Microway Technical Note

Visualization, Measurement, and Improvement of Linux Cluster Performance

Paul Howard

Chief Scientist, Microway, Inc.

A cluster of Linux compute nodes can be coupled with an interconnect fabric such as Ethernet or InfiniBand for High Performance Computing. Communication between compute nodes is commonly achieved using a library implementing the MPI message-passing standard. Tools for profiling and visualizing application performance on a single node are available to help the application programmer improve local code performance. Microway provides visualization tools that help identify and address global network and MPI performance issues, and in this note we use the tools to investigate the performance of existing and emerging interconnect technologies.

Visualizing system performance

In a Linux Cluster, application performance may be limited by interconnect performance and degraded by issues with the interconnect, such as loose cables or mismatched channel adapters. Microway's MPI Link-Checker™ can help in rapidly identifying and diagnosing system and network problems. Figure 1 shows a DDR (double data rate) InfiniBand-based cluster with four improved latency nodes (the light + in the left-hand chart, corresponding to four low-latency TriCom-X™ InfiniBand HCAs) and two reduced-bandwidth nodes (the darker + in the right-hand chart, corresponding to two SDR HCAs).
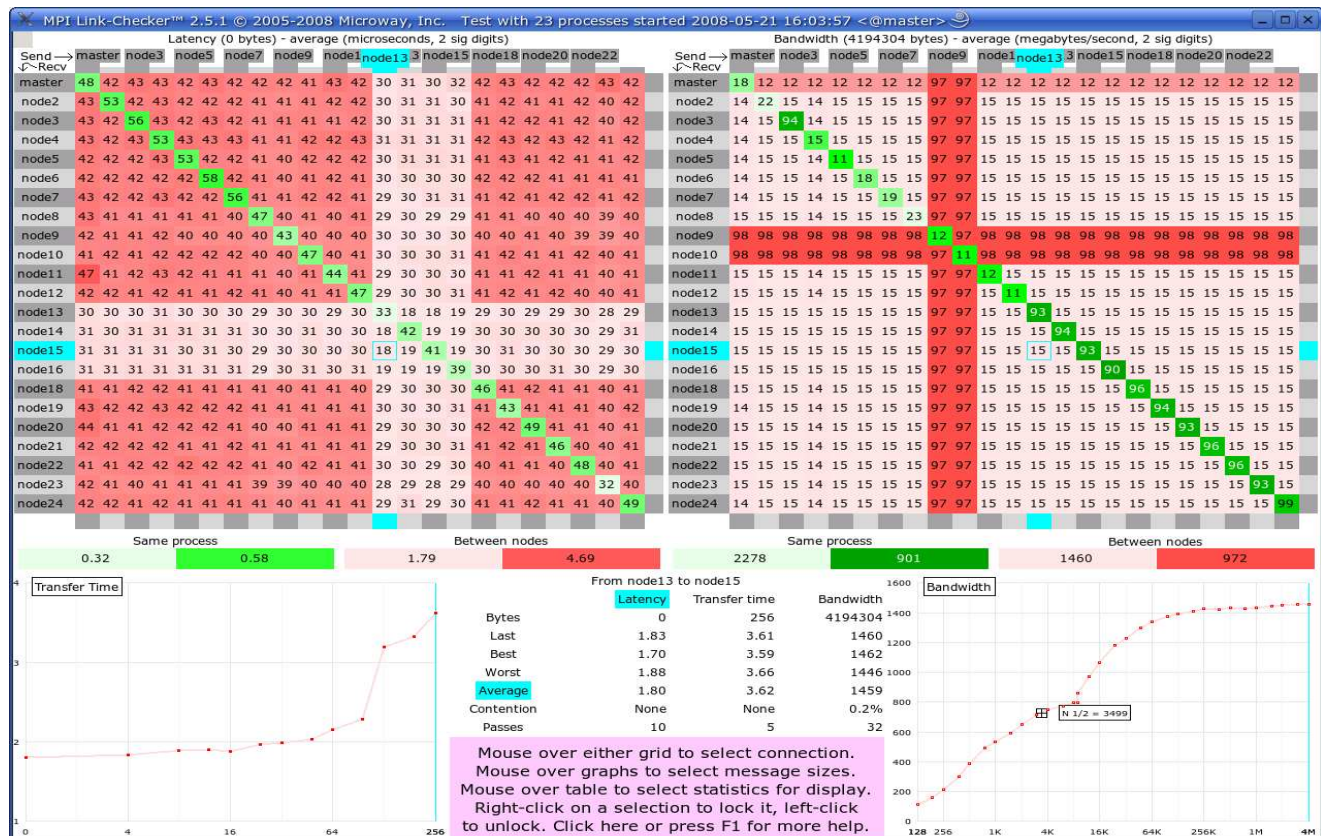

Figure 1: MPI Link-Checker shows latency and bandwidth.

The tool shows average and best-case performance for each connection, and can resolve message-size-dependent MPI issues. In addition, it simulates heavy interprocess traffic to detect potential network contention issues.

Figure 2 shows data for a 2-node PCI Express Gen 2 cluster with TriCom-X HCAs connected through a DDR InfiniBand switch. Latency is below 1.2 µs and bandwidth is over 1900 MB/s. Furthermore, the bandwidth half-power point is just 835 bytes, showing that there is not much small-message overhead.
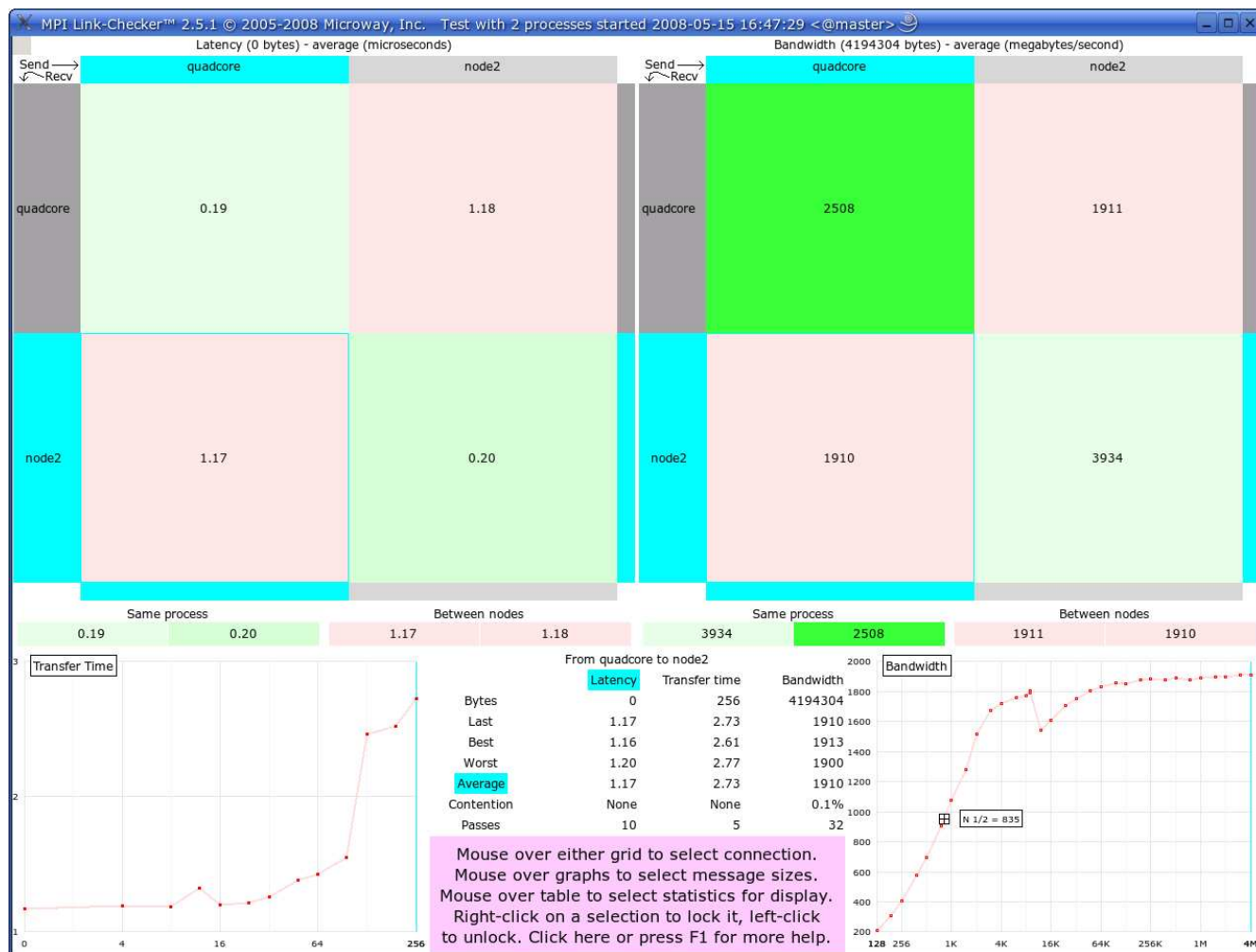

Figure 2: MPI Link-Checker shows latency below 1.2 microseconds in TriCom-X HCAs using PCI Express Gen 2.

Once your interconnect has been verified by MPI Link-Checker to be performing at full capability, you can visually inspect applications with Microway InfiniScope™. InfiniScope shows all connections between HCAs and switches in an InfiniBand network, and displays in real time the traffic passing through each port, as well as a historical record of traffic (at any desired time scale) for a selected port or switch, or for all hosts collectively. InfiniScope includes a Fabric Loading Program that can simulate a variety of traffic patterns, helping to answer architectural partitioning questions before code is written. The left part of the graph at the bottom of Figure 3 shows the traffic pattern for repeated parallel half-duplex transmission of 5 GB messages between randomly selected pairs of nodes. The right part of the graph shows the pattern for the MPI Link-Checker run shown in Figure 1.
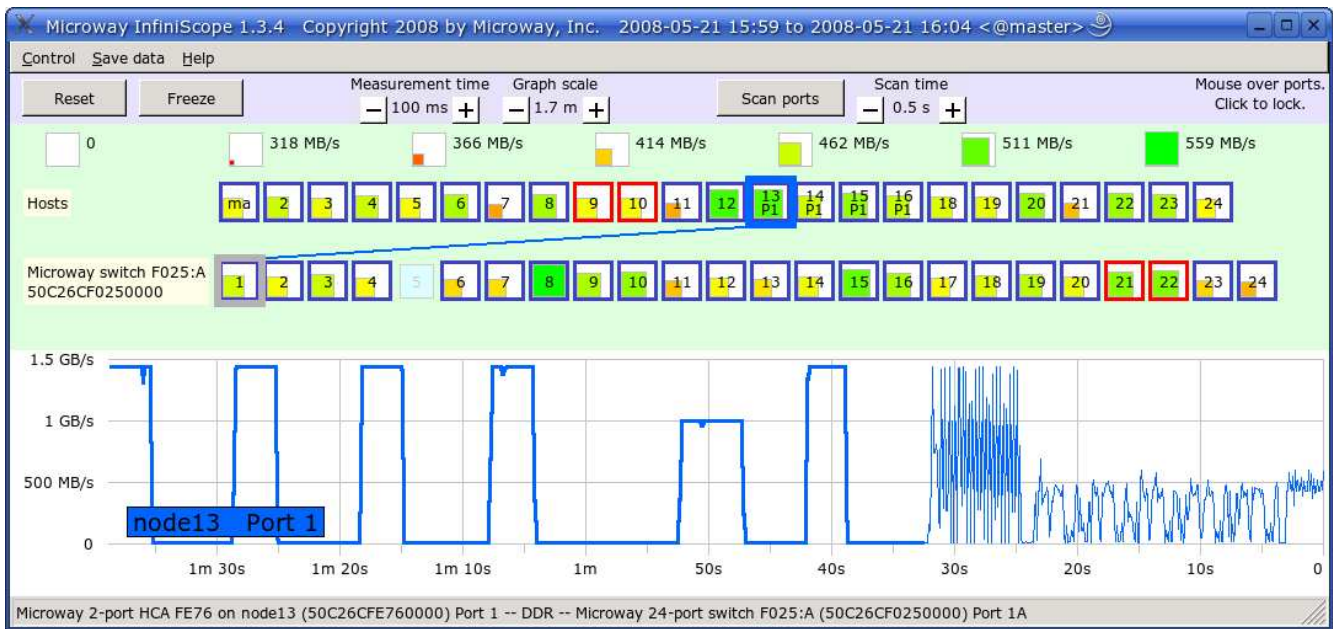
Figure 3: InfiniScope shows InfiniBand traffic.

Measured fabric performance

Choice of the interconnect fabric is dependent on the application and other system considerations. Many applications are latency-bound or bandwidth-bound, and cannot achieve full parallel speedup if either the interconnect bandwidth or the latency become a bottleneck. Table 1 shows the measured bandwidth and latency performance of a selection of fabric technologies.

Table 1: Interconnect Technologies Compared

Interconnect	Bandwidth (MB/s)	Latency (µs)	Cable length (m)
Gigabit Ethernet	115	30	100
10GigE	860	9	100
InfiniBand DDR	1455	3.6	100*
TriCom-X™ InfiniBand DDR	1910 †	1.2 †	100*

* using Intel cables.

† using PCI Express Gen 2. With PCI Express Gen 1, bandwidth is 1460 MB/s, latency is 1.8 µs.

How well does your cluster perform?

Microway will be happy to provide one free report of your cluster's interconnect performance. Just download the MPIcheck application from http://www.microway.com/MPIcheck, compile and run the program, and post the results back to our website. We will get back to you with a report containing MPI Link-Checker screenshots like Figure 1 and an analysis of the strengths and weaknesses of your cluster. Where appropriate we will also offer recommendations for improving performance. For continued monitoring and performance reporting, you can license the tools for your cluster.