NVIDIA® Tesla™

GPU Computing Technical Brief

Version 1.0.0

5/24/07

Table of Contents

Chapter 1.
High-Performance Computing on the GPU

1.1 High-Performance Parallel Computing

The CPU and operating system powering the modern PC solve an incredibly difficult problem in computing. As you use the computer, the operating system tracks all your activities, communicates in the background, and organizes the information you use while you're listening to music, browsing the Web, and reading e-mail. Even though the CPU works on separate tasks one at a time, it has enough speed so these serial tasks appear to operate simultaneously. With new multi-core CPUs, each core can handle an additional task with true simultaneity .

A different class of computing problem, parallel computing, has until recently remained the realm of large server clusters and exotic supercomputers. Standard CPU architecture excels at managing many discrete tasks, but is not particularly efficient at processing tasks that can be divided into many smaller elements and analyzed in parallel. This is exactly the type of problem solved by graphics processing units (GPUs).

The GPU has great potential for solving such problems quickly and inexpensively. GPU computing makes supercomputing possible with any PC or workstation and expands the power of server clusters to solve problems that were previously not possible with existing CPU clusters.

The goal of computing with GPUs is to apply the tremendous computational power inherent in the GPU to solve some of the most difficult and important problems in high performance computing.

1.2 GPU as a Parallel Computing Device

In just a few years, the graphics processor unit (GPU) has evolved into a computing workhorse. With up to 128 processors and very high memory bandwidth, GPUs offer incredible resources for both graphics and non-graphics processing.

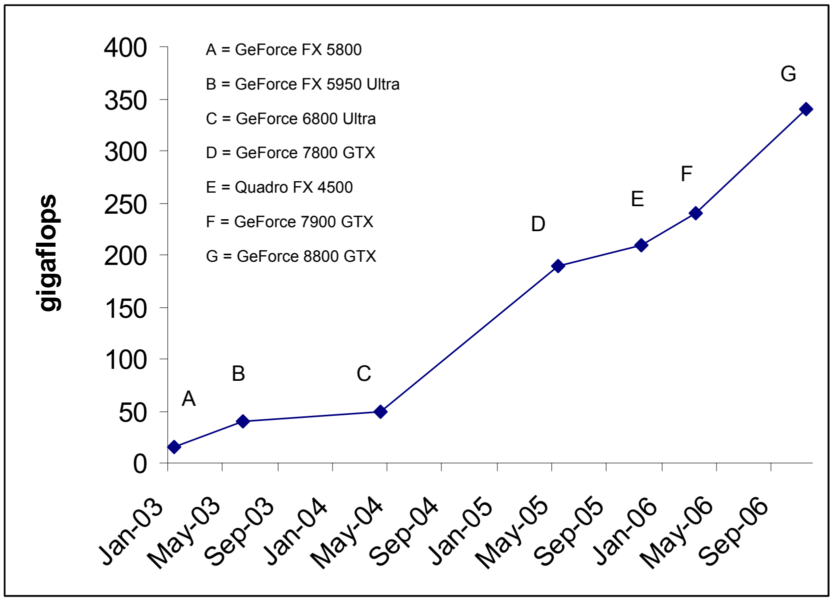


Figure 1-2. Floating-Point Operations/Second on the GPU

The main reason for this is that the GPU is architected for compute-intensive, highly parallel computation–exactly what is required for graphics rendering. Therefore on a GPU, more transistors are devoted to data processing than are to data caching and flow control

The GPU is especially well-suited to address problems that can be expressed as data-parallel computations with high arithmetic intensity–in other words when the same program is executed on many data elements in parallel with a high ratio of arithmetic to memory operations.

While processing with CPUs uses a single processing program to loop over data sequentially, 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 data-parallel programming model to speed up the computations. These applications include, for example:

- Seismic simulations
- Computational biology
- Option risk calculations in finance
- Medical imaging
- Pattern recognition
- Signal processing
- Physical simulation

Chapter 2.
NVIDIA® Tesla™ Solutions

2.1 GPU Computing Solution Set

NVIDIA is offering a complete line of GPU computing products, including system products and a new development environment. At the center of this line is the latest 8 Series GPU architecture from NVIDIA, coupled with the NVIDIA® CUDA™ Software Development Kit (SDK) and C compiler.

These different compute products offer a variety of compute power and density configurations that fulfill the spectrum of compute requirements. These solutions include:

- NVIDIA® Tesla™ [*] GPUs to bring high-performance computing to the desktop and workstation
- NVIDIA Tesla GPU Deskside Supercomputer that attaches to standard workstations to provide incredible computing density at the workstation
- NVIDIA Tesla GPU Server for very high GPU density in a 1U form factor
- NVIDIA CUDA development environment including FFT and BLAS libraries
- NVIDIA software developers kit with documentation and programming examples

2.2 Industry Standard Architecture

The NVIDIA compute solution is designed to fit seamlessly into existing IT infrastructures, relying on industry standards like:

- C compiler, providing a familiar and well-supported development environment.
- 128 independent IEEE 754 single-precision floating-point units with support for advanced floating-point features found in modern CPU floating-point units. (Double precision will be available in future versions.)

[*] Trademark pending

- Compatibility with x86 32-bit and 64-bit microprocessor architectures from Intel/AMD; Microsoft or Linux operating system.
- PCI Express bus architecture that delivers up to 4 GBps in both upstream and downstream data transfers.
- Standard industry form factors, for both desktop and rack-mounted configurations.
- NVIDIA unified driver architecture (UDA).

2.3 NVIDIA Tesla GPU

The new NVIDIA Tesla C870 GPUs are intended for use specifically in computing applications. GPU Computing board-level products do not have display connectors and are specifically designed for computing. Processor clocks, memory configuration and computing features will evolve differently than graphics board products.

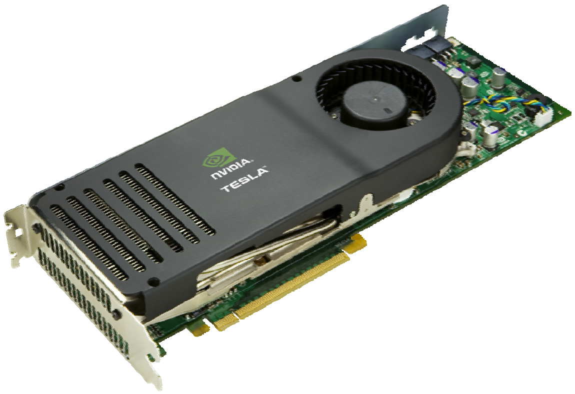


Figure 2-3: Tesla C870 GPU

The computing GPUs, while they lack the connectors to power the display, retain the full OpenGL and DirectX functionality of NVIDIA Quadro® graphics boards or NVIDIA® GeForce® GPUs, allowing them to power applications based on those APIs in addition to the CUDA SDK.

Tesla C870 GPU specifications:

- One GPU (128 thread processors)
- 518 gigaflops (peak)
- 1.5 GB dedicated memory
- Fits in one full-length, dual slot with one open PCI Express x16 slot

2.4 NVIDIA Tesla GPU Deskside Supercomputer

An NVIDIA Tesla GPU Deskside Supercomputer provides high computing density at the workstation and the flexibility of rack-mounted solutions. Each Tesla D870 provides dual GPUs for use alongside a workstation. Alternatively, two deskside systems can be rack-mounted for four Tesla GPUs in a 3U configuration.

Future versions of the deskside system will be able to provide up to four Tesla GPUs per system, or eight Tesla GPUs in a 3U rack mount.

Figure 2-4: Tesla Deskside Supercomputer

Tesla D870 specifications:

- ❑ Two GPUs (128 thread processors per GPU)
- ❑ 1.036 teraflops (peak)
- ❑ 3 GB system memory (1.5 GB dedicated memory per GPU)
- ❑ Quiet operation (40 dB; suitable for office environment)
- ❑ Connects to host via low-power PCI Express x8 or x16 adapter card
- ❑ Optional rack mount kit

2.5 NVIDIA Tesla GPU Server

For the ultimate in compute density, NVIDIA will be providing a 1U GPU computing server. With four to eight GPUs in a 1U form factor, GPU computing with the highest performance per volume per watt will be possible.

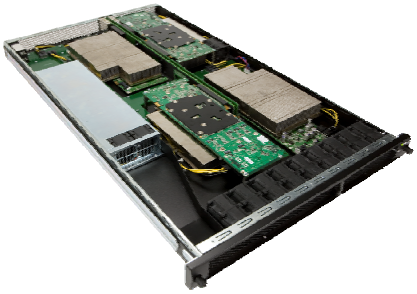


Figure 2-5: Tesla GPU Server

Tesla GPU server products are designed to optimize performance in a 1U rack mount volume at standard server power levels. Server products will be available in different configurations that offer maximum performance or lower power. The first server offering with four 8 Series GPUs dissipates a typical power of 550 Watts.

Initial Tesla GPU server specifications:

- Four GPUs (128 thread processors per GPU)
- 2.072 teraflops (peak)
- 6 GB of system memory (1.5 GB dedicated memory per GPU)
- Standard 19" (48.26 cm), 1U rack-mount chassis
- Connects to host via low-power PCI Express x8 or x16 adapter card
- Standard configuration: 2 PCI Express connectors driving two GPUs each (4 GPUs total)
- Optional configuration: 1 PCI Express connector driving four GPUs

2.6 NVIDIA CUDA GPU Computing Software

The NVIDIA CUDA technology is the new software architecture that exploits the parallel computational power of the GPU. When executing CUDA programs, the GPU operates as coprocessor to the main CPU. The GPU handles the core processing on large quantities of parallel information while the CPU organizes, interprets, and communicates information. Compute-intensive portions of applications that are executed many times, but on different data, are extracted from the main application and compiled to execute in parallel on the GPU.

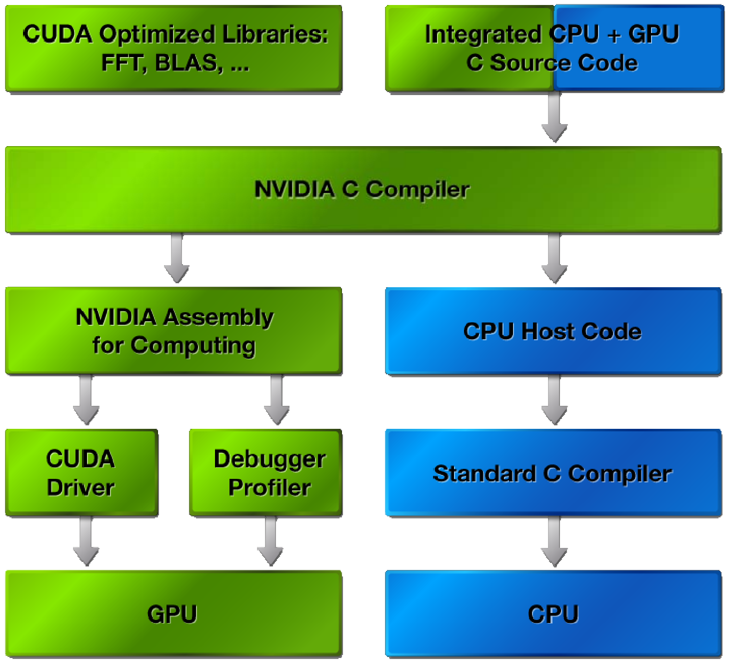

CUDA Optimized Libraries: FFT, BLAS, ...
Integrated CPU + GPU C Source Code
NVIDIA C Compiler
NVIDIA Assembly for Computing
CPU Host Code
CUDA Driver
Debugger Profiler
Standard C Compiler
GPU
CPU

Figure 2-6: CUDA Software Stack

CUDA includes three major components: new features on the 8 Series GPU to efficiently execute programs with parallel data; a C compiler to access the parallel computing resources on the GPU; and a runtime driver dedicated to computing.

The key to CUDA is the C compiler for the GPU. This first-of-its-kind programming environment simplifies coding parallel applications. Using C, a language familiar to most developers, allows programmers to focus on creating a parallel program instead of dealing with the complexities of graphics APIs. To simplify development, the CUDA C compiler lets programmers combine CPU and GPU code into one continuous program file. Simple additions to the C program tell the CUDA compiler which functions reside on the CPU and which to compile for the GPU. The program is then compiled with the CUDA compiler for the GPU, and then the CPU host code is compiled with the developer's standard C compiler.

Developers use a novel programming model to map parallel data problems to the GPU. CUDA programs divide information into smaller blocks of data that are processed in parallel. This programming model allows developers to code once for GPUs with more multiprocessors and for lower-cost GPUs with fewer multiprocessors.

When running a GPU computing program, developers simply run the program on the host CPU. The CUDA driver automatically loads and executes the programs on the GPU. The host-side program communicates with the GPU over the high-speed PCI Express bus. Data transfers, GPU computing function launches, and other interactions between the CPU and GPU are accomplished by calling special operations in the runtime driver. These high-level operations free developers from having to manually manage computing resources in the GPU.

The ability to interpret and manipulate massive amounts of information is the frontier of computing and with the widely available CUDA SDK any application can exploit the power of the GPU. With the combination of CUDA software and Tesla GPUs, developers now have the ability to bring the power of large-scale supercomputing to the desktop and dramatically increase the capability of server clusters.

Chapter 3.
Case Studies

3.1 GPU Computing Case Studies

The following are several examples of the power of GPU computing, high-performance computing applications in various industries that have been significantly accelerated using NVIDIA GPUs.

3.2 Medical Imaging: Digital Tomosynthesis

Digital tomosynthesis is a mammography technique that promises to make cancer lesions in breast tissue easier to see and earlier to detect. In this application, researchers at Massachusetts General Hospital used NVIDIA GPUs for the intense computations required to reconstruct images from the data provided by this technique.

To better visualize tumors and other abnormalities, tomosynthesis uses parallax, the effect in which a nearby object appears to move a greater distance against the background than does an object that is further away. Low-dose X-ray images are obtained from multiple angles while the breast remains motionless in the mammography system. Computers then electronically line up the images so that only the structures in a single plane of the breast are visible. The technique eliminates the overlapping structures that can obscure a cancer.

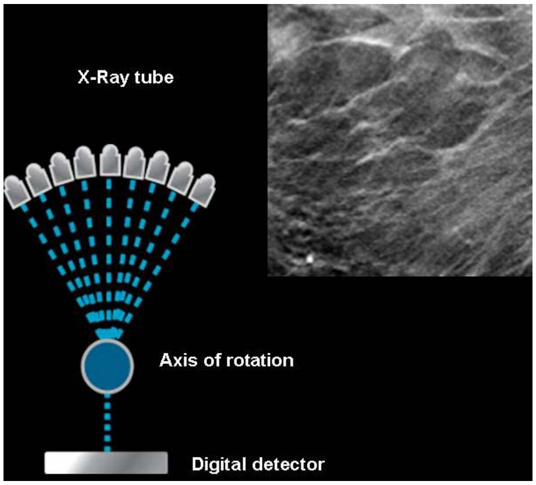


Figure 3-2: Digital Tomosynthesis

By analogy, traditional mammography is like reading a book with clear pages—all the letters are superimposed upon one another, producing a jumble that is almost impossible to read. But with tomosynthesis, the individual pages can be separated, producing images of "slices" of breast tissue only one millimeter thick.

Tomosynthesis has been a medical imaging concept since the 1960s, but it wasn't until the 1990s that digital detectors were sensitive enough to make it a reality. And even then, the processing power available was not up to the task. The first attempts to produce images from tomosynthesis data took a standard PC five hours to process a single patient's scans—too long for practical use. Attempts at using compute clusters of 34 PCs brought processing time down to 20 minutes, but clusters like this, while fine for research purposes, are not feasible for working hospital radiology labs.

Using NVIDIA GPUs in compute mode, the researchers at Massachusetts General have achieved a 100-fold speed-up in the reconstruction of the image, taking the computation time down from five hours to about five minutes on a single PC. This allows the radiologist to immediately review the image and the patient receives the results in the same visit. With fewer false positives, clearer images, and better visualization, digital breast tomosynthesis using NVIDIA compute technology will result in earlier detection of life-threatening cancers.

3.3 Simulation & Design: MRI-Safe Biomedical Implants

Boston Scientific designs and manufactures pacemakers and other biomedical implants that are safe for use with MRIs and other diagnostic imaging tools. The design simulations required for this task are compute-intensive and take considerable time to run on standard computer clusters.

Boston Scientific turned to Acceleware, which fields a proprietary simulation solution combining Schmid & Partner Engineering AG's SEMCAD X software and NVIDIA GPU computing technology. As a result, engineers at Boston Scientific were able to speed up their simulations by factors up to 25 times. Acceleware's solution allowed Boston Scientific to tap into the high-performance computing power of NVIDIA's parallel GPU architecture, dramatically reducing the time required to test various designs as each engineer's workstation now has the compute power of a cluster of CPUs.

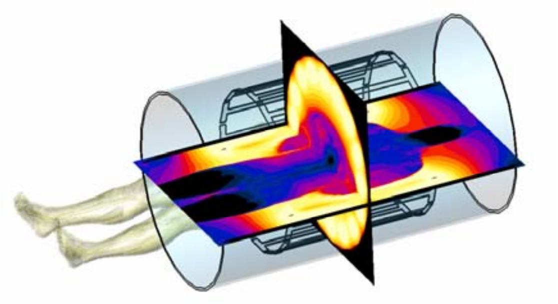


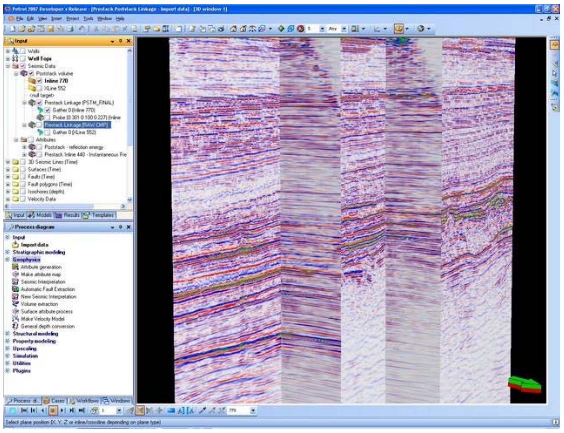
Figure 3-3: Boston Scientific Simulation: E-field distribution at 64 MHz

Using Acceleware's solution, engineers at Boston Scientific are better able to investigate the influence and mutual dependency of multiple design variables. Not only can the simulations be conducted faster, resulting in lower-cost devices, but these advances also foster the development of new algorithms for simulating biomedical processes inside the body.

3.4 Geosciences: Oil & Gas Exploration

Oil and gas deposits are becoming increasingly more difficult to find. Large reservoirs are now found at greater depths and in sediments that are much harder to analyze, like the recent Jack Field discovery in the Gulf of Mexico which was found at more than 20,000 feet under the sea floor. To interpret geologic data and discover these reservoirs it is necessary to acquire and process huge amounts of seismic data. And due to the complexity of the sediment layers, better resolution is needed in the images, which means acquiring even more data.

At a time when most people still thought of GPUs as consumer gaming technology, Houston-based Headwave, a company that specializes in geophysical data analysis, began developing a next-generation computing platform that could harness the parallel-processing power of the graphics card.



Select plane position (X, Y, Z or inline/crossline depending on plane type)

Figure 3-4: Headwave Analysis Program

Headwave solutions, implemented on NVIDIA GPUs using the CUDA SDK, allow geophysicists to apply advanced filters to their data and instantly see results even on multi-terabyte datasets. In addition, geophysicists can analyze the original acquired seismic ("pre-stack") data in multiple dimensions as part of their daily workflow.

Processing of terabyte datasets traditionally required months of manual preparation and more months of compute time. Headwave brings products to market that allow geophysicists to instantly see results even on multi-terabyte datasets. Using NVIDIA GPU computing technology, Headwave is able to reduce compute times and time spent in manual operations by 2000%.

Working with terabyte data sets in real time would not have been possible without the recent advances in NVIDIA's GPU computing technology. Oil and gas company workstations are already outfitted with GPUs, meaning that much of the hardware infrastructure to exploit this technology is already in place. As a result, oil and gas companies are poised to start taking advantage of this new technology almost immediately after its launch.

3.5 Computational Biology: Molecular Dynamics Simulation

The University of Illinois at Urbana-Champaign's (UIUC) Nanoscale Molecular Dynamics (NAMD) and Visual Molecular Dynamics (VMD) are powerful and widely used tools for simulating and visualizing biomolecular processes. But simulating complex molecular systems can be time consuming and require large, sophisticated clusters of computers.

To boost performance, the UIUC researchers ported "cionize" ion placement tool to an NVIDIA GPU computing solution. The goal was to use the GPU to accelerate the computationally intensive kernels for calculating the interaction of biological molecules and ions. In doing so, the UIUC researchers achieved speedups on ion simulations over 100 times that of an 18-CPU cluster (based on total CPU time v. total GPU time).

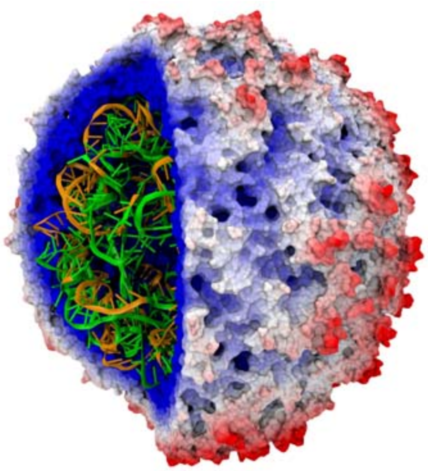


Figure 3-5: NAMD Molecular Simulation

With a three-GPU workstation, a similar calculation for time-averaged electrostatics in the VMD tool reaches 705 gigaflops of realized performance. This remarkable performance allows any bioscience researcher to have the equivalent of a computing cluster on their workstation.

With GPU computing, these molecular simulations are no longer restricted to clusters in server rooms. By running the simulations on workstations in individual labs and desktops, projects are no longer competing with one another for scarce computing resources and the researchers are getting the results when they need them, as opposed to when they can be scheduled.

Furthermore, with GPUs in large-scale server clusters new classes of problems can be addressed for which the necessary computing power was only a dream a year ago.

The combination of NAMD and NVIDIA compute technology is a marriage of cutting-edge research and software development, aimed at harnessing the nation's fastest supercomputers to decipher the tiniest components of living cells. These new computing tools are quickening the pace of drug discovery and other vital research in unraveling biological processes.

3.6 Scientific Computing: MathWorks MATLAB

MATLAB is a high-level programming language especially well suited to rapidly coding programs for scientific and mathematical algorithms. MATLAB is the language of choice for many application areas, including university-based research, signal and image processing, test and measurement, financial modeling and analysis, and computational biology.

MATLAB applications can be accelerated by the NVIDIA GPU using two methods. The first method does not require changes to the MATLAB code. By simply plugging in the CUDA FFT libraries underneath the MATLAB application, any calls to FFT or BLAS functions are simply executed on the GPU. For single precision applications, speed ups of up to three times have been observed on example research code. For greater acceleration of MATLAB applications, the CUDA MATLAB plug-in allows programmers to replace certain critical functions with optimized CUDA programs. The new CUDA function is then called by the original MATLAB application, decreasing the execution time. By replacing just the critical functions with CUDA functions, MATLAB users have an easier path to faster applications without having to rewrite an entire application in another programming language.

As an example, NVIDIA ported a MATLAB application found in published research, running an order of magnitude faster on CUDA than on a CPU. The algorithm, for pseudo-spectral simulation of 2D isotropic turbulence, went from 992 seconds on a CPU to 93 seconds using NVIDIA compute architecture, the new MATLAB plug-in, and a GPU.

3.7 Neural Circuit Simulation: Evolved Machines

Evolved Machines is reverse engineering brain circuits to develop a new paradigm for device technology. Their research work requires the large-scale simulation of biologically realistic neural circuits which require enormous parallel computing capacity. Simulation of a single neuron involves evaluation of 200 million differential equations per second, requiring approximately four gigaflops. A neural array engaged in sensory processing requires thousands of neurons and the detailed simulation of neural systems in real time requires more than ten

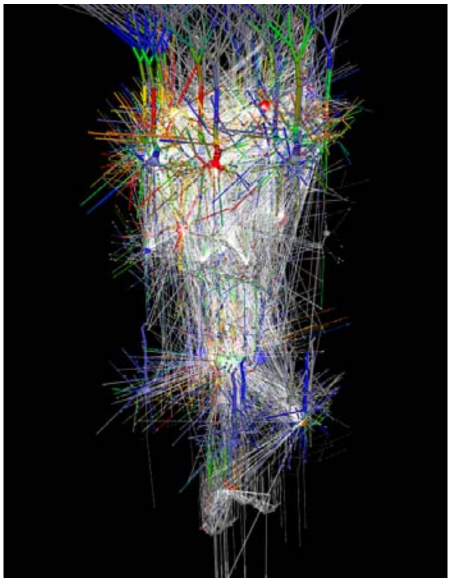


Figure 3-7: Neural Array Simulation

teraflops of computing power.

Evolved Machines started working with NVIDIA GPUs in September 2006. They achieved accelerations of approximately 130-fold against simulations on current-generation x86 microprocessors. They are now engaged in the design of a rack of GPUs which will rival the world's top systems, at 1/100 of the cost.

Applications under active development at Evolved Machines include visual object recognition and odorant recognition. To develop devices which can learn the characteristics of objects and then recognize those objects in real-world environments, the synthetic neural circuitry is allowed to gradually "wire" itself during exposure to sensory input—much as a baby learns to recognize objects in its environment over the first months of life.

With the GPU, devices with the ability to learn and sense odors to detect explosives in real environments or to monitor food for freshness become possible. Advances in robotic image detection can also exploit neural simulation processing to achieve a level of capability not currently possible.

Chapter 4.
Where to Go for More Information

4.1 NVIDIA Web Sites

Public web site:

http://www.nvidia.com

Press web site (login required):

http://newsroom.nvidia.com

NVIDIA developer site

http://developer.nvidia.com

NVIDIA developer forums including CUDA support forums

http://forums.nvidia.com

4.2 PR Contact

Andrew Humber
Senior PR Manager, Mobile & GPU Computing Products
NVIDIA Corporation
2701 San Tomas Expressway
Santa Clara, CA, USA 95050

T: +1 408 486 8138

M +1 408 416 7943

ahumber@nvidia.com

4.3 Case Study Information

For more information on the case studies, see the following:

- Improved Detection with Digital Breast Tomosynthesis, J. Daniel Janzen, *Scientific Computing*, March 2007.
- Headwave corporate web site
- NAMD Scalable Molecular Dynamics, University of Illinois at Urbana-Champaign
- Evolved Machines corporate web site

Notice

ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE.

Information furnished is believed to be accurate and reliable. However, NVIDIA Corporation assumes no responsibility for the consequences of use of such information or for any infringement of patents or other rights of third parties that may result from its use. No license is granted by implication or otherwise under any patent or patent rights of NVIDIA Corporation. Specifications mentioned in this publication are subject to change without notice. This publication supersedes and replaces all information previously supplied. NVIDIA Corporation products are not authorized for use as critical components in life support devices or systems without express written approval of NVIDIA Corporation.

Trademarks

Copyright