TCP/ATM Experiences in the MAGIC Testbed

Benjamin J. Ewy, Joseph B. Evans, Victor S. Frost, Gary J. Minden

Telecommunications & Information Sciences Laboratory
Department of Electrical Engineering & Computer Science
University of Kansas
Lawrence KS 66045-2228

Abstract

This paper describes performance measurements taken in the MAGIC gigabit testbed relating to the performance of TCP in wide area ATM networks. The behavior of TCP with and without cell level pacing is studied. In particular, we focus on results that indicate that the TCP rate control mechanism alone is inadequate for congestion avoidance and control in wide-area gigabit networks. We also present results showing that TCP augmented by cell-level pacing addresses these problems and allows the full bandwidth capacity to be utilized. These results demonstrate the viability of high performance distributed systems based on wide area ATM networks given the proper ATM traffic management infrastructure.

1: Introduction

The behavior of the Transmission Control Protocol (TCP) [9] has been studied by a variety of authors [3, 5, 7, 11, 12, 16]. Understanding TCP performance in networks based on SONET and ATM is of fundamental importance to the evolution of the existing national infrastructure to higher capacities, and hence to the growth of high performance distributed computing in wide area networks.

This paper presents results from experiments conducted on the MAGIC gigabit testbed. During the initial testing of the MAGIC terrain visualization application [10], observed throughput was less than that expected given the link capacity, processing capabilities of the hosts involved, and individual host to host throughput tests. The experiments we report in this paper were performed in an attempt to explain these anomalies and find solutions to the throughput limitations. We focus on results that indicate that the TCP rate control mechanism alone may be inadequate for congestion avoidance and control in wide-area gigabit networks. We also present results showing that TCP augmented by cell-level pacing addresses these problems and allows the full bandwidth capacity to be utilized.

1.1: Overview of Results

The default TCP/IP performance over congested ATM networks is poor. These congestion and buffer overflow conditions, which are caused by bandwidth mismatches or multiple sources contending for the same link, are extremely common in data networks with high performance. A number of simulation studies [5, 11] have predicted this behavior, and the measurements from MAGIC confirm this. Simulations [5] as well as preliminary experiments at the Digital Equipment Corporation Systems Research Center and in MAGIC indicate that strict link-level ATM flow control will avoid congestion-induced cell loss and provide excellent TCP throughput in local area networks, although this may be a extremely expensive solution in wide area networks.

The experimental results from MAGIC address several issues. First, is TCP rate control effective in wide area ATM networks as currently implemented? Second, if TCP rate control is not adequate, why is this the case? Third, can solutions at the ATM cell level, in particular traffic pacing, be used to improve performance?

2: Background

2.1: MAGIC Network

The Multidimensional Applications and Gigabit Internetwork Consortium (MAGIC) is a group of industrial, academic, and government organizations participating in gigabit network research. The MAGIC backbone network operates at 2.4 Gb/s and each site on the network includes LANs or hosts communicating at gigabit per second rates. The MAGIC network is depicted in Figure 1.

The University of Kansas (KU) has deployed an experimental gigabit LAN called the AN2, provided by Digital Equipment Corporation and developed by Digital's Systems Research Center [1]. Digital DEC 3000 AXP and DECStation-5000 hosts equipped with OTTO AN2 host adapter boards (which conform to standard OC-3c SONET/ATM specifications) have been attached to the switches. These hosts communicate locally via the AN2 switches, and with remote sites via the MAGIC backbone. Interoperability has been demonstrated between the Digital AN2 switch, Digital OTTO host interfaces, the Fore Systems ASX-100 switch, a variety of Fore Systems host interfaces, and Northern Telecom SONET network termination equipment.

2.2: Facilities Used for Experiments

The most unique facilities used for these experiments are the Digital OTTO OC-3c SONET/ATM interfaces, which were installed on the Turbochannel bus of several Digital DEC 3000/400 AXP workstations. This interface includes

The authors can be contacted via e-mail at evans@tisl.ukans.edu. This research is partially supported by ARPA under contract F19628-92-C-0080, Digital Equipment Corporation, the Kansas Technology Enterprise Corporation, and Sprint.

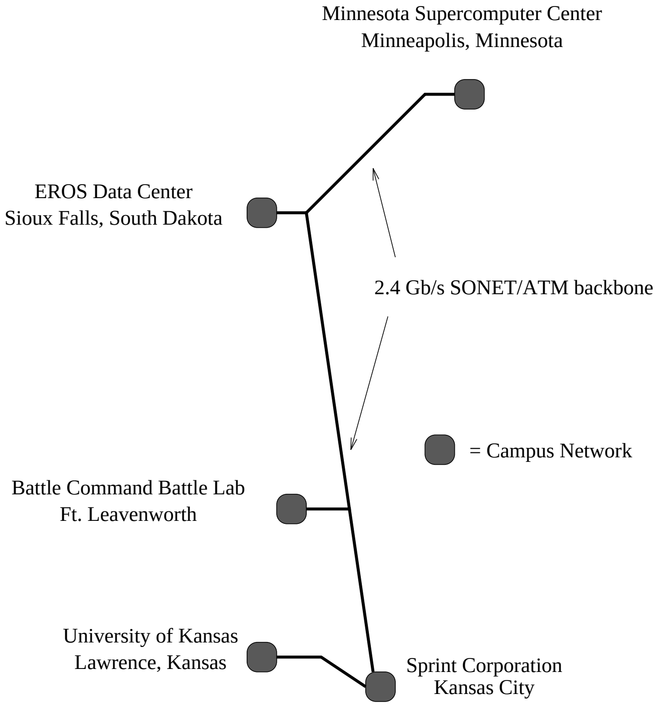

Figure 1. MAGIC Network

several features which make it interesting for network studies. The interface supports credit-based flow control, which can be used with similarly equipped interfaces or a Digital AN2 switch. This flow control guarantees no cell loss within a local area AN2 network. Cell level pacing is also implemented on the OTTO interface. This pacing was originally intended for sending constant bit rate (CBR) traffic. Pacing can be enabled on a per virtual circuit (VC) basis. It is implemented as a transmission schedule, in which a time slot is allocated to a particular VC. If no traffic needs to be sent on that VC during that time slot, the slot is available for use by other VCs. For example, VC_i in Figure 2 will use up to half of the available bandwidth if there is any traffic to send on that VC. The OTTO interfaces implement AAL5 [13], which was used exclusively in these tests, in hardware.

The other host interfaces were Fore Systems 100 Mb/s TAXI interfaces on Sun SPARCstation-10 and SGI Onyx workstations. These interfaces implement most AAL5 functionality in hardware.

The switches used for most of these studies were Fore Systems ASX-100 ATM switches equipped with a mixture of 155 Mb/s SONET OC-3c ports and 100 Mb/s TAXI ports. The ASX-100 versions used for our experiments were equipped with 256 cell output buffers per port.

The primary software tool used for our measurements was ttcp, a public domain tool originally created at the US Army Ballistics Research Lab and modified by several other authors. This application allows us to write TCP packets from local memory to memory on a remote host as quickly as the intervening operating systems, interfaces, and networks allow. This tool was extremely useful in initial network testing and for providing an upper bound on the performance of the MAGIC terrain visualization application. The correlation between ttcp results and the application performance has been shown to be excellent [14, 15]. In each of the experiments discussed in this work, ttcp used write buffers which were 64 kB each.

After accounting for ATM and physical layer (SONET or TAXI) overhead, the maximum throughput at the user level over SONET OC-3c links is approximately 134 Mb/s and 100 Mb/s TAXI links support slightly less than 90 Mb/s to the user.

3: Experiments

In order to understand TCP performance over ATM in the wide area environment, several experiments were conducted using the MAGIC testbed. Each of these trials address an issue which must be understood to determine the behavior of the system.

3.1: Experiment 1

Question: How is WAN performance effected by TCP window size?
Experiment: Transmit from a Digital DEC 3000 AXP with an OTTO OC-3c interface to another DEC 3000 AXP over a 600 km link with 8.8 ms round-trip delay.

The results are shown in Table 1. They are consistent with the theoretical limits caused by latency, that is,

$$B = \frac{W}{T}, \tag{1}$$

where W is the TCP rate control window size in bits and T is the round-trip propagation delay in seconds. Large TCP windows are necessary for acceptable throughput in this environment because of the large round-trip delay. The TCP windows extensions for sizes beyond 64 kB [8] are implemented in all of the machines used in these experiments. In this simple experiment there are no rate mismatches or switch contention problems in the network.

3.2: Experiment 2

Questions: Will high bandwidth TCP sources overrun ATM switch buffers at points of bandwidth mismatch, or will the TCP source pacing mechanism solve this problem? If TCP is not adequate, will performance be improved by cell-level pacing?
Experiment: A Digital DEC 3000 AXP (OC-3c) in Lawrence, Kansas transmits to a SPARCstation-10 (TAXI) at EDC in Sioux Falls, South Dakota (600 km), that is, a single host transmits to another host with a 155 Mb/s to 100 Mb/s bandwidth constriction in the path. The experimental conditions in this case included 128 kB TCP windows and 64 kB ttcp write buffers. The ATM pacing in this case was fixed at a bandwidth of 70 Mb/s.

The results are shown in Table 2. They demonstrate that cell-level pacing successfully compensated for rate mismatches in the ATM network, while TCP rate control alone was ineffective.

Table 2. Throughput with Rate Mismatches in an ATM WAN

No Pacing	Pacing
0.87 Mb/s	68.20 Mb/s

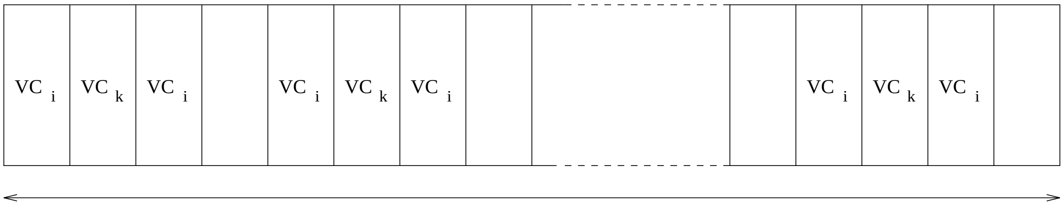

Figure 2. OTTO Pacing Scheme

Table 1. Throughput and TCP Window Size in an ATM WAN

TCP Window (bytes)	0.5k	1k	2k	4k	8k	16k	32k	64k	128k
Throughput (Mb/s)	0.47	0.93	1.8	3.7	7.4	14.9	29.8	59.6	119

Table 3. Combined Throughput with Switch Congestion in an ATM WAN

No Pacing	Pacing
1.66 Mb/s	52.36 Mb/s

Table 4. Combined Throughput with Switch Congestion in an ATM WAN

	No Pacing	Pacing
Scenario A	46.71 Mb/s	-
Scenario B	-	61.17 Mb/s

3.3: Experiment 3

Questions: Will multiple high bandwidth TCP sources overrun ATM switch buffers at multiplexing points? Will performance be improved by cell-level pacing?
Experiment: Two Digital DEC 3000 AXPs (OC-3c) in Lawrence, Kansas transmit to a SPARCstation-10 (TAXI) in South Dakota (600 km), to evaluate the effect of two hosts causing contention on the switch port to a third host. The experimental conditions included 128 kB TCP windows, 64 kB write buffers in ttcp, and ATM pacing at 35 Mb/s for each source.

The results of Experiment 3 are shown in Table 3. These indicate that pacing is very effective in improving throughput.

3.4: Experiment 4

Questions: Do the effects seen in the previous experiments arise due to interoperability problems between Fore System and Digital equipment? Can the effects of switch congestion be observed as packet losses at the TCP/IP level? What effect does pacing have in these cases?
Experiment: Two scenarios were used. In Scenario A, two SPARCstation-10s (TAXI) in Kansas and South Dakota were used to transmit to an SGI Onyx (TAXI) in Kansas City. This involved Fore interfaces and switches only. In Scenario B, two Digital DEC 3000 AXPs (OC-3c) in Lawrence, Kansas transmitted to the same SGI Onyx (TAXI) in Kansas City. In this case, the two transmitting hosts supported pacing.

The results of Experiment 4 are shown in Table 4. Two to four packet losses per second were observed in Scenario A. No packet losses were observed in Scenario B, where pacing was used. These observations indicate that better buffer utilization can be attained using cell-level pacing, which in turn leads to superior performance.

3.5: Experiment 5

Question: Will TCP rate control be more effective if the TCP segment size is small relative to switch buffer size?
Experiment: Transmit from a Digital DEC 3000 AXP (OC-3c) in Lawrence, Kansas to a Sun SPARCstation-10 (TAXI) in South Dakota (600 km), while varying the TCP segment size.

The results of Experiment 5 are shown in Figure 3. They indicate that the effectiveness of TCP rate control improves as the segment size decrease relative to the available buffers. Another effect which is illustrated by this graph is the effect of window-based pacing, that is, when the window size is sufficiently small, the throughput is limited by the round-trip delay, rather than switch congestion.

3.6: Experiment 6

Question: Does a TCP performance trade-off exist due to congestion limits versus machine processing limits?
Experiment: Experiment: Transmit directly from a Digital DEC 3000 AXP (OC-3c) in Lawrence, Kansas to another DEC 3000 AXP (OC-3c) at same location, while varying the TCP segment size.

The results of Experiment 6 are shown in Figure 4. These results indicate that TCP performance is heavily dependent on TCP segment size, at least for this experimental platform. This implies that decreasing the segment size to improve TCP rate control performance on ATM switches with small buffers must be balanced against the processing requirements of the smaller segments.

4: Extrapolation

In this section, we attempt to extrapolate from the results in the experiments to more general TCP/ATM networks. In particular, we focus on switch buffer sizes, which are currently small but can be expected to grow, and processing speed, which should continue to increase rapidly.

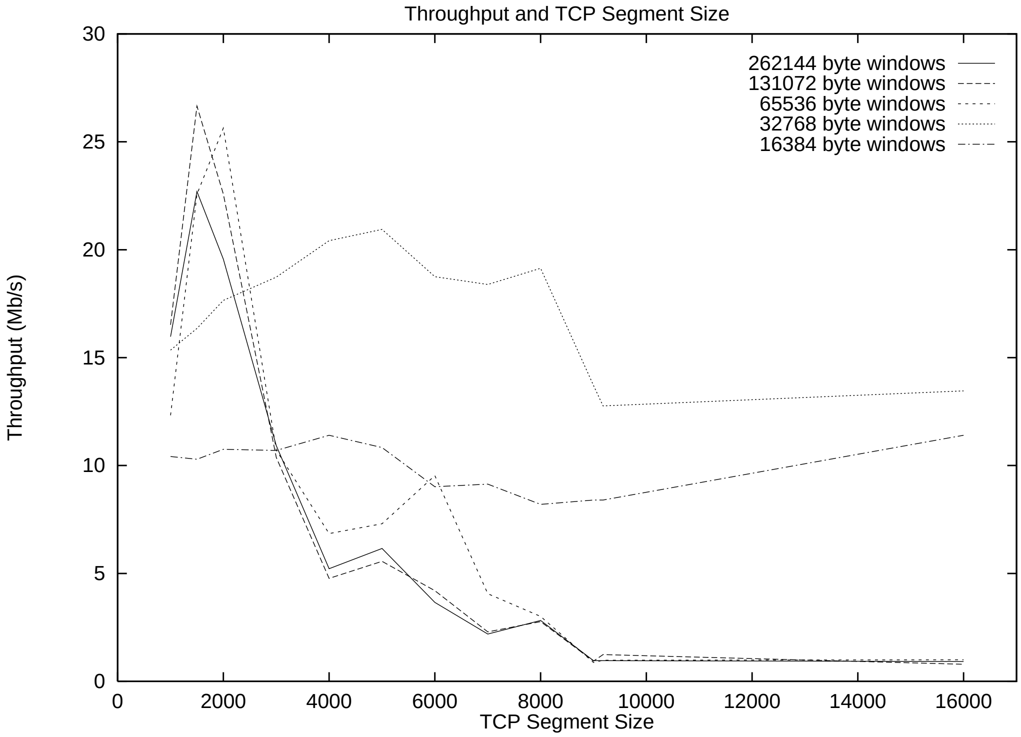

Figure 3. Throughput Variations with TCP Segment Sizes

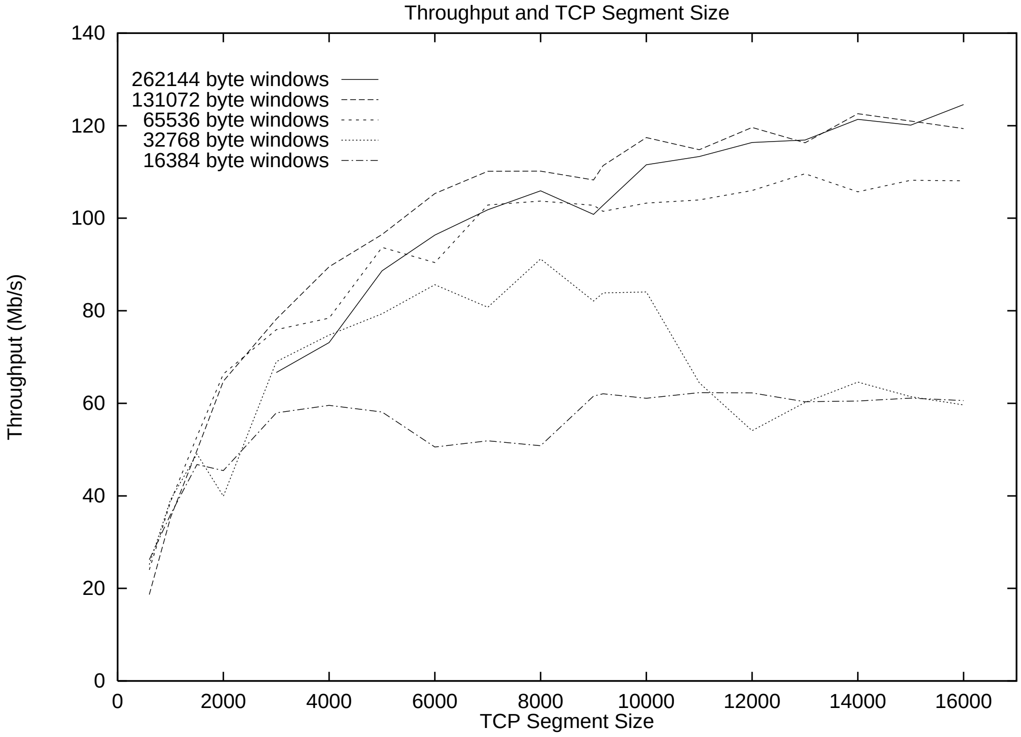

Figure 4. Throughput Variations with TCP Segment Sizes, Direct Connection

4.1: Buffer Size Effects with Bandwidth Contention

The Fore ASX-100 switches used in the majority of our experiments have only 256 cell (12288 byte) [6] buffers. Given that the standard IP MTU (maximum transmission unit) size and hence maximum TCP segment size for ATM networks is approximately 192 cells (9180 bytes) [2], not many contending segments are needed to overflow the switch buffers. Other first generation ATM switches exhibit similar buffer limitations.

We would like to determine if the effects we have seen occur only with large segment sizes relative to the available switch buffers, since larger buffers will tend to alleviate the problem. In order to understand this relationship, we assume given cell loss rate and then find the load that can be supported for a given buffer size. The motivation for this assumption is that the TCP "goodput", or throughput of error-free data, is strongly influenced by the ATM cell loss rate [5]. For example, simulations in [5] indicate that cell loss rates better than 1% are needed to provide reasonable goodput. To get an initial indication of this relationship, a simple Geometric/Geometric/1/N model and a given cell loss rate (1%) is assumed, from which we can find the number of "segment buffers" needed.

The effects of buffer size and segment size are shown in Figure 5. These results imply that TCP rate control alone may be inadequate for ATM networks, as fairly large switch buffers are needed to support contending output streams at moderate utilization.

4.2: Buffer Size Effects with Bandwidth Mismatches

A relationship was developed between buffer size, delay in the network, bandwidth mismatches and the maximum obtainable throughput for a given system [4]. This type of system is common when interfacing OC-3c ATM LANs to typical wide area networks such as 45 Mb/s DS3 connections, or LANs based on 100 Mb/s TAXI interfaces.

For a buffer of size Q bits, with an input capacity of C_i and an output capacity of C_o, and a round trip time of T seconds on the output link, the number of bits it takes to overflow a buffer can be found to be

$$\text{Bits to overflow queue} = Q(1 + \frac{C_o}{C_i - C_o}). \qquad (2)$$

This can then be used to find the maximum sustainable output rate

$$R_o = \frac{Q}{T} * (1 + \frac{C_o}{C_i - C_o}). \qquad (3)$$

Using Equation 2 and solving for a system with an OC-3 to TAXI mismatch, a switch with per port buffers of 256 cells such as the Fore ASX-100, and a network with delays such as those in Scenario B of Experiment 4, we find that the port buffers will overflow with a TCP window size greater than 34,630 bytes. Figure 6 shows the throughput for one of the Digital Alphas transmitting without cell level pacing to the SGI Onyx in Kansas City during Experiment 4. It can be clearly seen that with window sizes smaller than 34 kB the throughput reaches its CPU limited peak, but with larger windows the throughput quickly drops to very small levels due to buffer overflows. The experimental results thus match our predictions nicely.

4.3: TCP Processing Bounds

We would also like to understand the throughput limits due to segment size and machine processing speed. At least in the case of the Digital DEC 3000 AXP workstations used in these experiments, TCP appeared to be receiver CPU bound. As segment size decreases, the TCP rate control mechanism appears to be more effective at controlling congestion, as the segment size is a smaller fraction of the available buffers. On the other hand, smaller segment sizes imply more receiver processing due to the per-segment overhead (copy costs should not change). To determine if this effect is a long-term problem, we can extrapolate from measured data to higher machine capabilities. This is shown in Figure 7. Since this simple model does not model memory copy costs or other factors, it has limited predictive value, but it does illustrate the restrictive upper bound imposed by small segment sizes.

5: Conclusion

Performance in TCP/ATM wide area networks is a function of a number of factors, including switch congestion, window size limitations, and processing limitations. ATM traffic shaping (for example, as with the OTTO pacing) is critical when bottlenecks such as OC-3c to TAXI rate mismatches occur in the network, even when only single hosts are involved. ATM traffic shaping also substantially improves performance when traffic from multiple hosts is multiplexed across a single link; significant packet losses were observed even with relatively slow sources in our experiments, while pacing leads to no observed losses and higher throughput. Cell level pacing is necessary because the TCP rate control mechanism does not control traffic burstiness sufficiently to avoid congestion-induced cell losses in wide area networks, at least with the standard IP MTU size for ATM networks. This problem can be alleviated by smaller segment sizes, larger switch buffers, or strict link-level flow control, but all of these solutions have costs.

This study has illustrated some of the problems with TCP/ATM wide area networks, and demonstrated some of the solutions that will make wide area high performance distributed computing feasible.

Acknowledgements

We would like to express our thanks to Chuck Thacker, Hal Murray, and Mike Burrows of the Digital Equipment Corporation Systems Research Center for the use and support of prototype OTTO ATM OC-3c interfaces. We would also like to thank Johnny Malcom of Wayne State University for his help in the collection of the MAGIC network measurements.

References

[1] ANDERSON, T., OWICKI, C., SAX, J., AND THACKER, C. High speed switch scheduling for local area networks. In *Proc. ACM Int. Conf. Arch. Supp. Prog. Lang. and Op. Sys.* (Boston, Oct 1992).

[2] ATKINSON, R. Default IP MTU for use over ATM Adaption Layer 5 (AAL5). IETF draft-ietf-atm-mtu-07.txt, IETF, Feb 1994.

[3] COMER, D. E. *Internetworking with TCP/IP, Volume I.* Prentice-Hall, Englewood Cliffs, New Jersey, 1991.

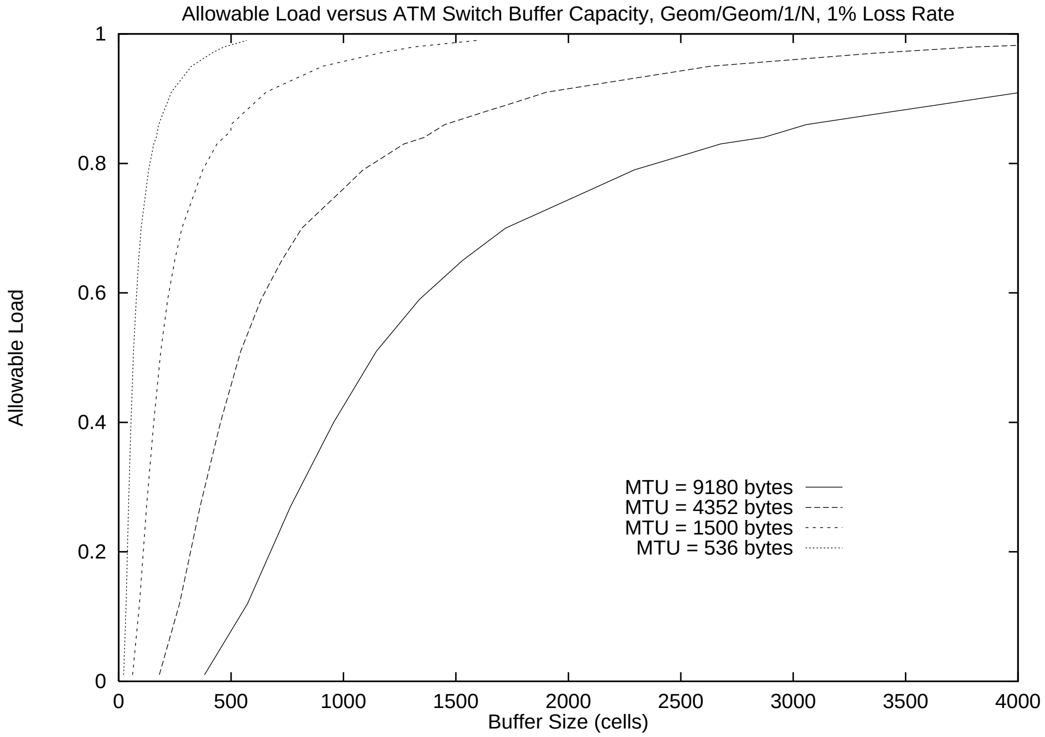

Figure 5. Allowable Load for Different Segment Sizes and Switch Buffer Sizes

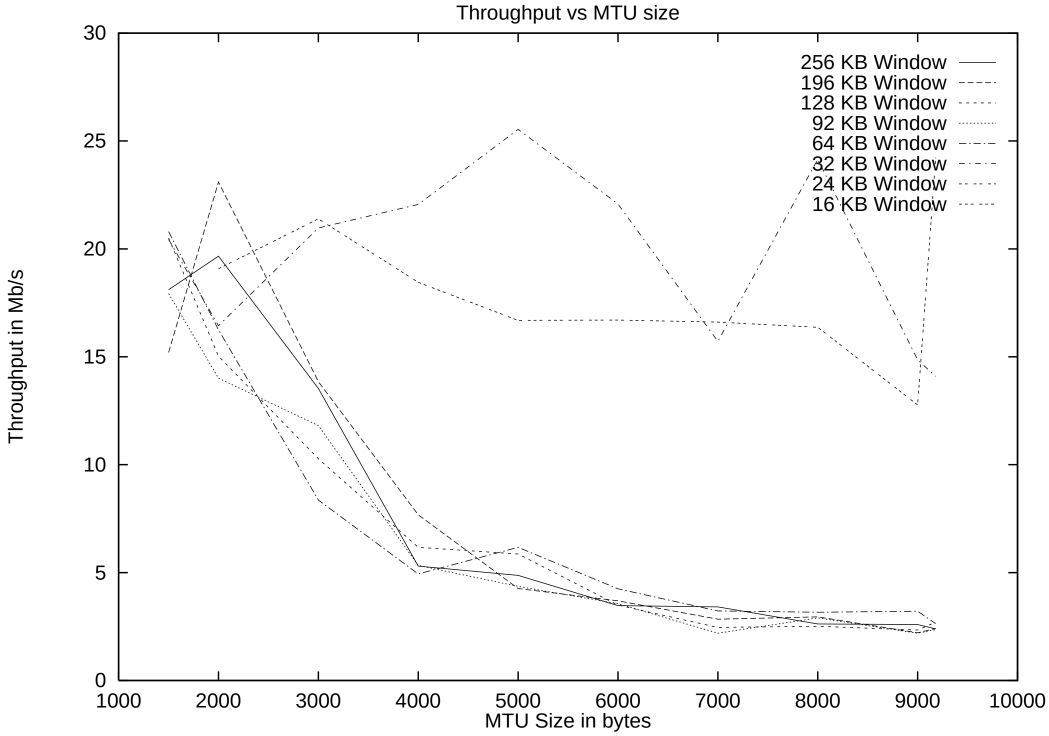

Figure 6. Throughput vs window size for various MTUs

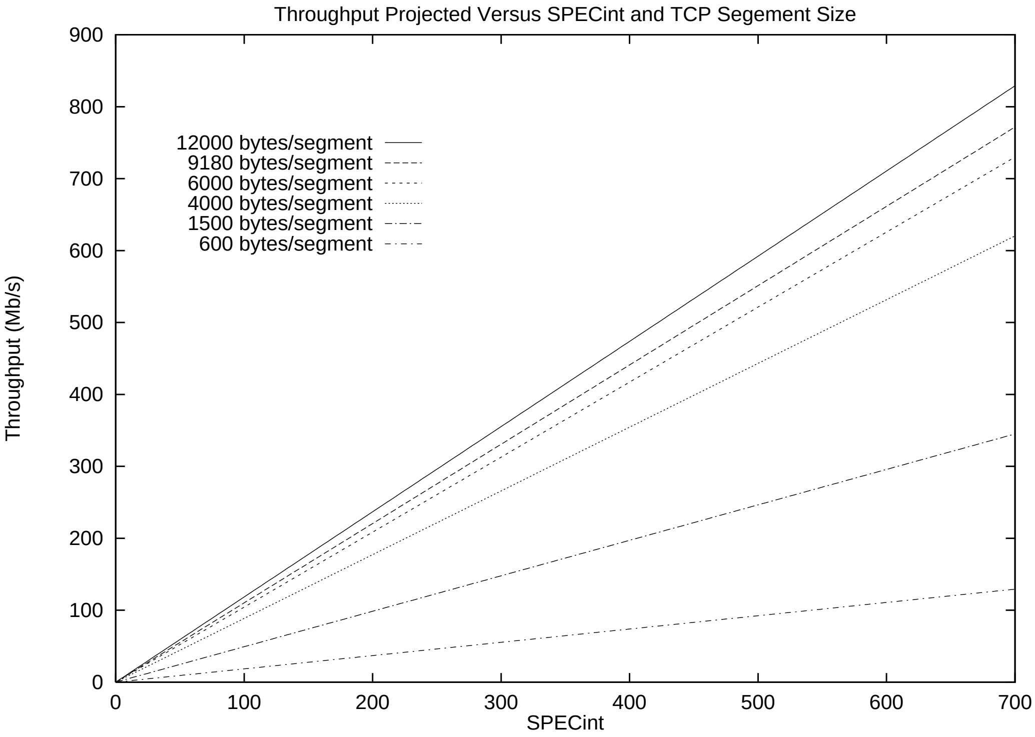

Figure 7. Throughput for Different Segment Sizes and Machine Performance

[4] EWY, B. J., EVANS, J. B., FROST, V. S., AND MINDEN, G. J. Wide area TCP/ATM measurements. In *Proc. 1994 Gigabit Testbed Maxijam* (Reston, Virginia, November 1994).

[5] FANG, C., CHEN, H., AND HUTCHINS, J. Simulation analysis of TCP performance in congested ATM LAN using DEC's flow control scheme and two selective cell-drop schemes. In *ATM Forum Contribution 94-0119* (1994).

[6] FORE SYSTEMS, INC. *ForeRunner ASX-100 ATM Switch Architecture Manual*, version 2.1 ed., 1993.

[7] JACOBSON, V. Congestion avoidance and control. In *Proc. of SIGCOMM '88* (Aug 1988), ACM.

[8] JACOBSON, V., BRADEN, R., AND BORMAN, D. TCP Extensions for High Performance. Internet Working Group Request for Comments 1323, IETF, May 1992.

[9] POSTEL, J. Transmission Control Protocol. Internet Working Group Request for Comments 793, USC/Information Sciences Institute, Marina del Rey, California, Sept 1981.

[10] RICHER, I. The MAGIC project. In *Proc. Fourth Gigabit Testbed Workshop* (Reston, Virginia, June 1993).

[11] ROMANOW, A., AND FLOYD, S. Dynamics of TCP traffic over ATM networks. In *Proc. of SIGCOMM '94* (Aug 1994), ACM.

[12] STEVENS, W. R. *TCPIP Illustrated, Volume 1*. Addison-Wesley, Readings, Massachusetts, 1994.

[13] T1S1.5/91-449, A. C. T. C. AAL 5 – A New High Speed Data Transfer AAL. Bellcore Technical Reference Issue 2, IBM et al, Dallas, Texas, Nov 1991.

[14] TIERNEY, B., JOHNSTON, W., HERZOG, H., HOO, G., JIN, G., AND LEE, J. The image server system. In *Proc. Networking '94* (Santa Fe, New Mexico, September 1994).

[15] TIERNEY, B., JOHNSTON, W., HERZOG, H., HOO, G., JIN, G., AND LEE, J. The image server system. In *Proc. 1994 Gigabit Testbed Maxijam* (Reston, Virginia, November 1994).

[16] ZHANG, L. Why TCP timers don't work well. In *Proc. of SIGCOMM '86* (Aug 1986), ACM.