10 Gigabit Ethernet

Helps Relieve Network Bottlenecks for Bandwidth-Intensive Applications

Los Alamos National Laboratory researchers configured standards-based Dell™ servers with Intel® PRO/10GbE LR Server Adapters to test the actual network throughput of 10 Gigabit Ethernet (10GbE)–based local area networks (LANs), metropolitan area networks (MANs), and wide area networks (WANs). This article provides an overview of those tests, highlighting the key components that can help create cost-effective, high-speed network connections for a wide range of server-centric applications.

BY MATT W. BAKER AND WU-CHUN FENG

As bandwidth-intensive applications propagate, 10 Gigabit Ethernet (10GbE)[1] adapters can provide a way to increase network throughput using cost-effective, standards-based servers that are already in place. The ratified 10GbE standard is the same as previous Ethernet standards in almost every respect. Ten Gigabit Ethernet is still Ethernet, ensuring interoperability with all existing Ethernet technologies. Although standards are being developed for moving 10GbE to copper wire, the current 10GbE standard—like early forms of previous Ethernet technologies such as Gigabit Ethernet[2]—runs over various types of fiber-optic media only. The current standard enables full-duplex 10GbE transmissions over distances up to 300 meters (0.16 mile) using multimode, 850-nanometer fiber-optic cable (SR); up to 10 kilometers (6 miles) using single-mode, 1,310-nanometer optical fiber (LR); and up to 40 kilometers (25 miles) using 1,550-nanometer fiber (ER).[3]

The recently ratified fiber-optic extensions to the Ethernet standard provide 10GbE with the capability to interoperate with Synchronous Optical Network (SONET) installations, paving the way to extend 10GbE backbones into metropolitan area networks (MANs) and wide area networks (WANs). At the same time, the 10GbE standard offers the potential to increase the performance and productivity of local area networks (LANs), as shown in Figure 1.

However, with each evolutionary step, the question of actual performance versus theoretical performance arises. If historical Ethernet developments are any guide, products based on the 10GbE standard will likely reach

[1] This term does not connote an actual operating speed of 10 Gbps. For high-speed transmission, connection to a 10GbE server and network infrastructure is required.

[2] This term does not connote an actual operating speed of 1 Gbps. For high-speed transmission, connection to a Gigabit Ethernet server and network infrastructure is required.

[3] For more information, see "10 Gigabit Ethernet Technology Overview," Intel Corporation, at http://www.intel.com/network/connectivity/resources/doc_library/white_papers/pro10gbe_lr_sa_wp.pdf.

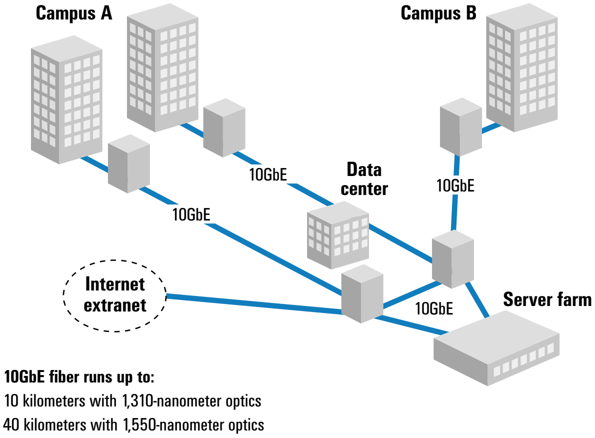

Figure 1. Configuring 10GbE links to expand LAN environments

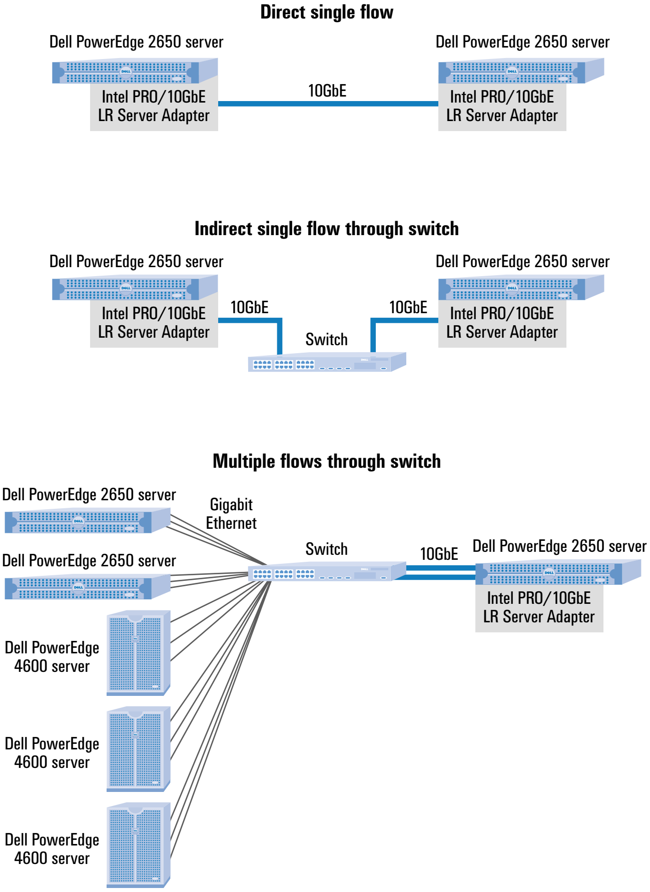
Direct single flow
Dell PowerEdge 2650 server
Dell PowerEdge 2650 server
Intel PRO/10GbE LR Server Adapter
10GbE
Intel PRO/10GbE LR Server Adapter
Indirect single flow through switch
Dell PowerEdge 2650 server
Dell PowerEdge 2650 server
Intel PRO/10GbE LR Server Adapter
10GbE
10GbE
Intel PRO/10GbE LR Server Adapter
Switch
Multiple flows through switch
Dell PowerEdge 2650 server
Gigabit Ethernet
Dell PowerEdge 2650 server
Switch
10GbE
Dell PowerEdge 2650 server
Intel PRO/10GbE LR Server Adapter
Dell PowerEdge 4600 server
Dell PowerEdge 4600 server
Dell PowerEdge 4600 server

Figure 2. Configuring various server-to-server connections in a LAN

their potential 10 Gbps transmission rate by 2005–2006. Meanwhile, IT managers pressed for more bandwidth must answer the immediate question: How fast is 10GbE right now?

Enhancing the productivity of server-based applications

Already, 10GbE network components are fast enough to satisfy many bandwidth-intensive applications. For example, to accelerate video rendering for *The Hulk*—a film released in 2003 that included a significant amount of computer-generated images—the filmmakers used a 10GbE trunk to create the conduit between servers in the artists' production network and servers in the production data center.

Meanwhile, researchers at Los Alamos and other national laboratories are using 10GbE components to power client/server data visualization applications. Los Alamos is using ParaView, an open source application designed to support distributed computing models that process large, complex data sets. Such applications help researchers analyze data sets in fields ranging from climatic modeling to DNA sequencing. Other applications enabled by 10GbE network components include *collaboratories*—which are based on a computing model that supports geographically dispersed collaborative research, particularly in scientific and engineering fields.[4]

Large-scale scientific applications benefit from 10GbE bandwidth because they typically require high-performance network connections for server-to-server as well as server-to-storage throughput. Using 10GbE bandwidth can help process distributed data up to six or seven times faster than Gigabit Ethernet in various applications. Another area in which 10GbE throughput can be beneficial is *bioinformatics*—collecting, classifying, storing, and analyzing biochemical and biological information. For example, in 2003 a worldwide research effort took several weeks to identify the Severe Acute Respiratory Syndrome (SARS) virus. Had researchers been able to take advantage of 10GbE-linked collaboratories, they could have shared data worldwide with greater speed and efficiency, quite possibly resulting in faster identification of the virus.

Putting 10GbE to the test

To evaluate the actual performance of 10GbE network components today versus their theoretical performance potential, researchers at the Los Alamos National Laboratory tested a range of LAN, MAN, and WAN configurations.

LAN test configurations

The test team created several configurations for LAN testing, including direct single flow between two servers, indirect single flow between two servers through a switch, and multiple flows between servers through a switch (see Figure 2).[5] Depending on the CPU, bus speed, and architecture of the servers used, Los Alamos

[4] For more information about collaboratories, visit http://www.accessgrid.org or http://www.scienceofcollaboratories.org.

[5] Standard Ethernet-compatible switches or routers can be used in any Ethernet configuration, including the configurations shown in Figure 2. However, these components must have 10GbE-compatible ports or adapters to make the necessary 10GbE server connections.

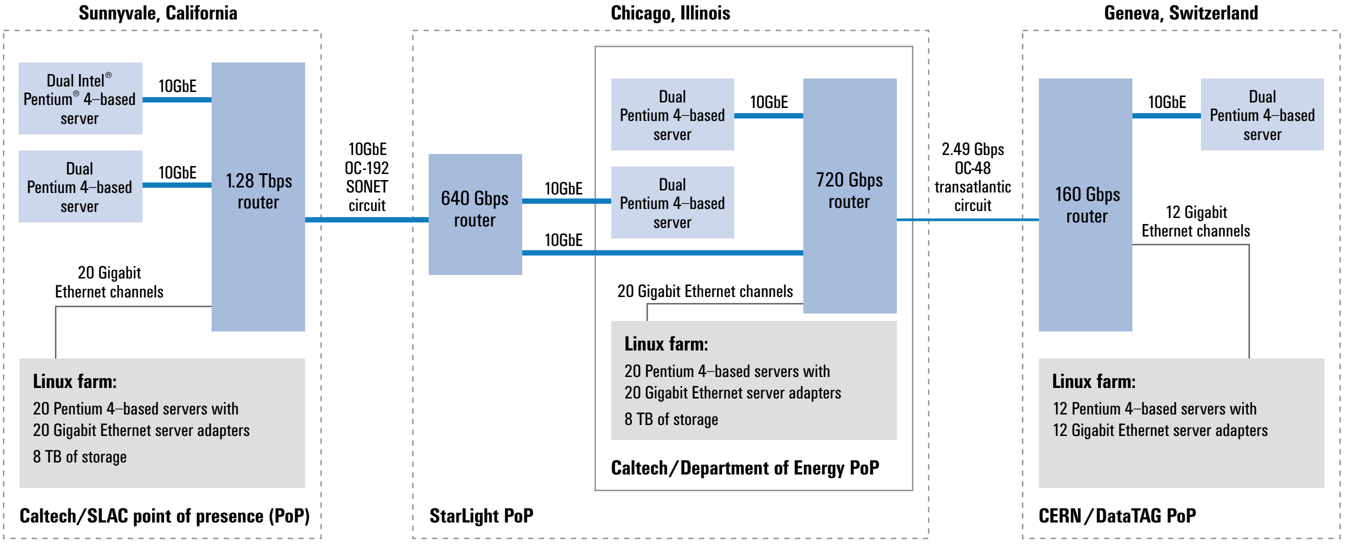
Sunnyvale, California
Chicago, Illinois
Geneva, Switzerland
Dual Intel® Pentium® 4–based server
10GbE
Dual Pentium 4–based server
10GbE
1.28 Tbps router
20 Gigabit Ethernet channels
Linux farm:
20 Pentium 4–based servers with
20 Gigabit Ethernet server adapters
8 TB of storage
Caltech/SLAC point of presence (PoP)
10GbE OC-192 SONET circuit
640 Gbps router
10GbE
10GbE
StarLight PoP
Dual Pentium 4–based server
10GbE
Dual Pentium 4–based server
720 Gbps router
20 Gigabit Ethernet channels
Linux farm:
20 Pentium 4–based servers with
20 Gigabit Ethernet server adapters
8 TB of storage
Caltech/Department of Energy PoP
2.49 Gbps OC-48 transatlantic circuit
160 Gbps router
10GbE
Dual Pentium 4–based server
12 Gigabit Ethernet channels
Linux farm:
12 Pentium 4–based servers with
12 Gigabit Ethernet server adapters
CERN/DataTAG PoP

Figure 3. Configuring 10GbE WAN environment stretching from Sunnyvale, California, to Geneva, Switzerland

researchers were able to achieve an end-to-end throughput of over 7 Gbps between applications running on different Linux®-based servers, with end-to-end latency as low as 12 microseconds.[6]

MAN and WAN test configurations

For the MAN and WAN tests, Los Alamos researchers worked with the California Institute of Technology (Caltech), the Stanford Linear Accelerator Center (SLAC), and CERN (European Organization for Nuclear Research). This technical alliance tested 10GbE performance over the WAN configuration shown in Figure 3—a true wide area network that spanned more than 9,600 kilometers (6,000 miles) between Sunnyvale, California, and Geneva, Switzerland. Even though the 2.49 Gbps OC-48 transatlantic circuit presented a significant bottleneck, the WAN test group was able to transfer more than 1 TB of data in less than an hour with a sustained Sunnyvale-to-Geneva throughput of 2.38 Gbps[6]—breaking the then-current Internet2 Land Speed Record (I2-LSR) by 2.5 times in February 2003.[7] More recently, in October 2003, a new I2-LSR record was set when Caltech and CERN researchers moved more than 1 TB of data across 7,000 kilometers (4,350 miles) in less than 30 minutes at a sustained throughput rate of 5.44 Gbps.[8]

Understanding key 10GbE network components

In both the Los Alamos LAN test configurations and the February 27, 2003, I2-LSR record-breaking configuration, the Intel® PRO/10GbE LR Server Adapter provided the Ethernet interface for Dell™ PowerEdge™ 2650 and PowerEdge 4600 servers. In the Los Alamos configurations shown in Figure 2, the PRO/10GbE adapter helped to create high-performance TCP/IP-over-Ethernet network connections without requiring any modifications to the application code.

The PRO/10GbE adapter achieved its high performance through the architecture shown in Figure 4, which is based on the Intel 82597EX single-chip controller. This controller provides capabilities including direct memory access (DMA) without register mapping, minimized programmed I/O read access, and minimized interrupts for device management. In addition, the controller offloads various TCP/IP tasks such as checksums and segmentation from the host CPU.

The host server accesses the network adapter chip through the Peripheral Component Interconnect Extended (PCI-X®) interface, which connects to a 33/66 MHz, 32-/64-bit Peripheral Component Interconnect (PCI®) bus or a 33/66/100/133 MHz, 32-/64-bit PCI-X bus. To the right of the Media Access Control (MAC) layer is an 8B/10B physical encoding sublayer and a 10 Gbps media-independent interface (XGMII) for the 1,310-nanometer serial fiber-optics module.

The fiber-optics module provides optical transmission over single-mode fiber to distances of 10 kilometers. Although the PRO/10GbE adapter can support a 20 Gbps bidirectional data rate, current host PCI-X bus bandwidth presents at least one limiting factor: the peak network bandwidth of a 133 MHz, 64-bit PCI-X bus is 8.5 Gbps. Thus, with no other bottlenecks, 8.5 Gbps would be the maximum throughput to be expected from such servers. Fortunately, moving forward, new system interconnect technologies such as PCI Express™ will remove this artificial bottleneck.

[6] "Optimizing 10 Gigabit Ethernet for Networks of Workstations, Clusters, and Grids: A Case Study" by Wu-chun Feng et al. in *Proceedings of ACM/IEEE SC 2003: High-Performance Networking and Computing Conference,* November 2003, http://www.sc-conference.org/sc2003/paperpdfs/pap293.pdf.

[7] For more information see "I2-LSR Timeline, 27 February 2003" online at http://lsr.internet2.edu/history.html.

[8] For more information, see "I2-LSR Timeline, 1 October 2003" online at http://lsr.internet2.edu/history.html.

First, the Los Alamos team tested unoptimized Dell PowerEdge 2650 servers, each with two 2.2 GHz processors, for single-flow TCP/IP throughput. The result using standard, 1500-byte maximum transfer units (MTUs) was 1.8 Gbps. When testers used a 9000-byte Jumbo frame MTU, they achieved network throughput of 2.7 Gbps. Although well short of 10 Gbps, 2.7 Gbps is a significant bandwidth advance over Gigabit Ethernet. Moreover, this throughput was accomplished using out-of-the-box PowerEdge 2650 servers.[9]

Fine-tuning network performance

The Intel PRO/10GbE LR Server Adapter enables existing network servers to enter the 10GbE performance realm. Furthermore, it is highly customizable to meet particular application requirements for performance, network stability, and quality of service. To optimize existing servers, the Los Alamos team tested several approaches to isolate bottlenecks and increase bandwidth. These approaches included tuning TCP window sizes, increasing PCI-X burst size, and tuning MTU size.

For example, on PowerEdge 2650 servers using a standard TCP stack with 9000-byte MTUs, testers improved network throughput by 33 percent—from 2.7 Gbps to 3.6 Gbps—by increasing the PCI-X burst size to 4096 bytes. Even better performance was achieved by adjusting MTU sizes. Using an 8160-byte MTU, testers achieved a peak bandwidth of 4.11 Gbps.[9] However, that result was accomplished running Linux and may be attributed to the Linux memory-allocation system; other operating systems may not achieve a comparable increase in throughput by using nonstandard MTUs.

To explore the network performance of high-end Linux-based servers, the Los Alamos test team ran preliminary trials on a uniprocessor server configured with an Intel® Itanium® 2 processor at 1.5 GHz and a PRO/10GbE LR Server Adapter. Using the same optimizations as with the PowerEdge 2650 system—an 8160-byte MTU and PCI-X burst size of 4096 bytes—the Itanium 2 processor–based server produced a unidirectional throughput of 7.2 Gbps with 12-microsecond socket-to-socket latency.[9,10]

Extending the reach of networked applications

High bandwidth and low latency are the primary features of 10GbE network performance, helping to expand application capabilities, increase productivity, and reduce time to result for large-scale, complex scientific and engineering applications. In addition, enterprises that manipulate or store vast amounts of data may benefit from 10GbE performance in various applications, including feature-length film production, publishing, high-end graphics, finance, and economic modeling.[10]

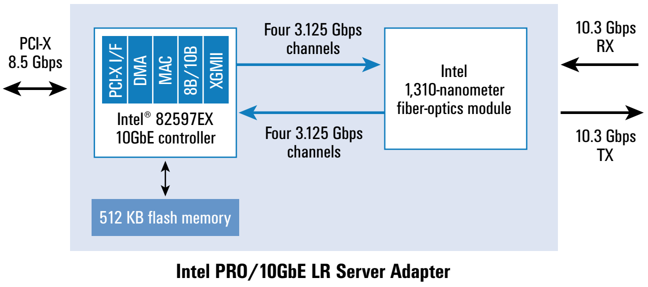

Figure 4. Exploring the Intel PRO/10GbE LR Server Adapter architecture

IT administrators also can benefit from a 10GbE deployment—for example, faster network throughput enables faster backups. At full wire speed, a backup that takes one hour with Gigabit Ethernet can be completed in six minutes with 10GbE. But that is only the beginning. Emerging 10GbE technology promises the potential to provide a unified fabric interconnect for storage that could enable administrators to minimize or possibly eliminate costly proprietary network interconnects, which are traditionally used in such areas. Using standards-based Ethernet components as the network fabric could enable administrators to reduce equipment and maintenance costs significantly.

Moreover, 10GbE networks allow connections over significantly longer distances—as far as 10 kilometers of fiber-optic cable with 1,310-nanometer optics and 40 kilometers with 1,550-nanometer optics. These transmission distances can enable administrators to extend LANs across larger campuses and move LAN data centers to more cost-effective locations.

Matt W. Baker (matt.w.baker@intel.com) is a technical marketing engineer for server network interface cards in the LAN Access Division at Intel.

Wu-chun Feng (feng@lanl.gov) leads the Research and Development in Advanced Network Technology (RADIANT) team at Los Alamos National Laboratory.

FOR MORE INFORMATION

Intel PRO/10GbE LR Server Adapter: http://www.intel.com/network/connectivity/products/pro10GbE_LR_server_adapter.htm

10 Gigabit Ethernet Technology Overview: http://www.intel.com/network/connectivity/resources/doc_library/white_papers/pro10gbe_lr_sa_wp.pdf

Internet2 Land Speed Records: http://lsr.internet2.edu/history.html

[9] "Optimizing 10 Gigabit Ethernet for Networks of Workstations, Clusters, and Grids: A Case Study" by Wu-chun Feng et al. in *Proceedings of ACM/IEEE SC 2003: High-Performance Networking and Computing Conference,* November 2003, http://www.sc-conference.org/sc2003/paperpdfs/pap293.pdf.

[10] For more information, see "Los Alamos National Lab Smashes Networking Records with Intel's 10 Gigabit Ethernet Server Adapter" online at http://www.intel.com/network/connectivity/case_studies/16832_LosAlamos_CS_r03.pdf.