Measurement, Visualization, and Improvement of Linux Cluster Performance

Paul G. Howard, Chief Scientist, Microway, Inc. (paulhoward@microway.com)
Bruce Schulman, Business Development, Microway, Inc. (bschulman@microway.com)
Stephen Fried, Chief Technology Officer, Microway, Inc. (steve@microway.com)

Introduction

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. We introduce visualization tools that help identify and address global network and MPI performance issues, and we use the tools to investigate the performance of several existing and emerging interconnect technologies.

Visualizing System Performance

In a Linux Cluster, application performance may be degraded by issues with the interconnect, such as loose cables or mismatched channel adapters. A maintenance tool like MPI Link-Checker™ can help in rapidly identifying and diagnosing system and network problems. Figure 1 shows an InfiniBand-based cluster with one reduced-bandwidth node (the darker + in the right-hand chart, corresponding to a single-data-rate HCA in a double-data-rate cluster) and one improved-latency node (the lighter + in the left-hand chart, corresponding to a low latency TriCom-X™ InfiniBand HCA).

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.

When the interconnect has been verified to be performing at full capability, an application can be visually inspected with a tool like 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. InfiniScope includes a Fabric Loading Program that can simulate a variety of traffic patterns, helping to answer architectural partitioning questions before code is written. Figure 2 shows the traffic pattern for repeated parallel half-duplex transmission of 5 GB messages between randomly selected pairs of nodes, highlighting the performance of a TriCom-X™ HCA.

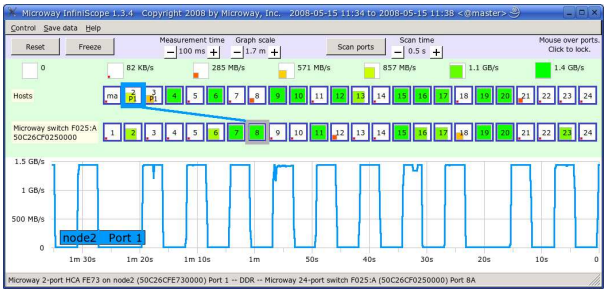

Figure 2: InfiniScope™ shows InfiniBand traffic.

Measured Interconnect 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	110	30	100
10GigE	860	9	100*
InfiniBand DDR	1455	3.6	100*
TriCom-X™ InfiniBand DDR	1800	1.8	100*

***** using Intel cables

Performance results to be presented

The presentation will include specific examples of:

1. Testing and validating network components;
2. Using visualization for rapid application optimization in MPI Networks;
3. Improving application performance with lower latency and higher bandwidth interconnects.

References

[1] S. Fried and P. Howard, "Test Your Backplane: MPI Link-Checker™", ClusterWorld Vol. 2, No. 5, May 2004, 16-22.

[2] P. Howard and S. Fried, "A Low Latency Modular Approach to Designing InfiniBand Fabrics Used in MPI Clusters", www.microway.com, *2007*.