



CSILabs Helps Researchers Get “More Science Per Dollar” With Supercomputing Cluster Built on Intel® Gigabit Adapters

Groundbreaking 256-node 3-D mesh cluster serves as prototype for systems that can perform scientific calculations an order of magnitude more cost-effectively

CASE HIGHLIGHTS

Profiled Organizations

CSILabs is the high-performance computing (HPC) division of Concentric Systems Inc., a large systems integrator based in Alpharetta, Georgia. CSILabs specializes in providing HPC solutions to universities and government research labs, including customers such as the Thomas Jefferson National Accelerator Facility (JLab), an experimental physics research facility in Virginia.

Challenge

Design a supercomputing solution for JLab that can perform challenging physics theory calculations at higher bandwidth and lower cost than currently available switched cluster interconnects.

Solution

Build a prototype 256-node 3-D Gigabit Ethernet mesh Linux* cluster using Intel® PRO/1000 MT Dual Port Server Adapters as the interconnect. Each Intel® Xeon™ processor-based compute node is wired point-to-point to six adjacent nodes using three Intel® cards, eliminating the need for a costly switch. Operational since fall of 2003, this cluster is believed to be the first production three-dimensional mesh computer system in the world.

Benefits

Intel's Gigabit Ethernet adapters deliver high bandwidth at a fraction of the cost of other solutions. JLab scientists have proven that their application sustains more than 300 gigaflops of performance on this prototype cluster, paving the way for larger and more powerful clusters that provide increasingly “more science per dollar.”

Summary

Since 2002, the CSILabs division of Concentric Systems Inc. (CSI) has served the computing needs of the scientific and academic communities by designing high-performance data and research supercomputing solutions. Custom high-performance computing (HPC) solutions draw on the experience that CSI has gained in a dozen years of selling desktops, notebooks, and servers to small and mid-sized businesses and K-12 education, as well as manufacturing products on contract.

The Thomas Jefferson National Accelerator Facility (JLab), a U.S. Department of Energy research lab in Newport News, Virginia, is typical of CSILabs customers. Through a contract competition, JLab came to CSILabs with a challenge. Scientists specified an inexpensive, high-speed interconnect to further their research work in an area of physics called lattice quantum chromodynamics (lattice QCD). Typically, QCD calculations run on supercomputers that can

cost hundreds of millions of dollars. Driven by the need to achieve extremely high bandwidth at much less cost, CSILabs and JLab worked to build a first-of-its-kind, three-dimensional Gigabit Ethernet (GbE) mesh cluster that effectively handles JLab's computationally demanding application. The interconnect chosen for the 256-node

cluster: the Intel® PRO/1000 MT Dual Port Server Adapter.

This case study profiles the prototype cluster and discusses why it was designed around Gigabit Ethernet—and in particular, Intel's Dual Port Gigabit adapter. It also explains how this grid-style approach offers a cost-effective platform for applications that can exploit simultaneous communications in multiple directions. In the QCD arena, the accepted performance metric is the cost per unit of sustained performance, stated as millions of dollars per teraflop¹. The cost of running QCD calculations on high-

end supercomputers today is typically \$10 million per teraflop; the JLab GbE cluster can do it for \$1.5 million per teraflop. JLab researchers are already designing larger clusters that should hit the \$1 million per teraflop mark this year, thereby achieving scientific calculations an order of magnitude more cost-effectively.

Complex Calculations Demand Extreme Performance

Like any facility under contract to the U.S. Department of Energy (DOE) or other government body, JLab is constrained by budget. Consequently, JLab researchers can look with envy at traditional supercomputers—like the 35-teraflop machine that recently sold for \$350 million—but they don't have anywhere near the \$100 million it would take to buy the 10-teraflop machine that would be ideal for their own purposes.

Those purposes entail solving complex equations for lattice QCD, an especially challenging physics theory that seeks to explain the interactions between various types of sub-nuclear particles. This research requires scientists to simulate a quantum background of space/time—essentially, a four-dimensional vacuum that isn't a vacuum, but rather a sea of quarks and anti-quarks—so they can take measurements and make predictions. To be sufficiently accurate and produce useful science, the theory calculations have to be enormous. And that means the

2,000 physicists around the world who use JLab resources could well be affected by the lattice QCD experiments and research solved there.

The need for extreme bandwidth on a less-than-extreme budget has prompted scientists in the QCD realm to explore architectures other than the traditional supercomputer model. Chip Watson, High Performance Computing Group leader at JLab, discussed his idea for a cluster solution with CSILabs in summer 2002.

"We were trying to come up with an approach that's an order of magnitude more cost-effective for our problem," explains Watson, pointing out that the prototyping and software development work for this JLab project falls under a DOE program called Scientific Discovery through Advanced Computing (SciDAC). A collaborative effort that includes JLab and other national laboratories, SciDAC aims to make highly advanced supercomputers that can make trillions of calculations per second routinely available to scientists across the country.

Watson had looked at some of the more exotic high-speed interconnect offerings for the project. But while they offered appealing performance, he didn't want to see JLab spend more money on interconnects than it did on processing power.

"Proprietary interconnect offerings are extremely expensive, and in most cases, they can be more expensive than the node itself, especially once you add the switch," explains Chris Cartrett, vice president of sales and marketing at CSI. "JLab needed to maximize the number of compute nodes—to have as many processors as possible. So the goal was to get the best product, with the best results, with the highest-speed interconnects, for the amount of money budgeted for this project."

Watson determined that the answer to cost-effectiveness lay in high-performance cluster computing with a mesh network. His think-outside-the-box approach called for a 3-D cluster built on commodity components and standards-based Gigabit Ethernet networking technology.

Designing for "More Science Per Dollar"

The hardware design for the prototype cluster derived from a working relationship between CSILabs, JLab, and Intel. The innovative design consisted of a 256-node 3-D GbE mesh Linux cluster, arranged in a 4x8x8 (torus) configuration, and using Intel PRO/1000 MT Dual Port Server Adapters as the interconnect. Each node in this "cube" of computers is wired directly to six adjacent nodes using three of the Intel® GbE cards (one per dimension). In contrast to most GbE clusters, this high-density configuration does away with the need for a switch, thereby decreasing latency, increasing bandwidth, and eliminating the cost of additional hardware.

"It [Intel's GbE adapter] delivers all the bandwidth that the specification is capable of—it runs flat out right at the limits of Gigabit Ethernet. And, the density of two ports to a single low-profile card enabled us to put six connections in each box."

Chip Watson

High Performance Computing Group Leader, Jefferson Lab

“At the time we did this [the cluster went live in fall of 2003], it was not only the first but also the largest cluster of its kind in the world,” says Joey Sims, national account manager at CSI. (A smaller two-dimensional mesh cluster had been prototyped earlier in Europe.) More important to JLab, a mesh cluster based on Gigabit Ethernet technology proved to be an optimal platform for solving lattice QCD problems.

“Because our application can use multiple links simultaneously, we achieve higher bandwidth with Gigabit Ethernet in the aggregate across the multiple links than we would with semi-commodity high-performance network offerings,” Watson says. “From a bandwidth point of view, it is extremely high-performance. We do make some tradeoffs to get this lower cost, such as slightly more software overhead and latency, but in the overall balance, we get more science per dollar.”

To assist JLab with its selection of a high-performance adapter, CSILabs provided several of the Intel GbE cards for evaluation purposes. Sims says he recommends Intel® adapters for their price/performance, quality, solid warranty, and superior driver support. Moreover, Intel offered a two-port solution in a low-profile configuration, which proved a perfect fit for the high-density requirements and space constraints of the JLab cluster topology.

“We did some performance evaluations and some quick preliminary software development, which showed that all the characteristics were good and it was a well-behaved card,” Watson says. “It delivers all the bandwidth that the specification is capable of—it runs flat out right at the limits of Gigabit Ethernet. And, the density of two ports to a single low-profile card enabled us to put six connections in each box.”

Sims says Intel's adapter also ended up offering a bonus on the software side. JLab developers had expected to write all their own software for using the mesh network—either in-house or in collaboration with university partners—but the quality of Intel's driver gave them a head start. JLab developers optimized Intel's open source Linux driver for their mesh network and their lattice QCD code, adding a layer called the Virtual Interface Architecture (VIA) and drawing upon other open source work to develop their software.

JLab's code for QCD calculations is also optimized to run on the Intel® Xeon™ processor with the SSE2 instruction set. All of the compute node motherboards in the 256-node GbE mesh cluster feature a 2.66 GHz Intel Xeon processor. The decision to build the cluster on Intel® CPUs was based on JLab's evaluations of processor technology.

JLab researchers discovered that the innermost numerical kernel of their QCD application, which is executed repetitively at very high speed, can take advantage of the

multimedia capability of an Intel® processor. In other words, the same processor technology that supports high-quality, interactive graphics on a PC helps perform the HPC group's physics theory calculations. JLab chose the Xeon processor specifically because its chipset supports PCI-X, the high-performance I/O bus.

Sustained, Scalable Performance for Many Applications

In the world of research labs, “peak” performance is a meaningless number. Thus, for benchmarking purposes, Watson and his HPC colleagues at JLab measure how many gigaflops² their QCD application can sustain on the GbE cluster—in essence, how much science they're getting out of it. On the 256-node mesh cluster, JLab is achieving more than 300 gigaflops of sustained performance, with expectations of reaching 400 gigaflops when the software is fully tuned. Later in 2004, JLab will deploy another prototype cluster that aims at more than 500 gigaflops.

Already in the planning phases, this next cluster is expected to include 512 nodes in a 3-D GbE mesh, and it will likely be built around Intel GbE adapters. One or two years down the road, Watson believes 10-Gigabit Ethernet may drop in price to where JLab could feasibly build a mesh using that next-generation technology.

In the meantime, the clusters that JLab and CSILabs build this year are only prototypes of much larger machines to come. Although the current 256-node cluster provides useful science, JLab actually needs a system that's probably eight times larger to begin doing significant comparisons with experiments. The roadmap to bigger systems will depend largely on funding levels, ongoing component evaluations, and price points, but each year, JLab's HPC group hopes to build a cluster that's at least twice as big as the one before.

“Because our application can use multiple links simultaneously, we achieve higher bandwidth with Gigabit Ethernet in the aggregate across the multiple links than we would with semi-commodity high-performance network offerings. From a bandwidth point of view, it is extremely high-performance. We do make some tradeoffs to get this lower cost, such as slightly more software overhead and latency, but in the overall balance, we get more science per dollar.”

Chip Watson
High Performance
Computing Group Leader,
Jefferson Lab

“Our goal,” Watson says, “is to be running clusters in the range of 2,000 to 4,000 processors.” And, he adds, although new architectures could lead JLab to alter course in the future, the effort will continue to be driven by science per dollar. “We pretty much have to follow the mass market, because that’s where the volume shipments are, and that’s what gives us the most cost-effective components.”

JLab scientists are confident they’re headed in the right direction with their grid-type machine approach. Given that a 3-D mesh cluster is architecturally similar to the next-generation high-end supercomputers being built today, Watson expects this model will be evaluated by numerous scientific applications as the concept becomes more mainstream. He’s convinced that a large class of applications can profit from high-performance cluster computing with a mesh network, especially homogeneous applications involving science problems in three dimensions. Examples include scientists simulating some volume, such as a finite element analysis, or performing weather simulations for long-range weather modeling.

Conclusion

Gigabit Ethernet clusters based on readily available, standardized building blocks offer high performance and excellent value for HPC applications that can exploit multiple network links in a mesh topology to solve scientific problems. JLab researchers, using hardware procured from CSILabs, are already successfully exploiting large-scale 3-D mesh clusters for lattice QCD, with performance expected to climb up as clusters grow larger and components continue to improve.

In the lattice QCD community, the most relevant performance metric for calculations is the cost per unit of sustained performance, stated as millions of dollars per teraflop. While the cost of running QCD calculations on high-end supercomputers today is on the order of \$10 million per teraflop, the GbE mesh cluster does the same for as little as \$1.5 million per teraflop. JLab scientists are already designing larger clusters that should hit the \$1 million per teraflop target this year, thereby achieving scientific calculations an order of magnitude more cost-effectively.

More information:

www.intel.com/network/connectivity/products/pro1000mt_dual_server_adapter.htm

¹A teraflop is 1 trillion floating point operations per second.

²A gigaflop is 1 billion floating point operations per second.

Information in this document is provided in connection with Intel products. No license, express or implied, by estoppel or otherwise, to any intellectual property rights is granted by this document. Except as provided in Intel's Terms and Conditions of Sale such products, Intel assumes no liability whatsoever, and Intel disclaims any express or implied warranty, relating to sale and/or use of Intel products including liability or warranties relating to fitness for a particular purpose, merchantability, or infringement of any patent, copyright or other intellectual property right. Intel products are not intended for use in medical, life saving, or life-sustaining applications. Intel may make changes to specifications and product descriptions at any time, without notice.

Information regarding third party products is provided solely for educational purposes. Intel is not responsible for the performance or support of third party products and does not make any representations or warranties whatsoever regarding quality, reliability, functionality, or compatibility of these devices or products.

Copyright © 2004 Intel Corporation. All rights reserved.

Intel, the Intel logo and Intel Xeon are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

*Other names and brands may be claimed as the property of others.

