

LA-UR- 01 - 5675

Approved for public release;
distribution is unlimited.

e.1

Title: High-Performance Networking

Author(s): RADIANT: Research And Development In Advanced Network
Technology

Submitted to: SC 2001



Los Alamos

NATIONAL LABORATORY

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the University of California for the U.S. Department of Energy under contract W-7405-ENG-36. By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

High-Performance Networking

RADIANT: Research And Development In Advanced Network Technology (<http://www.lanl.gov/radiant>)
Computer & Computational Sciences Division
Los Alamos National Laboratory

POSTER Description of “High-Performance Networking”

HIGH-PERFORMANCE NETWORKING

Our research in high-performance networking addresses the communication needs of Grand Challenge applications over a wide range of environments — wide-area network (WAN) in support of grids and 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 networks of workstations and 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 solutions specifically tailored to each of the aforementioned environments.

FLYER Description for “High-Performance Networking”

Our research in high-performance networking addresses the communication needs of Grand Challenge applications over a wide range of environments — wide-area network (WAN) in support of grids and 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 networks of workstations and 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 solutions specifically tailored to each of the aforementioned 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

This work was supported by the U.S. Dept. of Energy’s Laboratory-Directed Research & Development Program and the Los Alamos Computer Science Institute through Los Alamos National Laboratory contract W-7405-ENG-36. Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of DOE, Los Alamos National Laboratory, or the Los Alamos Computer Science Institute.

increases latency and decreases throughput and (2) the PCI I/O bus in today’s PC compute nodes 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.

For the latter problem, there exist many solutions on the horizon, e.g., PCI-X, InfiniBand, and 3GIO or Arapahoe from Intel, but none of these solutions currently exist. Thus, 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).

The former problem 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), respectively.

Our¹ home-grown, Quadrics-based cluster with Intel nodes running Linux produces user-level (MPI), unidirectional bandwidth and latency of 307 MB/s and 5 μ s, respectively, as shown in Figure 1. These numbers are over 50% better than any other technology available today.

For additional information on the architecture and performance of the Quadrics network, visit <http://www.lanl.gov/radiant> or peruse the following publication:

• 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: A Symposium on High-Performance 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.² 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, we invite the reader to visit <http://www.lanl.gov/radiant> for more information.

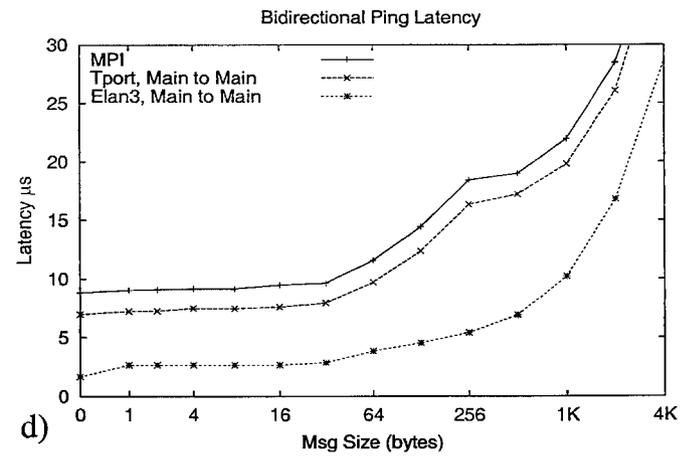
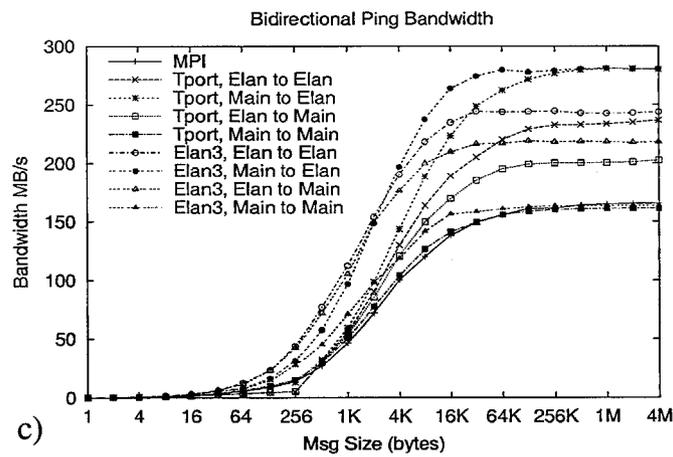
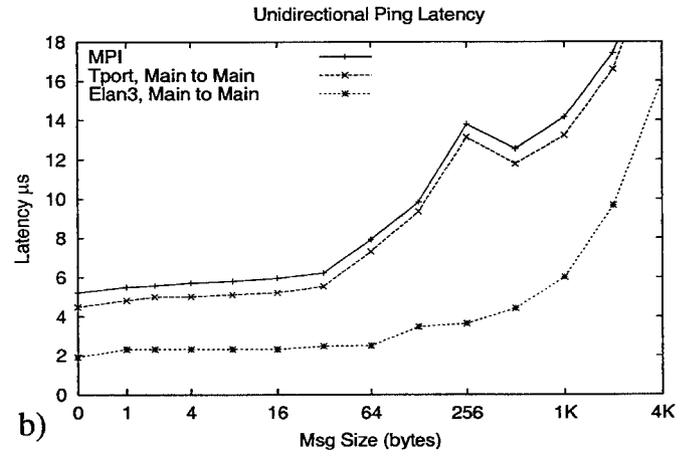
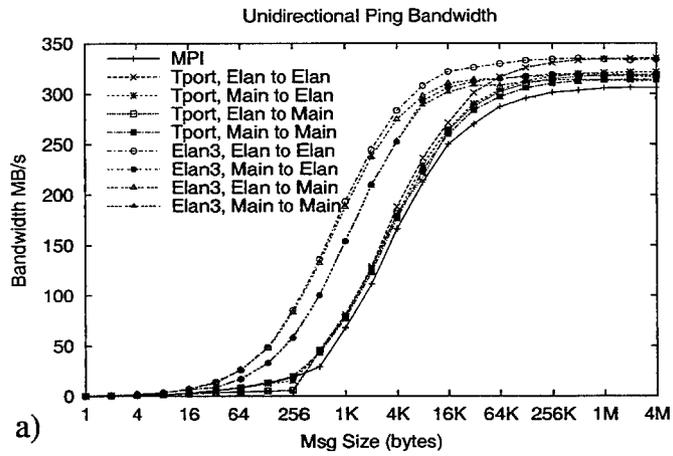
A. Dynamic Right-Sizing: Eliminating the Flow-Control Bottleneck

With the advent of computational grids, networking performance over the WAN has become a critical component in

¹i.e., Parallel Architectures team and the RADIANT team

²W. 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.

Fig. 1. Unidirectional and Bidirectional Pings



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. As a result, network gurus must undertake the tedious process of manually tuning system buffers to allow TCP flow control to scale to today's grid environments. And although recent research has shown how to set the size of these system buffers automatically at connection set-up, the buffer sizes are only appropriate at the beginning of the connection's lifetime. To address these problems, we present an automated, lightweight, and scalable technique called dynamic right-sizing (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 set-up. When implemented in the kernel, DRS produces order-of-magnitude speed-ups 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. The real-time performance of TCP with default flow-control windows and dynamically right-sized windows is shown in Figures 3 and 4.

The problem with running DRS in the kernel is that the DRS kernel patch must be installed in the operating systems of every pair of communicating hosts in a grid.³ The installation of our DRS kernel patch requires knowledge about adding modules to the kernel and *root* privilege to install the patch. Thus, the DRS functionality is not accessible to the typical end user (or developer). However, in the longer term, we anticipate that this patch will be incorporated into the kernel core so that its installation and operation are transparent to the end user.

In the meantime, end users still demand the better performance of DRS but with the pseudo-transparency of Enable and AutoNcFTP. Thus, we 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`. The differences between our DRS-`ftp` and AutoNcFTP are two-fold. First, AutoNcFTP relies on NcFTP (<http://www.ncftp.com/>), a non-standard version of `ftp` whereas DRS-`ftp` uses the standard `ftp`. Second, the buffers in AutoNcFTP are only tuned at connection set-up whereas DRS-`ftp` buffers are dynamically tuned over the lifetime of the connection, thus resulting in better adaptation 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 Adjustment of TCP Window Sizes," *Los Alamos Unclassified Report 00-3221*, July 2000.
- 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 Simula-

tion 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.

³Once installed, not only do grids benefit, but every TCP-based application benefits, e.g., `ftp`, multimedia streaming, WWW.

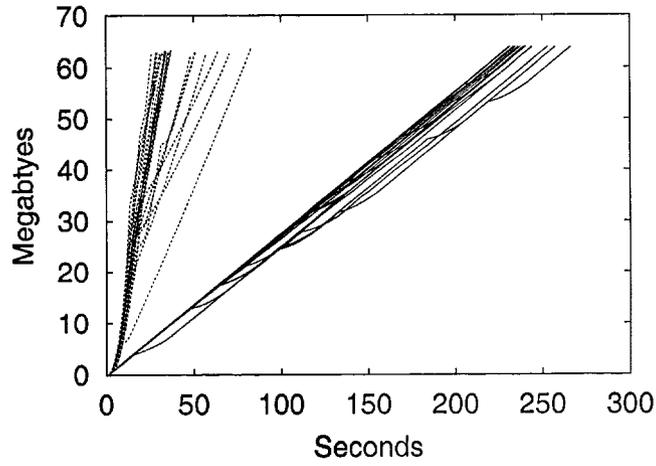


Fig. 2. Progress of Data Transfers

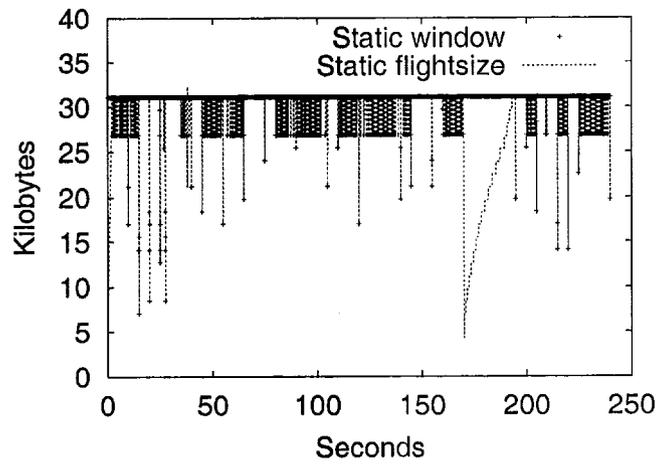


Fig. 3. Default (Static) Flow-Control Window: Flight & Congestion Window Sizes

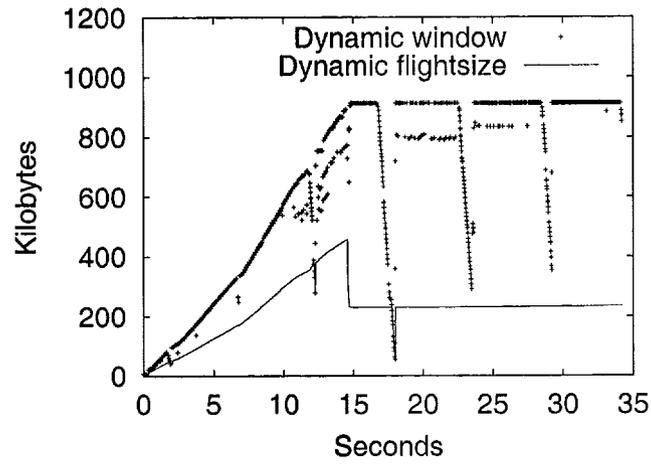


Fig. 4. Dynamic Right-Sized Window: Flight & Congestion Window Sizes