

Why TCP Is Broken in ATM WANs, and How Traffic Shaping Can Fix It

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

Gigabit Networking Workshop '94

Abstract

This note describes performance measurements taken in the MAGIC gigabit testbed relating to the performance of TCP in wide area ATM networks. In particular, the behavior of TCP with and without cell level pacing is studied.

This note presents results from experiments conducted on the MAGIC gigabit testbed. 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.

Table 1 illustrates some of the results obtained transmitting from a DEC Alpha with a DEC OTTO OC-3c interface to a similarly-equipped DEC Alpha with various TCP window sizes over a 600 km link with an 8.8 ms round-trip delay. These results are consistent with the theoretical limits caused by latency. No pacing is needed in this case because there is no rate mismatch.

Table 1: Throughput and TCP Window Size in a Wide Area Network

TCP Window Size	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

The important point to note is that the window size must be fairly large to obtain reasonable throughput. The implication of this is that small TCP window sizes cannot be used for rate control in wide area networks without sacrificing significant throughput. Previous results within MAGIC have shown that an interface such as the OTTO which can source data at 130 Mb/s can overrun switches when there is a rate mismatch, that is, a link in the path at lesser rates such as 100 Mb/s, and when switch buffers are too small to absorb the bursts. Extremely short window sizes (e.g., 512 bytes) can overcome this problem in the local area, but not in the wide area due to the latency effects.

Additional results demonstrate this effect, and the need for traffic pacing at the cell level. In each case, the workstations in these tests used 128 kB TCP windows, and write buffers were 64 kB each. The Alphas used the DEC OTTO cards, the other host interfaces were Fore Systems 100 Mb/s TAXI interfaces, and the switch through which the traffic was routed was a Fore Systems ASX-100. The scenarios that were investigated were:

- Scenario 1: Alpha (OC-3c) in Lawrence, Kansas to SPARC-10 (TAXI) in South Dakota (600 km) - a single host to another host
- Scenario 2: Two Alphas (OC-3c) in Lawrence, Kansas to SPARC-10 (TAXI) in South Dakota (600 km) - two hosts to a third host

- Scenario 3a: Two SPARC-10s (TAXI) in Kansas and South Dakota to SGI Onyx (TAXI) in Kansas City - two hosts using standard Fore interfaces to a third similarly equipped host
- Scenario 3b: Two Alphas (OC-3c) in Lawrence, Kansas to SGI Onyx (TAXI) in Kansas City - two hosts supporting pacing to a third host

The throughput observed at the receiver using no transmit pacing and using pacing based on scheduled cell transmissions are shown in Table 2. Note that in Scenario 3a, 2 to 4 packet errors per second

Table 2: Throughput and Cell Level Pacing

	No Pacing	Pacing
Scenario 1	0.87 Mb/s	68.20 Mb/s
Scenario 2	1.66 Mb/s	52.36 Mb/s
Scenario 3a	46.71 Mb/s	-
Scenario 3b	-	61.17 Mb/s

were observed, while no packet errors were observed with pacing in Scenario 3b.

Some of the conclusions that might be drawn are:

- Traffic shaping (for example, the OTTO scheduling) is critical when bottlenecks such as OC-3c to TAXI rate mismatches occur in the network, even when only single hosts are involved as in Scenario 1.
- Traffic shaping 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 Scenario 3, while pacing lead to no observed losses and higher throughput to the Onyx in this case.
- 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.