IJCSC International Journal of Computer Science & Communication Vol. 1, No. 2, July-December 2010, pp. 117-120

High Speed WAN- Approaches Towards Augmented Performance

[1]Ruchira Bhargava & Madhulika Arora[2]

[1]Head R&D Cell, Lecturer, Computer Science and Engineering Department, Marudhar Engineering College, Bikaner, Research Scholar Singhania University, Jhunjhunu (Raj.)
[2]Center for E-Governance, Govt. Engineering College Bikaner (KDC) Karni Nagar Industrial Area, Bikaner (Raj)
Email: [1]ekhvaabs09@gmail.com, [2]madhulika08.arora@gmail.com

ABSTRACT

This paper focuses on quality-of-service (QOS) provided during sessions in a high-speed wide area network and briefly survey research in this area. Four approaches towards providing Quality of Service guarantees are described and discussed, e.g. the tightly controlled approach, the approximate approach, the bounding approach, and the observation-based approach. A survey research aimed at resolving the challenges and identify open issues involved in providing Quality of Service guarantees is discussed. In second section, several different approaches defining Quality of Service and discuss the fundamental challenges involved in providing performance guarantees were identified. In third section, four basic approaches for providing Quality of Service guarantees have been overviewed. Forth section concludes this paper.

Keywords: High Speed WAN, Quality of Service, Service Guarantees

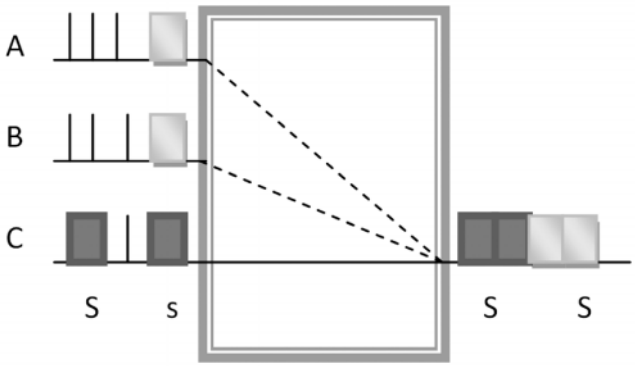
1. INTRODUCTION

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, unlike traditional data networks. The need to provide end-to-end Quality of Service 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.

A lot of problems comes when Quality of Service 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 Quality of Service, without violating existing Quality of Service guarantees made to on-going calls?" To answer this question, the network must be able to compute (or provide) not 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. 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.

2. QUALITY OF SERVICE GUARANTEES

There can be two types of Quality of Service guarantees; they are 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. With cell loss as a performance metric, the deterministic guarantee might be that no cell loss occurs. With cell loss, a statistical guarantee might be that no more than x% of the cells in the session can lost. In the case of 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 the guarantee would hold with certainty. However, if the session duration is finite, there is some probability that the guarantee would be violated.



of length I." The relationship between steady-state and interval-based Quality of Service guarantees is discussed at length in. An example which illustrates this relationship in 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} However, more than 40% of the calls would experience a packet loss probability exceeding this value if calls were five minutes in length. 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 Quality of Service guarantees have also recently been proposed as part of a flow specification in an Internet Request for Comment. Characterizing 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 (Quality of Service) 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 switch point. 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 s is output at s+2 and the session C cell arriving in time s+3 is immediately output at time s + 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.

3. FOUR APPROACHES TOWARDS PROVIDING QUALITY OF SERVICE GUARANTEES

3.1. Tightly Controlled Approaches

In tightly controlled approaches a non-work conserving multiplexing (queuing) 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. An example of a tightly-controlled approach is the so-called "stop-and-go" queuing discipline. 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 queuing further requires that a cell always be transmitted in next output frame starting after the arriving cell's input frame ends. With a mechanism such as stop-and-go queuing, 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 queuing 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. 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. But if the peak-to-average ratio is not large (i.e., the traffic itself is circuit-like in nature), tightly-controlled approaches towards providing Quality of Service guarantees could be quite attractive.

3.2. Approximate Approaches

In approximate approaches toward providing Quality of Service guarantees 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). In order to determine whether or not the multiplexed sources will receive their required Quality of Service, the queuing behavior of the multiplexed traffic streams is then analyzed. The Quality of Service calculate of interest is packet loss; in the measure of interest is maximum delay.

Approximate approaches to Quality of Service guarantees have both pros and cons over the tightly controlled approaches. Perhaps their most important benefit is their simplicity, which makes them well-suited for real-time, on-line implementation. Finally, unlike the tightly controlled approaches described above, approximate approaches are also able to take benefit of statistical multiplexing gains.

Beside this, there are several potential disadvantages. The first is the fact that the approaches are, as their name suggests, "approximate." Empirical evidence in indicates that the Quality of Service computations for certain well-defined sources tend to be conservative, thus providing a guarantee on performance. While traffic at the edge of the network may be reasonably well-approximated by such models, 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 but rather take the input traffic characteristics as a given.

In the coming subsections, a final open research issue that arises both with approximate approaches as well as with many of the approaches is that the guarantees provided are local guarantees, i.e., performance guarantees provided to a session at a single multiplexing point. User-specified Quality of Service requirements are based on an end-to-end performance requirement.

3.3. Bounding Approaches

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

Two types of verifiable performance bounds may be identified – those that provide deterministic guarantees and those which provide statistical guarantees. 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:"

Note that the bound is quite loose compared with actual delay distribution. Recently, Parekh and Gallager have developed a methodology for providing deterministic guarantees under a work-conserving queuing discipline known as packetized generalized processor sharing, PGPS (referred to as the weighted fair queuing discipline in). 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 (although quite different in what is actually computed during each step). 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.

The main 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 queuing analysis) that the queues themselves are indeed all "stable".

3.4. Observation-Based Approaches

The final set of approaches toward providing Quality of Service guarantees are the "observation-based" approaches. The 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, 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, no firm Quality of Service guarantees can be made since call admission (and thus the Quality of Service "guarantee") is based on measured traffic loads at call entrance 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. 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 some issues and methods involved in providing Quality of Service guarantees and briefly surveyed research in this area. We identified four approaches toward providing Quality of Service guarantees: the tightly controlled approaches, the approximate approaches, the bounding approaches, and the measurement based approaches. The necessitate to provide Quality of Service 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 quite far away.

On-going research in the area of Quality of Service guarantees represents a significant step in helping to close that gap.

REFERENCES

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

[2] 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.

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

[4] Demers, S. Keshav, and S. Shenker, "Analysis and Simulation of a Fair Queuing Algorithm," *Internetworking: Research and Experience*, **1**, 1990, pp. 3-26.

[5] Elwalid and D. Mitra, "Effective Bandwidth of General Markovian Traffic Sources and Ad- mission Control of High-Speed Networks," *Submitted to IEEE/ACM Transactions on Networking*, 1992.

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

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

[8] 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.

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

[10] 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*, **9**, No. 7 (Sept. 1991), pp. 968-991.

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

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

[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.