Supercomputing Hits the Desktop

Beth Stackpole, Contributing Editor

1/23/2009 1 Comment

NO RATINGS

inShare

Much like how the advent of the PC forever altered the work habits and productivity level of everyday corporate employees, the emerging personal supercomputer platform promises to have a similar impact on how engineers, scientists and researchers tackle complex design and simulation work.

The recent wave of new personal supercomputers packs on a single desktop the processing horsepower - in some cases, four teraflops and beyond - that typically has only been available from dedicated, multi-million dollar, high-performance computer (HPC) and cluster offerings. Fueling the additional computing muscle is a new breed of hybrid system that harnesses both CPU and GPU (graphics processing unit) technology, amassing a performance punch that's on par with traditional, dedicated HPC systems.

"The difference between a regular computer and a supercomputer is the scale of work you can do on it," says Andy Keane, general manager for Tesla supercomputing, at NVIDIA, the GPU manufacturer, which is pushing its new Tesla C1060 Computing Processor as the heart of this new computing paradigm. "Someone doing computer tomography or electrodynamics needs a certain level of computation to do useful work. Over time, the amount of computing horsepower needed to solve the problems in a useful time has outgrown what people are doing on the desktop."

NVIDIA and its customers have been circumventing the performance roadblock by piecing together custom systems that combine multiple multicore CPUs and Tesla GPUs. MIT and Harvard, for instance, have 16 GPU systems cranking through complex medical imaging research. The University of Antwerp, Belgium operates an eight GPU system in its FASTRA GPU Super PC project for tomography, a technique used in medical scanners to create three dimensional views of patient organs based on a collection of X-rays that explore a range of angles. Antwerp claims its Tesla GPU-powered system performs these medical reconstructions in only a few hours - a rate nearly 100 times as fast as with traditional HPC systems.

Based on success stories like these, NVIDIA, late last year, decided to take its grassroots personal supercomputer effort to the masses. In November, it released the Tesla C1060, a packaged version of its GPU platform that will allow mainstream OEMs and custom systems builders to more easily build off-the-shelf personal supercomputer systems. The Tesla C1060 offers 960 cores for a total of four teraflops or 250X the performance of traditional desktop workstations and is priced under $10,000, according to NVIDIA officials. To date, 24 partners, including supercomputer giant Cray and Dell Computer, as well as some custom computer makers such as Colfax International and Western Scientific, have shipped or announced plans for Tesla GPU-based desktop supercomputers.

The benefits of a desktop-class supercomputer are many. On one hand, more affordable and accessible HPC horsepower allows engineers to do virtual reality studies on design concepts and solve problems around computational fluid dynamics (CFD) and structural mechanics far earlier on in the process, prior to building costly prototypes. Desktop machines of this class also have a convenience factor, allowing engineers to perform these complex calculations without having to wait for precious computing resources to become available and without injecting unnecessary time delays into the development cycle. This is particularly important in industries such as high-tech and electronics where demand for innovation and speed to market is acutely tied to a company's success.

"Computer chips are no longer just microprocessors - they're complete systems," says Drew Gude, director of U.S. high tech and electronics industry for Microsoft, which has put its stake in HPC territory with the release of Windows HPC Server. "The engineering and testing required to test these systems becomes incrementally complicated and access to this kind of computing power allows individuals creating those systems to compile software more quickly, run a variety of tests and do CFD analysis from one system."

A Hybrid Approach

While the personal supercomputer as a marketing concept has been percolating for years, the hybrid CPU/GPU architecture is new and has the potential to really advance the market, according to Addison Snell, vice president and general manager for Tabor Research, a research firm solely focused on the HPC market. Other types of accelerators, including FPGAs, have been used to advance workstation-class HPC systems, but NVIDIA's GPU approach appears to have the momentum, Snell says. "The hybrid has the potential of reincarnating the classic market space of the technical workstation, which has been on the decline," he says. "This now enables you to have a powerful workstation on the desktop, combining several processors and graphics co-processors to either enhance the computational aspects or post-process rendering."

Snell says one of the longstanding obstacles for a hybrid approach is the issue of programmability since software has to be rewritten to take advantage of the new parallel architecture. "At some level, software needs to contain instructions that tell the system to execute a given calculation on a GPU rather than a CPU," he says. "The approach can be made more straightforward, but it doesn't eliminate the step that software has to be rewritten."

NVIDIA tackles that problem with CUDA, its C++-like programming environment used to tune applications to the GPU's parallel computing architecture. CUDA has been widely embraced among leading software vendors, including Mathematica and National Instruments, giving NVIDIA's GPU hybrid a jump on competing accelerator models, Snell says. However, CUDA has limitations in that it ties developers to the NVIDIA platform, according to Microsoft's Ryan Waite, product unit manager for Windows HPC Server. What the market requires, Waite says, is a general-purpose language for GPU programming that works across several platforms. To that end, he says Microsoft is planning a set of extensions to its Direct X language to enable developers to tap into the parallel processing engines in GPUs.

In addition to Microsoft's efforts, The Khronos Group has formed an industry working group to create royalty-free, open standards for programming parallel computing across GPUs and CPUs. The so-called Compute Working Group includes participation from such leaders as IBM, Apple, ARM, AMD, NVIDIA, Freescale and others. Apple has proposed the Open Computing Language (OpenCL) specification as a basis for the standard and the group is soliciting companies to provide input.

There are other software-related advances helping to propel HPC's march onto the desktop, according to Microsoft. Windows HPC Server, particularly the latest 2008 version, has done its part to help make these very complex systems easier to use and manage, especially for engineers and scientists who aren't necessarily versed in computer administration tasks, according to Waite. "If you're an engineer or scientist using Mathlab, you want to stay in that application," he says. "A personal supercomputer supports this idea instead of having centralized resources that only a limited number of people have access to and which require a lot of care and feeding."

John Stone, a senior research programmer at the University of Illinois at Urbana-Champaign , has long been a user of HPC clusters, but admits a personal supercomputer brings a level of convenience not seen with traditional systems. Stone, part of the university's Theoretical and Computational Biophysics Group, taps HPC technology to create a mathematical model that simulates the behavior of biological molecules at an atomic level - something that's not visible even through highly sophisticated microscopes. A couple years ago, running these highly complex calculations meant tapping some far-off supercomputer for days. Last year, Stone's group began working with CUDA and a custom-built hybrid CPU/GPU system and was able to perform the same work on a single GPU machine in their own lab in only 27 minutes. Stone says early tests show the new Tesla hybrid platform drastically improves even that level of performance.

"Being able to do this work locally without having to transfer files to a remote machine managed by someone else eliminates hassles and is a big time saver," he says. "The same reason why you'd want a personal computer in the first place now applies to a problem that previously required a small cluster or supercomputer."