Big Iron Moves Toward Exascale Computing

Neal Leavitt

Developers and researchers face many challenges in trying to produce exascale supercomputers, which would perform a thousand times faster than today's most powerful systems.

Supercomputing is entering a new frontier: the exaflops era, in which high-performance machines could run a thousand times faster than today's petaflops systems.

Some experts say that someone could build an exaflops machine—capable of performing 10^{18} floating-point operations per second—by the end of this decade.

More speed would be welcome, as supercomputing is used in many areas—including nuclear-weapon testing simulations, analyzing the geologies of various areas for possible oil deposits, astronomy, astrophysics, financial services, life sciences, and climate modeling—that could benefit from higher performance.

"This will necessitate new hardware and software paradigms," said Arend Dittmer, director of product marketing for Penguin Computing, a high-performance computing (HPC) services provider.

But for the first time in decades, computing-technology advances might be threatened, said John Shalf, Computer Science Department head in the US Lawrence Berkeley National Laboratory's Computing Research Division.

"While transistor density on silicon is projected to increase with Moore's law, the energy efficiency of silicon is not," he noted. "Power [consumption] has rapidly become the leading design constraint for future high-performance systems."

Thus, said Dittmer, software developers will have to optimize code for power efficiency rather than just performance.

Researchers are also looking into a number of disruptive hardware technologies that could dramatically increase efficiency, including new types of memory, silicon photonics, stacked-chip architectures, and computational accelerators, explained Dimitrios S. Nikolopoulos, professor and director of research at Queen's University Belfast.

The US, China, Japan, the European Union, and Russia are each investing billions of dollars in supercomputer research.

Achieving HPC improvements could even help those who don't use supercomputers. "As always," explained Intel Labs Fellow Shekhar Borkar, "the technology will trickle down to mainstream computing."

However, building exascale machines faces some significant challenges.

BACKGROUND

University of Illinois at Urbana-Champaign researchers started building supercomputers in the early 1950s and parallel supercomputers in the early 1960s.

Seymour Cray, who founded Cray Research—the forerunner of today's Cray Inc.—in the 1970s, is considered the father of commercial supercomputing.

Early supercomputers were designed like mainframes but adapted for higher speed.

In the 1980s, the next wave of HPC machines used custom processors. During the 1990s, general-purpose commercial processors began offering good performance, low prices, and reduced development costs, which made them attractive for use in supercomputers, said Stanford University

 Published by the IEEE Computer Society

Table 1. World's fastest supercomputers as ranked by Top500 (www.top500.org), June 2012.

Rank	Site	Manufacturer	Computer	Country	Cores	Maximum throughput (petaflops)	Power (megawatts)
1	US Lawrence Livermore National Laboratory	IBM	Sequoia	USA	1,572,864	16.30	7.89
2	RIKEN Advanced Institute for Computational Science	Fujitsu	K computer	Japan	795,024	10.50	12.66
3	US Argonne National Laboratory	IBM	Mira	USA	786,432	8.16	3.95
4	Leibniz Rechenzentrum	IBM	SuperMUC	Germany	147,456	2.90	3.52
5	National Supercomputer Center in Tianjin	National University of Defense Technology	Tianhe-1A	China	186,368	2.57	4.04
6	US Oak Ridge National Laboratory	Cray	Jaguar	USA	298,592	1.94	5.14
7	CINECA	IBM	Fermi	Italy	163,840	1.73	0.82
8	Forschungszentrum Juelich	IBM	JuQUEEN	Germany	131,072	1.38	0.66
9	Commissariat a l'Energie Atomique	Bull	Curie thin nodes	France	77,184	1.36	2.25
10	National Supercomputing Center in Shenzhen	Dawning	Nebulae	China	120,640	1.27	2.58

Source: Professor Jack Dongarra, University of Tennessee, Top500 project

professor William Dally, who is also chief scientist and senior vice president of research at GPU maker Nvidia.

The first machine to break the petaflops barrier was IBM's Roadrunner in 2008.

Getting to exascale computing is critical for numerous reasons. Faster supercomputers could conduct calculations that have been beyond reach because of insufficient performance, noted IBM Research director of computing systems Michael Rosenfield.

Another issue is that many complex problems have a large number of parameters. The only way to deal with such problems is to simultaneously run multiple sets of calculations using different combinations of parameters, which requires tremendous computational resources.

Bill Kramer, deputy director for the National Center for Supercomputing Applications' (NCSA's) Blue Waters petascale computing project, said research teams are working on difficult problems in areas such as solar science, astrophysics, astronomy, chemistry, material science, medicine, social networks, and neutron physics.

"All are far more complex to solve than what has been done in the past, and it's only now, with petascale going to exascale, that we can begin to solve these in less than a lifetime," he explained.

TOMORROW'S BIG IRON

While some aspects of supercomputing—such as the traditional forms of security it uses—are unlikely to change to enable exascale computing, others will.

For example, developers are placing processing engines inside memory, rather than outside, to overcome the bottlenecks of today's memory-to-processor connections. They are also working with alternate programming languages that optimize, enhance, and simplify parallelism, as well as communications and control approaches that improve performance.

Fastest Supercomputer: IBM's Sequoia

Sequoia, an IBM supercomputer at the US Lawrence Livermore National Laboratory, is part of the company's BlueGene/Q HPC line and performs 16.3 Pflops.

The Top500 project, in which several academic and research experts rank the world's nondistributed supercomputer systems, placed Sequoia at the top of its list in its recent semiannual report, as Table 1 shows. Ranking second was Fujitsu's K computer, which performs 10.5 Pflops.

Sequoia, which runs Linux and is primarily water cooled, consists of 96 racks, 98,304 16-core compute nodes, 1.6 million total cores, and 1.6 petabytes of RAM.

Despite being so powerful, the system at peak speeds is 90 times more energy efficient than ASC Purple and eight times more than Blue Gene/L, two other very fast IBM supercomputers.

Power consumption

Sequoia uses 7.89 megawatts at peak performance. At that rate, a one-exaflops machine would consume 400 MW, about one-fifth of Hoover Dam's entire generation capacity, said Nathan Brookwood, research fellow with semiconductor consultancy Insight 64.

Without improvements, an exascale computer "might need its own nuclear power plant or large hydroelectric dam," said Carl Claunch, vice president and distinguished analyst with market research firm Gartner Inc.

When an early Univac powered up in the 1950s, the lights dimmed in the surrounding neighborhood, Brookwood noted. "Imagine the impact of powering up a 400-MW supercomputer and watching the lights dim in the Southwest," he continued. "Future systems must improve their performance per watt by a factor of 40 or more to deliver exascale results within reasonable power envelopes."

The US Department of Energy has set a goal that supercomputers use no more than 20 MW of power, which would require radical redesigns of processors, interconnects, and memory, noted Alan Lee, vice president of research and advanced development for chipmaker AMD.

Supercomputer designers now routinely incorporate energy-conserving features that turn off idle elements within their chips when possible. Modern chips, noted Stanford's Dally, use both *power gating*—in which power to parts of a chip is shut off—and *clock gating*—in which the power is left on but the clock is turned off.

Stacked DRAM placed close to the processor increases memory bandwidth while requiring significantly less power to transfer data than current designs, he noted.

Supercomputers could become more energy efficient by using low-power memory and also components that run at lower frequencies, as well as reducing the amount of data movement, added Lee.

In the future, said University of California, San Diego (UCSD) professor Michael Taylor, using highly specialized, low-speed application-specific coprocessors could help decrease energy consumption.

Processing

Processor performance will be a key factor in exascale computing. And parallelism, created via multiple cores on a chip working on different tasks simultaneously, is driving processor-performance improvements.

Future supercomputers will have more cores per chip and each core will run many threads to hide latency, according to Stanford's Dally.

There is an emerging consensus that future supercomputers will be heterogeneous multicore computers, with each processing chip having different types of cores specialized for different tasks, he said.

There could be an exascale computer by the end of this decade.

For example, Dally explained, the majority of the cores would be throughput-optimized to execute parallelized work quickly and with minimum energy consumption, as is the case with GPUs.

A small number of cores would be latency-optimized, like those in CPUs, for use in critical serial tasks.

For parallel tasks, said AMD's Lee, GPUs are more energy efficient than CPUs because they use a single instruction to perform many operations and because they run at lower voltages and frequencies.

Each type of processor provides distinct advantages, said Tony King-Smith, vice president of marketing for Imagination Technologies, which designs and licenses multimedia and communications semiconductor cores.

However, using different kinds of cores would not come without challenges. For example, programming multiple types of processors to take advantage of their distinct characteristics can be complex and time consuming.

Investigating such matters is the Heterogeneous System Architecture (HSA) Foundation consortium of chipmakers, academics, equipment manufacturers, software companies, and operating-system vendors.

According to King-Smith, supercomputer developers will face a challenge in balancing performance, power, and chip size.

Moreover, noted Dally, an exascale machine will have to run over a billion parallel threads at any time to keep busy—many times more than today's fastest computers—and this will require new programming approaches.

Designers of exascale computers could turn to 3D processors with various layers of circuitry stacked on top of one another. Integrating a large processing system this way improves memory access and performance, added Dally.

However, this also creates challenges for cooling and power supply. For example, noted Georgia Institute of Technology assistant professor Richard Vuduc, heat builds up between layers, making them harder to cool. In addition, he said, the interlayer connections are difficult to design, and there are few tools for developing and testing 3D circuits.

Memory and interchip communications

Two crucial limitations that exascale computing faces are the increasing speed disparity between a CPU and external memory, and the relative slowdown in interconnect support.

Processing speeds have increased exponentially, but the connecting fabric and memory controllers are still working to keep up.

Introducing solutions such as data compression, as well as optimizing memory organization and usage by adding localized caches and thereby keeping more of the processing on

chip, would improve some memory- and interconnect-related issues, said King-Smith.

Optical links and 3D chip stacking could improve interchip communications and lower power consumption, but further research in this area is necessary, said the UCSD's Taylor. He predicted that optical connections will be built onto chips during the next few years.

However, Stanford's Dally noted, some major technical hurdles must be cleared for this to happen, such as reducing the cost and power dissipation of optical links.

Internode networking

Many supercomputers use high-speed Ethernet for communication between processing nodes. However, the technology's deep router queues and tendency to occasionally drop packets could create high and unpredictable latencies unsuitable for exascale computing.

Proprietary networking technologies like those used in Cray machines and other supercomputers have lower-latency routers, topologies with fewer hops, and more efficient buffer management, said Bill Blake, Cray's senior vice president and chief technology officer. This approach provides low internode latency and high bandwidth.

However, proprietary technologies are expensive.

Cooling

Supercomputers generate huge amounts of heat. If the heat is not either cooled or moved away from chips, connectors, and the machine's many other heat-sensitive components, they—and the entire system—will fail.

In the past, supercomputers have used liquid cooling and/or air cooling via fans and heat sinks. Liquid cooling is highly effective, but it can be expensive and would become much more so in exascale systems.

Thus, exascale systems might implement hybrid cooling systems using both liquid and air cooling, said Cray's Blake.

A current trend, noted IBM's Rosenfield, is to use room-temperature water to efficiently conduct heat away from sensitive components without requiring water chillers, which would increase energy consumption.

According to Dally, the individual components of an exascale machine could be reliable and have low failure rates, but combining so many into one large computer increases the chance a failure will occur somewhere.

"Concerted government and industry investment and collaboration are needed to overcome the challenges [of exascale computing]. Leadership is necessary … as evidenced by sovereign strategic commitments to HPC in Japan, China, Russia, and Europe," said IBM's Rosenfield.

Added King-Smith, industry consortiums such as the Khronos Group and the HSA Foundation should continue pushing mainstream adoption of technologies like heterogeneous processing and GPU computation.

In the past 20 years, US expenditures on HPC have steadily grown, but it's not certain that the country will have the first practical exascale system, noted the NCSA's Kramer. The technical challenges and uncertainties; the complexity of industrial, government, and national laboratory partnerships; and budget problems might mean there won't be the focused US effort necessary for such an expensive and technically challenging undertaking.

Current plans for a US system by 2018 are no longer likely to bear fruit, and the country may not have an exascale machine until between 2023 and 2025, Kramer said. "China, Europe, even Russia may arrive at some type of exascale system first," he added.

If exascale systems aren't built and computing performance stalls at today's levels, said the Lawrence Berkeley National Laboratory's Shalf, the information-technology industry will shift from a growth industry to a replacement industry, and future societal impacts of computing will be limited.

Neal Leavitt *is president of Leavitt Communications (www.leavcom.com), a Fallbrook, California-based international marketing communications company with affiliate offices in Brazil, France, Germany, Hong Kong, India, and the UK. He writes frequently on technology topics and can be reached at neal@leavcom.com.*

Editor: Lee Garber, *Computer*; l.garber@computer.org