

High-Performance Interconnects in Cluster Environments

*A series about
Windows high-
performance
cluster planning
and administration*

By Cornell Theory Center

Network interconnects significantly affect a cluster system's ability to achieve required performance levels. The applications' communication requirements should influence the design of high-performance clusters. This article discusses performance issues, presents some basic performance data, and briefly describes some future interconnects.

Network interconnects may be the one component in a high-performance cluster configuration that most significantly affects the system's ability to achieve required performance levels. High-performance interconnects can also become the most expensive part of the cluster configuration. This makes it important to understand the communications requirements of the applications.

Many applications perform well with a standard 100BaseT switched Ethernet. Others require the high bandwidth and low latency of a specialized network such as Emulex® (formerly Giganet) or Myrinet™. Several interconnects currently under development hold promise for the future.

Understand the applications' requirements

It is important to understand the requirements of the applications that will run on the cluster to ensure that the network configuration can handle them. For example, some applications require very little data, run on a single node, and produce very little output. At the other extreme, another application may require significant data traffic over the network before, during, and after its execution, but runs on a single cluster node.

Or, an application runs in a tightly coupled parallel fashion, often using Message Passing Interface (MPI) or Parallel Virtual Machine (PVM). The worst-case scenario combines these extreme examples.

Data requirements

Applications that have large data requirements but run on single processors typically need a network with high-bandwidth capability. Gigabit Ethernet networks, which provide approximately 30 MB/sec bandwidth, often address this need. However, because

Gigabit Ethernet uses the standard TCP/IP protocol, the network latency may not be optimal for a latency-sensitive application.

If Gigabit Ethernet performance does not provide the bandwidth required, the alternative is a network that supports a high-performance protocol such as Virtual Interface Architecture (VIA) provided by Emulex, and the Grand Message (GM) message passing system provided by Myricom. These networks can provide more than 100 MB/sec bandwidth with very low latency since they do not use the standard TCP/IP protocol and they have specialized switching hardware.

**Applications that
have large data
requirements but run
on single processors
typically need a network
with high-bandwidth
capability.**

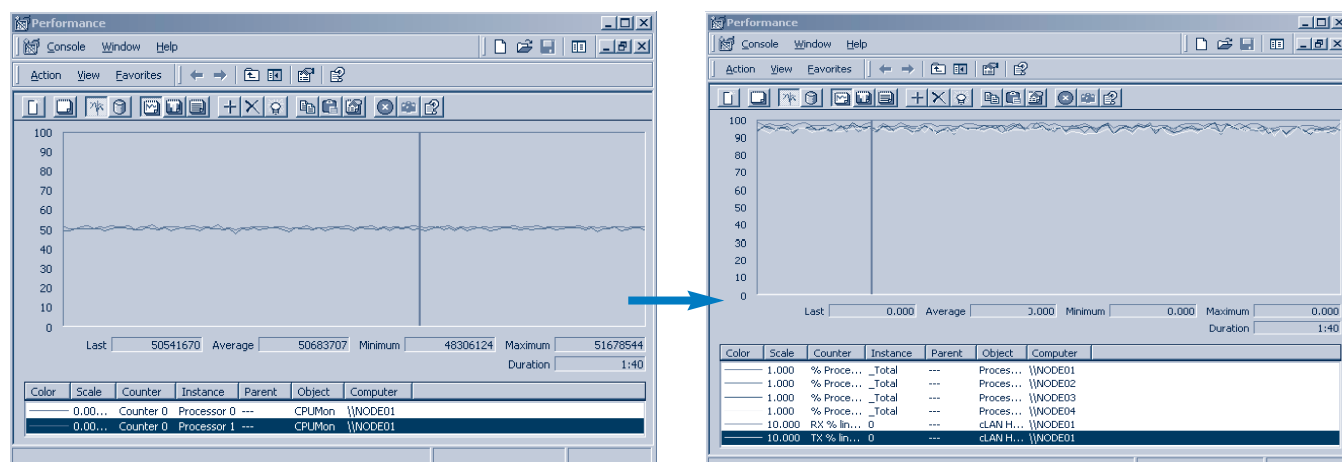


Figure 1. Message-passing code over VIA network

Interprocess communication patterns

Applications that use message passing for interprocess communication can be very sensitive to bandwidth and latency. Such applications typically are computation-intensive, written in FORTRAN, C, or C++, and communicate using MPI or PVM. Today some commercial databases including Microsoft® SQL Server, Oracle8i, and IBM® DB2® use high-performance clusters for distributed database work.

Codes that send many small messages should have a network optimized for low latency; codes that send large messages need high-bandwidth capability. Many codes need both. In addition, many applications must run in environments that have a variety of users with a broad range of requirements.

Importance of choosing the right solution

The right interconnect can make the difference between optimal and very poor performance, as the following example demonstrates. Consider a latency-sensitive code that performs sparse matrix calculations and sends a large number of messages using MPI. Run this code on four servers over a standard 100BaseT switched Ethernet network and over an Emulex cLAN™ switch using the VIA protocol. If the processors must wait for data in transit, they stop computing and the overall execution time increases.

Figure 1 shows two Windows Performance Monitor graphs produced when the code

communicates using the Emulex interface. The graph on the left displays the megaflops (MFLOPS) rate across each processor; the graph on the right shows CPU utilization across each processor. The MFLOPS hold steady at approximately 50, and the CPU utilization consistently tops 90 percent.

The same code run over standard 100BaseT switched Ethernet creates the graphs in Figure 2. The high latency of the TCP/IP-based 100BaseT network produces inconsistent CPU utilization. Since the processors must wait for data, they do not stay busy. This delay significantly affects the sustainable MFLOPS rates.

Applications that use message passing for interprocess communication can be very sensitive to bandwidth and latency. Such applications typically are computation-intensive, written in FORTRAN, C, or C++, and communicate using MPI or PVM.

Basic performance tests

In a basic performance test, a simple MPI code¹ allocates a buffer and sends it from one processor to another and then back again. Each processor “signs” the buffer to confirm the message was received and sent properly. When the message returns to the original sender, a timer indicates the roundtrip travel time. This simplistic test illustrates the basic performance capabilities of the various communication methods.

The test also sends a message through shared memory from one processor to another on the same SMP-based server and compares the performance. Figure 3 shows the results of running this code with messages in 5 MB increments (1 MB to 50 MB). The performance of all interconnects is similar with small messages, but as the message sizes increase, the performance rapidly diverges.

¹ This code, called `roundtrip.c`, is available with the online version of this article at www.dell.com/powersolutions.

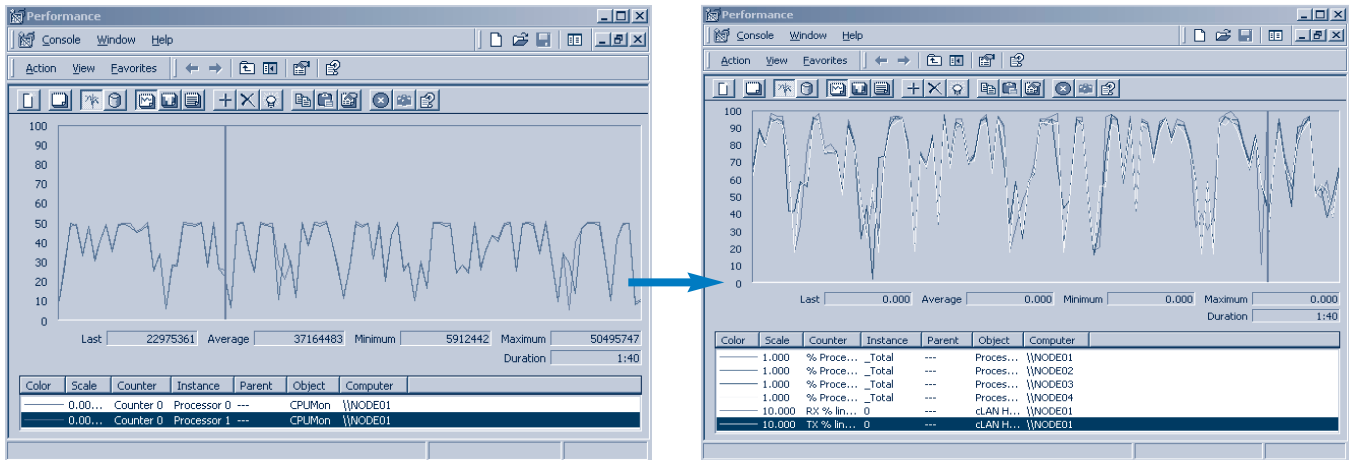


Figure 2. Message-passing code over 100BaseT network

Note that the VIA performance over Emulex is extremely close to the shared-memory performance of an SMP server. A common observation in real code is that VIA can actually outperform shared memory. That is, sending a message from a processor on one server to a processor on another can be faster over VIA than if the two processors were inside the same SMP server.

Future interconnects

Networking technology is progressing quickly. Just as industry-standard clusters are bringing better price/performance, so too

are the networks that connect them. Several new technologies anticipated within the next year should bring new levels of performance and scalability and drive down the cost of all networking equipment.

VIA over Gigabit Ethernet

Currently, VIA is available only on proprietary specialized hardware that tends to be expensive. Several companies might use special software drivers to provide VIA over standard Gigabit Ethernet equipment. This approach would provide the low latency and high

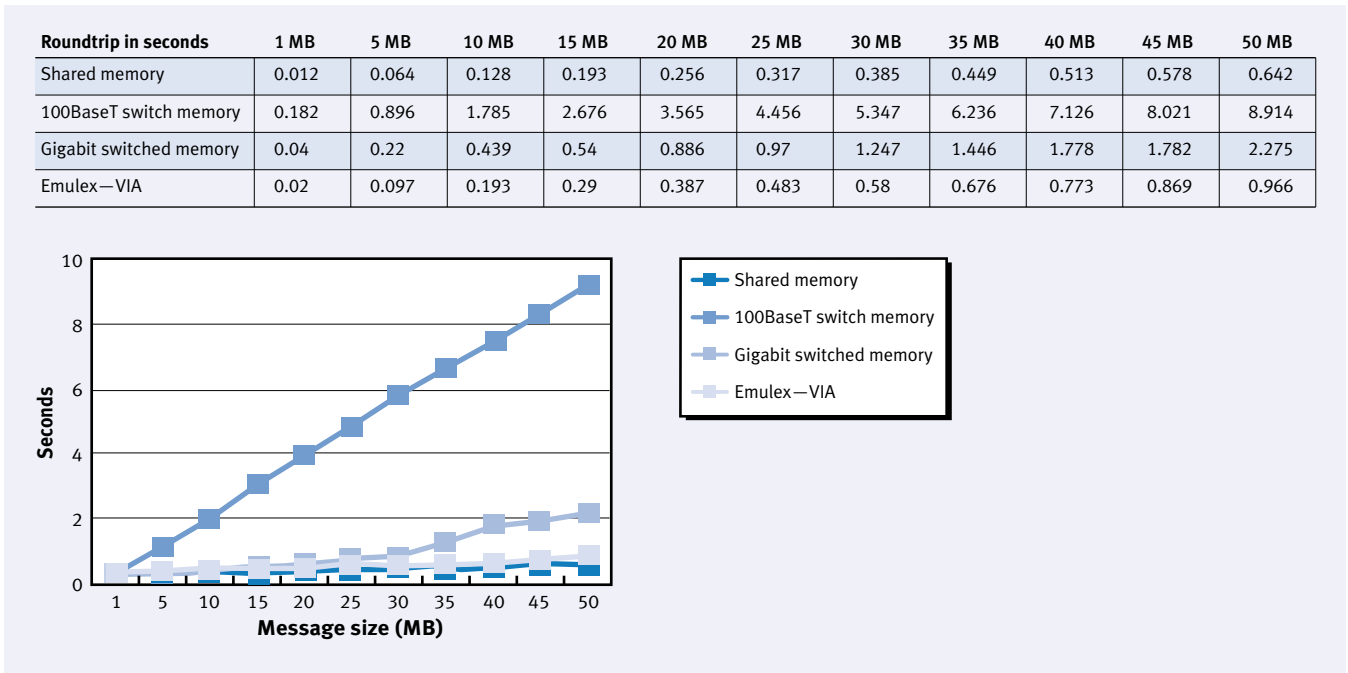


Figure 3. Roundtrip performance results (in seconds)

bandwidth of hardware such as Emulex with the cost-performance benefits of industry-standard networking hardware.

VI-IP

Emulex recently announced VIA over IP (VI-IP), which would provide the cost benefits of industry-standard Gigabit Ethernet switches and the added performance of Emulex specialized network interface cards (NICs). VI-IP will allow greater scalability compared to Emulex cLAN and still provide the same high-bandwidth, low-latency performance.

Network attached storage (NAS) from Network Appliance also supports VI-IP, providing a high-bandwidth path to storage from all servers on the network. Currently, many high-performance clusters have high-performance interconnects between compute nodes, but lower performance networks (100BaseT Ethernet) from the nodes to file servers.

Scalable Coherent Interface

Work is underway, particularly in Europe, to develop implementations of the Scalable Coherent Interface (ANSI®/IEEE® Standard 1596-1992) specification for a high-performance cluster interconnect. Published papers and books about this technology promise great performance.


Infiniband

The Infiniband™ specification for a high-performance interconnect leverages the new Intel® IA-64-based machines. Infiniband promises great performance by allowing data to flow directly from the network to memory. For optimal performance using IA-64, the processors should not wait for data. Because these new processors are so much faster, they require an I/O architecture different from the IA-32 bus architecture.

Interconnect selection influences performance

The communication needs for applications should be examined carefully when designing a high-performance cluster.

The communication needs for applications should be examined carefully when designing a high-performance network. The right network interconnects can significantly influence the system's ability to achieve required performance levels.

The right network interconnects can significantly influence the system's ability to achieve required performance levels. 

Cornell Theory Center (www.tc.cornell.edu) is a high-performance computing and interdisciplinary computational research center located at Cornell University. Researchers associated with the center work in fields such as genomics, digital materials, drug design, and financial risk analysis. CTC supports faculty and staff from more than 100 different research areas as well as corporate clients that require leading-edge computational resources.

FOR MORE INFORMATION

Infiniband:

<http://www.intel.com/technology/infiniband>

Network interconnects:

<http://wwwip.emulex.com/ip/index.html>

<http://www.myri.com/>

MPI:

<http://www.mpi-softtech.com/>

<http://www-unix.mcs.anl.gov/mpi/mpich/>

Scalable Coherent Interface:

<http://www.bode.cs.tum.edu>

<http://hsi.web.cern.ch/HSI/sci/sci.html>

<http://www.dolphinics.com/>

VIA:

<http://www.viarch.org>

VI-IP:

wwwip.emulex.com/ip/products/clan5000.html