CASE STUDY

Virginia Tech Taps 40Gb/s InfiniBand for Research Cluster

Introduction

For nearly ten years, Virginia Tech has been a leader in high-end computer systems research. In the university's Center for High-End Computing Systems (CHECS), faculty and a staff of graduate students investigate key hardware and software architecture challenges and design solutions that address them. Through full and partial ownership, CHECS has access to several compute clusters that handle research as well as production tasks.

As the director of CHECS, Dr. Srinidhi Varadarajan directs his teams in specific research into distributed shared memory systems (which improve cluster performance) and scalable software development tools (which allow for applications to be written and compiled on desktop systems and then run on clusters without modification). In addition, Dr. Kirk Cameron, director of the scalable performance lab (SCAPE), is using CHECS systems to head up research into power-aware systems that can reduce power consumption in large compute facilities by throttling back CPU and other components when possible.

Although CHECS has operated so-called "production" compute clusters for several years, these have been primarily used to support academic computing applications, while researchers were forced to use subsets of nodes in these production clusters. To facilitate his and Dr. Cameron's work, Dr. Varadarajan built a new cluster exclusively dedicated to research. Thanks to Mellanox InfiniBand technology, the cluster meets Dr. Varadarajan's stringent performance requirements and offers plenty of headroom for future growth.

The Importance of Interconnect

The challenge was not only to build a powerful computing resource, but to enable the fastest possible node-to-node interconnect. This was important due to the distributed computing nature of the team's research.

"The research into shared memory systems in particular is what drove the need for the fastest possible interconnect," says Varadarajan. "In order to effectively predict application performance, we needed an interconnect that had enough bandwidth overhead and low enough latency for us to reliably predict the speed at which data could be accessed from memory on one node by another node. Ideally, we wanted to be able to use a model that simply added another step to the progression of performance levels, from Level 1 to Level 2 cache and from Level 2 cache to main memory within a node, and that step had to be within an order of magnitude of the Level 2-to-main memory step."

The choice of an InfiniBand interconnect was based on both prior experience and current benchmarks. Since 2003, Virginia Tech has used 10 gb/s Mellanox InfiniHost™ PCI-X adapters and Mellanox InfiniScale-based switches in an 1100-node cluster designed to support production applications. Varadarajan liked the technology's robust and stable performance. Still, he did due diligence and tested several interconnect solutions for the new cluster. The testing compared the latest 40gb/s Mellanox InfiniScale IV 40gb/s switch technology and ConnectX 40Gb/s adapters against other industry offerings.

"We benchmarked Mellanox ConnectX 20Gb/s adapters, InfiniScale II and InfiniScale IV switches, Myrinet interconnect, and InfiniBand products from QLogic to see which had the highest bandwidth and the lowest la-

tency," says Varadarajan. "InfiniScale IV was clearly the winner. It gave us extremely low latency with the largest amount of headroom for future performance demands, and the performance is very stable, so it makes it much easier for us to predict application performance."

Easy Deployment, Rapid Results

In the summer of 2008, Varadarajan and a dozen students built a new cluster, called System G, which is based on 324 Intel®-based Apple® Mac® Pro servers (a total of 2592 CPU cores), each of which has a Mellanox 40gb/s quad data rate (QDR) adapter, interconnected with 19 Mellanox InfiniScale IV 36-port 40gb/s switches. It was one of the first QDR InfiniBand interconnects deployed in the world, and significantly, the 40gb/s products from Mellanox were based on silicon that had arrived only a month before.

Once the servers and other components had arrived, it took only a few days to assemble the cluster. "After we brought the first server nodes up, we did our first benchmark in only four hours," says Varadarajan, "which is a real statement about the stability of the Mellanox silicon and the overall robustness of the cluster."

So far, the System G cluster has delivered a Linpack benchmark of 22.8 teraflops with 80 percent efficiency, which should place it easily among the fastest 100 clusters in the world in the November 2008 Top500 ranking. With just 324 nodes, the system is almost twice as fast as Virginia Tech's previous performance champ, a 1100-node cluster that tops out at 12.25 teraflops. And in another testament to the stability of the System G cluster, Varadarajan's team was able to achieve the 10.3 teraflops performance of the original System X within 90 seconds after initial benchmarking began.

"Unlike most of the clusters I have ever used, we have never had a Linpack run failure with this cluster, not one," says Varadarajan. "This is really remarkable when you're dealing with hundreds of nodes, the interconnect, the cluster management software, and everything else that makes a system like this run."

Headroom for Current Needs, Future Growth

With this new system in place, Virginia Tech's CHECS finally has a dedicated, high-performance compute cluster for pure research. "Since it's a pure research system," says Varadarajan, "we will be able to do a lot of things we can't do with a production machine."

In addition to conducting distributed memory research, the cluster will support Dr. Cameron's power systems research. The cluster is equipped with 11,000 thermal sensors and 5000 power sensors, so it will be the first cluster in the world that can produce highly specific live data on power consumption and heat dissipation as it runs. And, as one of the first clusters to use Mellanox 40gb/s technology, System G will also be used by Mellanox researchers to test for advanced ConnectX features such as congestion control and adaptive routing as they pursue more advanced features for future products.

With 40 gb/s on tap, there is plenty of headroom to evolve the system over time. "We expect to be adding nodes, upgrading processors, and upgrading the interconnect when 100 Gb/s InfiniBand becomes available," says Varadarajan. "We have even designed the system to support side-by-side testing of 100gb/s and 40gb/s I/O adapters."

In the end, having the fastest interconnect delivers a lot of flexibility and future scalability, both of which are vital to ongoing research. Relying on Mellanox InfiniBand technology has enabled CHECS to conduct distributed memory and power-aware systems research in ways that simply weren't possible with other interconnect technologies, and Dr. Varadarajan is confident in his choice. "Most high-performance computing centers are coming to prefer InfiniBand because it delivers the highest bandwidth with the lowest latency and the most predictable performance," he says, "and we have found that when it comes to InfiniBand, Mellanox is ahead of

Mellanox TECHNOLOGIES

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