IDC EXECUTIVE BRIEF

Heterogeneous Computing: A New Paradigm for the Exascale Era

November 2011

Adapted from *IDC HPC End-User Study of Processor and Accelerator Trends in Technical Computing* by Earl C. Joseph and Steve Conway, IDC #228098

Sponsored by NVIDIA

The Rise of the Heterogeneous Computing Paradigm

The worldwide high-performance computing (HPC) market is already more than three years into the petascale era (June 2008–present) and is looking to make the thousandfold leap into the exascale era before the end of this decade. This pursuit is global in scope. IDC expects the United States, the European Union, Japan, China, and Russia to vie with each other to reap exascale computing's anticipated substantial benefits for scientific advancement, industrial-economic competitiveness, and the quality of human life.

But as many HPC experts have noted, achieving reasonable exascale performance in this compressed time frame presents an array of daunting challenges that cannot be met only through evolutionary extrapolations from existing technologies and approaches. These challenges include, but are not limited to, the following:

- **System costs (flops/dollar).** Twenty years ago, the world's leading HPC sites spent $25 million to $30 million for the most powerful supercomputers available. Today's single-digit petaflops supercomputers often cost over $100 million. Early exaflop systems could cost $500 million to $1 billion each. This cost escalation will be difficult to sustain. Anything that can increase the flops/dollar ratio will be welcome.
- **Application performance (time/solution).** This perennial challenge grows continually as HPC users seek to scale their applications to new, larger systems. With clock rates stalled, future performance gains must come almost entirely from increased parallelism, resulting in tremendous concurrency requirements for exascale computing. A 1GHz machine would need to perform a billion independent operations every clock tick. Over time, many large science problems will be able to scale to this level. Other problems will lack the required concurrency for single runs but may make use of extreme-scale systems to run ensemble calculations. Automotive design engineers, for example, have greatly increased the number of parametric runs — along with the resolution of each run — that can occur in their allotted phase of the design cycle.
- **Space and compute density requirements (flops/square foot).** A worldwide IDC study revealed that most HPC sites are struggling mightily with datacenter space limitations. Two-thirds of the sites were planning to expand or build new HPC datacenters. Half of the sites planned, or had already begun, to distribute their HPC resources to multiple locations.
- **Energy costs for computation and data movement (flops/watt, bytes/watt).** Last but not least, power has become both a significant design constraint and a major contributor to cost of ownership. With voltage scaling slowing dramatically, power is no longer holding constant as we

grow the transistor count with Moore's law, resulting in processor designs that are power constrained today and becoming more so with each new IC generation. Performance in this era is determined largely by power efficiency, so the great challenge in system design is making processors and data movement more energy efficient without overly compromising performance. The rapid growth in HPC system sizes has elevated energy requirements. Today's largest HPC datacenters consume as much electricity as a small city, and multi-petascale and exascale datacenters promise to devour even more. Energy prices have risen substantially above historic levels, although prices have moderated from their 2008 highs. Another element in this "perfect storm" is that HPC datacenter power and cooling developments are occurring at a time of growing sensitivity toward carbon footprints and global climate change. Finally, some of the biggest HPC datacenters worry that their local power companies may balk at fully supplying their future demands. One such site, already seeing the need for a 250-megawatt datacenter, may have to go off the power grid and build a small nuclear reactor.

The Heterogeneous Computing Paradigm

During the past decade, clusters leveraging the economies of scale of x86 processors became the dominant species of HPC systems — doubling the size of the global HPC server market from about $5 billion in the early 2000s to $9.5 billion in 2010. The reigning paradigm has been to advance peak performance by deploying larger and larger clusters containing more and more standard x86 CPU cores.

But x86 processors were never designed to handle all HPC applications well, and x86 single-threaded performance started to hit a heat and power wall half a dozen years ago. It is becoming increasingly clear that although x86 processor road maps lay out substantial advances, the paradigm of sole dependency on x86 processors will not suffice to meet the challenges associated with achieving exascale computing in this decade.

In recent years, an alternative paradigm, "heterogeneous computing," has gained market momentum for addressing these challenges. This emerging paradigm augments x86 CPUs with accelerators, primarily GPGPUs (henceforth to be called GPUs), so that each processor type can do what it does best. GPUs are particularly adept at handling the substantial number of codes, and portions of codes, that exhibit strong data or thread-level parallelism. That makes GPUs the heirs apparent to vector processors, except that GPUs benefit from far greater economies of scale and related competitive advantages. IDC research shows that the worldwide PC market for discrete graphics processing units alone was worth about $4 billion in 2010.

The heterogeneous computing paradigm is ramping up nicely across the HPC market as a whole. IDC's 2008 worldwide study on HPC processors revealed that 9% of HPC sites were using some form of accelerator technology alongside CPUs in their installed systems. Fast-forward to the 2010 version of the same global study and the scene has changed considerably. Accelerator technology has gone forth and multiplied. By this time, 28% of the HPC sites were using accelerator technology — a threefold increase from two years earlier — and nearly all of these accelerators were GPUs. Although GPUs represent only about 5% of the processor counts in heterogeneous systems, their numbers are growing rapidly.

Heterogeneous computing is making its greatest impact at the high end of the HPC market. GPUs first appeared on the TOP500 list of the world's supercomputer sites (www.top500.org) in 2008. By June 2011, three of the top 10 systems on the list employed GPUs. And in October 2011, the U.S. Department of Energy's Oak Ridge National Laboratory unveiled plans to upgrade the number one U.S. supercomputer to a successor system ("Titan") with a planned peak performance of 20–30 petaflops by complementing more than 18,000 x86 CPUs with an equal number of GPUs. Following that, the Texas Advanced Computing Center revealed plans for a heterogeneous supercomputer

("Stampede") that will initially target 10 peak petaflops by combining two petaflops of x86 CPUs with eight petaflops of MIC accelerator processors.

The adoption of heterogeneous computing by these and other bellwether HPC sites indicates that GPUs are moving out of the experimental phase and into the phase where they will be increasingly entrusted with appropriate production-oriented, mission-critical work.

Definitions

- **Cluster:** IDC defines a cluster as a set of independent computers combined into a unified system through systems software and networking technologies. Thus, clusters are not based on new architectural concepts so much as new systems integration strategies.
- **Heterogeneous processing:** Heterogeneous processing and the synonymous term heterogeneous computing refer to the use of multiple types of processors, typically CPUs in combination with GPUs or other accelerators, within the same HPC system.
- **High-performance computing:** IDC uses the term high-performance computing to refer to all technical computing servers and clusters used to solve problems that are computationally intensive or data intensive. The term also refers to the market for these systems and the activities within this market. It includes technical servers but excludes desktop computers used for technical computing.

Heterogeneous Computing Benefits for the Exascale Era

The benefits of the heterogeneous computing paradigm for HPC are interrelated and address some of the most important exascale challenges:

- **System costs.** GPUs and related accelerators can provide a lot of peak and Linpack flops for the money. Especially for HPC sites seriously pursuing the upper ranges of the TOP500 supercomputers list, bolting on GPUs can provide a kind of flops warp drive, rocketing Linpack performance to where almost no one has gone before. Witness China's Tianhe-1A supercomputer, which supplemented x86 processors with GPUs to seize the number one spot on the November 2010 TOP500 list. To achieve this feat, Tianhe-1A employed 14,336 x86 CPUs and 7,168 GPUs. NVIDIA suggested at the time that it would have taken "50,000 CPUs and twice as much floor space to deliver the same performance using CPUs alone." In any case, by June 2011, three of the top five systems on the list employed GPUs. And in October 2011, as noted earlier, the U.S. Department of Energy's Oak Ridge National Laboratory unveiled plans to upgrade the number one U.S. supercomputer to peak performance of 20–30 petaflops by complementing more than 18,000 x86 CPUs with an equal number of GPUs.
- **Speed.** HPC users have reported excellent speedups on GPUs, frequently in the 3–10x range, especially for codes, or portions of codes, that exhibit strong data parallelism. GPUs are already enabling real-world advances in HPC domains, especially the life sciences, financial services, oil and gas, product design and manufacturing domains, and digital content creation and distribution. GPUs are a particularly promising fit for molecular dynamics simulations, which extend across multiple applications domains.
- **Space and compute density.** At a time when many HPC datacenters are approaching the limits of their power and space envelopes, GPUs promise to deliver very high peak compute density. A contemporary GPU may contain as many as 512 compute cores, compared with 4 to 16 cores for a contemporary CPU. Keep in mind, however, that heterogeneous computing is heterogeneous for a reason: Each processor type, CPU, and accelerator is best at tackling a different portion of the problem-solving spectrum.

- **Energy costs.** The rapid growth in HPC system sizes has caused energy requirements to skyrocket. Today's largest HPC datacenters consume as much electricity as a small city, and exascale datacenters promise to devour even more — an estimated 120 megawatts or more, if they were implemented using existing technologies. The Department of Energy's exascale goal is to bring that number down to no more than 20 megawatts for an exascale system deployment. This is desired to avoid massive increases in energy costs, to ensure the availability of adequate energy supplies from local utilities, and to keep datacenter spatial requirements within reason. GPUs can be valuable partners with CPUs in heterogeneous computing configurations by significantly improving energy efficiency on the substantial subset of codes (and portions of codes) exhibiting strong data parallelism.

Adoption Barriers

As a relatively new technology — at least as used for computing — GPUs have encountered adoption barriers that IDC expects to ease over time. HPC buyers report the following main barriers to more extensive GPU deployment:

- **Ease of programming.** Despite the availability of useful tools such as CUDA, OpenCL, and The Portland Group's directives-based compiler that's designed to transform Fortran or C source code into GPU-accelerated code, HPC buyers and end users generally report that programming GPUs remains more challenging than the more familiar approaches to programming standard x86 processors. This barrier will likely continue to drop over time as familiarity with GPU programming methods grows — through the more than 450 universities offering GPU curricula today and as the GPU programming methods advance.

- **Mediated communication.** Another issue frequently cited by HPC users is the fact that GPUs today are typically implemented as coprocessors that need to communicate with x86 or other base processors via PCI Express channels that are comparatively slow — at least when weighed against implementing the CPU and GPU on the same die. This mediated communication affects some applications more than others. It has not prevented HPC users from achieving impressive time-to-solution speedups on a growing number of application codes.

- **Waiting for future CPU generations.** Some HPC users believe that waiting to see what improvements future-generation x86 processors deliver is a risk worth taking, compared with the effort of learning how to program GPUs and adapting portions of their codes to run on GPUs. And because GPUs are still relatively new devices for high-performance computing, some users worry that the substantial effort to rewrite their codes could be wasted if GPU architectures evolve in a new direction or if GPUs are not an enduring phenomenon in the HPC market. This wait-and-see group has been declining as GPUs have increased their influence in the global HPC market and as directive-based programming of GPUs has become more prevalent.

Trends

Heterogeneous computing, which today typically couples x86 processors with GPUs implemented as coprocessors, is an important new paradigm that is increasingly taking its place alongside the existing paradigm of pure-play x86-based HPC systems.

An important sign of GPUs' momentum is the spread of GPU-related academic offerings. NVIDIA, which supplies educational materials for parallel programming, reports that its CUDA parallel programming language is being taught at 478 universities in 57 countries. The list includes MIT, Harvard, Stanford, Cambridge, Oxford, the Indian Institutes of Technology, National Taiwan University, and the Chinese Academy of Sciences.

For reasons stated earlier (refer back to the Benefits section), heterogeneous computing is proving especially attractive to large HPC sites that are pushing up against the boundaries of computational science and engineering, as well as energy and spatial boundaries. Hence, heterogeneous computing looks particularly attractive as a new paradigm for the exascale computing era that will begin later in this decade. Heterogeneous computing involving GPUs is also making inroads into smaller research sites and industrial organizations.

Keep in mind, however, that x86 processor technology is not standing still and promises to remain the HPC revenue leader through 2015, the close of IDC's current HPC forecast period. In addition, accelerator technology will be available from an increasing number of vendors and in a growing array of "flavors," giving users more options.

Conclusion

Heterogeneous computing that today typically couples x86-based processors with GPUs implemented as coprocessors is an important emerging paradigm in the worldwide HPC market — especially as a strategy to help meet the daunting challenges of the coming exascale computing era. IDC believes that heterogeneous computing will be indispensable for achieving exascale computing in this decade.

GPUs are rapidly emerging from the experimental phase and are often used today for production-oriented tasks such as seismic processing, biochemistry simulations, weather and climate modeling, computational finance, computational fluid dynamics, and data analysis. They have tripled their worldwide footprint at HPC sites in the past two years alone, they have become more indispensable for attaining prominence on the closely watched TOP500 supercomputers lists, and they have enabled a growing number and variety of real-world achievements.

The adoption of heterogeneous processing involving GPUs by some of the world's leading HPC sites indicates that this paradigm is moving beyond the experimental phase, and GPUs are increasingly being entrusted with appropriate portions of production-oriented, mission-critical workloads.

As GPU hardware and software technologies advance, as more university students and others learn how to exploit GPUs, and as more GPUs become available to the world's most creative scientific, engineering, and computational minds, IDC believes GPUs will play an increasingly important role in the global HPC market, complementing x86 processors within the HPC ecosystem.

ABOUT THIS PUBLICATION

This publication was produced by IDC Go-to-Market Services. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis independently conducted and published by IDC, unless specific vendor sponsorship is noted. IDC Go-to-Market Services makes IDC content available in a wide range of formats for distribution by various companies. A license to distribute IDC content does not imply endorsement of or opinion about the licensee.

For more information on IDC, visit www.idc.com. For more information on IDC GMS, visit www.idc.com/gms.

Global Headquarters: 5 Speen Street Framingham, MA 01701 USA P.508.872.8200 F.508.935.4015 www.idc.com