CASE STUDY

Mellanox InfiniBand Powers Dual-Boot HPC Cluster at Holland Computing Center

CHALLENGE

- *Build one of the most powerful supercomputing clusters in the world*
- *Enable dynamic Linux and Windows application support*
- *Provide the fastest, lowest-latency interconnect for current and future needs*

SOLUTION

- *150 Dell PowerEdge server nodes*
- *Rocks+MOAB+CCS management and scheduling software allows dynamic node assignment*
- *Cisco InfiniBand SDR HCAs and DDR switches deliver 10 and 20 Gb/s node-to-node throughput with 1.2 microsecond latency*

When the University of Nebraska Omaha (UNO) decided to build one of the world's largest cluster computer systems, it combined InfiniBand interconnect technology with cutting-edge servers, storage, management, and power conditioning systems. The result was a system that easily placed among the 50 most powerful supercomputers in the world in the fall, 2007 Top500 test, and has the scalability to drive far more performance in the near future.

Funded with a budaget of $20 million and located in a 2000 square-foot glass-walled enclosure in the Holland Computing Center (HCC) at the Peter Kiewit Institute at UNO, the cluster represents a major milestone toward increasing Omaha and UNO's stature as a leading US research center. HCC has attracted cluster users from the Department of Defense, the Gallup organization, Microsoft, Milliman (an actuarial firm), and several leading consumer products companies. It also provides a major resource for UNO faculty, students, and research scientists from around the country.

Technology Choices

HCC's staff, which includes Jim Skirvin, the president, along with HPC system administrators Chris Cox and Patrick Sutton, brought extensive experience in building Linux- based clusters. However, the group wanted to also support Windows-based applications to offer a resource for the growing number of enterprises that are leveraging HPC systems for mission-critical research and analysis.

The goal of the project was not only to build a world-class compute cluster, but to enable provisioning of dedicated nodes for several simultaneous Windows and Linux users with minimal staff intervention. The HCC team selected Dell as the project manager. Dell initially specified 800 PowerEdge SC 1435 1U servers as compute nodes, with a Dell PowerEdge 6950 server as a master node. (This was later supplemented with an additional 350 compute nodes for today's total of 1150.)

Each compute node was configured with 2.8 GHz dual-core AMD Santa Rosa processors, 8 GB of RAM, and 80GB of local storage. HCC chose AMD processors for the servers because they would be pin-compatible with AMD's new Barcelona quad-core Opteron processors.

"With the dual-core Santa Rosa processors," says Skirvin, "we could deliver 21.5 sustained Teraflops of performance in the first Top500 run, but the real driver was that we

could upgrade to quad-core processors by simply swapping them in and updating the BIOS. Since we could upgrade to quad-core in the same sockets, we could design the cluster for dual-core systems and then upgrade without changing our power or cooling requirements." Since the initial deployment, the HCC team has swapped in quad-core Opteron processors in 256 compute nodes, and will eventually upgrade the entire cluster.

When it came to selecting an interconnect for the server network, InfiniBand was the only technology that Dell and HCC seriously considered. "10 Gb/s Ethernet switches were coming into production, but our choice was more about the latency, where InfiniBand has a huge advantage," said Aaron Rhoden, systems consultant for Dell.

Of course, bandwidth is another advantage to running Mellanox InfiniBand in its native RDMA mode. HCC achieves 10 Gb/s throughput in native RDMA mode compared with 4 Gb/s when running IP over InfiniBand. "Essentially," says Rhoden, "InfiniBand's RDMA interface gave us a way to connect server motherboards at bus speeds with about four microseconds of latency."

With message passing interface (MPI) applications in particular (where nodes communicate with each other while they are computing), InfiniBand's low latency is critical because it reduces the wait time between servers. On a TCP/IP network with a 10 Gb/s connection, there would be only 4 Gb/s of bandwidth and the messaging traffic would go up significantly.

Configuration and Tweaking

The servers are configured with Mellanox-based Cisco InfiniBand PCIE dual-port adapters with 4XIB SDR chips and 128MB of RAM. At the top of each 32-node rack, a pair of Cisco 7000p 24-port "leaf" switches uses Mellanox-based InfiniBand 4X 12-PORT DDR line cards to forward InfiniBand traffic to a pair of 244-port Cisco SFS concentrators.

For the Linux side of the cluster, the Dell team initially chose LSF scheduling software and OCS cluster management software from Platform Computing, and it was an early adopter of Windows Compute Cluster Server 2003 (CCS) for Windows applications. In the fall of 2007, HCC it was the largest Windows Compute Cluster Server 2003 deployment in the world.

Initially, the Holland Computing Center wanted a fully automated deployment of the Windows operating system as needed by each node. However, the Remote Installation Services that were used by Windows Compute Cluster Server 2003 did not support a zero-touch deployment for machines that had already been partitioned with the LINUX operating system. To remedy the situation, the HCC team switched to Clustercorp's Rocks+MOAB, which was leveraged to implement a more seamless solution.

"The cluster was originally broken into two or more clusters, one side Linux and the other under Windows CCS," says Tim McIntire, president of Clustercorp. "To dedicate a group of nodes for one particular user, they would have to set aside a new cluster for dedicated use, and that adds another layer of management complexity."

To address the issue, Clustercorp collaborated with Cluster Resources to add support for hybrid clusters into Rocks+MOAB. The new system, informally dubbed Rocks+MOAB+CCS, provides a single point of management and dynamic OS-switching capability for the entire cluster.
"Now," says McIntire, "if they have a user who needs 100 Windows nodes, they can simply execute the job from the front-end and it automatically boots those nodes into Windows CCS on demand."

Throughout the setup and transition to new management software and the processor upgrades, Mellanox InfiniBand technology has not only provided outstanding performance, but it has been rock-solid. "From a management standpoint, there's little or no management," says Chris Cox, HPC systems administrator at HCC. "Once you have your subnet manager set up, InfiniBand just goes."

Higher Performance Ahead

Today, HCC is busy handling service requests from a growing variety of government, university, and corporate users. "Our upgrade path necessitates migrating all of the nodes over to quad-core, and we're going to have to double down on our memory per node, going from 8 to 16 gigabytes," says Skirvin. "As application requirements rise, we will also look at going to DDR InfiniBand HCAs."

But even though 10 Gb/s Ethernet switches and interface cards are now available in quantity, the HCC team has never looked back on its choice of Mellanox InfiniBand. According to HPC systems administrator Chris Cox, "InfiniBand adds a great deal of value to what we're doing. We were getting 88 percent of theoretical peak performance out of a single node by itself. When we scaled up to 1150 nodes over InfiniBand with RDMA, we only lost three percent, and the overall cluster performance as posted on Top500.org site is 85 percent. If we had tried that over Gigabit Ethernet, it would probably have been only 60-65 percent of peak. That's a huge difference."

Joining a rapidly-growing percentage of the world's supercomputing sites, the Holland Computing Center and the University of Nebraska Omaha chose Mellanox InfiniBand's scalability, performance, and rock-solid reliability as the foundation for a world-class cluster. With 20 Gb/s and 40 Gb/s InfiniBand available for upgrades, technical organizations like HCC have all the runway they need to continue pushing the performance envelope without worrying about interconnect bottlenecks.

Mellanox TECHNOLOGIES

2900 Stender Way, Santa Clara, CA 95054
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com