High-Performance Networking

RADIANT: Research And Development In Advanced Network Technology
http://www.lanl.gov/radiant
Parallel Architectures & Performance Team
http://www.c3.lanl.gov/par_arch

Our research in high-performance networking addresses the communication needs of Grand Challenge applications over a range of environments—(1) wide-area network (WAN) in support of grids, and (2) local-area network (LAN) and system-area network (SAN) in support of network of workstations and clusters.

While the high-performance computing (HPC) community generally groups clusters and grids together as commodity supercomputing infrastructures, the networking aspects of clusters and grids are fundamentally different. In clusters, the primary communication bottleneck is the host-interface bottleneck, whereas in grids, the bottlenecks are adaptation bottlenecks; in particular, flow control and congestion control. To address these problems, we offer a set of general solutions for each of these environments.

I. Host-Interface Bottleneck in SANs & LANs

Two factors contribute to the host-interface bottleneck found in SANs and LANs: (1) software overhead that substantially increases latency and decreases throughput, and (2) the PCI I/O bus in today's PCs that artificially throttles throughput to a theoretical maximum of 4.2 Gb/s (assuming a 64-bit, 66-MHz PCI bus), or more realistically, 2.5 Gb/s due to the scheduling of the PCI bus.

Software overhead has been widely addressed with OS-bypass protocols, also known as user-level network interfaces. The OS-bypass protocols for the Quadrics and HiPPI-6400/ GSN interfaces are the Elan OS-bypass and Scheduled Transfer (ST).

The current incarnation of the PCI I/O bus simply cannot keep up with today's high-speed interconnects such as Quadrics (3.2 Gb/s), HiPPI-6400/GSN (6.4 Gb/s), or prototypical 10 Gigabit Ethernet (10.0 Gb/s).

Our[1] current solution (a home-grown, Quadrics-based cluster with Intel nodes running Linux) produces unidirectional bandwidth and latency for user-level MPI of 307 MB/s and 5 μs, respectively, as shown in Figure 1. These numbers are more than 50% better than any other technology available today. For additional information on the architecture and performance of the Quadrics[2] network, visit *http://www.lanl.gov/radiant* or peruse the following publication:

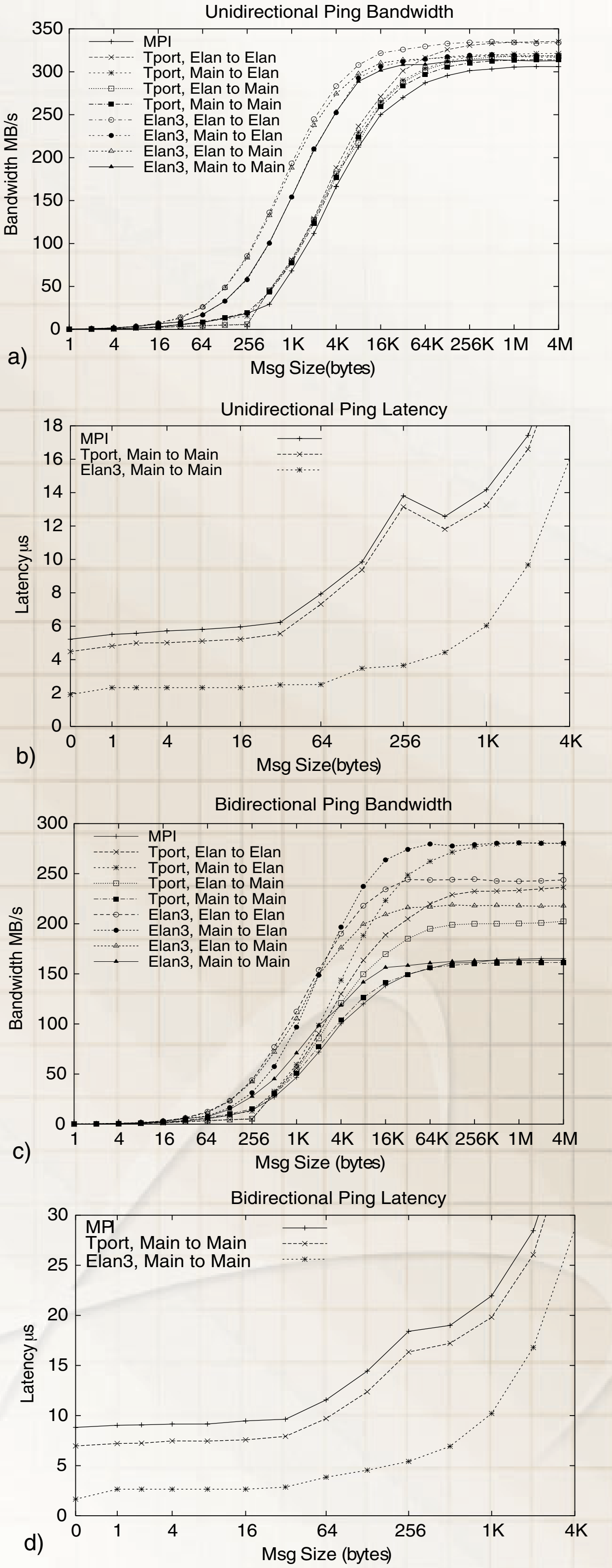

Fig. 1. Unidirectional and Bidirectional Pings.

F. Petrini, W. Feng, A. Hoisie, S. Coll, and E. Frachtenberg, “The Quadrics Network (QsNet): High-Performance Clustering Technology,” Proc. of the 9th IEEE Hot Interconnects, August 2001.

II. Adaptation Bottlenecks in WANs

WANs in support of computational grids suffer from two adaptation bottlenecks: (1) flow-control adaptation, and (2) congestion-control adaptation.[3] In this flyer, we focus on the former bottleneck by proposing a technique called dynamic right-sizing (DRS); this technique can be implemented either in kernel or user space. For our research in congestion-control adaptation, please visit *http://www.lanl.gov/radiant* for more information.

Dynamic Right-Sizing (DRS): Eliminating the Flow-Control Bottleneck

With the advent of computational grids, networking performance over the WAN has become a critical component in the grid infrastructure. Unfortunately, many high-performance grid applications only use a small fraction of their available bandwidth because operating systems are still tuned for yesterday’s WAN speeds. To address the many problems of system buffers, we present an automated, lightweight, and scalable technique called DRS, which can increase realizable throughput by an order of magnitude while abiding by TCP semantics.

DRS automatically tunes the size of system buffers over the lifetime of the connection, not just at connection setup. When implemented in the kernel, DRS produces order-of-magnitude speedups over a high-end WAN grid as shown in Figure 2—median transfer times for TCP with default flow-control windows and dynamically right-sized windows are 240 seconds and 34 seconds, respectively.

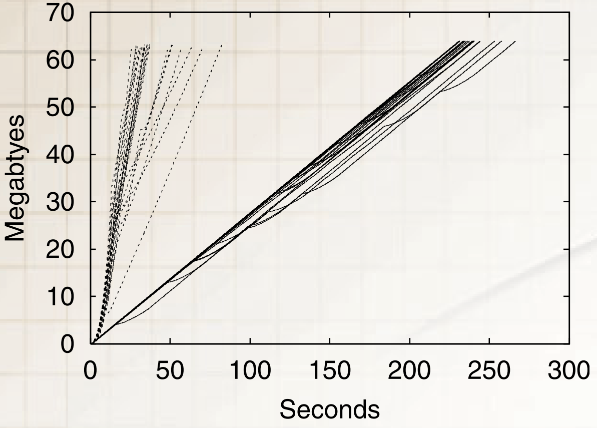

Fig. 2. Progress of Data Transfers.

We also propose a coarser-grained but more portable implementation of DRS in user space that is transparent to the end user. Specifically, we integrate our DRS technique into ftp (drsFTP). The primary difference between our drsFTP and AutoNcFTP is how buffer sizes are tuned. The buffers in AutoNcFTP are only tuned at connection setup, whereas drsFTP buffers are dynamically tuned over the lifetime of the connection, thus resulting in better adapation and better overall performance.

For additional information on dynamic right-sizing, visit *http://www.lanl.gov/radiant* or peruse the following publications:

M. Fisk and W. Feng, “Dynamic Right-Sizing in TCP,” Proc. of the 2nd Annual Los Alamos Computer Science Institute Symposium, October 2001.

E. Weigle and W. Feng, “Dynamic Right-Sizing: A Simulation Study,” Proc. of the 10th Int’l Conf. on Computer Communications and Networks, October 2001.

M. Fisk and W. Feng, “Dynamic Right-Sizing: TCP Flow-Control Adaptation (Poster),” Proc. of SC 2001: High-Performance Network and Computing Conference, November 2001.

Contact Information

radiant-info@lanl.gov

RADIANT: Research And Development in Advanced Network Technology (http://www.lanl.gov/radiant) and the Parallel Architectures & Performance Team (http://www.c3.lanl.gov/par_arch)

Computer & Computational Sciences Division
Los Alamos National Laboratory
Los Alamos, NM 87545

Sponsored in part by ASCI (Accelerated Strategic Computing Initiative) DisCom (Distance and Distributed Computing and Communication), DOE Office of Science, and DOE Los Alamos Computer Science Institute.

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36.

November 2001

LALP-01-247

A US DEPARTMENT OF ENERGY LABORATORY

Los Alamos
NATIONAL LABORATORY
Los Alamos NM 87545

[1] *i.e., Parallel Architectures & Performance team and the RADIANT team at Los Alamos National Laboratory.*

[2] *http://www.quadrics.com*

[3] *W. Feng and P. Tinnakornsrisuphap, “The Failure of TCP in High-Performance Computational Grids,” Proc. of SC 2000: High-Performance Networking and Computing Conference, November 2000.*