

IEEE DISTRIBUTED SYSTEMS ONLINE 1541-4922 © 2005 Published by the IEEE
Computer Society
Vol. 6, No. 1; January 2005

US Supercomputing Receives Multifaceted Boost

Greg Goth

For several years, notably since Japan's NEC Earth Simulator earned the top spot on the list of the world's fastest computers in 2002, the US-based high-performance computing community has warned that the federal government's lagging interest and investment could cause serious long-term consequences for both the public and private sectors in the US.

Apparently, those warnings haven't gone entirely unheeded. In November 2004, US-based supercomputing received both psychological and financial reinforcement. IBM's Blue Gene/L wrested the top spot on the global Top 500 supercomputer list (www.top500.org) from NEC with a performance rate of 70 teraflops per second. A second US-based computer, a cluster built by SGI and owned by NASA, took the second-place spot with 51 teraflops. In addition, federal officials demonstrated a willingness to invest across the range of high-performance architectures, from cluster interconnect technology to top-end shared-memory computers.

No room for complacency

One leading computer scientist says the recent good news should be just the first step in a sustained strategy that might or might not come to fruition.

"There is a risk for complacency," says Marc Snir, head of the Computer Science Department at the University of Illinois at Urbana-Champaign. Snir also cochaired the National Research Council's Committee on the Future of Supercomputing (www7.nationalacademies.org/cstb/project_supercomputing.html) and coedited *Getting Up to Speed: The Future of Supercomputing*, a comprehensive report issued by the NRC calling for more robust federal leadership in sustaining high-performance computing (HPC) research and deployment. "The point we make in our report is 'What is a top machine in a Linpack benchmark [the standard by which the Top 500 are evaluated]?' is an interesting point of information, but it's really not what makes a strong supercomputing infrastructure."

Snir says Blue Gene's pioneering design shouldn't veil the state of the art's other shortcomings.

"Blue Gene is a very interesting design," he says, "and certainly I won't belittle [it], especially since I was involved in some of the Blue Gene work, but it clearly is not a universal solution for all the needs in supercomputing. There is no universal solution."

The industry still needs large investments in software, Snir says. "Yes, it's nice the US is now leading the Top 500 list, but there are still significant problems with our supercomputing infrastructure that are not resolved."

HPC efforts win funding

Two recent investments could be signals that federal agencies are reawakening to the benefits blue-sky research might yield someday.

On 30 November, President George W. Bush signed into law the Department of Energy High-End Computing Revitalization Act of 2004, allocating US\$165 million for the DoE to

conduct advanced scientific and engineering research and development on supercomputing, develop potential advancements in HPC system hardware and software, and provide access to the systems arising out of this research on a competitive, merit-reviewed basis to researchers in US industry, institutions of higher education, national laboratories, and other federal agencies.

The funding for the law (<http://thomas.loc.gov/cgi-bin/bdquery/z?d108:h.r.04516:>) runs through the 2007 fiscal year.

Additionally, in what might have a far more wide-ranging effect on "street level" HPC clusters at universities, incubator labs, and enterprise grids, the DoE awarded the Open InfiniBand Alliance (www.openib.org) a grant to develop an open source InfiniBand software stack for the Linux operating system, a popular OS for cluster HPC architectures. The alliance comprises three companies involved in the high-speed InfiniBand interconnect technology: Voltaire, Topspin, and Intel.

One veteran of the HPC community is eagerly awaiting a more standardized InfiniBand stack.

"We can only benefit from that," says Steve Woods, principal systems architect at MCNC, a North Carolina-based technology incubator who's pioneering a statewide enterprise grid that links educational institutions and, eventually, commercial customers. "We keep running into issues here."

For example, Woods says, MCNC uses a file system named Lustre, manufactured by Cluster File Systems, on its 64-node Linux cluster. The file system includes software that's predeveloped and custom-built at MCNC.

"[W]e have a problem because we have InfiniBand," Woods says, "because their drivers may not work with our built kernel for Lustre. Whatever they could do to build it into the kernel would be wonderful."

Woods says such obstacles become a headache. "[Y]ou have to go back to the vendor, in our case it's Topspin, and say, 'Hey, we need a new kernel driver and new load module.' Or even something as subtle as when an emergency security kernel patch comes out: You get the patch, put it in, and guess what? Your InfiniBand doesn't work anymore because it won't load the module."

Is InfiniBand near the promised land?

The effort to create the new InfiniBand stack is one of several being undertaken by the

research and vendor community to standardize interfaces at various layers of HPC software. When InfiniBand was introduced in October 2000, it was promoted as a high-speed, low-latency technology that would replace the PCI bus as the predominant I/O technology in server architectures and data centers.

However, some say the early reality didn't live up to its publicity. "There was a lot of hype but nothing really spectacular," Woods says. "It kind of went away, came back, and went away again. It took a while for it to happen."

Although InfiniBand offered a more elegant I/O architecture than the PCI bus, its early incarnations weren't a significant improvement over familiar technologies such as Gigabit Ethernet in HPC deployments. "When they came out with 1x (with a base data rate of 2.5 gigabits per second), it was kind of 'Whoopdee . . . what's so exciting? Yeah, it's better than GigE, but I've got to deal with special drivers, and so on.' GigE is a follow-on to 10/100, so it was familiar. So at that point, it just wasn't that exciting."

But, Woods says, with 4x InfiniBand technology (with a base rate of 10 Gbits per second) being deployed widely and 12x (30 Gbits per second) on the way, plus expanded storage and interconnect capabilities with other technologies such as GigE and Fibre Channel, the HPC community is taking to InfiniBand like a long-lost friend.

Executives from Voltaire and Topspin say the OpenIB grant, financial details of which haven't been released, is formal recognition that the time for widespread InfiniBand deployment is here.

"As InfiniBand deployment started to happen, as the DOE started getting serious with it, they realized this is good stuff," says Arun Jain, Voltaire's vice president of marketing. "They realized clusters are the way to go these days. . . . They realized they would like to promote a true open source any-to-any InfiniBand environment. The OpenIB effort is moving along, but not moving along fast enough—the players are mostly small, new companies. DoE wanted to expedite the process to make sure InfiniBand software drivers are in place. That was the impetus."

Jain says the economics of InfiniBand in HPC are quickly driving further deployments.

"Up until a couple of years ago, high-performance computing, typically defined as the province of specialized mainframes, cost about \$20 million per teraflop," Jain says.

"When InfiniBand came around a year to a year and a half ago, it immediately brought

the cost to about \$1 million per teraflop. Today, it's about \$500,000 per teraflop and heading south. So many things are happening and each one is making it better in terms of performance for price."

Jain also says the OpenIB grant will be a foundational move toward commoditizing Linux for HPC.

"The other idea here is that by having a true open source stack, this is also going to be available as part of Red Hat and SuSE; and once that happens, then it truly becomes open source, and it becomes sort of a snowball effect."

Off-the-shelf hardware already prevalent

Even without the availability of the fully standardized open source stack, InfiniBand-enabled commodity clusters have proven to be among the elite of the Top 500 list. In 2003, the InfiniBand-connected Apple Macintosh cluster built by computer scientists and students at Virginia Tech for about \$5 million was the third-fastest computer on Earth. This year, NASA's Columbia, named after the space shuttle that disintegrated on reentry into the Earth's atmosphere in February 2003, was the second-fastest computer on the Top 500 list, after Blue Gene. Columbia is a 10,240-CPU SGI Altix cluster, powered by Intel Itanium 2 processors and using Voltaire's ISR 9288, a 288-port, 4X InfiniBand switch between all the cluster's nodes.

The new cluster's possibilities seem enormous. For instance, using NASA's previous top-performing computer, a single 512-processor Altix, computer modeling of decades of ocean circulation could be done in a matter of a few days. Using its best computer prior to the 512-node machine, simulations showing five years' worth of changes in ocean temperatures and sea levels were taking a year to complete.

"Everybody we talk to—private companies, universities, DoE—their mission seems to be very clear," Voltaire's Jain says. "If they can run whatever they need to run on a cluster, they'll do it on a cluster. That doesn't mean people are seeking 100 percent cluster-high-performance computing everywhere. But look at Columbia. It reached 50 teraflops with commodity components of Intel processors and Voltaire switches."

And by many accounts, InfiniBand hardware is also reaching that commodity stage.

"We're really at the point where InfiniBand is starting to ramp quickly," says Stu Aaron, Topspin's vice president of marketing. The company shipped 10,000 server switch ports in the second quarter of 2004, Aaron says. "The rate at which we ramped from zero to 10,000 was just over five quarters in shipping product, about two times faster than Fibre Channel in its early days and roughly equal to Gigabit Ethernet."

Software's on the way

The OpenIB grant isn't the first attempt software researchers have made to simplify InfiniBand's use. In 2002, Ohio State University professor Dhabaleswar Panda and Peter Wyckoff, a scientist at the Ohio Supercomputer Center, developed a technology called MPI for InfiniBand on VAPI Layer, or MVAPICH (<http://nowlab.cis.ohio-state.edu/projects/mpi-iba/>). MVAPICH bridged a chasm between InfiniBand and MPI, the message-passing interface that sits higher on the software stack, and which was developed to give researchers more portability in writing HPC programs. Prior to MPI's introduction, researchers had to write separate programs for specific HPC systems. However, the existing MPI stack didn't work with InfiniBand until Panda and Wyckoff developed MVAPICH .

Today, MVAPICH is in use in more than 160 HPC facilities, including the Virginia Tech cluster, which has slipped from third to seventh place on the Top 500 list. Panda says the MVAPICH site has logged more than 1,100 direct downloads of the technology.

Panda says he thinks the recent surge in awareness of HPC's needs might not subside. As the performance-to-price ratio of clustering continues to improve, scientists worldwide will strive to develop ever faster technologies; no one can afford to take a breather.

"The second largest InfiniBand cluster using MVAPICH is in Belarus," Panda says. "The next largest is in China. This is a global economy."

However, one researcher also says reliance on the academic and research community for software advances might be hindering their current and future work. Srinidhi Varadarajan, a computer science professor at Virginia Tech, oversaw the construction of the university's Mac cluster. Varadarajan says the popularity of Panda's MVAPICH stack demonstrates the need for a wider distribution of effort.

"Panda's stack is really the only one that runs well today," Varadarajan says, "and his

group is a research group. The last thing they want to see is that kind of pressure that comes along with having to support it. Perhaps the work of the OpenIB Alliance will anticipate that kind of problem."

Varadarajan himself is pushing the frontier of HPC research. The Virginia Tech cluster's low cost remains a topic of awe in some circles. The next problem his team is researching, in concert with the DoE, might also obliterate the conventional wisdom that cluster computing is fine for some problems but shared-memory architectures are required for others.

"InfiniBand hardware is evolving, evolving to the point that they are going to be putting processing power right on the card itself."

Such an architecture, Varadarajan says, may make it possible for distributed-memory clusters to go head-to-head with the much more expensive SMPs (symmetric multiprocessing systems).

If all goes well, he says, that work may be complete "in a couple years."

CONCLUSION

And, while such a machine may be able to perform like an SMP machine on the Top 500, Topspin's Aaron says the real payoff comes in the aftermath of beating the benchmark.

"InfiniBand is a connection-oriented technology, which means you physically set up and tear down connections," Aaron says, "so you can carve a large cluster into virtual smaller clusters. Any of the Top 500 may be deployed initially as a vanity project to get on the list, but the long-term value comes in carving it up and selling it as individual services to research departments or the commercial sector within your geography."

Cite this article: Greg Goth, "US Supercomputing Receives Multifaceted Boost," *IEEE Distributed Systems Online*, vol. 6, no. 1, 2005.