End-to-end guaranteed QoS with statistical multiplexing for ATM networks*

Sanjeev Rampal, Douglas S. Reeves, and Dharma P. Agrawal†

Departments of Electrical and Computer Engineering and Computer Science
North Carolina State University
Box 7911
Raleigh, NC 27695 USA

Abstract
We investigate a method for supporting diverse quality-of-service requirements in broadband networks based on ATM technology. The method uses deterministic bandwidth reservation at the Virtual Path (VP) level and statistical multiplexing within each VP. A deterministic server such as a Weighted Round Robin (WRR) server is used to enforce bandwidth reservations among the VPs. We develop a connection admission algorithm which accounts for end-to-end delay and loss guarantees for Virtual Circuits which traverse a single VP. We show that under certain conditions the amount of network bandwidth required by a VP is minimized by incurring all the allowable loss at the first link of a VP. Achievable utilization is demonstrated using simulation. The effect of the parameters of the WRR server (*i.e.*, the vacation time) on the cell loss probability is also studied using simulation.

Keyword Codes: C.2.1, C.2.2, C.4
Keywords: Network Architecture and Design, Network Protocols, Performance of Systems

1 INTRODUCTION

Broadband Integrated Services Digital Networks (BISDNs) of the near future will be based on the Asynchronous Transfer Mode (ATM) standard. These networks are being designed to support a wide variety of traffic types including voice, video and data. These traffic types vary widely in their bandwidth requirements, and tolerance to network delay and cell/packet loss (i.e. *Quality-of-Service* or QoS requirements).

The Connection Admission Control (CAC) function admits a new connection only if its QoS can be satisfied while continuing to meet the QoS needs of currently-admitted connections. Estimating the QoS that a connection (also referred to as a "call", "con-

*This work was supported by the Air Force Office of Scientific Research Under Grant F49620-92-J-0441.

†E-mail:{sdrampal,reeves,dpa@eos.ncsu.edu}

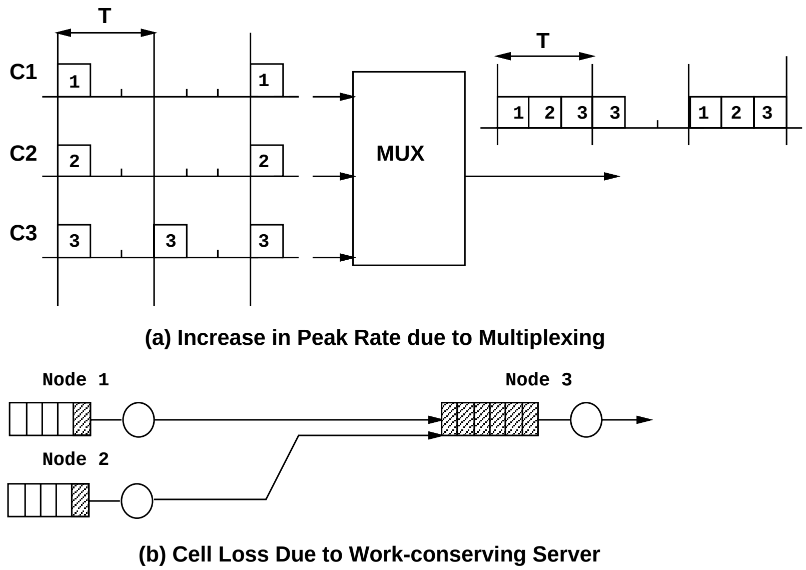

Figure 1: Problems in determining QoS for multi-hop case

nection", or "virtual circuit" in this paper) will experience is a difficult problem. Several CAC and flow control schemes for end-to-end QoS guarantees exist but require peak bandwidth allocation [10], [4]. Schemes employing statistical bandwidth allocation exist for the single link case [5],[6]. but have not been extended to the multi-hop case so far.

Figure 1 illustrates the difficulty of extending admission control based on statistical multiplexing to the multi-hop case. In Figure 1a, connection 3's minimum inter-cell arrival time changes from 3 slots to 1 slot because of the work-conserving multiplexor (*Note:* a server is work-conserving if it does not idle as long as there is a cell waiting for transmission). The user-specified parameters of a connection change as it goes through several switches and multiplexors in a multi-hop path. Figure 1b shows that in a tandem queuing environment, we can encounter cell loss because of a work-conserving cell transmission policy. In the figure node 1 and node 2 both transmit a cell to node 3 which has a full buffer, leading to cell loss. This cell loss could have been avoided by the use of feedback between adjacent nodes or by the use of a non-work-conserving cell service policy. Feedback-based schemes will be used to a limited extent in the high-speed network environment mainly because of the high delay-bandwidth product and low loss requirements. Cruz [2] has shown that with simple work-conserving FIFO, as loss tolerance approaches 0, the buffer requirement grows exponentially with the number of hops (H) in the path. Equivalently, if we use smaller buffers, we may have to allocate more than the peak bandwidth to ensure zero loss. In comparison, the use of a non-work-conserving scheduler can always guarantee zero losses, while requiring only $O(H)$ buffer requirements and peak bandwidth allocation [4]. Clearly, for the multi-hop case there are situations when it is better to be non-work-conserving than to be work-conserving.

We would like to statistically multiplex all the different traffic types together to obtain the maximum multiplexing gain. However, several studies have shown that different traffic types which vary widely in burstiness do not multiplex very well. As an example, Bae et al [1] have shown that the QoS specification for a set of heterogeneous sources may have to be set more stringent than the most stringent of any of the individual QoS requirements to ensure all sources receive their required QoS. Hence a mechanism for separating out different traffic types seems necessary.

We now describe a method which attempts to resolve some of these problems of providing guaranteed end-to-end QoS support for heterogeneous traffic. We first introduce the proposed architecture. Next we develop the theoretical basis for this approach. Finally some simulation results are presented to demonstrate the achievable utilization.

2 AN END-TO-END QOS CONTROL SCHEME

The architecture we propose for providing guaranteed QoS in ATM networks involves a reservation scheme at the virtual path (VP) level, and complete resource sharing within each VP. Flow control is completely preventive in nature, thereby avoiding problems associated with feedback-based approaches. A common problem with preventive techniques however is that they tend to be conservative and link utilization may suffer. One of the goals of our investigations is quantifying this utilization through simulations of standard traffic models. In the proposed scheme, each VP is *guaranteed* a bandwidth, which is statistically shared by all the virtual circuits (VCs) within it. In general a VP traverses multiple physical hops and a VC may also traverse multiple VPs in going from its origin to its destination.

VP bandwidth guarantees are enforced using a deterministic scheduler. In this paper we investigate the use of a Weighted Round Robin (WRR) type scheduler (also called multi-rate time-division multiplexing). The WRR scheduler is very simple to implement and analyze. Many other deterministic schedulers, such as Stop&Go [4] or Weighted Fair Queuing [10] could also be used. However, these are significantly more complicated than WRR, and with our method WRR is good enough to achieve high utilizations (as we will show). The idea of round-robin type service of different traffic classes for ATM has been suggested by others also (see Sriram's Dynamic Time Slice Scheme in particular [12]). However, to our knowledge the performance achievable by this approach has not been quantified so far, particularly for the multi-hop case.

2.1 Operation of the WRR Server

Let the length of a server cycle be T time units (we assume the time unit is the transmission time of a single cell). Let there be $K+1$ VPs being served by this server, denoted $\mathcal{V}_0$, $\mathcal{V}_1$, $\mathcal{V}_2$, $\ldots \mathcal{V}_K$. $\mathcal{V}_0$ denotes a VP carrying best-effort type traffic e.g. data files, network management traffic etc. Such traffic is not normally delay sensitive and we assume that provision of an appropriate long-term average bandwidth for such traffic is sufficient. Let the number of slots reserved for $\mathcal{V}_j$ in each server cycle be denoted n_j. Buffer space at each output link is logically partitioned such that $\mathcal{V}_j$ is allotted a buffer of size B_j^h at the h'th physical hop. In each cycle, $\mathcal{V}_j$ is served for exactly n_j slots. If $\mathcal{V}_j$ does not have a cell to transmit, a cell from B_0 (the best-effort queue) is transmitted instead. If B_0 is also

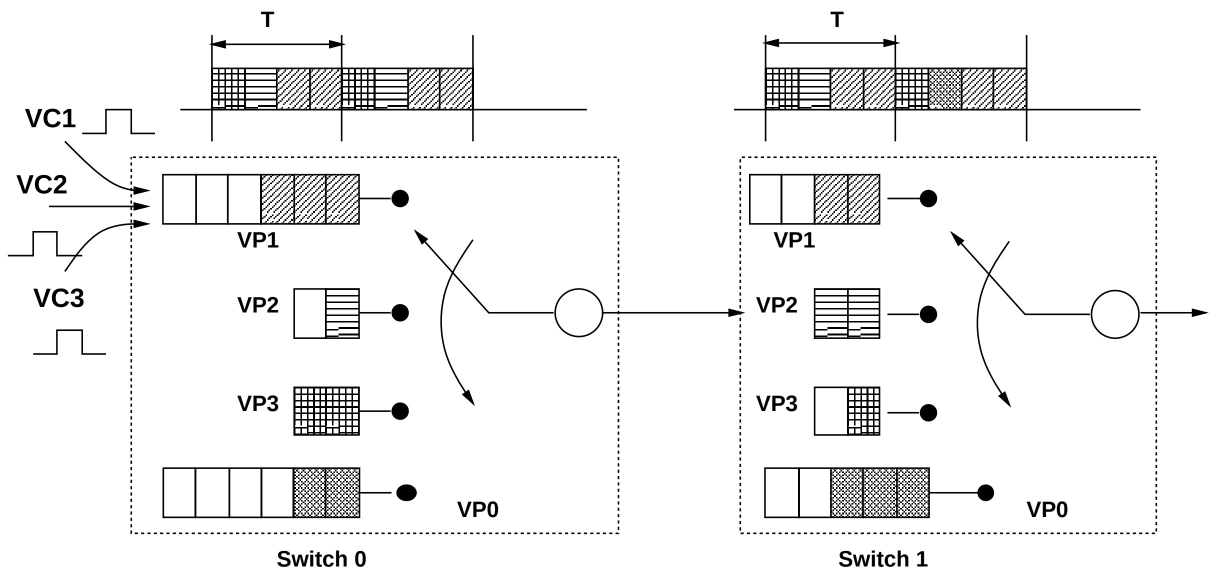

Figure 2: WRR Scheduling at the VP Level in a Multi-hop Environment

empty, no cell is transmitted and the server is idle.

These definitions are illustrated for 4 VPs in Figure 2, where $\mathcal{V}_1$ is assigned 2 slots, and $\mathcal{V}_2$ and $\mathcal{V}_3$ are assigned one slot each ($n_1 = 2$, $n_2 = 1$,$n_3 = 1$, $T = 4$). $\mathcal{V}_0$ is not assigned any slots in the cycle and only gets to transmit when a VP does not have a cell to transmit during its slot. $\mathcal{V}_1$ originates at the first node and itself consists of several bursty VCs. In general each VP sees a service "window" followed by a server "vacation" while other VPs are served.

Using this notation, we state the following theorem.

Theorem 1 *Consider a VC i assigned to a VP $\mathcal{V}_j$, which traverses H physical hops. Let each node in the path employ the WRR server described above, with identical cycle lengths of T slots. Denote by p_i^h the Cell Loss Probability of i due to overflow of the buffer B_j^h reserved for VP $\mathcal{V}_j$ preceding its h'th physical hop. If $B_j^h \geq 2 * n_j, h = 2, 3, \ldots H$, then $p_i^h = 0, h = 2, 3 \ldots H$.*

Proof : The proof follows from the non work-conserving nature of the WRR server [11] $\Box$.

If adequate buffer space is reserved and a deterministic server is used, no losses are incurred by statistical multiplexing of VCs onto a single VP at the second and successive nodes along the VP. To avoid losses at the first hop, however, (B_j^1) must be significantly greater than B_j^2, B_j^3, as we show below.

2.2 A Call Admission Procedure for a VC over a Single VP

An upper bound on the maximum end-to-end delay experienced by any cell of $\mathcal{V}_j$ which traverses H physical hops, is given by

$$D_{max} = \lfloor B_j^1/n_j \rfloor * T + (B_j^1 - \lfloor B_j^1/n_j \rfloor * n_j)/C + (H-1)*T + \sum_{h=1}^{H} T_{prop}^h \quad (1)$$

where C is the physical link capacity (assumed same for all hops), and T_{prop}^h is the propagation delay for the h'th hop [11].

Denote by $\mathcal{F}(i,j)$, the CAC function used within VP $\mathcal{V}_j$ to admit or deny admission to requesting VC i. The exact nature of the CAC function is not critical to our discussion. We assume that it determines whether the cell loss probability for i along this path will be within the user-specified allowable loss bounds. $\mathcal{F}$ is a function of the set S_j of VCs currently admitted within $\mathcal{V}_j$, the bandwidth C_j reserved for $\mathcal{V}_j$, the multiplexing buffer size B_j^1 at the first hop of $\mathcal{V}_j$ and the QoS requirements and traffic model of the requesting VC. This could be based on bandwidth tables computed offline, or achievable performance / space calculations [6] (if VPs are allowed to carry heterogeneous traffic). $\mathcal{F}(i,j)$ results in the value TRUE if the CLP is acceptable, otherwise it results in the value FALSE.

The delay specification for a voice or video connection is typically specified as $Prob(d_i > D) \leq \epsilon$, where d_i is the actual end-to-end delay for any cell of i and D is a delay bound. Equation (1) specifies a delay bound which is guaranteed 100% of the time for cells when the method of this paper is used. Thus the QoS requirements of a call are met when $\mathcal{F}(i,j) =$ TRUE and $D_{max} < D$.

To summarize, the main advantages of the proposed approach are:

- Simplified call admission since traffic within each VP is isolated from traffic in other VPs.
- Strong end-to-end guarantees on cell delay and cell loss probability can be provided.
- Network management is simplified, since by changing just the (logical) configuration of VPs and their bandwidth assignments, the service provider can control call-level grade-of-service (in terms of call acceptance probabilities) while ensuring per-call QoS (in terms of delays and CLPs).
- Fairness constraints for network accessibility (call acceptance probability) can be easily implemented, by appropriate assignment of traffic to VPs.

The drawbacks of our approach are the modest expense of implementing the WRR server, and the need for servers to cooperate, or synchronize. We believe this cooperation is not unduly constraining if it is done at a sufficiently coarse level (at the VP level).

2.3 The Maximal Gain First QoS Allocation Policy

In any scheme, the end-to-end QoS specifications are split into per-link specifications; if the "per-link" QoS specifications are met, the end-to-end QoS specification can be met [9]. We refer to these as "QoS allocation policies". In this section, we examine a particular policy for allocating an end-to-end CLP specification into per-link CLP specifications.

Consider a VP $\mathcal{V}_j$ and its route in a network along links L_1, L_2, $\ldots L_H$. The physical bandwidth of each link is C. Let us assume that $\mathcal{V}_j$ only carries homogeneous traffic i.e. all the VCs which use $\mathcal{V}_j$ have identical characteristics and QoS requirements. Let the tolerable end-to-end CLP for any VC i using $\mathcal{V}_j$ be P. Let , represent an arbitrary QoS allocation policy. Denote by P_1^Γ, P_2^Γ, $\ldots P_H^\Gamma$ the per-link CLP allowances resulting from policy , . Denote by C_h^Γ the bandwidth requirement of $\mathcal{V}_j$ at the h'th hop under QoS allocation policy , , assuming a FCFS cell service for the cells of VCs assigned to $\mathcal{V}_j$. The total network bandwidth requirement of VP $\mathcal{V}_j$ under QoS allocation policy , is just the sum of the link bandwidth requirements, i.e., $\sum_{h=1}^{H} C_h^\Gamma$.

We now justify a QoS allocation policy which we shall refer to as the Maximal Gain First, or MGF, policy. According to this policy, P is split such that $P_1^{MGF} = P$, $P_h^{MGF} = 0, h = 2, 3, \ldots H$. We first make the following explicit assumptions.

A1 Bandwidth allocation of each VP in the network is performed independently of other VPs. Under this assumption, statistical multiplexing is not performed across VPs. Assumption A1 has been applied by others to the VC level where bandwidth allocation is required to be linear [3], [5].

A2 The nodal independence assumption applies. Under this assumption, the same traffic model can be used for a VC at an intermediate node as that used at the network edge. While not true in general, empirical results supporting this assumption under some conditions have been derived by Lau [7].

A3 The traffic characteristics of each VC are such that the bandwidth requirement is monotonic wrt tolerable loss. That is, the more loss that can be tolerated by a VC, the smaller is its bandwidth requirement. This assumption is supported by the following theorem.

Theorem 2 *Given a G/D/1/K queue with server rate μ and a sample path of arrivals A_t in $[0, t)$, the steady state average cell loss probability varies monotonically with μ i.e. the higher the server rate, the lower the cell loss probability.*

Proof: This is proved using sample path techniques in [11]. □.

Admittedly, a WRR server does not strictly conform to the $G/D/1$ model because of the vacations. However, as the simulation results in section 3 will indicate, when the server cycle length is "sufficiently" small the effect of vacations is negligible.

A4 The average CLP of VCs belonging to a VP $\mathcal{V}_j$ is assumed not to be affected by the vacation times of the WRR server and depends only on the multiplexing buffer size (B_j^1) and *average* service rate (C_j). As in assumption A3, this is true if the server cycle is "sufficiently" small and is supported by the simulation results in section 3.

Based on these assumptions, we state the main theorem of this section.

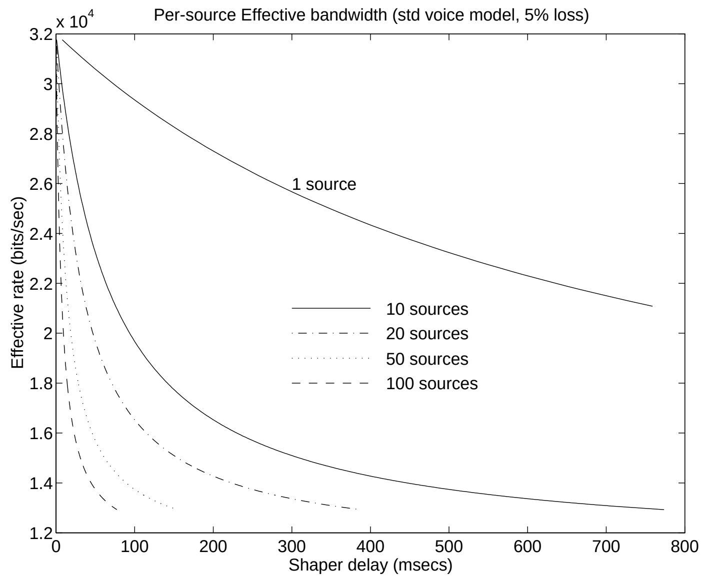

Figure 3: Per-Source Effective Bandwidth Requirement for Multiplexed Voice Sources (Fluid-flow model, peak 32Kb/s, mean 11.24 Kb/s)

Theorem 3 *Under conditions in which assumptions A1 through A4 apply, no loss allocation policy , implemented with FCFS scheduling requires less total network bandwidth than the MGF policy implemented with WRR scheduling.*

Proof: The proof is in [11]. Assumption A1 is used to examine $\mathcal{V}_j$ in isolation from other VPs in the network. The monotonicity relation in assumption A3 then yields the desired result. $\square$

This theorem shows that MGF is an optimal QoS allocation policy for our method. Total network bandwidth is minimized by forcing all the cell loss to occur at the first hop. Intuitively, the allowable end-to-end loss can be used to "buy" bandwidth reduction. If there is no additional multiplexing potential downstream, allocating the entire loss to the first hop will buy the maximal bandwidth reduction.

2.4 Multiplexing Potential of Realistic Sources

As described above, we advocate exploiting all the multiplexing potential at the first hop of a VP. We now examine the utilization achievable with a single multiplexing operation for a typical traffic model. Figures 3 and 4 show the per-source effective bandwidth requirement for a multiplexed set of voice sources and a multiplexed set of video sources, respectively. These curves are obtained by using the equivalent bandwidth approximation [5], [3]. The delay shown in the curves is the maximum queuing delay that would be encountered in the first hop of the VP onto which the sources are multiplexed. The standard On-Off model with exponentially distributed On and Off durations was used for the voice sources [12]. The video model used was a well-known Markov-modulated fluid model with 10 states [8].

From this curve the multiplexing potential of both video and voice sources can be

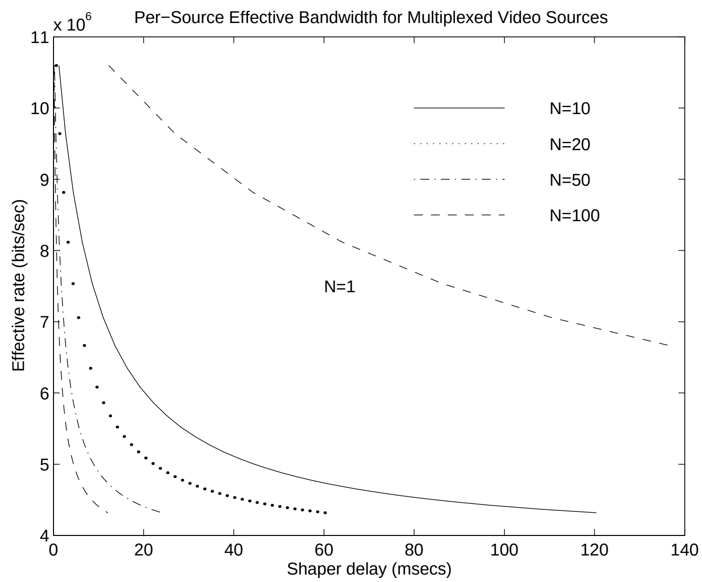

Figure 4: Per-Source Effective Bandwidth Requirement for Multiplexed Video Sources (Fluid-flow model, peak 11.7 Mb/s, mean 3.85 Mb/s)

clearly seen. For example, for 20 voice sources and an allowable queuing delay of 100 msecs the per-source effective rate is about 16 Kb/s; this represents an average utilization of 70%. For 20 video sources and an allowable delay of 100 msecs, the utilization is even better (almost 90%). For 20 voice sources and 200 msecs of delay, 80% utilization can be achieved. Typical end-to-end acceptable delay limits are up to 250-300 msecs. The equivalent capacity formulas are in fact conservative over-estimates and thus in practice even higher utilizations will be achieved [3]. Clearly, for both video and voice traffic a single level of multiplexing is able to achieve high utilization. The isolation of VPs from one another and the use of a WRR server for VPs provides the end-to-end QoS guarantees that are not possible with purely statistical approaches.

3 VARIATION OF CELL LOSSES WITH WRR PARAMETERS

In Theorem 3, we stated the optimality of the MGF policy when loss decreases monotonically as bandwidth increases. We have not, however, proved that this is always the case for the WRR service policy, In this section we present some simulation results on the variation of cell loss probability with the parameters of the WRR server.

The simulation is simplified by two observations. Due to the fact that VP bandwidths are guaranteed, we can safely analyze a single VP in isolation (i.e., without considering influence of other VPs). Also, since losses occur only at the first hop, we need only analyze the losses at the multiplexing buffer B_j^1 of $\mathcal{V}_j$ to measure the end-to-end VP loss.

We simulated a single VP $\mathcal{V}_j$, which was deterministically guaranteed a bandwidth of C_j bits/sec using a WRR server as described earlier. We denote by T_{on} and T_{off} the On and Off periods of the WRR server as seen by $\mathcal{V}_j$; the server period is $T = T_{on} + T_{off}$.

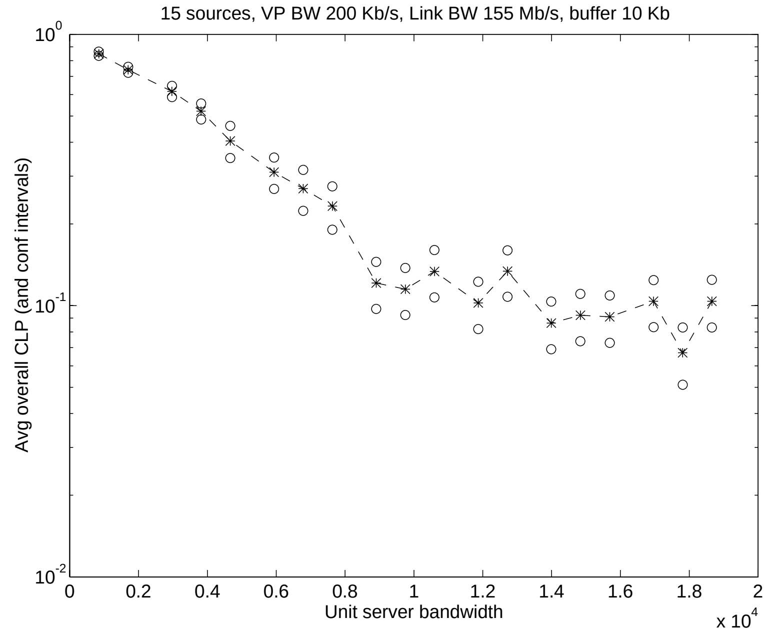

Figure 5: Variation of CLP with Unit Bandwidth ($B_j^1 = 10$ Kb)

Note that $C_j = C * T_{on}/(T_{on} + T_{off}) = C/(1 + T_{off}/T_{on})$. When T_{off} and T_{on} are varied, the bandwidth C_j remains the same if the ratio T_{off}/T_{on} remains constant. However the server "burstiness" varies as T changes. As T becomes smaller and smaller, the server performance approaches that of a uniform deterministic server. A larger value of T results in a much more "bursty" service, which may result in higher losses. However, the granularity of bandwidth allocation (which we assume is 1 cell per T time units) decreases as T increases. Since a smaller granularity results in more precise allocation of bandwidth (with less wasted bandwidth), this is desirable.

Our simulations measured the cell loss probability as a function of the server period. As mentioned, we only needed to measure the losses in the first buffer, due to the MGF policy. We used as a traffic model the standard two-state IBP model of a voice source. In all cases, 15 voice sources were multiplexed into a single VP of bandwidth $C = 200$Kb/s. The assumed link speed was 155 Mb/s.

3.1 CLP Variation with Buffer Size

Figures 5, 6 and 7 show the variation in average CLP with the unit bandwidth of the WRR server (i.e., 1 cell / T time units). For these experiments the multiplexing buffer size B_j^1 was set to 10 Kb, 20 Kb and 40 Kb respectively. Note that tolerable CLP for voice traffic is 5-10%[9].

Two distinct regions of behavior are observed. The losses vary linearly in each of these regions. For small unit bandwidths, (large values of T) the slope is high. We refer to this as region 1. For large unit bandwidths (small T), the slope is nearly zero; this is region

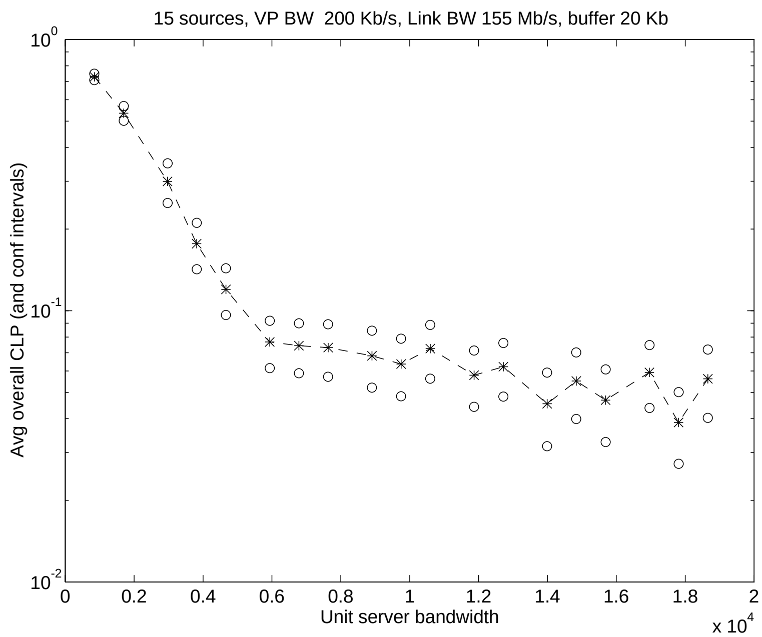

Figure 6: Variation of CLP with Unit Bandwidth ($B_j^1 = 20$ Kb)

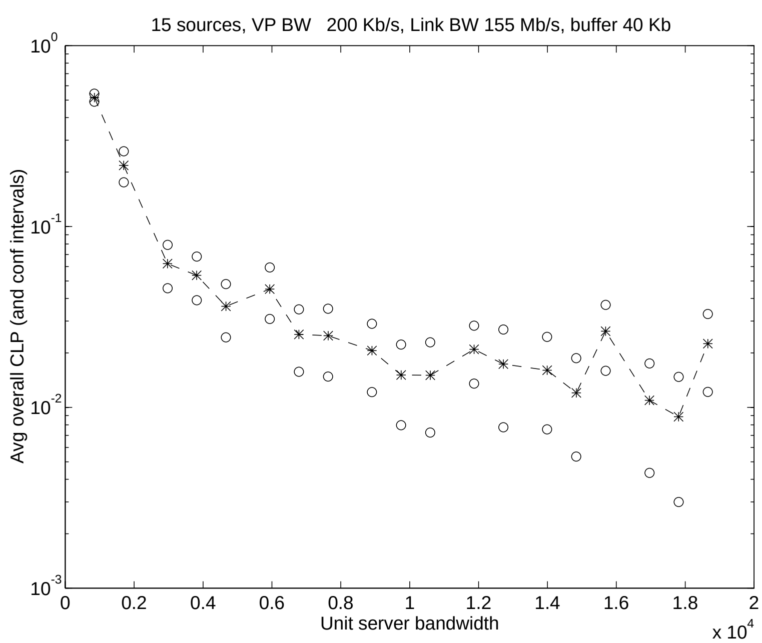

Figure 7: Variation of CLP with Unit Bandwidth ($B_j^1 = 40$ Kb)

2. The location of the transition point between the two regions is sensitive to the buffer size B_j^1. Some inferences that we can make from the above plots are as follows.

- Let $Q_{cycle} = C_j * T_{on}$ denote the amount of data of $\mathcal{V}_j$ that can be transmitted by the WRR server in one cycle. The transition point corresponds to the unit bandwidth at which $Q_{cycle} \approx B_j^1$. In region 1, $Q_{cycle} > B_j^1$, and the server completely empties the buffer during each On period. In region 2, $Q_{cycle} < B_j^1$, implying that the WRR server cannot empty the entire buffer in one On period. Once the buffer has been emptied, the remainder of the slots assigned to $\mathcal{V}_j$ will be wasted, except for cells that arrive during the On time. Hence in region 1, even the average service rate offered to $\mathcal{V}_j$ decreases linearly with the unit bandwidth (ignoring the arrivals during the On time), resulting in the exponential increase in cell loss.

- In region 2, CLP is nearly constant. This implies that as long as $Q_{cycle} < B_j^1$, the CLP is independent of the WRR cycle length. Thus the server cycle length can be chosen to optimize delay or quantization effects, as long as it is short enough to satisfy this constraint on Q_{cycle}.

- By multiplexing just 15 voice sources, we observe a CLP of less than 10% (which is acceptable for voice) while reserving 200 Kb/s. This yields a payload utilization of 93%. Clearly, in this case there is very little additional multiplexing gain to be had from these sources. The small sacrifice in utilization from VP isolation seems well worth the gain in predictability of end-to-end QoS which is obtained by our scheme.

We also simulated other choices of number of sources, bandwidths, link speeds, buffer sizes, etc. The results were similar to those reported above [11].

4 CONCLUSIONS AND FUTURE WORK

We proposed a method for providing QoS support in an ATM networks. The method uses deterministic bandwidth reservation at the VP (Virtual Path) level and statistical multiplexing of VCs within each VP. We have illustrated this method using a Weighted Round Robin (WRR) type cell scheduler to enforce the bandwidth reservations. The advantage of this method is predictable end-to-end QoS (delay and average cell loss) for Virtual Circuits. Theoretical analysis and simulations using voice traffic indicate that high utilizations are achievable.

We showed that under certain assumptions the network bandwidth which is needed by a VP to achieve a given QoS is minimized by allocating all the cell loss to the first hop. We called this the Maximal Gain First QoS allocation policy.

The effect of the parameters of the WRR server on the losses seen by individual VPs was studied using simulation. Our results indicated that the cell loss probability is insensitive to the "On" and "Off" times of the WRR server as long as they are in a certain range which is dependent on the multiplexing buffer size.

One issue for future investigation is how to allocate loss when a single VC traverses multiple VPs. This may occur when the number of VCs with the same source and destination is inadequate to achieve good utilization from statistical multiplexing. Our work also indicates renewed attention to single-hop multiplexing potential of realistic traffic sources is warranted. In particular, the issue of multiplexing vs. isolation of VCs with differing QoS specifications and traffic characteristics needs further investigation.

References

[1] J.J. Bae, T. Suda and R. Simha, "Analysis of a Finite Buffer Queue with Heterogeneous Markov Modulated Arrival Processes: A Study of the Effects of Traffic Burstiness on Individual Packet Loss," in *Proc. IEEE INFOCOM '92*, pp.219-230, 1992.

[2] R.L. Cruz, "A Calculus for Network Delay, Part II: Network Analysis," *IEEE Trans. Information Theory*, vol.37, no.1, pp.132-141, Jan 1991.

[3] A.I. Elwalid and D. Mitra, "Effective Bandwidth of General Markovian Traffic Sources and Admission Control of High Speed Networks," *IEEE/ACM Trans. Networking*, Vol.1, No.3, June 1993, pp.329-343.

[4] S.J. Golestaani, "A Framing Strategy for Congestion Management," *IEEE Journal on Selected Areas in Communications*, Vol.9, No.7, Sep 1991, pp.1064-1077.

[5] R. Guerin, H. Ahmadi and 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, Sep 1991, pp. 968-981.

[6] Y.H. Jeon and I. Viniotis, "Achievable Loss Probabilities and Buffer Allocation Policies in ATM Nodes with Correlated Arrivals," *Proc. Intl Conference on Communications*, pp.365-369, 1993.

[7] W.C. Lau and S.Q. Li, "Traffic Analysis in Large-Scale High-Speed Integrated Networks: Validation of Nodal Decomposition Approach," in *Proc. IEEE INFOCOM '93*, pp.1320-1329, 1993.

[8] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson and J. Robbins, "Performance Models of Statistical Multiplexing in Packet Video Communications," *IEEE Trans. on Communications*, Vol. 36, No. 7, July 1988, pp.834-844.

[9] R. Nagarajan, J.F. Kurose and D. Towsley, "Local Allocation of End-to-end Quality-of-Service Resources in High-Speed Networks ," *proc. IFIP Workshop on Performance Analysis of ATM Systems*, Martinique, Jan 1993.

[10] A.K. Parekh and R.G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks-the multiple node case," *Proc. IEEE INFOCOM '93*. pp.521-530, 1993.

[11] S. Rampal, D.S. Reeves and D.P. Agrawal, "A Reservation-Based Method for Providing Guaranteed QoS in ATM Networks," CCSP Technical Report, N. C. State University, Raleigh, NC 27695. Available by ftp from ftp.csc.ncsu.edu, directory /pub/rtcomm.

[12] K. Sriram, "Methodologies for bandwidth allocation, transmission scheduling, and congestion avoidance in broadband ATM networks," *Computer Networks and ISDN Systems*, vol.26, pp.43-59, 1993.