Editors: Volodymyr Kindratenko, kindr@ncsa.uiuc.edu
Pedro Trancoso, pedro@cs.ucy.ac.cy

# Trends in High-Performance Computing

By Volodymyr Kindratenko and Pedro Trancoso

**HPC system architectures are shifting from the traditional clusters of homogeneous nodes to clusters of heterogeneous nodes and accelerators.**

We can infer the future of high-performance computing (HPC) from the technologies developed today to showcase the leadership-class compute systems—that is, the supercomputers. These machines are usually designed to achieve the highest possible performance in terms of the number of 64-bit floating-point operations per second (flops). Their architecture has evolved from early custom design systems to the current clusters of commodity multisocket, multicore systems. Twice a year, the supercomputing community ranks the systems and produces the Top-500 list (www.top500.org), which shows the world's 500 highest performing machines. As we now describe, the technologies used in the top-ranked machines give a good indication of the architecture trends.

## The Top-500

The November 2010 Top-500 list of the world's most powerful supercomputers stressed two noticeable developments in HPC: the advent of HPC clusters based on graphical processing units (GPUs) as the dominant petascale architecture, and the rise of China as a dominant player in the supercomputing arena.

To build more powerful machines, system architects are moving away from traditional clusters of homogeneous nodes to clusters of heterogeneous nodes (CPU+GPU). As the sidebar, "Architecture of Current Petaflops Systems" describes, three out of seven systems that achieved over one quadrillion flops (petaflops) on the standard benchmark tool Linpack (www.top500.org/project/linpack) are Nvidia GPU-based, and two out of the three largest systems are deployed in China.

The number one system on the November 2010 Top-500 list is Tianhe-1A, deployed at the National Supercomputing Center in Tianjin. It achieves about 2.57 petaflops on Linpack with a theoretical peak performance of 4.7 petaflops. The number three system is Nebulae, deployed at the National Supercomputing Centre in Shenzhen. It achieves 1.27 petaflops on the Linpack benchmark with a theoretical peak of almost 3 petaflops. The highest performing US system, Jaguar, deployed at the Oak Ridge National Laboratory, is number two on the Top-500 list, achieving 1.75 petaflops with theoretical performance of 2.3 petaflops. The number four system on the November 2010 Top-500 list is the Japan-built Tsubame 2.0 system, which is a GPU-based system. Europe's most powerful system is the Tera-100, which is ranked at number six and deployed at the French Atomic and Alternative Energies Commission.

The Tianhe-1A is composed of 7,168 nodes—each containing two Intel Xeon X5670 hex-core (Westmere) processors and one Nvidia Tesla M2050 GPU. The Jaguar is a more traditional Cray XT system consisting of 18,688 compute nodes containing dual hex-core AMD Opteron 2435 (Istanbul) processors. The highest-performing GPU-based US-built system, deployed at Lawrence Livermore National Laboratory, is number 72 on the Top-500 list. We've yet to see any substantially larger GPU-based HPC systems deployed in the US. Instead, we see many research centers deploying small and mid-range GPU systems, such as the US National Science Foundation (NSF) Track 2D Keeneland system, which is in its initial deployment phase at the Georgia Institute of Technology (number 117 on the Top-500 list), or the Forge GPU HPC cluster currently being developed at the US National Center for Supercomputing Applications at the University of Illinois. The Keeneland project in particular is interesting in that a part of the effort is devoted to developing the software infrastructure necessary to utilize the GPUs in an HPC

1521-9615/11/$26.00 © 2011 IEEE

## Architecture of Current Petaflops Systems

The November 2010 Top-500 list included seven petaflops systems:

1. Tianhe-1A, deployed at the National Supercomputing Center in Tianjin, China, achieves about 2.57 petaflops on Linpack with a theoretical peak performance of 4.7 petaflops. Tianhe-1A is composed of 7,168 nodes, each containing two Intel Xeon EM64 X5670 six-core (Westmere) processors and one Nvidia Tesla M2050 GPU. The system uses a proprietary interconnect.
2. Jaguar, deployed at the Oak Ridge National Laboratory, US, achieves 1.75 petaflops with theoretical performance of 2.3 petaflops. Jaguar is a Cray XT system consisting of 18,688 compute nodes containing dual six-core AMD Opteron 2435 (Istanbul) processors, or 224,256 processor cores total. The system uses a proprietary interconnect.
3. Nebulae, deployed at the National Supercomputing Centre in Shenzhen, China, achieves 1.27 petaflops on the Linpack benchmark with a theoretical peak of nearly 3 petaflops. Nebulae is composed of about 4,700 compute nodes containing Intel EM64 Xeon six-core X5650 (Westmere) processors, or 120,640 processor cores total, and Nvidia Tesla C2050 GPU. The system used Quad Data Rate Infiniband interconnect.
4. Tsubame 2.0, deployed at the Tokyo Institute of Technology, Japan, achieves 1.19 petaflops on the Linpack benchmark with a theoretical peak of about 2.3 petaflops. It consists of 1,442 compute nodes containing (mostly) six-core Intel Xeon X5670 (Westmere) CPUs, or 73,278 processor cores in total, and three Nvidia Tesla M2050 GPUs per node. The system used Quad Data Rate Infiniband interconnect.
5. Hopper, deployed at the US National Energy Research Scientific Computing Center, achieves just over a petaflops on Linpack, with a theoretical peak of almost 1.3 petaflops. Hopper is a Cray XE6 system consisting of 6,384 nodes containing 12-core AMD Operon processors, or 153,216 processor cores in total. The system uses a custom interconnect.
6. Tera-100, deployed at the Alternative Energies and Atomic Energy Commission, France, achieves 1.05 petaflops on Linpack with a theoretical peak of 1.25 petaflops. The system consists of 4,300 bullx S series server nodes based on eight-core Intel EM64 Xeon 7500 (Nehalem) processors, or 138,368 processor cores total. The system used Quad Data Rate Infiniband interconnect.
7. Roadrunner, deployed at Los Alamos National Laboratory, US, was the first supercomputer in the world to achieve over one petaflops. It currently stands at 1.04 petaflops on Linpack, with theoretical peak of 1.38 petaflops. Its main workforce is a nine-core IBM PowerXCell 8i experimental chip. The system consists of 12,960 PowerXCell 8is and 6,480 dual-core AMD Opterons, or 122,400 processor cores total, and uses Voltaire Infiniband interconnect.

environment. It's also NSF's first system funded under the Track 2 experimental/innovative design program.

The Chinese HPC community's approach to designing high-end systems is certainly worth note. In terms of raw performance, adding a GPU to a conventional HPC cluster node can quadruple its peak performance, or even increase it by an order of magnitude when using 32-bit arithmetic. But the increased peak performance doesn't necessarily translate into sustained application performance. As an example, take Tianhe-1A, with sustained 2.57 petaflops on Linpack and theoretical peak performance of 4.7 petaflops. Its efficiency in terms of sustained versus peak performance is 2.57/4.7 = 0.55. Jaguar, on the other hand, achieves 1.75 petaflops on Linpack with theoretical performance of 2.3 petaflops. Thus, its efficiency is 1.75/2.3 = 0.76, or 38 percent above Tianhe-1A. The difference is even more pronounced when considering real scientific workloads.

Also, software availability for large-scale GPU-based systems is limited. While numerous applications have been ported to GPU-based systems over the past two years, many of the widely used scientific supercomputing codes that have been developed over the past 20 years have yet to be rewritten for GPUs. Many such applications have outlived several generations of computer architectures; it would be irrational to scrap all the prior work each time a new and drastically different architecture is presented, even if its performance is significantly higher. Not surprising, we have yet to see any applications that can take advantage of Tianhe-1A's computing power.

## The Green and Graph Lists

As the "Metrics and Benchmarks" sidebar describes, organizations use various metrics and applications to evaluate computing systems. In the past, supercomputers were developed for the single highest performance goal (as measured by Linpack), but current system development is driven by two additional factors: power consumption and complex applications. Traditional approaches to achieving performance have reached an unbearable power consumption cost. As such, the community has proposed a second list—the Green-500 list (www.green500.org)—that ranks systems according to their performance/power metric. The Green-500 list sorts systems from the Top-500 list by their power efficiency; it also shows the trend toward heterogeneous

## Metrics and Benchmarks

Supercomputer systems are evaluated according to different metrics using different applications. We show here the three most relevant ranking lists and present the benchmarks and metrics they use to rank systems.

### The Top-500

To rank systems, the Top-500 uses the Linpack benchmark program to solve a dense system of linear equations using the LU decomposition method. Because this application is regular, it performs well and thus gives a good indication of the system's peak performance. The metric used to rank the systems in the Top-500 list is the number of floating-point operations per second (flops), which is measured from the execution of the Linpack benchmark.

### The Green-500

The application used to rank the Green-50 list is also Linpack. In this case, we measure both performance and power consumption for executing the benchmark on the system. It thus uses flops per watt (total power consumption for the program's execution).

### The Graph-500

A graph-based application ranks the systems for this list. The authors are planning to use three versions—one for shared-memory systems, one for distributed-memory systems, and one for cloud systems using the map-reduce model. This benchmark represents 3D physics simulation applications. In contrast to Linpack, which is a compute-intensive application, this graph-based application is data intensive. It's also composed of two kernels: one for building a graph out of the raw data and another to operate on the graph. The structure is a weighted undirected graph; the second kernel operates on it based on a breadth-first search (BFS) algorithm. The metric used to rank this list's systems—the number of traversed edges per second—maps closely to the application domain.

## Upcoming 10+ Petaflops System Architectures

Based on recent announcements, we've identified several upcoming systems with aggregated peak performance of more than 10 petaflops:

- Sequoia, an IBM Blue Gene/Q-based 20-petaflops system at Lawrence Livermore National Laboratory, US;
- Titan, a GPU-based 20-petaflops system at Oak Ridge National Laboratory, US;
- Blue Waters, an IBM Power7-based 10-petaflops system at the US National Center for Supercomputing Applications;
- Pleiades, an SGI ICE 10-petaflops system deployed at NASA, US, is expected to be upgraded to reach 10-petaflops peak; and
- K supercomputer, a Fujitsu Sparc64-based 10-petaflops at the Riken Research Institute, Japan.

systems. Four of the top 10 systems are GPU-based (both Nvidia and ATI/AMD). Furthermore, the Green500 list's top system is the IBM BlueGene/Q, which is built from simple, power-efficient processors originally developed for embedded systems. These processors also include several task specific acceleration engines.

The Linpack application used to benchmark the Top-500 list systems is somehow unrealistic when compared to the real-world applications executed on real systems. Linpack itself isn't an all-inclusive performance measure because it measures the speed of performing floating-point operations—certainly a crucial component of most computation—but ignores system characteristics such as data transfer speed between memory, CPU, disk, and compute nodes, which is critical for most scientific applications to achieve high, sustained performance.

To take that into account, the community has proposed using graph-based applications to evaluate the best supercomputers for complex data-intensive applications. With the exception of the number one system, which is a BlueGene/P architecture composed of power-efficient processors, the top 10 positions in the Graph-500 list (www.graph500.org) are systems composed of the most powerful single-chip multicore processors. This shows that the most demanding HPC applications still need high-performance multicore processors.

### Future Trends

No doubt, we'll see more GPU-based large-scale systems. However, as the sidebar "Upcoming 10+ Petaflops System Architectures" shows, most of the upcoming very large-scale systems actually aren't based on GPUs.

In the US, two 20-petaflops systems are currently under construction: the Sequoia system being built at Lawrence Livermore National Laboratory and the Titan system to be deployed at Oak Ridge National Laboratory. Sequoia is the IBM Blue Gene/Q architecture, whereas Titan will include GPUs. Blue Waters, which is expected to go online at the US National Center for Supercomputing Applications this year, will be the first system with peak performance of 10 petaflops.

Also, NASA's Pleiades system is expected to be upgraded to reach the 10-petaflops peak in 2012. Blue Waters in based on IBM's Power7 architecture and Pleiades is an SGI Altix ICE architecture.

Japan's first 10-petaflops system, the Sparc-based K supercomputer, will become operational in 2012. In Europe, a 3-petaflops system is being constructed in Germany with expected operation by 2012 as well. The K supercomputer is based on the Fujitsu Sparc64 architecture; Europe's supercomputer, SuperMUC, is based on Intel Xeon processors running in IBM System x iDataPlex servers. The forthcoming Dawning 6000 supercomputer under development in China will provide one petaflops. Although not the fastest system, Dawning 6000 will use Chinese-made Loongson microprocessor.

Given current computing technology trends, projections indicate that we'll reach 1 exaflop (one quintillion flops) peak performance by 2019. But a growing number of scientists are starting to question this trend. Today's approaches to achieve 10- to 20-petaflops range, coupled with decreasing transistor size approximately every two years, are likely to be scalable up to 100 petaflops. But moving forward past a 100-petaflops range will require developing new ways of computing.

Technologies currently used to produce integrated circuits don't scale in terms of power consumption; a 100-petaflops computer built with today's technology will require a nuclear power plant to supply the electricity needed to power and cool it. Some technologies currently under development—such as integrated CPU and GPU processors, large-scale multicore processors, 3D stacking technology, and very-low-power processors—can help. But without major breakthroughs in integrated circuits technologies, we're heading to a time when the cost of building and operating increasingly larger systems won't provide the payback to society necessary to justify the expense.

***Moving forward past a 100-petaflops range will require developing new ways of computing.***

**Volodymyr Kindratenko** is a senior research scientist at the US National Center for Supercomputing Applications at the University of Illinois. His research interests include high-performance computing and special-purpose computing architectures. Kindratenko has a DSc in analytical chemistry from the University of Antwerp. He is a senior member of the IEEE and the ACM. Contact him at kindr@ncsa.illinois.edu.

**Pedro Trancoso** is an assistant professor at the Department of Computer Science at the University of Cyprus, Cyprus. His research interests include computer architecture, multicore architectures, memory hierarchy, parallel programming models, database workloads, and high-performance computing. Trancoso has a PhD in computer science from the University of Illinois at Urbana-Champaign. He is a member of the IEEE, the IEEE Computer Society, and the ACM. Contact him at pedro@cs.ucy.ac.cy.

*Continued from p. 96*

To retrace your steps through the vast column of books, you'd have the option of leaving a bookmark-like tag. And if you wanted help navigating, you could turn to one of several critics, the crowd-sourced opinions of other readers, or you could ask the bookcase to infer your likely preferences from your previous browsing sessions.

Could used bookstores deploy the online browser? Yes, but they'd need an as-yet-uninvented piece of technology: software that can populate a virtual bookcase from video footage of a store's inventory.

The rotating bookshelf in my godfather's living room was one source of inspiration. The other was Jorge Luis Borges' short story "The Library of Babel," which first appeared in his 1941 collection *The Garden of Forking Paths*. The library in the fantastical story was infinite, full of useful and useless information, and populated by people trying to make sense of it. Not unlike the Internet itself.

**Charles Day** is *Physics Today*'s online editor. He isn't a professional GUI designer, but he hopes the bookcase idea catches on.

*Selected articles and columns from IEEE Computer Society publications are also available for free at http://ComputingNow.computer.org.*