

Why TCP Will Not Scale for the Next-Generation Internet

Eric Weigle, Wu-chun Feng, and Mark Gardner
{ehw, feng, mkg}@lanl.gov

Research & Development in Advanced Network Technology
Los Alamos National Laboratory
Los Alamos, NM 87545



Introduction



Goal:

- Full utilization of network end-to-end.

Problem:

- Commercial off-the-shelf computers are no longer powerful enough to fully utilize available network technologies.

Why?:

- CPU speeds double every 18 months, but network speeds double every 12.
- Bus speeds are growing slower than CPU speeds.
 - Gigabit Ethernet and 10 Gigabit ethernet in the LAN: 1 and 10Gbps
 - Dense Wave Division Multiplexing for the WAN: 6.4Tbps
 - HiPPI-6400/GSN for interconnects/LAN: 6.4 Gbps
 - PCI bus: 32bit/33MHz: 1.056 Gbps, 64bit/66MHz: 4.224Gbps.

A standard PC cannot fully utilize available bandwidth.

Setup:

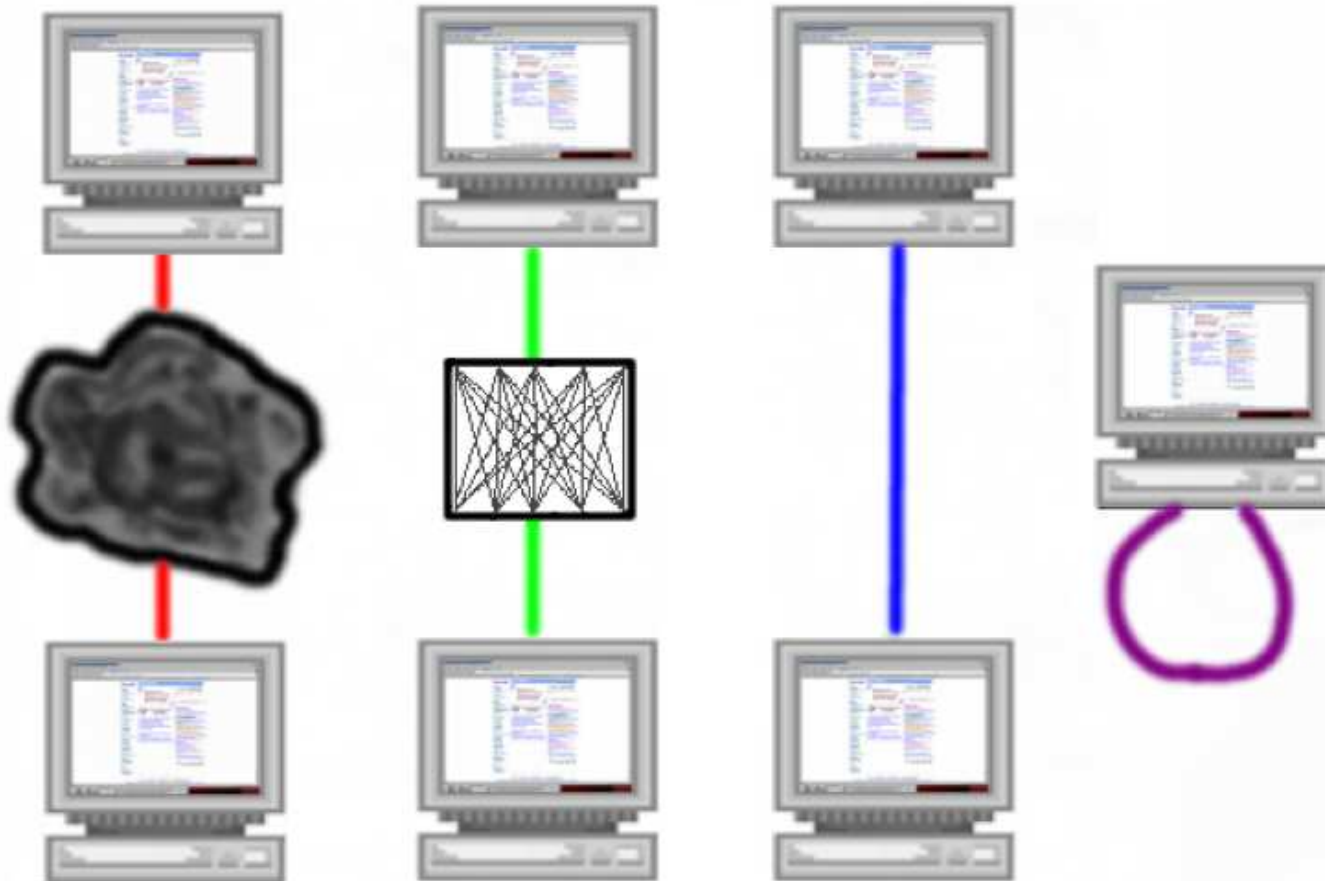
- Red Hat Linux 6.2 (kernel 2.2.17).
- Alteon AceNIC GigE cards on a 32bit/33MHz PCI bus.

Performance:

- 400MHz Pentium-II gives 335Mbps.
- 733MHz Pentium-III gives 420Mbps
 - CPU speed increased by 83%, rate by only 25%.
- Best 'general' case with tuned machines: 500Mbps.
 - Higher performance can be achieved locally.

One-half the bandwidth is **not** good enough.

Network configuration



WAN vs. LAN vs Loopback

Additional Problems



TCP (Reno) flow- and congestion-control mechanisms

- Induce bursty and chaotic behavior in the network.
- Flow control:
 - May throttle performance for inadequate window sizes.
- Default flow-control window:
 - 16 bit sequence field = 64KB window.
- Congestion control:
 - Uses only 75% of the available bandwidth.

Other versions (Vegas, etc.) address some of these issues.

1500-byte Ethernet MTUs:

- A fully utilized 10Gbps Ethernet produces 830,000 packets/sec.
- Assuming 2000 cycles/interrupt need a 1.6GHz CPU just to handle the interrupts!

Tuning:



General approach:

- 5% better by increasing buffer sizes from 64KB to 512KB.
- 10% via interrupt coalescing
- 5% via low-level tweaks
 - single-user mode, no virtual memory, realtime process

Other ways to increase bandwidth:

- Jumbo packets.
- Reduce copies in network stack-
 - 64bit/66MHz memory bus=4.2Gbps.

The Problem-- Even using loopback interface we received:

- 485Mbps for 400MHz and 590Mbps for 733MHz machines.
- software bottlenecks (TCP/IP), will persist even if hardware bottlenecks are removed.

General vs. Local Solutions



Jumbo packets, interrupt coalescing, and related approaches have problems.

Jumbo packets:

- Only effective in LANs due to fragmentation or drops in WAN.
 - Only 'guarantee' we have is that 576 byte packets will survive.
- Induce blocking on networks which do not allow out-of-band data.
 - Small, high-priority packets can be enqueued behind jumbo packets and have to wait.

Switched networks which either have a smaller MTU (e.g. ATM) or support out-of-band data (e.g. myrinet) do not suffer from this problem.

General vs. Local Solutions (2)



Interrupt coalescing:

- Forces a tradeoff between bandwidth and latency.

Protocol offloading:

- Not scalable.
 - Can not place all user network buffers on the card.
- Require more expensive cards.
 - Memory is a significant percent of the cost of a NIC.
- NIC CPUs will likely always be slower and simpler than central CPUs.

Security Implications



What happens when Bad People have fast, always-on connections?

- Trivial Denial-of-Service attacks.
 - A few PCs have comparable power to the average high-end server.
If all that power is dedicated to send packets to a server, and the network does not throttle their speeds...

TCP contains no effective security mechanisms.

Proposed but failed 'Solutions'



Use a reliable UDP over networks with large bandwidth-delay products.

- We need congestion control.
- Users want features found in TCP.
 - Nagel algorithm, urgent pointer, etc.

Use smart queueing to punish non-conformant flows

- Drop packets from UDP streams which ignore congestion, or use more than their 'fair share' of bandwidth.
- Will not scale- requires stateful routers.
- Can be circumvented via forged IP headers.
- Requires large adoption before becoming effective.

Summary



Several issues must be addressed:

- Host-interface bottlenecks (hardware and software).
- Protocol off-loading to the network interface card.
- Next-generation flow-control and congestion-control mechanisms.
- High-performance networking in a secure environment.
- Smart routers to punish non-conforming flows, e.g., UDP streams or streams from hosts generating distributed denial-of-service attacks.