synchronization problems, but the combined session needs a QOS as good as the best component of the combination. Further, if the components of the combination are to be handled by separate processors at the receiver, demultiplexing is a problem. Thus, the more general solution is to create appropriate mechanisms to synchronize separately transmitted multimedia sessions.

The last two talks in the workshop dealt with the interface between a network that provides QOS and the applications that use it. Abel Weinrib of Bellcore described the application interface provided by the Touring Machine project at Bellcore. One of the aims of this project is to present programmers with a simple interface to the network that hides the details of the underlying network. Programmers are presented with two abstractions: connectors and ports. By registering a connector and attaching it to some number of ports, a videoconference multicast is automatically set up. This data stream abstraction seems adequate for current single-service analog networks, but if the network is to become digital, it is not clear how the user should specify QOS requirements to the network. An ideal interface would be simple to use, yet give the user flexibility to exploit the capabilities of the digital network. Perhaps a parameterized interface with sensible default values is the right answer.

The last speaker of the workshop, David Feldmeier of Bellcore, talked about the TP++ transport protocol project at Bellcore. TP++ refers to the protocol, its environment and also the hardware implementing it. One key idea in TP++ is that VCIs are visible at all points in the network; there is no multiplexing in the transport and network layers. This allows the network to get the maximum information about user QOS requirements, that would be hidden if the transport protocol were to multiplex connections. Second, the protocol tries to separate control and data. For example, it distinguishes between packets and protocol data units and explicitly supports out-of-order receipt of packets. Both CRC-like checksumming and DES-like encryption are supported over packets that can be received out-of-order. Third, the host implements protocols with the same technology as is used in the network. This includes techniques such as error detection, forward error correction etc. Finally, the protocol assumes that the network will do flow and congestion control and so it does none of that on its own. Thus, flow and error control are separated.

The speaker observed the interrelationship of coding and network design: coding schemes assume some kind of network behavior and networks expect coders to produce a certain kind of traffic. The problem needs to be jointly optimized. Secondly, the notion of averaging interval is valid only to the extent that the traffic shows small autocorrelation intervals. If the autocorrelation interval is large, as some experiments show, then the notion of averaging interval needs to be re-thought. Perhaps some kind of fractal description of the input would be useful.

4. *Acknowledgements*

This workshop would not have been possible without the support of Sandy Fraser, Chuck Kalmanek and Sam Morgan. My sincere thanks to them. I would also like to thank the program committee and the session chairs for their time and patience. Finally, I would like to thank all the participants for making the workshop a success.

This report is based on the abstracts submitted by the participants, their slides, and a videotaped record of the workshop. Any errors or omissions are mine, and for these, my apologies.