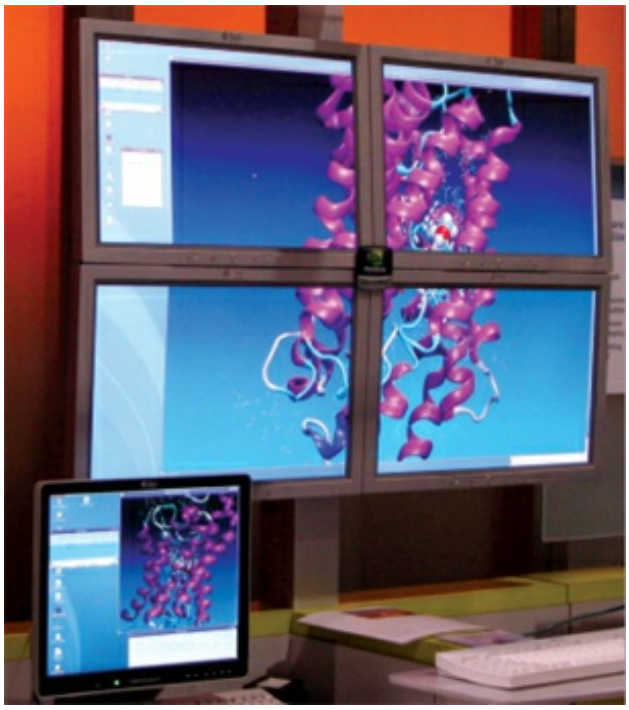
NVIDIA's Tesla Borrows from Games to Advance High-Performance Computing

By David Hoff

This multiscreen Sun machine used multiple GPUs to compute the inter-atomic forces that govern the dynamics of how certain materials pass through cell membranes. In this photo, a simulated aquaporin, a protein,becomes a channel for glycerol. ***Image courtesy Theoretical and Computational Biophysics Group at UIUC***



A single human cell out of billions starts to divide out of control and cascades into a process that leads to cancer. To find out how it happens and how it can be stopped, some scientists are using new computing techniques and high-powered arrays of computers to simulate basic cell behavior.

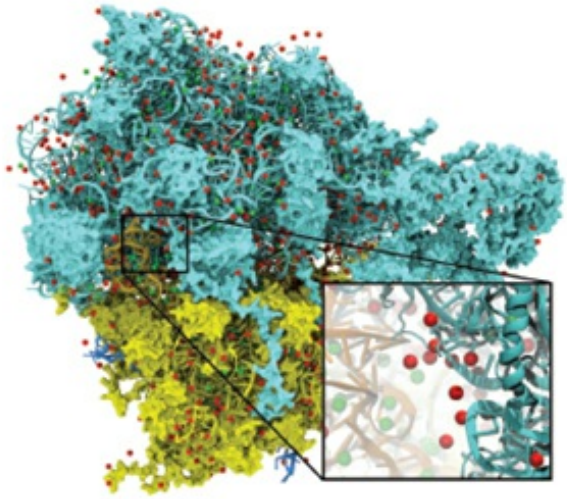
Until recently the sheer scale of simulating the parts of a cell, all their interactions, and how they respond to external factors that cause disease required some of the most powerful computer clusters in the world. But a change in the world of high performance computing (HPC) is bringing the power of supercomputers to the desktop, and the change is coming from a quarter that until recently has been traditionally associated with gaming. In fact, the graphics processing unit (GPU) is fundamentally transforming the way that scientists are able to work.

GPUs for HPC

While it may be surprising to some that a video game processor could be used for supercomputing, the potential of the GPU for computation is not surprising to those familiar with the technology. A GPU is built for compute-intensive, highly parallel computation, exactly what is required for calculating the graphics on the screen and extremely useful in all sorts of computational environments. Unlike a CPU, a GPU includes more transistors devoted to data processing rather than to data caching and flow control. This makes the GPU especially well suited for executing a single program on lots of data in parallel with high arithmetic intensity, or a high ratio of arithmetic-to-memory operations.

"The GPU is a powerful, programmable platform that is perfect for computing applications such as seismic processing for oil and gas exploration, computing in bioscience, and financial modeling," says Andy Keane, general manager of

This image illustrates the placement of ions in the space surrounding a molecular structure, reproducing biological conditions.The GPU was used to compute a 3D electrostatic field surrounding the molecule. Once calculated, this field is used to calculate the correct locations for placement of ions in the molecular model, in preparation for running a molecular dynamics simulation to study the structure and function of the molecule. *Image courtesy Theoretical and Computational Biophysics Group at UIUC*



the GPU computing business at NVIDIA, a pioneer in using GPUs for HPC. “The GPU will change the way engineers and researchers approach these problems.”

Traditional computing with a high-end, quad-core CPU may use up to four processing threads to loop over data sequentially, while data-parallel processing with a GPU maps data elements to thousands of parallel-processing threads. Many applications that process large data sets, such as arrays or volumes, can use a parallel programming model to greatly accelerate the computations.

Gigaflops of Performance

This shift in HPC began a year ago when NVIDIA launched CUDA, a C language development tool for the GPU. Unlike previous attempts to use the GPU to handle these kinds of computing tasks with graphics APIs, researchers and engineers can now use the familiar programming environment of C and port their code to use the parallel architecture of the GPU in a matter of hours, in some cases. As a result, applications that originally took days and weeks to run and deliver results are now taking hours or minutes.

The pace of the change quickened six months ago, with the launch of Tesla, a GPU product line designed specifically for the HPC market and for supercomputing installations, boasting gigaflops of performance in a fraction of the space required by current solutions.

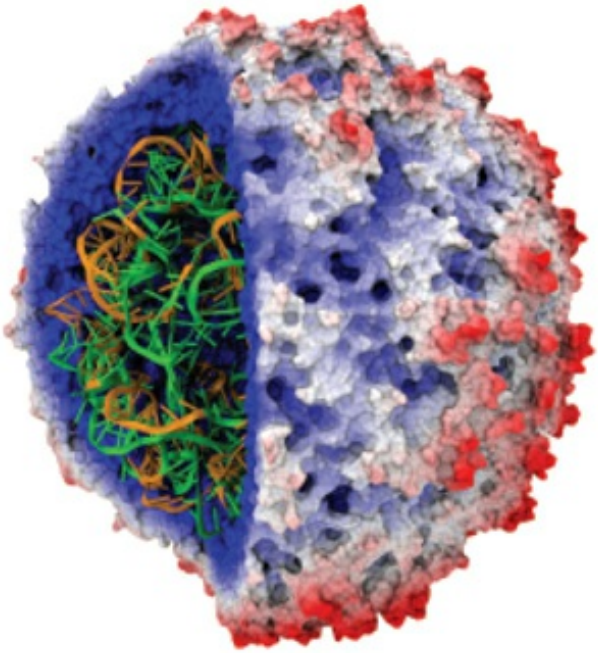
GPUs for computing are now available in a variety of form factors. In addition to GPUs in the traditional configuration of a workstation, multiple GPUs are now available in deskside units that can be attached to any PC workstation to add a performance boost. And servers as thin as 1U or 1.75 in. are available with as many as four GPUs, bringing the technology to back-room server clusters.

“The launch of a CUDA C-compiler and development environment was the key to unlocking the potential of the GPU as a computational engine," says Keane. "With an installed base of over 45 million CUDA-capable GPUs already in the market, downloads of the SDK increasing every day, and user-generated code examples appearing on our forums, it’s clear that GPU computing is answering the needs of the HPC industry. A large number of computing tools, ranging across the spectrum of applications, are being reengineered to harness the power of the GPU.”

Advances Due to GPUs

A pair of these applications is being developed at the University of Illinois at Urbana-Champaign (UIUC) for simulating biological molecular dynamics. UIUC researchers have been using their nanoscale molecular dynamics (NAMD) and visual molecular dynamics (VMD) software running on NVIDIA GPUs to conduct simulations of nano-devices that can

This is a cut-away view of a protein capsid (outer shell) and internally contained RNA for the Satellite Tobacco Mosaic Virus,one of the largest biological molecules simulated on a supercomputer. By viewing the virus structure in detail, researchers were able to study the mechanics of the virus capsid, an important component of understanding viral infections. ***Image courtesy Theoretical and Computational Biophysics Group at UIUC***



be used to sequence DNA in real-time and help reduce the cost of genomic medicine. They are seeing 100x to 240x speed increases and, more importantly, can now run these calculations at their desks, rather than queuing up to use large clusters in remote server rooms and waiting weeks for the results. With a GPU and the software from UIUC, any biologist now has the power to solve new problems, quite literally, at their fingertips.

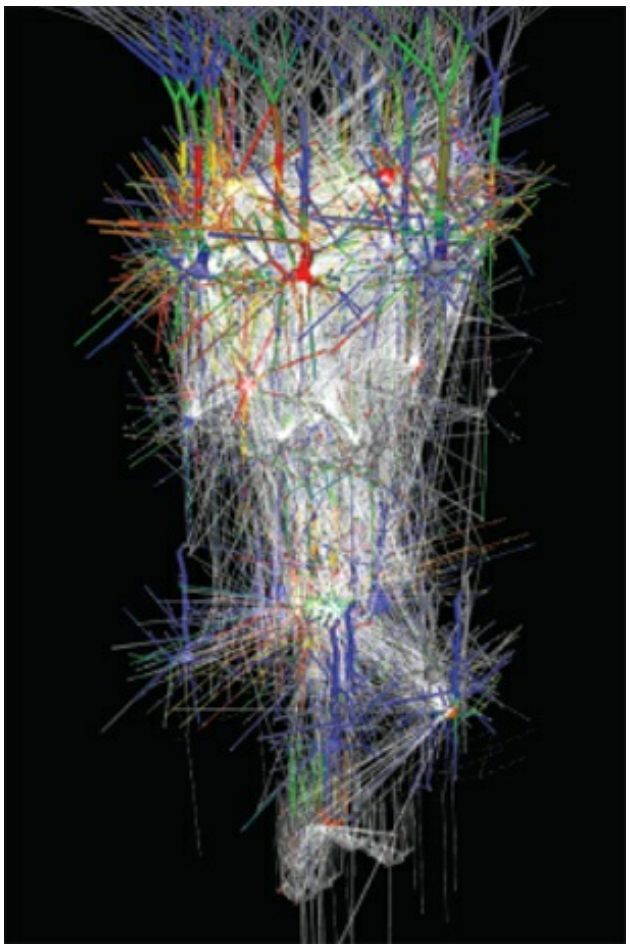
And UIUC is hardly alone. Researchers at Massachusetts General Hospital have been working to create medical images using digital tomosynthesis, a mammography technique that promises to make detection of cancer lesions in breast tissue easier and earlier. The technique has existed for some decades, but the computational power required to reconstruct images from the data has made it impractical until now. By using GPUs in a compute mode, researchers at Massachusetts General Hospital have achieved a 100-fold speedup in the reconstruction of the image, reducing the computation time from five hours to about five minutes on a single PC, a footprint that will allow digital tomosynthesis systems to be installed in any radiology suite.

The Palo Alto-based company Evolved Machines is using GPUs to reverse-engineer the circuits of the human brain. The goal is first to understand how the neural circuitry works and then to use those principles to create machines capable of similar functions. The company has, to date, achieved speed increases of more than 100 times and claims

that a single desktop system containing two GPUs runs as fast as a 200-core cluster, at a fraction of the cost and power.

Other Applications

In the field of finance, Hanweck Associates has developed a real-time options implied-volatility engine, Volera. Using a single PC containing three GPUs, Volera can evaluate 150,000 options per second. Two such systems networked together can evaluate the entire US-listed, equity option market in less than a second.



A neural network as visualized using NVIDIA Tesla. ***Images courtesy Evolved Machines***

This is just a sampling of the HPC applications that are using the GPU to solve large-scale, critical problems. By providing the processing power, software development tools, and building a GPU-computing community where developers can reach outside their industries and fields of research to consult and generate new ideas, NVIDIA is pioneering this new type of computing.

“GPUs aren’t ideal for every problem,” says Keane. “Traditional CPU-based server clusters are not going to go away. But for large numbers of applications, the GPU can move the computation out of the back room and onto the lab bench. And it can power new applications for which the needed computing power just wasn’t available before.”

More Info:

NVIDIA
Santa Clara, CA

Evolved Machines
Palo Alto, CA

Hanweck Associates
New York, NY

Massachusetts General Hospital
Boston, MA

University of Illinois Urbana-Champaign (UIUC)
Urbana, IL

David Hoff is the CUDA product manager at NVIDIA. Send e-mail about this article to DE-Editors@deskeng.com.