Open Issues and Challenges in Providing Quality of Service Guarantees in High-Speed Networks[1]

Jim Kurose
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
kurose@cs.umass.edu

Abstract

In this paper we identify the challenges and open issues involved in providing quality-of-service (QOS) guarantees to sessions in a high-speed wide area network and briefly survey research in this area. Four approaches towards providing QOS guarantees are described and discussed: the tightly controlled approach, the approximate approach, the bounding approach, and the observation-based approach.

1 Introduction

Unlike traditional data networks, future broadband-ISDN (BISDN) wide-area networks will be required to carry a broad range of traffic classes ranging from bursty, variable-rate sources, such as voice and variable-rate coded video, to smooth, constant bit rate sources [28]. Moreover, these networks will have to do so while providing a guaranteed performance or quality-of-service (QOS) to many of these traffic classes. The ability to guarantee performance is particularly important in networks supporting real-time applications [26]. The actual QOS performance metrics of interest are likely to vary from one application to another, but are projected to include such measures as cell loss, delay, and delay jitter guarantees. The need to provide end-to-end QOS guarantees while still taking advantage of the resource gains offered by a statistically multiplexed transport mechanism remains an important, yet largely unsolved problem facing BISDN architects.

The difficulties that arise in networks with QOS guarantees can be most easily illustrated by considering the "simple" question that must be answered each time a call/session arrives to such a network: *"Can the requested call be accepted by the network at its requested QOS, without violating existing QOS guarantees made to on-going calls?"* To answer this question, the network must be able to compute (or provide) not

[1]This paper is based on a panel presentation made by the author at the 3rd International Workshop on Network and Operating System Support for Digital Audio and Video, Nov. 12 - 13, 1992, San Diego, CA. The author's work is supported in part by the National Science Foundation under grant NCR-911618 and the Defense Advanced Research Projects Agency under contract NAG2-578.

only the guaranteed end-to-end quality-of-service to be received by the arriving session, but also determine the performance impact of admitting this session on the already-accepted sessions in the network. The problem here then is one of *performance-oriented call admission control* [28]. We note that this problem is distinct from (but in some respects dependent upon) the lower-level controls (such as flow/congestion control and cell discarding policy) which are applied to the individual cells or bursts within the traffic streams of already-accepted calls. We refer to the call admission process as being "performance-oriented" since in order to answer the call admission question, the performance realized by an admitted session and its impact on the performance of existing sessions must be explicitly considered. Performance concerns, once traditionally relegated to the off-line tasks of dimensioning and design (e.g., determining link bandwidths, buffer capacities and processing capacities at network switchpoints), now become an *on-line* concern given this need for performance-driven traffic control mechanisms.

The introduction of QOS requirements into high-speed wide-area networks poses a large number of technical challenges. In this paper we briefly survey research aimed at resolving these challenges and identify open issues involved in providing QOS guarantees. The remainder of this paper is structured as follows. In section 2, we identify several different approaches towards defining QOS and discuss the fundamental challenges involved in providing performance guarantees. In section 3, we overview four basic approaches for providing QOS guarantees. Section 4 concludes this paper.

2 QOS guarantees: Definitions and fundamental challenges

Several different ways of categorizing QOS guarantees may be identified. In [27], a distinction is made between *deterministic* guarantees and *statistical* guarantees. In the deterministic case, guarantees provide a bound on the performance of all cells (packets) within a session. For example, with real-time traffic, a deterministic guarantee might be that no cells would be delayed more that D time units on an end-to-end basis [29]. With cell loss as a performance metric, the deterministic guarantee might be that *no* cell loss occurs [16, 13, 14, 6, 7, 15]. Statistical guarantees promise that no more than a specified fraction of cells will see performance below a certain specified value. With real-time communication, a statistical guarantee might promise that no more than $x\%$ of the cells would experience a delay greater than D [18]. With cell loss, a statistical guarantee might be that no more than $x\%$ of the cells in the session are lost [1].

Among statistical guarantees, a distinction is also made between guarantees made in a "steady state" sense, and guarantees which are defined over specific intervals of time. In the former case, if a call were to hold for an infinitely long period of time (an "infinite horizon"), the guarantee would hold with certainty. However, if the session duration is finite, there is some probability that the guarantee would be violated. Markovian performance models, such as those in [28, 10, 1], typically fall within the category of statistical guarantees based on an infinite horizon model.

When guarantees are defined over intervals of time, the QOS guarantee specifies that a session will not experience a performance level worse than a given value in more than a given fraction of these intervals. For example, an interval-based QOS guarantee might state that "the session will have a packet loss fraction greater than y in no more than $x\%$ of the intervals of time of length I." The relationship between steady-state and interval-based QOS guarantees is discussed at length in [20]. An example which illustrates this relationship in [20] is that of multiplexed voice sources over a single T1-rate line. It is shown that when call holding times are infinite and at approximately 85% link utilization, for the voice model and buffer size considered, all calls see a packet loss probability of less than 10^{-3} (that is, *all* calls would meet a QOS loss requirement of 10^{-3}). However, more than 40% of the calls would experience a packet loss probability exceeding this value if calls were five minutes in length. (Equivalently, more than

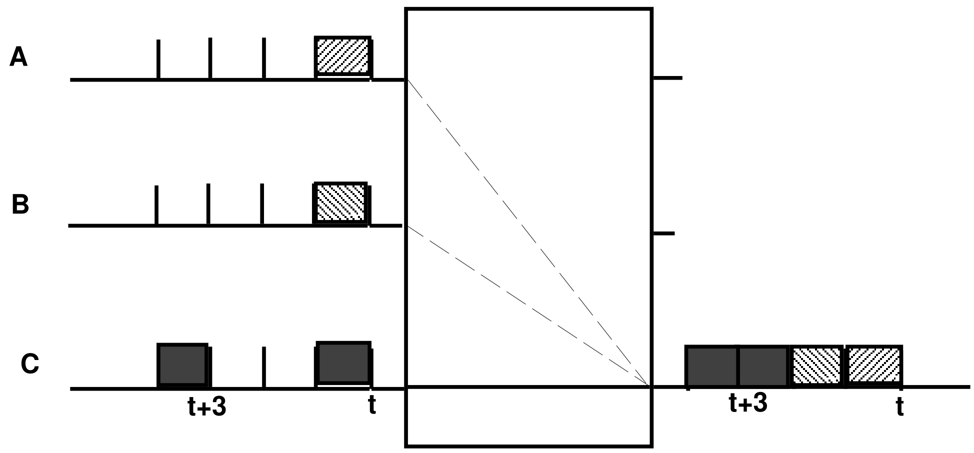


Figure 1: Change in peak rate as a result of multiplexing

40% of five minute intervals in the steady state model have a loss probability exceeding 10^{-3} even though the steady state loss probability is less than 10^{-3}.) Interval-based QOS guarantees have also recently been proposed as part of a flow specification in an Internet Request for Comment (RFC) [24].

Characterizing performance (and thus providing QOS guarantees) in BISDN networks poses a considerable challenge for several reasons. First, sources of traffic such as packetized voice and video exhibit correlated, time-varying behavior that is significantly more complex than that of traditional data network sources. Second, since performance (QOS) requirements are defined on an individual, per-session basis, it is no longer sufficient to simply determine the performance of the aggregated network traffic. Instead, performance must be characterized at the finer-grained, per-session basis. Finally, and most importantly, performance must be evaluated in a multi-hop *network* setting. Thus, the complex interactions among sessions must be considered as they interfere with each other as they pass through various network nodes. In this case, one must consider both intra-session and inter-session packet 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.

Figure 1 illustrates this last issue. In this figure, three sessions (A, B, and C) arrive to the network at some switch and are multiplexed over a common outgoing link to some downstream switchpoint. In order to evaluate a session's performance at a downstream node, we must clearly be able to characterize that session's departure process from the upstream switch. Note that on input to the switch shown in Figure 1, each session has a peak rate of 1 cell every 3 time slots. Let us assume a work-conserving FCFS multiplexing discipline at the output switch ports (i.e., a cell is transmitted whenever any cells are queued). Let us further assume that among cells arriving in the *same* time slot, session A cells are transmitted before session B cells, and session B cells before session C cells. As a result of this (not unrealistic) multiplexing discipline, the session C cell arriving at t is output at $t + 2$, and the session C cell arriving in time $t + 3$ is immediately output at time $t + 3$. As a result of multiplexing, session C traffic (which had a peak rate of 1 cell every 3 slots on input) now has a peak rate of 2 cells every 3 slots on output. *The simple act of multiplexing session C with two other sessions has thus resulted in a doubling of session C's peak rate* (defined over a three time-unit interval).

Clearly then, if we are to provide performance guarantees, we must be able either to

- provide multiplexing mechanisms which avoid such changes in traffic characteristics when sessions are multiplexed together, or
- characterize the increases in peak rates, as well as the other changes in a session's traffic characteri-

stics, which occur as a result of multiplexing.

In subsection 3.1 we outline techniques for providing QOS guarantees using the first of these two alternatives, while the techniques outlined in subsections 3.2 - 3.4 employ the second alternative.

3 Four approaches towards providing QOS guarantees

3.1 Tightly controlled approaches

In the "tightly controlled" approaches toward providing QOS guarantees, a non-work conserving multiplexing (queueing) discipline insures that an individual session's output traffic characteristics (i.e., after being multiplexed with other sessions at a switch output port) are the same as that session's input traffic characteristics (i.e., before multiplexing) [16, 27, 13, 14, 7]. An example of a tightly-controlled approach is the so-called "stop-and-go" queueing discipline [13, 14]. This mechanism defines time "frames" and insures that a cell arriving in one "frame" at a switch's input is never transmitted over an output link during the same time frame in which it arrived. Note that in order to satisfy this constraint, a cell may have to be held in the switch's output buffers while the output link is purposefully allowed to go idle. Stop-and-go queueing further requires that a cell *always* be transmitted in next output frame starting after the arriving cell's input frame ends. In order to insure this condition is always possible, arriving sessions may be blocked from entering the network.

With a mechanism such as stop-and-go queueing, a session's traffic characteristics (e.g., its peak rate over a frame interval) are preserved as it passes through the network and consequently performance bounds, such as a deterministic guarantee on the maximum delay experienced by a cell on an end-to-end basis, can be computed in a simple manner. There is, however, a price to be paid for this simplicity of computation. First, a fairly sophisticated, non-work-conserving queueing discipline must be implemented which tracks each individual session's timing requirements on a per-session basis. A second potential disadvantage is that a session admitted to the network essentially "reserves" bandwidth based on its *peak* rate – effectively resulting in a form of circuit-switching. As such, classes of traffic with high peak-to-average traffic rates will only utilize the links for a small fraction of their "reserved" amount of time, potentially leaving the links significantly underutilized. As noted in [15], however, it may be possible to utilize reserved, but unused, portions of time to transmit cells from traffic classes which do not require a guaranteed quality of service. Furthermore, if the peak-to-average ratio is not large (i.e., the traffic itself is circuit-like in nature), tightly-controlled approaches towards providing QOS guarantees could be quite attractive.

Figure 2 shows a network of switches connected by T1 rate lines. Call routing in this network is such that each link is carrying 48 32Kbps voice calls, with each call generating a packet every 16 ms during a talkspurt. Talkspurt lengths are exponentially distributed with a mean of 352 ms and silence periods (during which no packets are generated) are exponentially distributed with a mean of 6.5 ms. [The silence period is significantly shorter than that of typical voice calls and was chosen to provide a heavy network load of 98% utilization.] Each input link at the "edge" of the network (i.e., a line which is incident to the left side of a switch and does not connect to an upstream switch) thus brings 48 such sources into the network. At each switch, each output port receives one third of the incoming calls from each of the three input lines.

Figure 2 also shows simulation results for the end-to-end distribution of delay for those sessions which pass through the four hops beginning at the lowest switch in the figure and pass through the rightmost switch in the figure. FCFS scheduling is performed at all multiplexing points in the simulation.

Propagation delays and per-packet processing delays are not considered. The plotted results represent the average of ten independent simulation runs, each of which simulated approximately a half a million packets traversing those sessions passing through the four hops described above. Also indicated in the plot are the point-valued upper and lower bounds on delay under stop-and-go queueing. The results indicate that even at high utilization, the delays provided by stop and go queueing can be significantly larger than under the uncontrolled case.

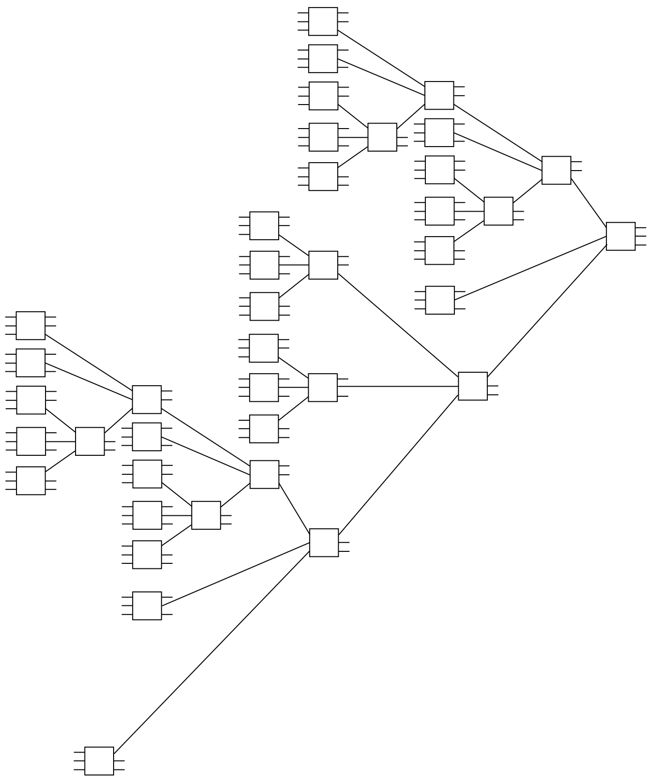
3.2 Approximate approaches

In approximate approaches toward providing QOS guarantees [28, 10, 1, 9], traffic sources at the network's edge (and within the network) are characterized by relatively "simple" models. An example of such a source model is the on/off source, which alternates between on-periods (during which cells are typically generated periodically) and off-periods (during which no cells are generated) [28, 1, 19]. In order to determine whether or not the multiplexed sources will receive their required QOS, the queueing behavior of the multiplexed traffic streams is then analyzed. In [1, 9], the QOS measure of interest is packet loss; in [10] the measure of interest is maximum delay.

Approximate approaches toward QOS guarantees have both advantages and disadvantages over the tightly controlled approaches. Perhaps their most important advantage is their simplicity, which makes them well-suited for real-time, on-line implementation. For example, the call admission control mechanism based on the approximate QOS scheme described in [1] can make a QOS computation (to determine whether or not to admit a call, given its QOS requirements and existing QOS guarantees) with a very small number of additions and multiplications. As noted above, tightly controlled approaches require a more complex scheduling discipline; approximate guarantees can be made using simple disciplines such as FCFS. Finally, unlike the tightly controlled approaches described above, approximate approaches are also able to take advantage of statistical multiplexing gains, potentially carrying traffic whose peak rate exceeds link capacity.

There are, however, several potential disadvantages. The first is the fact that the approaches are, as their name suggests, "approximate." Empirical evidence in [1] indicates that the QOS computations for certain well-defined sources tend to be conservative, thus providing a guarantee on performance. However, the conservative nature of the approximations has yet to be formally demonstrated, except in certain limiting cases [9]. Another open question is the extent to which more complicated sources can be characterized by the relatively simple approximate models considered thus far. A final potential concern is that traffic models, whether at the source or deep within the network, require some form of Markovian assumptions. While traffic at the edge of the network may be reasonably well-approximated by such models [19], it is still unknown whether this is also true for a session's traffic when it is "deep" within the network, having passed through several multiplexers. Note that the approximate approaches do not seek to characterize the changes in a session's traffic as a result of multiplexing with other sessions (e.g., as in figure 1) but rather take the input traffic characteristics as a given. The extent to which these interactions must be considered and the extent to which a reliable guarantee can be provided without taking such considerations into account remains an important question for future research. A study of this issue in the case of a tandem network with interfering traffic is performed in [21].

A final open research issue that arises both with approximate approaches as well as with many of the approaches described in the following subsections is that the guarantees provided are *local* guarantees, i.e., performance guarantees provided to a session at a single multiplexing point. User-specified QOS requirements, however, are based on an end-to-end performance requirement. The manner in which these



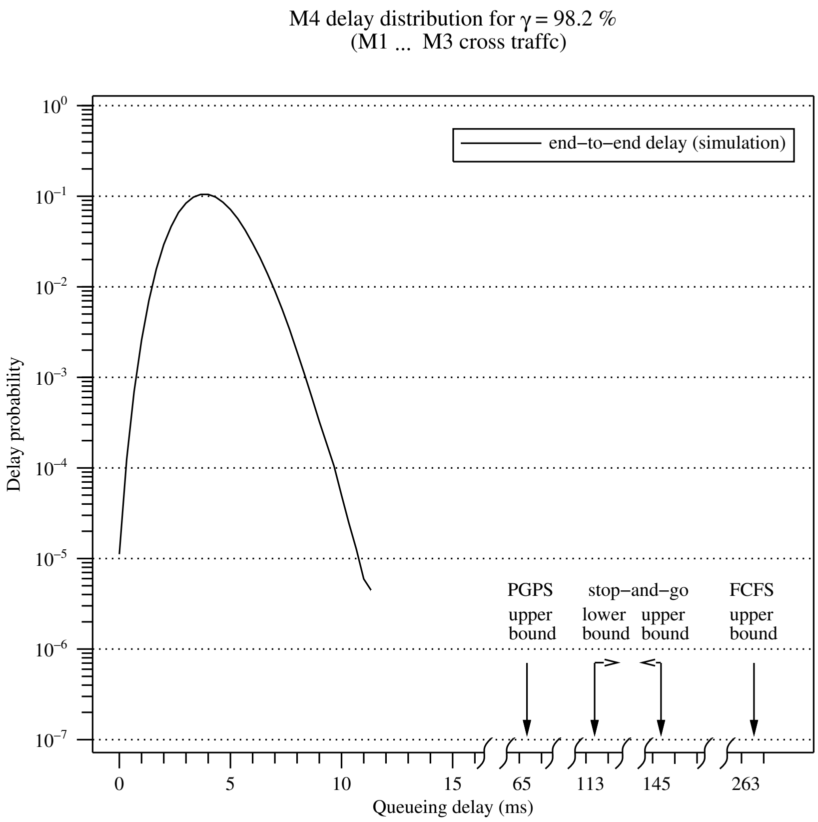


Figure 2: Delay distribution and bounds for a 4-deep network

end-to-end requirements are to be divided into local performance requirements (e.g., possibly having more congested nodes provide a poorer QOS guarantee while more lightly congested nodes provide a more stringent performance guarantee) which together satisfy the end-to-end requirement remains another important open research issue.

3.3 Bounding approaches

The third approach toward providing QOS guarantees explicitly accounts for the fact that a session's traffic does indeed change each time it passes through a work-conserving multiplexer.

Two type of provable performance bounds may be identified – those that provide deterministic guarantees [5, 6, 7, 22, 2, 23] and those which provide statistical guarantees [18, 25, 3]. As noted in section 2, the approaches which provide deterministic guarantees can be used to make statements such as "The delay of every cell from session i is less than x at queue j." The approaches which provide statistical bounds can be used to make statements such as "The probability that a cell from session i has a delay greater than y is guaranteed to be less than z at queue j."

The work of Cruz [5, 6, 7] has proven seminal both for bounding approaches which provide deterministic guarantees [22, 23] as well as bounding approaches which provide statistical guarantees [18, 25, 2]. The point-valued bound provided by Cruz's method for the network shown in figure 2 is also plotted in Figure 2 with the label "FCFS upper bound." Note that the bound is quite loose compared with actual delay distribution. Recently, Parekh and Gallager [22, 23] have developed a methodology for providing deterministic guarantees under a work-conserving queueing discipline known as packetized generalized processor sharing, PGPS (referred to as the weighted fair queueing discipline in [8]). In a network setting, this queueing discipline permits the analysis of the performance bound for a multi-hop session to be collapsed to the analysis of a single hop, considerably simplifying the analysis and also significantly tightening the bounds obtained via [7]. The bound obtained for the network in figure 2 under PGPS service at each multiplexer is also plotted in Figure 2.

We illustrate the statistical bounding approach by briefly considering the methodology described in [18]. In [18], no assumptions are made about the actual cell interarrival times, as is done in traditional queueing theory. Rather, for each session, a stochastic bound on the number of arrivals in *any* interval of time of length k is specified (typically, for a set of values for k). Given these stochastic bounds on traffic at the edge of the network, bounds can then be computed for each session's traffic after it passes through each multiplexer along its path in the network. Given a characterization of all sources at the "edge" of a given network, and given the routing of sessions, the process of computing performance bounds on a session-level basis is a two-step process. In the first step, all session flows are characterized at each multiplexer; in the second step performance bounds are computed. The two-step procedure is similar in spirit to [6, 7] (although quite different in what is actually computed during each step). In [18], performance bounds on the per-session distribution of delay are computed for a sample 27-session 13-node network, and are shown to be tight for some traffic parameter values but quite loose for others. An important outstanding research issue for statistical bounding approaches is the extent to which traffic can be characterized by (or policed to conform to) the form of the distributional bounds required by [18, 25, 3].

Perhaps the most important outstanding research issue for both the statistical and the deterministic bounding approaches is their reliance on the ability to bound the maximum length of each queue's busy period for a given set of traffic specifications. If this condition is not satisfied, no bound can be computed, even though it may be known (via traditional queueing analysis) that the queues themselves are indeed all "stable" (e.g., the expected delay at all queues is finite).

3.4 Observation-Based Approaches

The final set of approaches toward providing QOS guarantees are the "observation-based" approaches [11, 17, 12, 4]. In [11, 12], previously-made measurements of certain types of traffic sources are used to characterize an arriving call and in determining the call acceptance decision. We note that this has the advantage of not requiring that the call specify its traffic parameters, but that the call must belong to one of a predefined set of classes and its traffic presumably corresponds to the traffic characteristics of that class if the guarantees are to be reliable.

In the on-line approach described in [17, 4], the bandwidth requirements of already admitted, token bucket-controlled, sessions are determined from the *current, measured behavior* of these sessions rather than the traffic parameters declared by these calls when they first arrived to the network. This measured behavior, together with the declared parameters of an arriving call, are then used in making the call acceptance/rejection decision for the incoming call. Note, however, that with the observation-based approach of [17, 4], no firm QOS guarantees can be made since call admission (and thus the QOS "guarantee") is based on measured traffic loads at call admission time – loads which may change once the call is admitted. For this reason, call receiving guarantees based on observation are referred to as receiving "predictive service."

A potential advantage for offering predictive service rather than guaranteed service based on declared worst case traffic characterization is that the network may be more fully utilized. Returning to the example of Figure 2, for example, if an arriving call required an absolute delay bound of 50 ms, the bounding approaches would reject the call since in the worst case this QOS could not be guaranteed. However, since the tail of the delay distribution is negligible at 50 ms, an observation-based approach would likely admit the call and the newly admitted call would indeed be likely to meet its QOS guarantee. In [17], a simulation study of a two-hop network with predictive service also indicates that the approach may indeed provide relatively reliable guarantees. A number of open research issues remain to be addressed, however, including the effects of different measurement/estimation techniques on the protocol, the overhead involved in measurement, the influence of the number of multiplexed sessions on the reliability of the guarantees, and a thorough study of the mechanism in a larger network environment.

4 Summary

In this paper we have identified issues involved in providing QOS guarantees and briefly surveyed research in this area. We identified four approaches toward providing QOS guarantees: the tightly controlled approaches, the approximate approaches, the bounding approaches, and the measurement-based approaches.

The need to provide end-to-end QOS guarantees, while still taking advantage of the resource gains offered by a statistically multiplexed transport mechanism remains an important, yet largely unsolved problem facing BISDN architects. In a broader sense, although many of the basic hardware technological capabilities for high-speed networks are now becoming available in the laboratory, our understanding of the network's traffic, network control mechanisms, and their performance ramifications is still far behind. On-going research in the area of QOS guarantees represents a significant step in helping to close that gap.

Acknowledgments

The author would like to acknowledge the help of David Yates, who performed the simulations and calculations reflected in Figure 2. The author would also like to thank Henning Schulzrinne and Lixia Zhang for their comments on an earlier draft of this paper.

References

[1] R. Guerin, H. Ahmadi, M. Naghshineh, "Equivalent Capacity and its Application to Bandwidth Allocation in High-Speed Networks," *IEEE Journal on Selected Areas in Communications,* Vol. 9, No. 7 (Sept. 1991), pp. 968 - 991.

[2] C.S. Chang, "Stability, Queue Length and Delay, Part I: Deterministic Queueing Networks," IBM Research Report RC 17708, IBM TJ Watson Research Center, (Feb. 1992).

[3] C.S. Chang, "Stability, Queue Length and Delay, Part II: Stochastic Queueing Networks," IBM Research Report RC 17709, IBM TJ Watson Research Center, (Feb. 1992).

[4] D. Clark, S. Shenker, L. Zhang, "Supporting Real-Time Applications in an Integrated Services Packet Network: Architecture and Mechanism," *Proc. ACM SIGCOMM92,* (Aug. 1992, Baltimore, MD), pp. 14 - 26.

[5] R. Cruz, "A Calculus for Network Delay and a Note on Topologies of Interconnection Networks," Technical Report UILU-ENG-87-2246, Coordinated Science Lab., Univ. of Illinois, Urbana, IL, July 1987.

[6] R. Cruz, "A Calculus for Network Delay, Part I: Network Elements in Isolation," *IEEE Trans. on Info. Theory,* Vol. 37, No. 1 (Jan. 1991), pp. 114-131.

[7] R. Cruz, "A Calculus for Network Delay, Part II: Network Analysis," *IEEE Trans. on Info. Theory,* Vol. 37, No. 1 (Jan. 1991), pp. 132-141.

[8] A. Demers, S. Keshav, and S. Shenker, "Analysis and Simulation of a Fair Queueing Algorithm," *Internetworking: Research and Experience,* Vol. 1 (1990), pp. 3-26.

[9] A. Elwalid and D. Mitra, "Effective Bandwidth of General Markovian Traffic Sources and Admission Control of High-Speed Networks," submitted to *IEEE/ACM Transactions on Networking,* 1992.

[10] D. Ferrari and D. Verma, "A Scheme for Real-Time Channel Establishment in Wide-Area Networks," *IEEE Journal on Selected Areas in Comm.,* Vol. 8, No. 3 (April 1990), pp. 368-379

[11] J. Hyman, A. Lazar, G. Pacifici, "Real-Time Scheduling with Quality of Service Constraints," *IEEE Journal on Selected Areas in Communications,* Vol. 9, No. 7 (Sept. 1991), pp. 1052 - 1063.

[12] J. Hyman, A. Lazar, G. Pacifici, "Joint Scheduling and Admission Control for ATS-Based Switching Nodes," *Proc. ACM SIGCOMM92,* (Aug. 1992, Baltimore, MD), pp. 223 - 234.

[13] S.J. Golestani, "Congestion-Free Transmission of Real-Time Traffic in Packet Networks," *Proc. IEEE Infocom'90,* (San Francisco, June 1990), pp. 527-536.

[14] S.J. Golestani, "A Stop and Go Queueing Framework for Congestion Management," *Proc. ACM SIGCOMM'90,* (Philadelphia PA, Sept. 1990), pp. 8-18.

[15] J. Golestani, "Congestion-Free Communication in High-Speed Packet Networks," *IEEE Transactions on Communications,* Vol. 39, No. 1 (Dec. 1991), pp. 1802- 1812.

[16] P.M. Gopal and B.K. Kadaba, "Network Delay Considerations for Packetized Voice," *Performance Evaluation,* Vol. 9, No. 3 (June 1989), pp. 167-180.

[17] S. Jamin, S. Shenker, L. Zhang, and D. Clark, "An Admission Control Algorithm for Predictive Real-Time Service (extended abstract)," Proc. 3rd Int. Workshop on Network and Operating System Support for Digital Audio and Video, (Nov. 1992, San Diego CA), pp 308 - 315.

[18] J.F. Kurose, "On Computing Per-session Performance Bounds in High-Speed Multi-hop Computer Networks," *Proc. 1992 ACM SIGMETRICS/ IFIP Performance'92 Conf.* (Newport, RI, June 1992), pp. 128-139.

[19] R. Nagarajan, J. Kurose, D. Towsley, "Approximation Techniques for Computing Packet Loss in Finite-Buffered Voice Multiplexers," *IEEE J. on Selected Areas in Comm.,* Vol. 9, No. 4 (April 1991), pp. 368-377.

[20] R. Nagarajan and J. Kurose, "On Defining, Computing, and Guaranteeing Quality-of-Service in High-Speed Networks," *Proceedings of INFOCOM'92*, (Annual joint conference of the IEEE Computer and Communications Societies), pp. 2016-2025, May 1992. An extended version of this paper is available as a Technical Report 90-123, Dept. of Computer Science, U. of Mass., Amherst MA, 1992.

[21] Y. Ohba, M. Murata, H. Miyahara, "Analysis of Interdeparture Processes for Bursty Traffic in ATM Networks," *IEEE Journal on Selected Areas in Communications,* Vol. 9, No. 3 (April 1991), pp. 468 - 476.

[22] A. Parekh, R. Gallager, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks - the Single Node Case," *Proceedings of INFOCOM'92*, (Annual joint conference of the IEEE Computer and Communications Societies), May 1992, pp. 915-924.

[23] A. Parekh, R. Gallager, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks - the Multiple Node Case," to appear in *Proceedings of INFOCOM'93*, (Annual joint conference of the IEEE Computer and Communications Societies).

[24] C. Partridge, "A Proposed Flow Specification", Internet Request for Comments RFC-1363, 1992.

[25] O. Yaron and M. Sidi, "Calculating Performance Bounds in Communication Networks," to appear in *Proceedings of INFOCOM'93*, (Annual joint conference of the IEEE Computer and Communications Societies).

[26] J. Stankovic and K. Ramamritham, *Hard Real-Time Systems,* IEEE Press, 1988.

[27] D. Verma, H. Zhang, and D. Ferrari, "Guaranteeing Delay Jitter Bounds in Packet-Switching Networks," *Proc. TriComm'92,* (Chapel Hill, NC, April 1991).

[28] G. Woodruff and R. Kositpaiboon, "Multimedia Traffic Management Principles for Guaranteed ATM Network Performance," *IEEE J. on Selected Areas in Communications,* Vol. 8, No. 3 (April 1990), pp. 437-446

[29] H. Zhang and S. Keshav, "Comparison of Rate-Based Service Disciplines," *Proc. 1991 ACM Sigcomm Conference,* (Zurich, Sept. 1991), pp. 113 - 211.