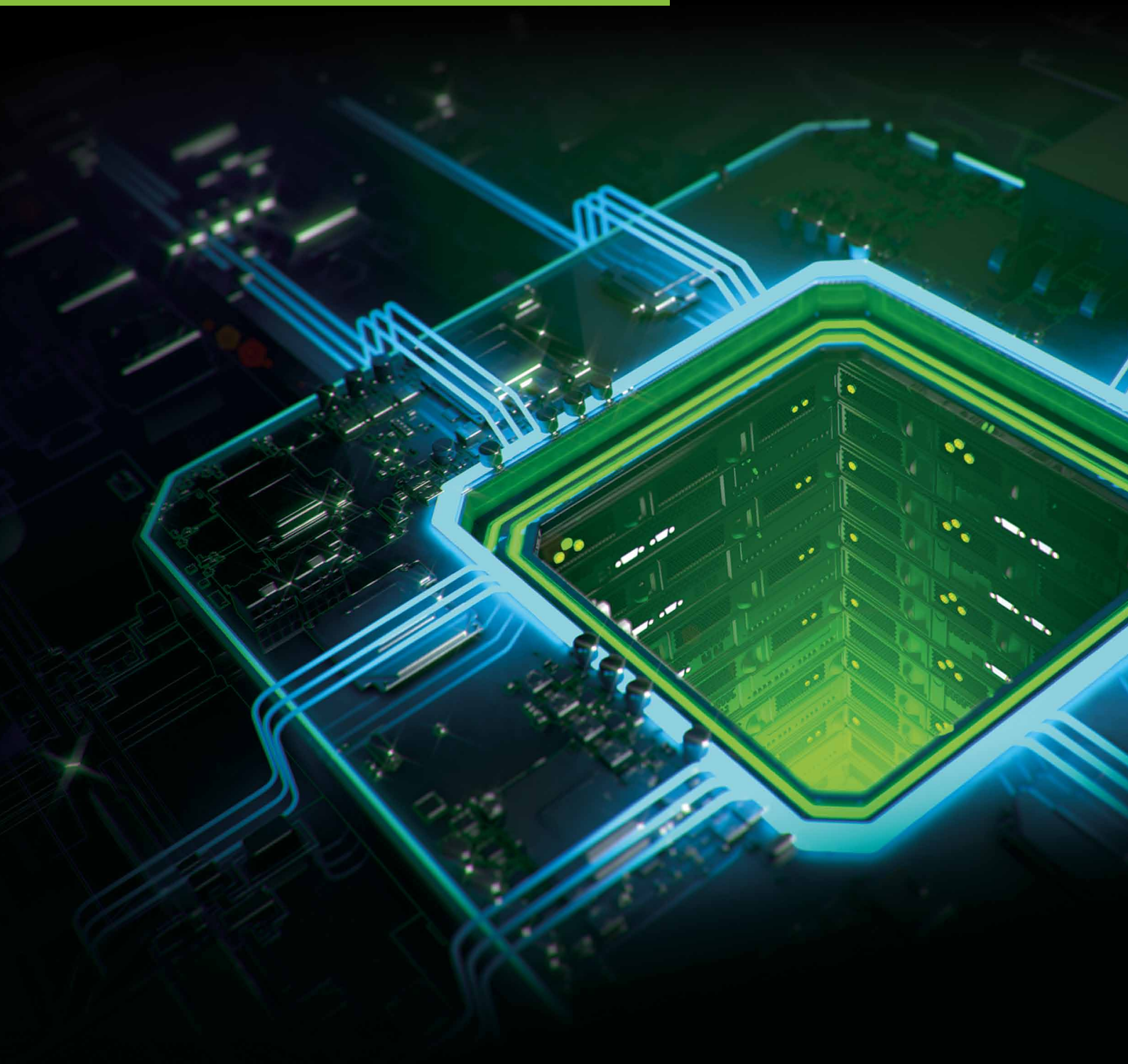
NVIDIA

REVOLUTIONIZING HIGH PERFORMANCE COMPUTING

NVIDIA® TESLA™



REVOLUTIONIZING HIGH PERFORMANCE COMPUTING NVIDIA® TESLA™

The high performance computing (HPC) industry's need for computation is increasing, as large and complex computational problems become commonplace across many industry segments. Traditional CPU technology, however, is no longer capable of scaling in performance sufficiently to address this demand.

TABLE OF CONTENTS

NVIDIA® TESLA™ GPU COMPUTING

REVOLUTIONIZING HIGH PERFORMANCE COMPUTING

"We not only created the world's fastest computer, but also implemented a heterogeneous computing architecture incorporating CPU and GPU, this is a new innovation."

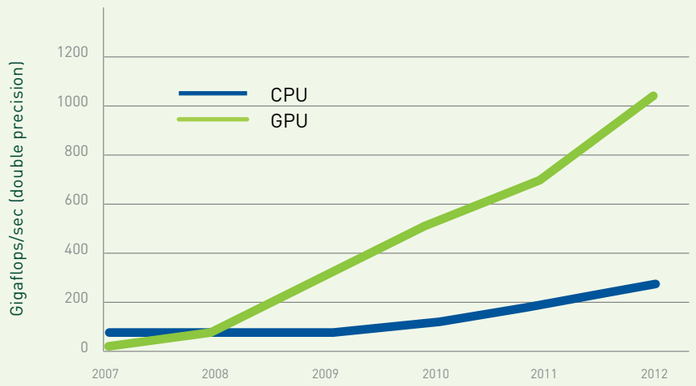
Premier Wen Jiabao
People's Republic of China

The high performance computing (HPC) industry's need for computation is increasing, as large and complex computational problems become commonplace across many industry segments. Traditional CPU technology, however, is no longer capable of scaling in performance sufficiently to address this demand.

The parallel processing capability of the GPU allows it to divide complex computing tasks into thousands of smaller tasks that can be run concurrently. This ability is enabling computational scientists and researchers to address some of the world's most challenging computational problems up to several orders of magnitude faster.

GPUs DELIVER SUPERIOR PERFORMANCE IN PARALLEL COMPUTATION.

GPUs deliver superior performance in parallel computation.



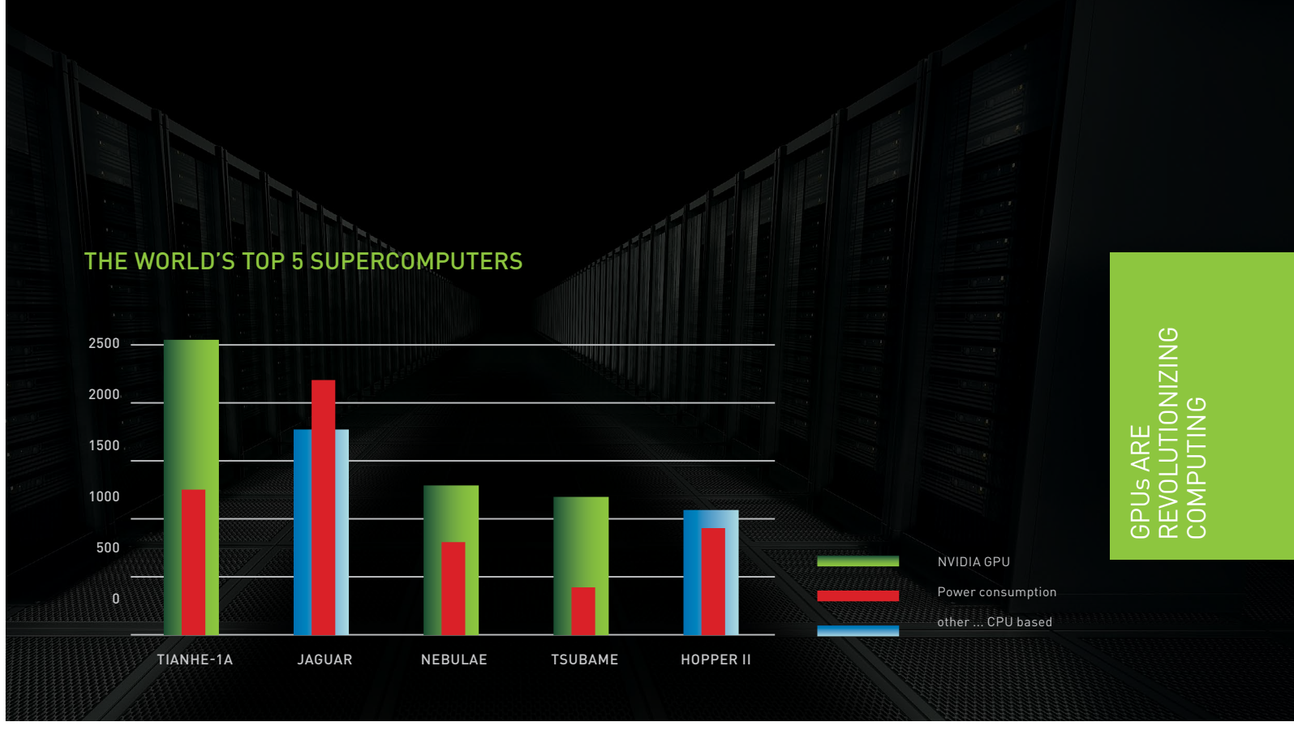
"The rise of GPU supercomputers on the Green500 signifies that heterogeneous systems, built with both GPUs and CPUs, deliver the highest performance and unprecedented energy efficiency,"

said Wu-chun Feng,
founder of the Green500 and
associate professor of
Computer Science at Virginia Tech.

This advancement represents a dramatic shift in HPC. In addition to dramatic improvements in speed, GPUs also consume less power than conventional CPU-only clusters. GPUs deliver performance increases of 10x to 100x to solve problems in minutes instead of hours–while outpacing the performance of traditional computing with x86-based CPUs alone.

From climate modeling to advances in medical tomography, NVIDIA® Tesla™ GPUs are enabling a wide variety of segments in science and industry to progress in ways that were previously impractical, or even impossible, due to technological limitations.

Figure 1: Co-processing refers to the use of an accelerator, such as a GPU, to offload the CPU to increase computational efficiency.



"I believe history will record Fermi as a significant milestone."

Dave Patterson, Director, Parallel Computing Research Laboratory, U.C. Berkeley Co-author of Computer Architecture: A Quantitative Approach

WHY GPU COMPUTING?

With the ever-increasing demand for more computing performance, systems based on CPUs alone can no longer keep up. The CPU-only systems can only get faster by adding thousands of individual computers – this method consumes too much power and makes supercomputers very expensive. A different strategy is parallel computing, and the HPC industry is moving toward a hybrid computing model, where GPUs and CPUs work together to perform general purpose computing tasks.

As parallel processors, GPUs excel at tackling large amounts of similar data because the problem can be split into hundreds or thousands of pieces and calculated simultaneously.

GPU SUPERCOMPUTING - GREEN HPC

GPUs significantly increase overall system efficiency as measured by performance per watt. "Top500" supercomputers based on heterogeneous architectures are, on average, almost three times more power-efficient than non-heterogeneous systems. This is also reflected on Green500 list – the icon of Eco-friendly supercomputing.

CORE COMPARISON BETWEEN A CPU AND A GPU

As sequential processors, CPUs are not designed for this type of computation, but they are adept at more serial based tasks such as running operating systems and organizing data. NVIDIA believes in applying the most relevant processor to the specific task in hand.

Figure 2: The new computing model includes both a multi-core CPU and a GPU with hundreds of cores.

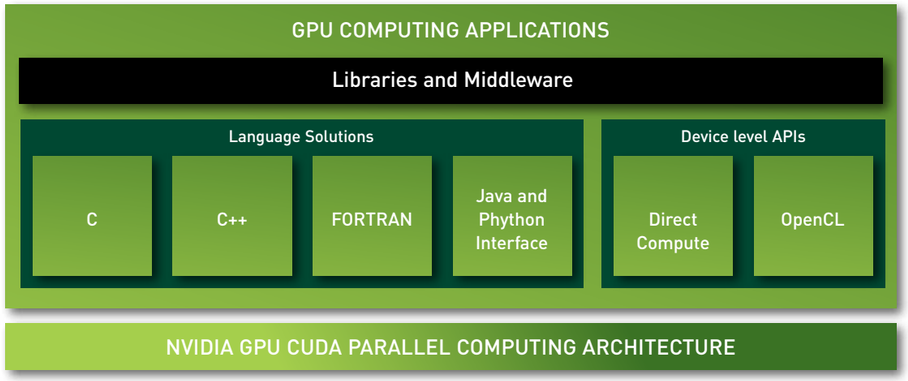
CPU GPU

CUDA PARALLEL COMPUTING ARCHITECTURE

CUDA® is NVIDIA's parallel computing architecture and enables applications to run large, parallel workloads on NVIDIA GPUs. Applications that leverage the CUDA architecture can be developed in a variety of languages and APIs, including C, C++, Fortran, OpenCL, and DirectCompute. The CUDA architecture contains hundreds of cores capable of running many thousands of parallel threads, while the CUDA programming model lets programmers focus on parallelizing their algorithms and not the mechanics of the language.

The current generation CUDA architecture, codenamed "Fermi", is the most advanced GPU computing architecture ever built. With over three billion transistors, Fermi is making GPU and CPU co-processing pervasive by addressing the full-spectrum of computing applications. With support for C++, GPUs based on the Fermi architecture make parallel processing easier and accelerate performance on a wider array of applications than ever before.

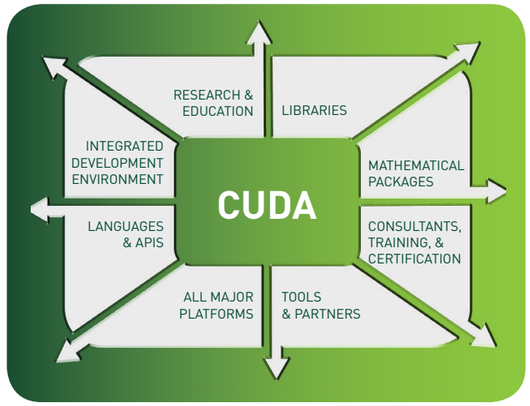
Just a few applications that can experience significant performance benefits include ray tracing, finite element analysis, high-precision scientific computing, sparse linear algebra, sorting, and search algorithms.



The CUDA parallel computing architecture, with a combination of hardware and software.

DEVELOPER ECOSYSTEM

In just a few years, an entire software ecosystem has developed around the CUDA architecture – from more than 400 universities worldwide teaching the CUDA programming model, to a wide range of libraries, compilers, and middleware that help users optimize applications for GPUs. This rich ecosystem has led to faster discovery and simulation in a wide range of fields including mathematics, life sciences, and manufacturing.



Locate a CUDA teaching center near you:
research.nvidia.com/content/cuda-courses-map

Order a custom on-site CUDA education course at:
www.parallel-compute.com

Microsoft® Visual Studio®

NVIDIA PARALLEL NSIGHT DEVELOPMENT ENVIRONMENT FOR VISUAL STUDIO

NVIDIA Parallel Nsight software is the industry's first development environment for massively parallel computing integrated into Microsoft Visual Studio, the world's most popular development environment for Windows-based applications and services. It integrates CPU and GPU development, allowing developers to create optimal GPU-accelerated applications.

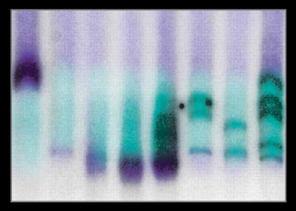
Parallel Nsight supports Microsoft Visual Studio 2010, advanced debugging and analysis capabilities, as well as Tesla 20-series GPUs.

For more information, visit: developer.nvidia.com/object/nsight.html.

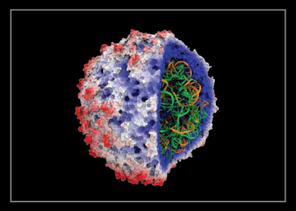
Request hands-on Parallel Nsight training:
www.parallel-compute.com

NVIDIA Parallel Nsight software integrates CPU and GPU development, allowing developers to create optimal GPU-accelerated applications.

100X
Astrophysics
RIKEN

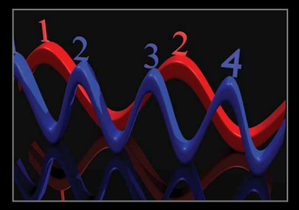


30X
Gene Sequencing
University of Maryland

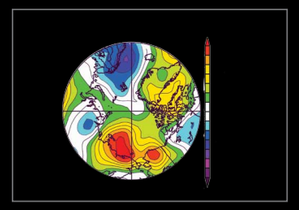


36x
Molecular Dynamics
University of Illinois,
Urbana-Champaign

"NVIDIA's Tesla C2050 Fermi card includes almost everything what scientists ever wanted" c't magazine (Germany)

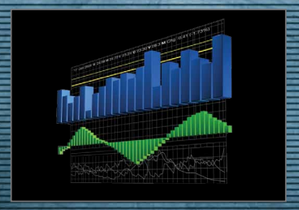


50x
MATLAB Computing,
AccelerEyes



80x
Weather Modeling
Tokyo Institute of
Technology

146x
Medical Imaging
U of Utah



149x
Financial Simulation
Oxford University

MATLAB ACCELERATIONS ON TESLA GPUs

MATLAB PERFORMANCE WITH TESLA

MATLAB® & SIMULINK®

MathWorks®

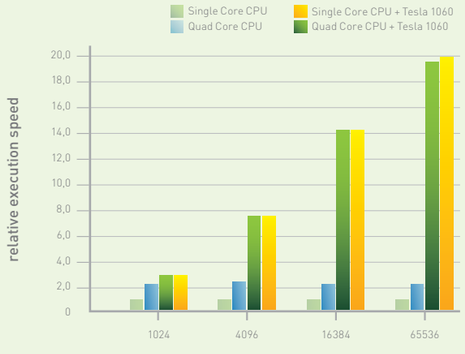
NVIDIA and MathWorks have collaborated to deliver the power of GPU computing for MATLAB users. Available through the latest release of MATLAB 2010b, NVIDIA GPU acceleration enables faster results for users of the Parallel Computing Toolbox and MATLAB Distributed Computing Server. MATLAB supports NVIDIA® CUDA™-enabled GPUs with compute capability version 1.3 or higher, such as Tesla™ 10-series and 20-series GPUs. MATLAB CUDA support provides the base for GPU-accelerated MATLAB operations and lets you integrate your existing CUDA kernels into MATLAB applications.

The latest release of Parallel Computing Toolbox and MATLAB Distributed Computing Server takes advantage of the CUDA parallel computing architecture to provide users the ability to

- Manipulate data on NVIDIA GPUs
- Perform GPU accelerated MATLAB operations
- Integrate users own CUDA kernels into MATLAB applications
- Compute across multiple NVIDIA GPUs by running multiple MATLAB workers with Parallel Computing Toolbox on the desktop and MATLAB Distributed Computing Server on a compute cluster

RELATIVE PERFORMANCE, POINT-IN-POLYGON DEMO

Compared to Single Core CPU Baseline



Core 2 Quad Q6600 2.4 GHz, 6 GB RAM, Windows 7, 64-bit, Tesla C1060, single precision operations.

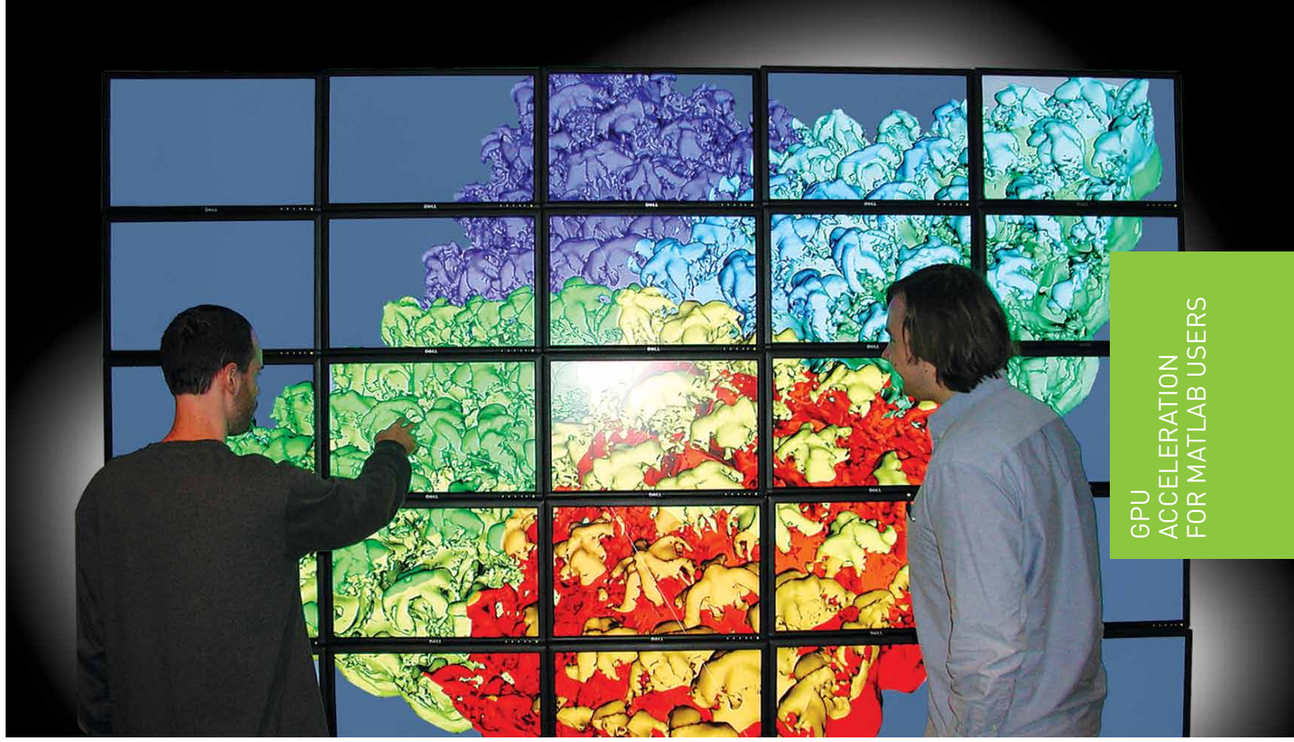
http://www.mathworks.com/products/distriben/demos.html?file=/products/demos/distribtb/MapDemo/MapDemo.html

TESLA BENEFITS

Highest Computational Performance
- High-speed double precision operations
- Large dedicated memory
- High-speed bi-directional PCIe communication
- NVIDIA GPUDirect™ with InfiniBand

Most Reliable
- ECC memory
- Rigorous stress testing

Best Supported
- Professional support network
- OEM system integration
- Long-term product lifecycle
- 3 year warranty
- Cluster & system management tools (server products)
- Windows remote desktop support



GPU ACCELERATION FOR MATLAB USERS

Image - courtesy of the Maryland CPU-GPU Cluster team.

ACCELEREYES®

GPU COMPUTING IN MATLAB WITH ACCELEREYES JACKET

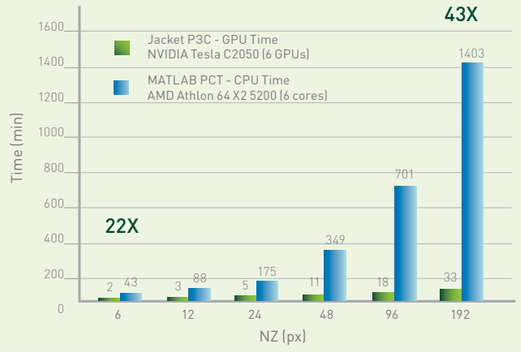
Jacket includes many key features to deliver results on full applications:

- Over 500 functions, including math, signal processing, image processing, and statistics
- Specialized FOR loops to run many iterations simultaneously
- An optimized runtime to optimize memory bandwidth and kernel configurations
- Integrate users' own CUDA kernels into MATLAB via the Jacket SDK
- Compute across multiple NVIDIA GPUs via Jacket MGL and HPC

With Jacket programming, the MATLAB community can enjoy GPU-acceleration with an easy, high-level interface.

POTENTIAL FIELD EXTRAPOLATION ON MAGNETOGRAM

Size: NX=994px, NY=484px



RECOMMENDED TESLA & QUADRO CONFIGURATIONS

High-End Workstation
- Two Tesla C2050 or C2070 GPUs
- Quadro NVS 295
- Two quad-core CPUs
- 12 GB system memory

Mid-Range Workstation
- Tesla C2050 or C2070 GPU
- Quad-core CPU
- 8 GB system memory

Entry Workstation
- Quadro 4000 GPU
- Single quad-core CPU
- 4 GB system memory

ENABLING NEW SCIENCE

TESLA™ BIO WORKBENCH

The NVIDIA Tesla Bio Workbench enables biophysicists and computational chemists to push the boundaries of life sciences research. It turns a standard PC into a "computational laboratory" capable of running complex bioscience codes, in fields such as drug discovery and DNA sequencing, more than 10-20 times faster through the use of NVIDIA Tesla GPUs.

It consists of bioscience applications; a community site for downloading, discussing, and viewing the results of these applications; and GPU-based platforms.

Complex molecular simulations that had been only possible using supercomputing resources can now be run on an individual workstation, optimizing the scientific workflow and accelerating the pace of research. These simulations can also be scaled up to GPU-based clusters of servers to simulate large molecules and systems that would have otherwise required a supercomputer.

Applications that are accelerated on GPUs include:

- **Molecular Dynamics & Quantum Chemistry**
  AMBER, GROMACS, HOOMD, LAMMPS, NAMD, TeraChem (Quantum Chemistry), VMD
- **Bio Informatics**
  CUDA-BLASTP, CUDA-EC, CUDA-MEME, CUDASW++ (Smith-Waterman), GPU-HMMER, MUMmerGPU

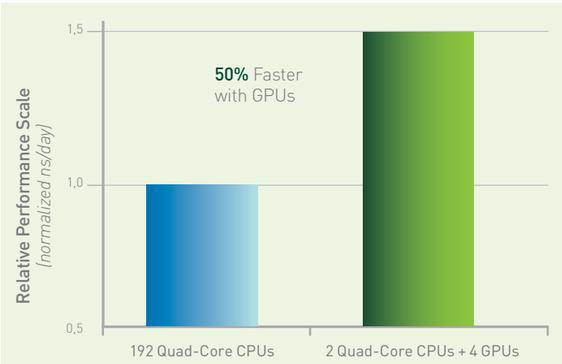
For more information, visit:
www.nvidia.com/bio_workbench

LEVERAGE SUPERCOMPUTER-LIKE PERFORMANCE FOR YOUR AMBER RESEARCH WITH TESLA GPUs

AMBER

Researchers today are solving the world's most challenging and important problems. From cancer research to drugs for AIDS, computational research is bottlenecked by simulation cycles per day. More simulations mean faster time to discovery. To tackle these difficult challenges, researchers frequently rely on national supercomputers for computer simulations of their models.

GPUs offer every researcher supercomputer-like performance in their own office. Benchmarks have shown four Tesla M2090 GPUs significantly outperforming the existing world record on CPU-only supercomputers.

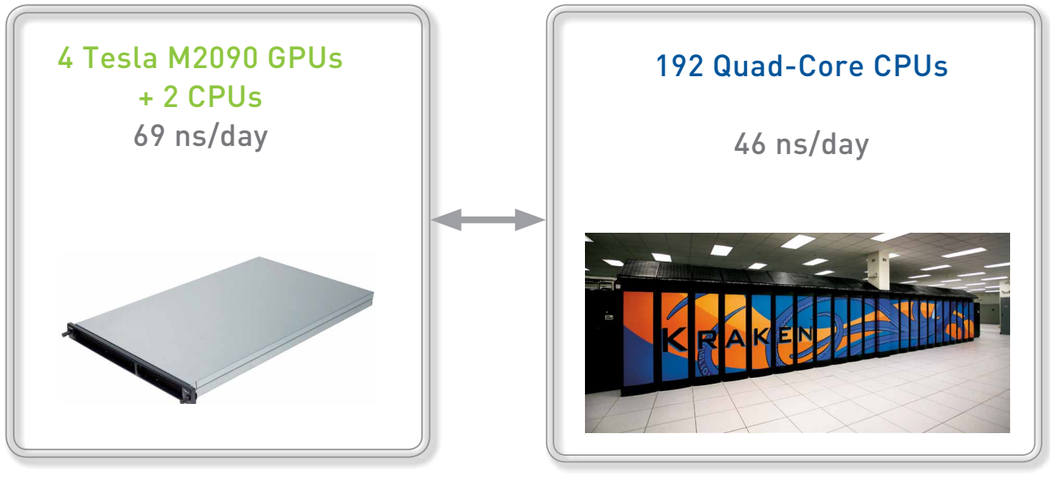


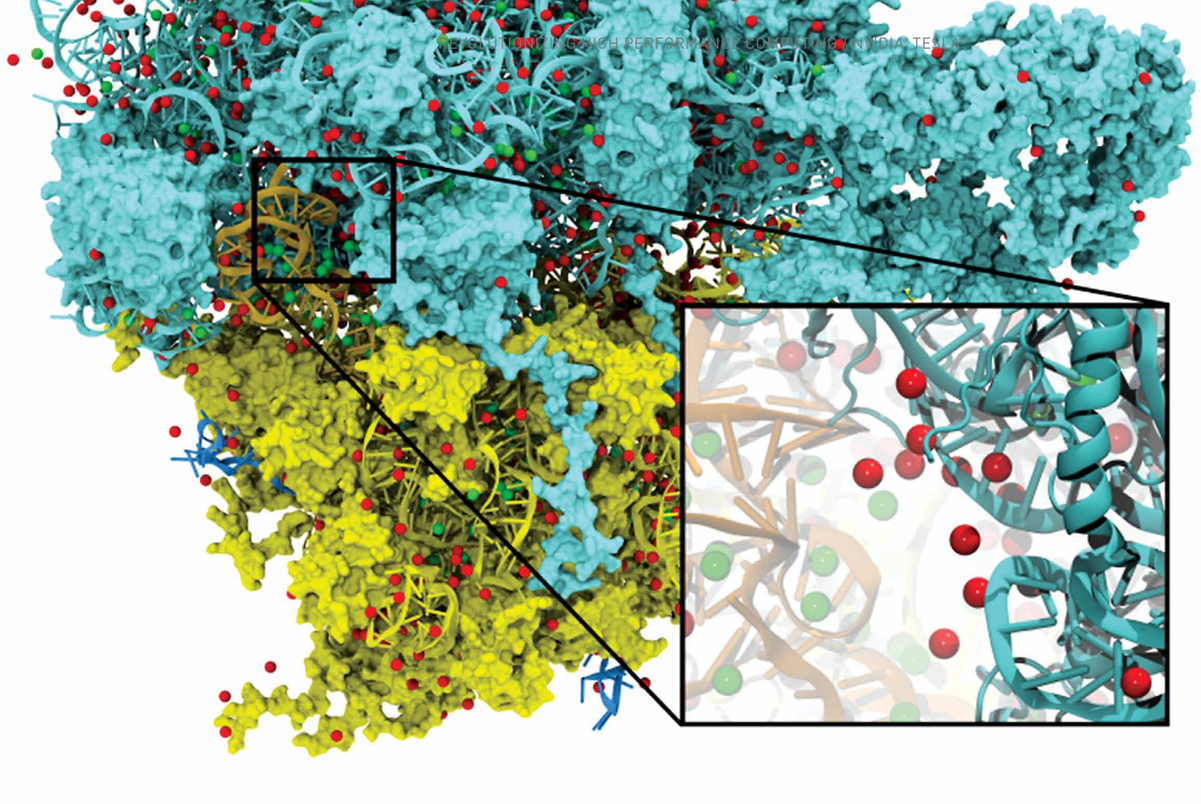
JAC NVE Benchmark
(left) 192 Quad-Core CPUs simulation run on Kraken Supercomputer
(right) Simulation 2 Intel Xeon Quad-Core CPUs and 4 Tesla M2090 GPUs

RECOMMENDED HARDWARE CONFIGURATION

Workstation
- 4xTesla C2070
- Dual-socket Quad-core CPU
- 24 GB System Memory Server
- Up to 8x Tesla M2090s in cluster
- Dual-socket Quad-core CPU per node
- 128 GB System Memory

Server
- 8x Tesla M2090
- Dual-socket Quad-core CPU
- 128 GB System Memory





CUDA-BLASTP

CUDA-BLASTP is designed to accelerate NCBI BLAST for scanning protein sequence databases by taking advantage of the massively parallel CUDA architecture of NVIDIA Tesla GPUs. CUDA-BLASTP also has a utility to convert FASTA format database into files readable by CUDA-BLASTP.

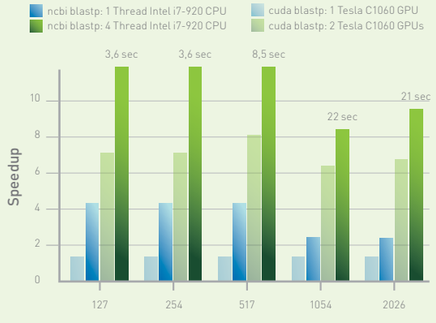
GPU-HMMER

GPU-HMMER is a bioinformatics software that does protein sequence alignment using profile HMMs by taking advantage of the massively parallel CUDA architecture of NVIDIA Tesla GPUs. GPU-HMMER is 60-100x faster than HMMER (2.0).

CUDA-BLASTP running on a workstation with two Tesla C1060 GPUs is 10x faster than NCBI BLAST (2.2.22) running on an Intel i7-920 CPU. This cuts compute time from minutes of CPUs to seconds using GPUs.

CUDA-BLASTP vs NCBI BLASTP Speedups

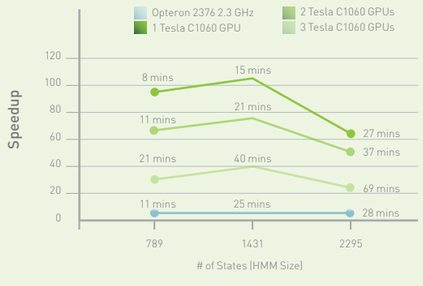
Data courtesy of Nanyang Technological University, Singapore



GPU-HMMER accelerates the hmmsearch tool using GPUs and gets speed-ups ranging from 60-100x. GPU-HMMER can take advantage of multiple Tesla GPUs in a workstation to reduce the search from hours on a CPU to minutes using a GPU.

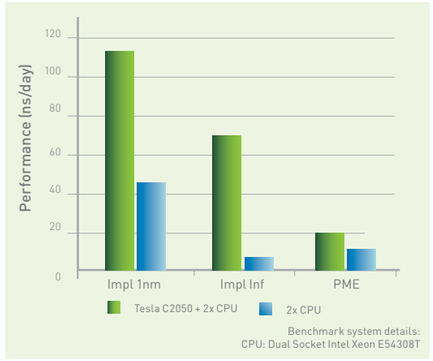
HMMER: 60-100X FASTER

Data courtesy of Nanyang Technological University, Singapore



GROMACS

GROMACS is a molecular dynamics package designed primarily for simulation of biochemical molecules like proteins, lipids, and nucleic acids that have a lot complicated bonded interactions. The CUDA port of GROMACS enabling GPU acceleration supports Particle-Mesh-Ewald (PME), arbitrary forms of non-bonded interactions, and implicit solvent Generalized Born methods.



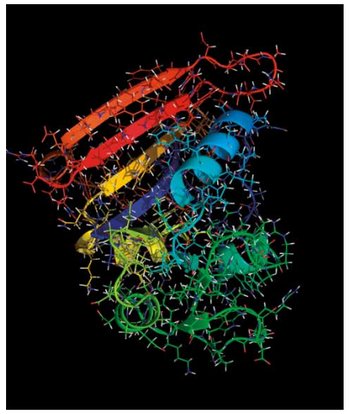
RECOMMENDED HARDWARE CONFIGURATION

Workstation

- 1xTesla C2070
- Dual-socket Quad-core CPU
- 12 GB System Memory

CUDA-MEME

CUDA-MEME is a motif discovery software based on MEME (version 3.5.4). It accelerates MEME by taking advantage of the massively parallel CUDA architecture of NVIDIA Tesla GPUs. It supports the OOPS and ZOOPS models in MEME.



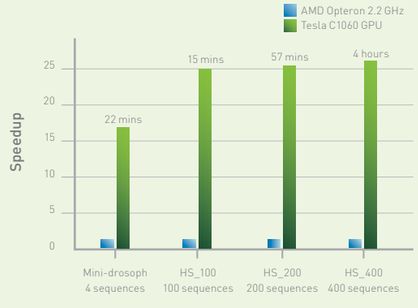
CUDA-EC

CUDA-EC is a fast parallel sequence error correction tool for short reads. It corrects sequencing errors in high-throughput short-read (HTSR) data and accelerates HTSR by taking advantage of the massively parallel CUDA architecture of NVIDIA Tesla GPUs. Error correction is a preprocessing step for many DNA fragment assembly tools and is very useful for the new high-throughput sequencing machines.

CUDA-MEME running on one Tesla C1060 GPU is up to 23x faster than MEME running on a x86 CPU. This cuts compute time from hours on CPUs to minutes using GPUs. The data in the chart below are for the OOPS (one occurrence per sequence) model for 4 datasets.

CUDA-MEME VS MEME SPEEDUPS

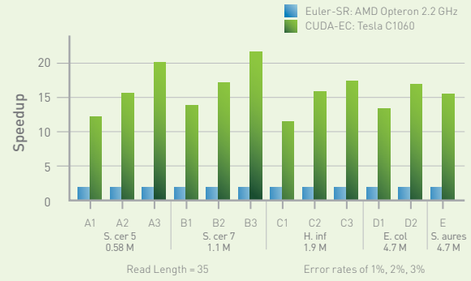
Data courtesy of Nanyang Technological University, Singapore



CUDA-EC running on one Tesla C1060 GPU is up to 20x faster than Euler-SR running on a x86 CPU. This cuts compute time from minutes on CPUs to seconds using GPUs. The data in the chart below are error rates of 1%, 2%, and 3% denoted by A1, A2, A3 and so on for 5 different reference genomes.

CUDA-EC vs Euler-SR Speedups

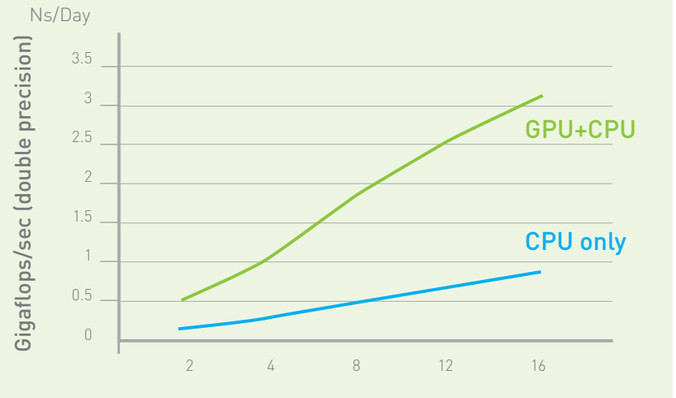
Data courtesy of Nanyang Technological University, Singapore



NAMD

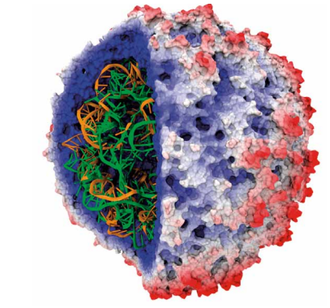
The Team at University of Illinois at Urbana-Champaign (UIUC) has been enabling CUDA-acceleration on NAMD since 2007, and the results are simply stunning. NAMD users are experiencing tremendous speed-ups in their research using Tesla GPU. Benchmark (see below) shows that 4 GPU server nodes out-perform 16 CPU server nodes. It also shows GPUs scale-out better than CPUs with more nodes.

Scientists and researchers equipped with powerful GPU accelerators have reached new discoveries which were impossible to find before. See how other computational researchers are experiencing supercomputer-like performance in a small cluster, and take your research to new heights.



NAMD 2.8 B1, STMV Benchmark
A Node is Dual-Socket, Quad-core x5650 with 2 Tesla M2070 GPUs

NAMD courtesy of Theoretical and Computational Bio-physics Group, UIUC

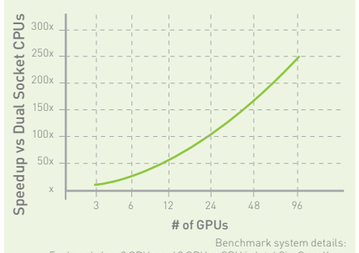


LAMMPS

LAMMPS is a classical molecular dynamics package written to run well on parallel machines and is maintained and distributed by Sandia National Laboratories in the USA. It is a free, open-source code.

LAMMPS has potentials for soft materials (biomolecules, polymers) and solid-state materials (metals, semiconductors) and coarse-grained or mesoscopic systems.

The CUDA version of LAMMPS is accelerated by moving the force calculations to the GPU.



Benchmark system details:
Each node has 2 CPUs and 3 GPUs. CPU is Intel Six-Core Xeon
Speedup measured with Tesla M2070 GPU vs without GPU
Source: http://users.nccs.gov/~wb8/gpu/kid_single.htm

RECOMMENDED HARDWARE CONFIGURATION

Workstation

- 4xTesla C2070
- Dual-socket Quad-core CPU
- 24 GB System Memory

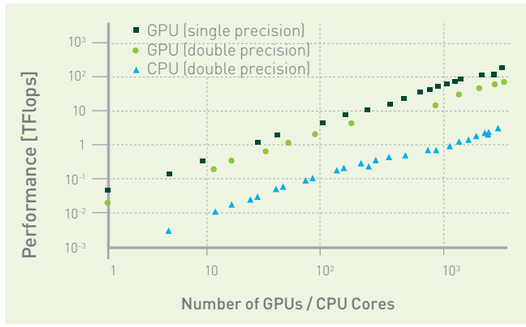
Server

- 2x-4x Tesla M2090 per node
- Dual-socket Quad-core CPU per node
- 128 GB System Memory per node

ASUCA (WEATHER MODELING) JAPAN'S TERASCALE WEATHER SIMULATION

Regional weather forecasting demands fast simulation over fine-grained grids, resulting in extremely memory-bottlenecked computation. ASUCA is the first high- resolution weather prediction model ported fully to CUDA.

ASUCA is a next-generation, production weather code developed by the Japan Meteorological Agency, similar to WRF in the underlying physics (non-hydrostatic model).

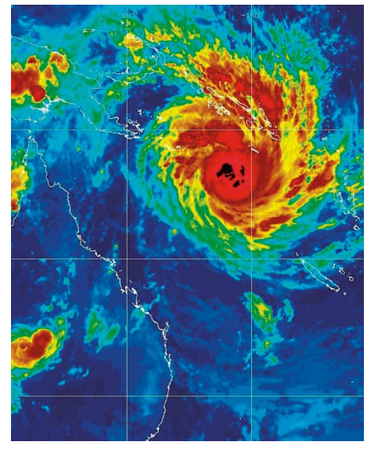


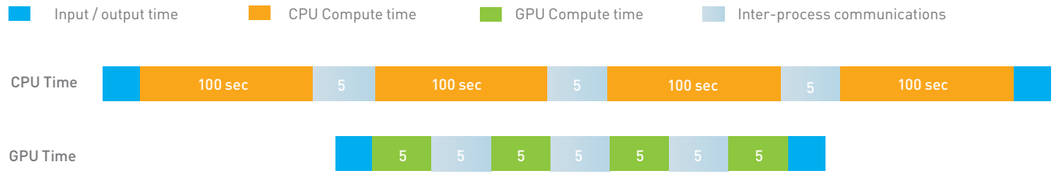
NOAA NIM EXPLORING GPU COMPUTING TO REFINE WEATHER FORECASTING

Earth System Research Lab in the National Oceanic & Atmospheric Administration (NOAA) of the United States has developed a next generation global model to more accurately and efficiently forecast weather. The Non-hydrostatic Icosahedral Model (NIM) uses the icosahedral horizontal grid and is designed to run on thousands of processors including Tesla GPUs.

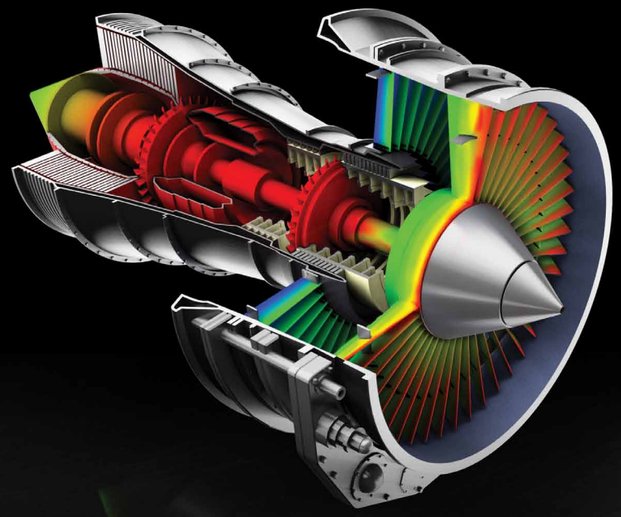
NIM Dynamics package has been ported over to CUDA for single GPU implementation. NOAA is actively working on the code to run parallel jobs on multiple GPUs.

The benchmark below shows ~20x improvement in computation time by using GPUs compared to a 1 core Nehalem Intel CPU. Due to significant computational speedup with GPUs, inter-process communication now becomes the bottleneck in simulation.





source: http://www.esrl.noaa.gov/research/events/espc/8sep2010/Govett_ESRL_GPU.pdf



ANSYS®

"A new feature in ANSYS Mechanical leverages graphics processing units to significantly lower solution times for large analysis problem sizes."

By Jeff Beisheim,
Senior Software Developer, ANSYS, Inc.

SPEED UP ANSYS SIMULATIONS WITH A GPU

With ANSYS® Mechanical™ 13.0 and NVIDIA® Tesla™ GPUs, you can:

- Improve product quality with 2x more design simulations
- Accelerate time-to-market by reducing engineering cycles
- Make high fidelity models easier to solve

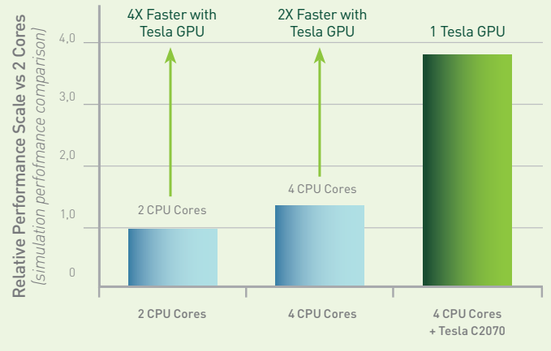
RECOMMENDED TESLA CONFIGURATIONS

Workstation
- Tesla C2070
- Dual-socket Quad-core CPU
- 48 GB System Memory

Server
- 2x Tesla M2090
- Dual-socket Quad-core CPU
- 128 GB System Memory

Bringing a breakthrough product ahead of competition that is far superior in design and quality needs faster turnaround times for complex engineering simulations.

Engineers working on models that use complicated multiple physics and greatly refined meshes — should investigate use Tesla GPUs for product design simulations.



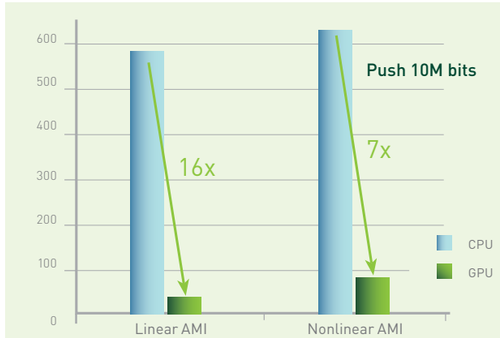
Benchmark: V13sp5 Model- Turbine Geometry, 2,100k DOF, Static nonlinear, Direct Sparse, CPU: Quad-core Xeon 2.67 GHz.

ANSYS NEXXIM

ENABLING STATE-OF-THE-ART CIRCUIT DESIGN WITH GPUs

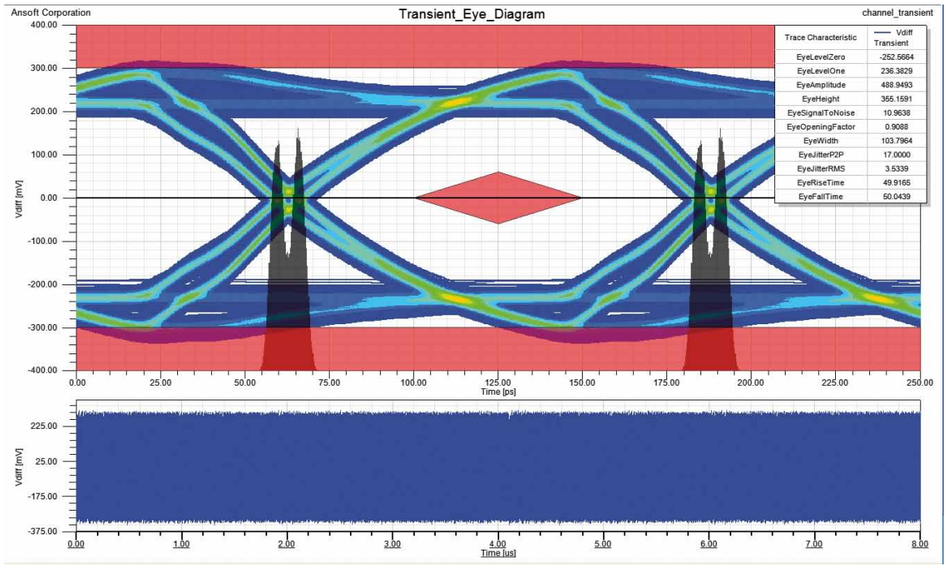
ANSYS Nexxim is a circuit simulation engine and analysis software delivering transistor-level simulation for complex RF/analog/mixed-signal IC design and verification. It provide transistor-level accuracy required for simulation of modern mixed-signal integrated circuit designs.

With GPUs, ANSYS Nexxim users can simulate large design models, which typically would take days, in merely hours. 7x to 15x faster simulation enables engineers to design superior products with more confidence, such as extremely low bit-error-rate (BER) channel design.



RECOMMENDED TESLA CONFIGURATIONS

Workstation
- 1x Tesla C2070
- Dual-socket Quad-core CPU
- 32 GB System Memory

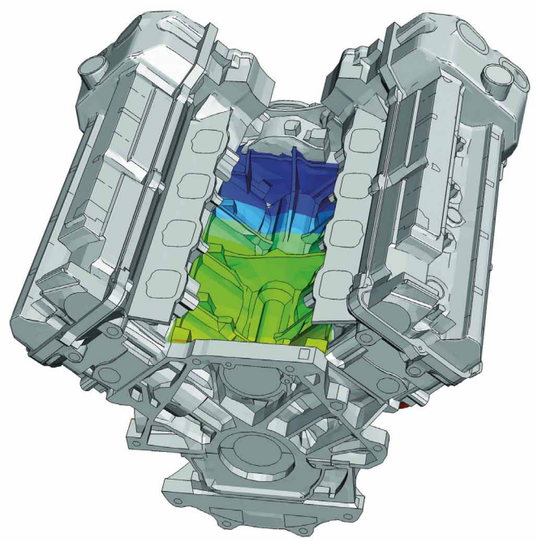


REDUCE ENGINEERING SIMULATION TIMES IN HALF
WITH SIMULIA ABAQUS / STANDARD

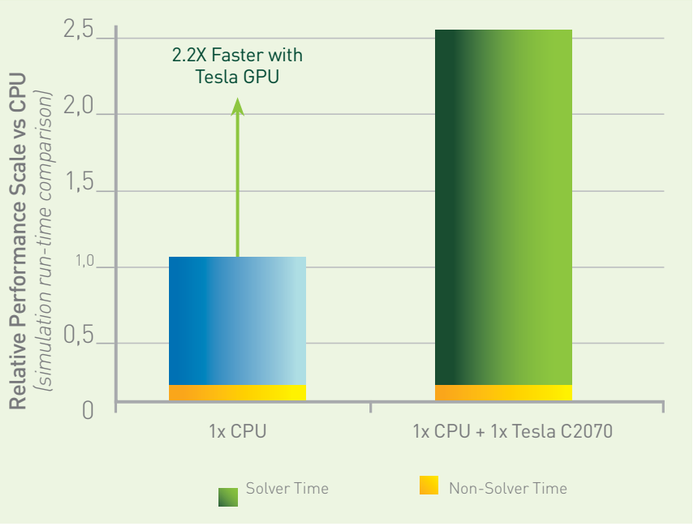
As products get more complex, the task of innovating with more confidence has been ever increasingly difficult for product engineers. Engineers rely on Abaqus to understand behavior of complex assembly or of new materials.

With GPUs, engineers can run Abaqus simulations twice as fast. Volkswagen, a leading automaker, reduced the simulation time of an engine model from 90 minutes to 44 minutes with GPUs. Faster simulations enable designers to simulate more scenarios to achieve, for example, a more fuel efficient engine.

Engineers at Volkswagen are looking into using NVIDIA GPU-acceleration in Abaqus to tackle multiple engineering challenges in car design and production. Speedups provided by the computational power of GPUs can take them to new levels of modeling realism, enabling confident innovation as cars are being optimized for lower fuel consumption and CO2 emissions.



SIMULIA



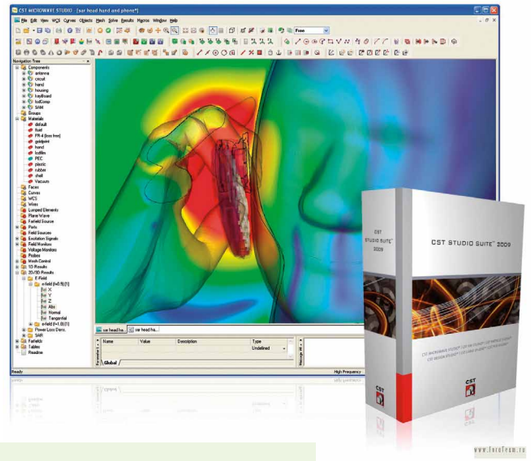
RECOMMENDED TESLA CONFIGURATIONS

Workstation
- Tesla C2070
- Dual-socket Quad-core CPU
- 48 GB System Memory

Server
- 2x Tesla M2090
- Dual-socket Quad-core CPU
- 128 GB System Memory

20X FASTER SIMULATIONS WITH GPUs
DESIGN SUPERIOR PRODUCTS WITH CST MICROWAVE STUDIO

CST



What can product engineers achieve if a single simulation run-time reduced from 48 hours to 3 hours? CST Microwave Studio is one of the most widely used electromagnetic simulation software and some of the largest customers in the world today are leveraging GPUs to introduce their products to market faster and with more confidence in the fidelity of the product design.

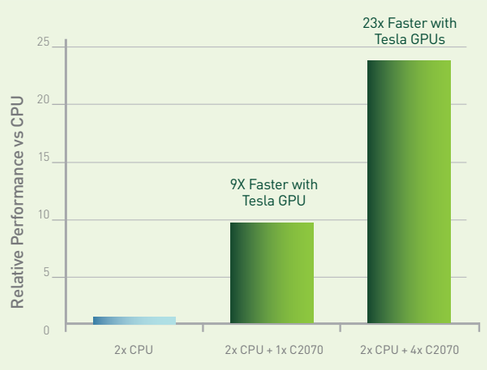
RECOMMENDED TESLA CONFIGURATIONS

Workstation
- 4x Tesla C2070
- Dual-socket Quad-core CPU
- 48 GB System Memory

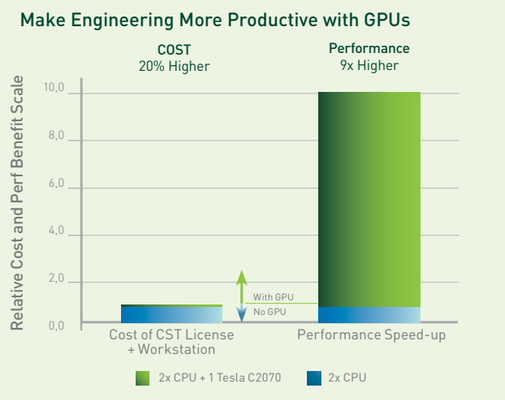
Server
- 4x Tesla M2090
- Dual-socket Quad-core CPU
- 48 GB System Memory

NVIDIA GPUs SPEED UP SIMULIA ABAQUS AND CST MICROWAVE STUDIO

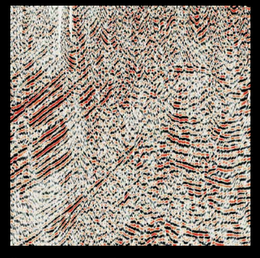
TO GET THE MOST OUT OF CST MICROWAVE STUDIO, SIMPLY ADD TESLA GPUS TO YOUR WORKSTATION OR CLUSTER AND INSTANTLY UNLOCK THE HIGHEST LEVEL OF SIMULATION PERFORMANCE.



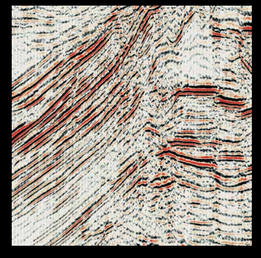
Benchmark: BGA models, 2M to 128M mesh models.
CST MWS, transient solver.
CPU: 2x Intel Xeon X5620.
Single GPU run only on 50M mesh model.



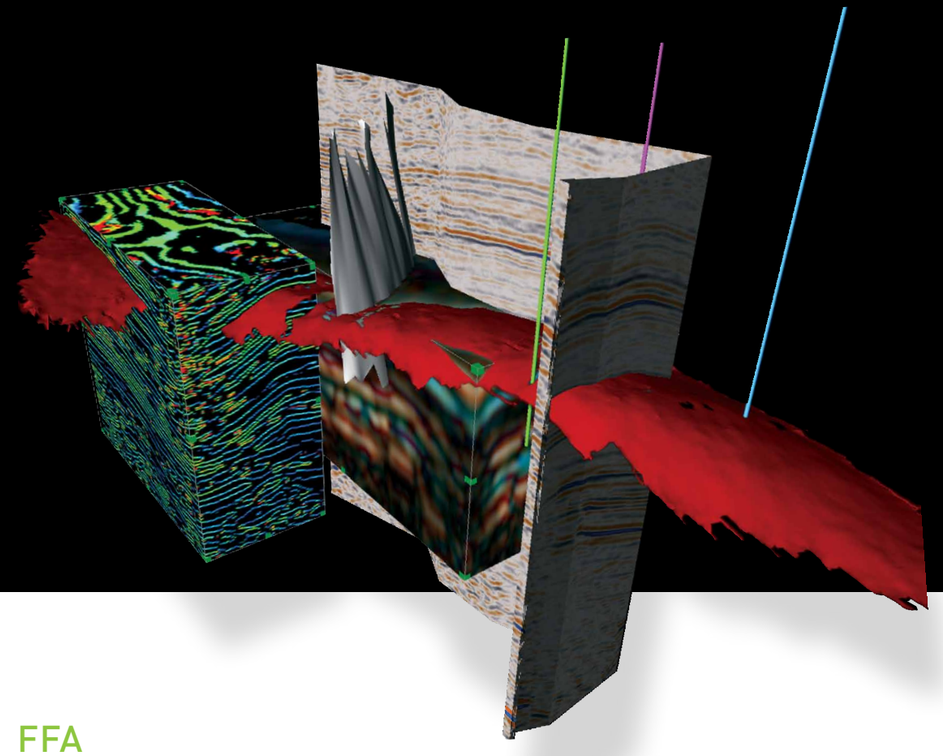
Cost Analysis includes CST Standart License + 1 GPU Acceleration Token, Workstation.
CPU: 2x Intel Xeon X5620.



Conventional Processing



Geomage Multifocusing™ Processing



GEOMAGE

A NEW APPROACH FOR CLUSTER BASED SEISMIC PROCESSING SOLUTIONS

THE CHALLENGE

Oil and gas companies have long been challenged by the need to visualize the Earth's subsurface in great detail to maximize efficiency when drilling new wells and producing from existing sources. Extremely expensive seismic surveys are conducted to achieve this, producing many terabytes of information that require processing.

Over time seismic survey output has grown significantly as companies seek to pinpoint reserves more accurately, making its analysis and computation more laborious, time consuming and costly.

The new generation of seismic processing methodologies requires tremendous computational power in order to cope with huge datasets. A cluster of several thousand central processing unit (CPU) cores is no longer sufficient, and can usually only handle one algorithm running one specific survey at a time. As a result, traditional CPU-based data centers are no longer sufficient.

THE SOLUTION

Geomage's proprietary MultiFocusing technology has enabled the development of specialized applications that provide a cutting-edge alternative to conventional seismic processing for the oil and gas industry. Geomage's MultiFocusing technology is currently one of the most powerful tools in the industry. However, it is also one of, if not the most, time-consuming seismic processing algorithms ever applied.

Geomage evaluated the available solutions in the market and, after extensive research, selected the NVIDIA CUDA framework along with the NVIDIA Tesla product line as its solution for replacing CPU clusters. Geomage's flagship algorithm was chosen to be migrated to the GPU and the core business code was rewritten using CUDA.

Most of the algorithms ported to the GPU have shown a performance boost of up to 100X, as compared to a high-end, single-CPU core.

THE IMPACT

"NVIDIA's Tesla GPUs have given Geomage a leading edge over other companies by enabling us to produce results in an acceptable time frame and develop new and even more compute intensive algorithms", said Tamir Tai, COO of Geomage.

"Our clients can now explore and use their data in a way that was not possible only a short time ago. The plan now is to move more and more algorithms to the GPU, to further improve their performance on the GPU and ultimately to move the entire processing department to use GPUs exclusively. An entire brand of software/hardware solutions is now available for exploration with the use of these Tesla solutions."

FFA

CHANGING THE WAY 3D SEISMIC INTERPRETATION IS CARRIED OUT

THE CHALLENGE

In the search for oil and gas, the geological information provided by seismic images of the earth is vital. By interpreting the data produced by seismic imaging surveys, geoscientists can identify the likely presence of hydrocarbon reserves and understand how to extract resources most effectively. Today, sophisticated visualization systems and computational analysis tools are used to streamline what was previously a subjective and labor intensive process.

Today, geoscientists must process increasing amounts of data as dwindling reserves require them to pinpoint smaller, more complex reservoirs with greater speed and accuracy.

THE SOLUTION

UK-based company ffA provides world leading 3D seismic analysis software and services to the global oil and gas industry. Its software tools extract detailed information from 3D seismic data, providing a greater understanding of complex 3D geology. The sophisticated tools are extremely compute-intensive.

With the recent release of its CUDA enabled 3D seismic analysis application, ffA users routinely achieve over an order of magnitude speed-up compared with performance on high end multi-core CPUs.

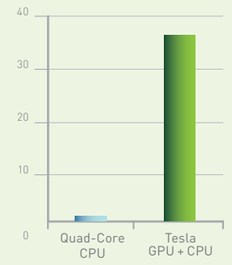
This step significantly increases the amount of data that geoscientists can analyze in a given timeframe. Plus, it allows them to fully exploit the information derived from 3D seismic surveys to improve subsurface understanding and reduce risk in oil and gas exploration and exploitation.

THE IMPACT

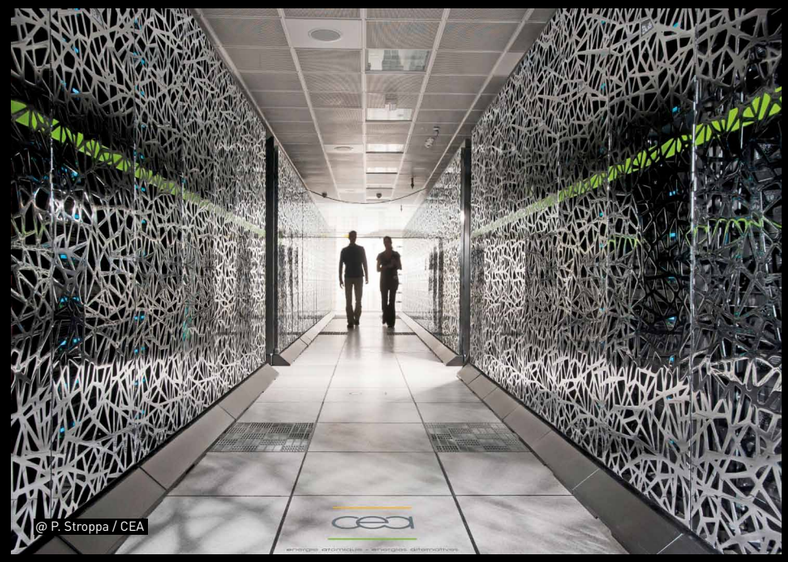
NVIDIA CUDA is allowing ffA to provide scalable high performance computation for seismic data on hardware platforms equipped with one or more NVIDIA Quadro FX and NVIDIA Tesla GPUs.

The latest benchmarking results using Tesla GPUs have produced performance improvements of up to 37x versus high end dual quad core CPU workstations.

STREAMLING WITH GPUs FOR BETTER RESULTS



The latest benchmarking results using Tesla GPUs have produced performance improvements of up to 37x versus high end dual quad core CPU workstations.



GPU COMPUTING IN FINANCE – CASE STUDIES

BNP PARIBAS ACCELERATES COMPUTATIONS AND REDUCES POWER CONSUMPTION

Challenge: BNP Paribas is one of the leading global players in credit, commodity and equity derivatives. It's subdivision - GECD manages portfolios of most complex financial instruments with non-stop massive calculations where computational time really matters.

Solution: Implementation of a new Tesla GPU-based architecture which simultaneously reduces power consumption and accelerates calculation times.
A significant portion of the calculations performed for GECD is transferred to this architecture.

Impact:
- Reducing electrical consumption by a factor of 190
- Overall reduction of response times by a factor of 15
- Improved legibility of market prices by traders.
- Significant reduction in total cost of operation

BLOOMBERG: GPUs INCREASE ACCURACY AND REDUCE PROCESSING TIME FOR BOND PRICING

Challenge: Bloomberg, one of the world's leading financial services organizations, prices loan baskets for its customers by running powerful algorithms that model the risks and determine the price. This technique requires calculating huge amounts of data, from interest rate volatility to the payment behavior of individual borrowers. Such data-intensive calculations can take hours to run with a CPU-based computing grid.

Solution: Bloomberg implemented an NVIDIA Tesla GPU computing solution in their datacenter. By porting this application to run on the NVIDIA CUDA parallel processing architecture Bloomberg received dramatic improvements across the board. Large calculations that had previously taken up to two hours can now be completed in two minutes.

The capital outlay for the new GPU-based solution was one tenth the cost of an upgraded CPU solution.

Impact: As Bloomberg customers make CDO/CMO buying and selling decisions, they now have access to the best and most current pricing information, giving them a serious competitive trading advantage in a market where timing is everything.

DRIVING THE HPC REVOLUTION ON FRANCE'S LARGEST HYBRID RESEARCH CLUSTERS

GPU EXTENSION OF THE TERA 100 SYSTEM AT CEA

CEA acquired from Bull a GPU-based extension to the TERA 100 petaflop-scale system. This extension includes 22 bullx blade chassis, each housing 9 bullx B505 accelerator blades, i.e. a total of 396 GPU processors.

Each bullx B505 blade integrates two new-generation NVIDIA® Tesla™ M2090 processors, along the two CPUs. They are designed so as to provide maximum bandwidth between each GPU and its host CPU, thanks to two chipsets and a double interconnect on each blade.

About CEA: The French Alternative Energies and Atomic Energy Commission (CEA) leads research, development and innovation in four main areas: low-carbon energy sources, global defence and security, information technologies and healthcare technologies. With its 15,600 researchers and collaborators, it has internationally recognized expertise in its areas of excellence and has developed many collaborations with national and international, academic and industrial partners.

GPU EXTENSION OF THE CURIE SUPERCOMPUTER AT GENCI

About GENCI: GENCI, Grand Equipement National de Calcul Intensif, is owned for 49 % by the French State, for 20 % by CEA, 20 % by CNRS, 10 % by French Universities and 1% by INRIA. GENCI's role is to set the strategic direction and to make France's most important investments in High-Performance Computing (HPC), in particular as the custodian of France's commitment to PRACE.

As part of its CURIE supercomputer that is currently being installed by Bull, GENCI has acquired 16 bullx blade chassis, housing a total of 144 B505 accelerator blades, i.e. 288 NVIDIA Tesla M2090 GPUs (147,456 GPU cores). The 1.6-petaflops CURIE system will be entirely dedicated to the use of the European PRACE partnership (the Partnership for Advanced Computing in Europe).

TESLA GPU COMPUTING SOLUTIONS

The Tesla 20-series GPU computing solutions are designed from the ground up for high-performance computing and are based on NVIDIA's latest CUDA GPU architecture, code named "Fermi". It delivers many "must have" features for HPC including ECC memory for uncompromised accuracy and scalability, C++ support, and 7x double precision performance compared to CPUs. Compared to the typical quad-core CPUs, Tesla 20-series GPU computing products can deliver equivalent performance at 1/10th the cost and 1/20th the power consumption.

• Superior Performance

Highest double precision floating point performance and large on-board memory to support large HPC data sets

• Highly Reliable

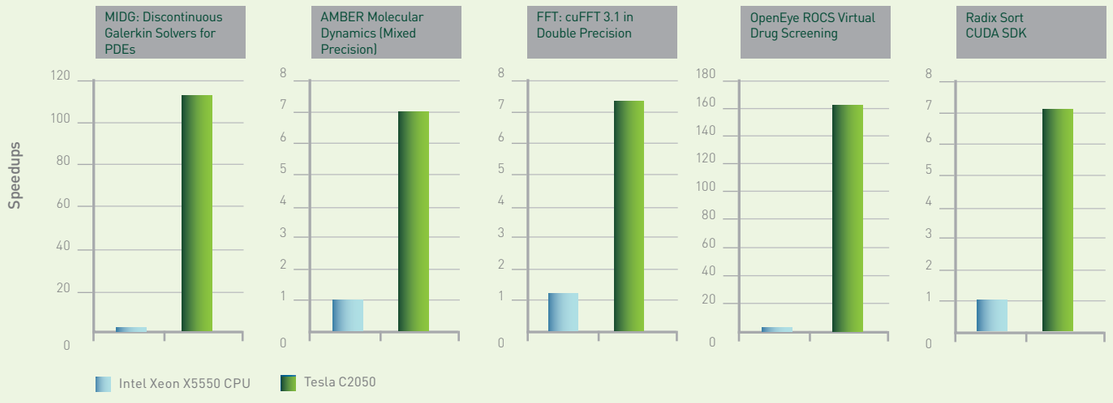
Uncompromised data reliability through ECC protection and stress tested for zero error tolerance

• Designed for HPC

For more information on Tesla GPU computing products and applications, visit www.nvidia.eu/tesla.

Highest Performance, Highest Efficiency

Workstations powered by Tesla GPUs outperform conventional CPU-only solution in life scince applications.



TESLA DATA CENTER PRODUCTS

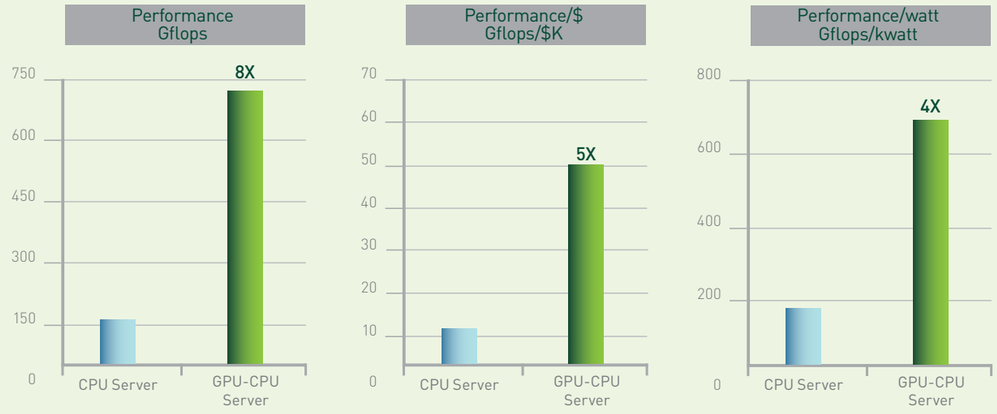
Available from OEMs and certified resellers, Tesla GPU computing products are designed to supercharge your computing cluster.

TESLA WORKSTATION PRODUCTS

Designed to deliver cluster-level performance on a workstation, the NVIDIA Tesla GPU Computing Processors fuel the transition to parallel computing while making personal supercomputing possible — right at your desk.

Highest Performance, Highest Efficiency

GPU-CPU server solutions deliver up to 8x higher Linpack performance.



CPU 1U Server: 2x Intel Xeon X5550 (Nehalem) 2.66 GHz, 48 GB memory, $7K, 0.55 kw

GPU-CPU 1U Server: 2x Tesla C2050 + 2x Intel Xeon X5550, 48 GB memory, $11K, 1.0 kw

Tesla M2050/M2070/M2090 GPU Computing Modules enables the use of GPUs and CPUs together in an individual server node or blade form factor.

Tesla C2050/C2070 GPU Computing Processor delivers the power of a cluster in the form factor of a workstation.

TESLA C-CLASS GPU COMPUTING PROCESSORS
THE POWER INSIDE PERSONAL MULTI-TERAFLOPS SYSTEMS

NVIDIA CUDA TECHNOLOGY UNLOCKS THE POWER OF TESLA'S MANY CORES

The CUDA C programming environment simplifies many-core programming and enhances performance by offloading computationally-intensive activities from the CPU to the GPU. It enables developers to utilise NVIDIA GPUs to solve the most complex computation intensive challenges such as protein docking, molecular dynamics, financial analysis, fluid dynamics, structural analysis and many others.

The NVIDIA® Tesla™ Personal Supercomputer is based on the revolutionary NVIDIA® CUDA™ parallel computing architecture and powered by up to thousands of parallel processing cores.

YOUR OWN SUPERCOMPUTER

Get nearly 4 teraflops of compute capability and the ability to perform computations 250 times faster than a multi-CPU core PC or workstation.

NVIDIA CUDA UNLOCKS THE POWER OF GPU PARALLEL COMPUTING

The CUDA parallel computing architecture enables developers to utilise C or FORTRAN programming with NVIDIA GPUs to run the most complex computationally intensive applications. CUDA is easy to learn and has become widely adopted by thousands of application developers worldwide to accelerate the most performance demanding applications.

ACCESSIBLE TO EVERYONE

Available from OEMs and resellers worldwide, the Tesla Personal Supercomputer operates quietly and plugs into a standard power strip so you can take advantage of cluster level performance anytime you want, right from your desk.

PETASCALE COMPUTING WITH TERAFLOP PROCESSORS

The NVIDIA Tesla computing card enables the transition to energy efficient parallel computing power by bringing the performance of a small cluster to a workstation. With hundreds of processor cores and a standard C compiler that simplifies application development, Tesla cards scale to solve the world's most important computing challenges more quickly and accurately.

TESLA PERSONAL SUPERCOMPUTER

Cluster Performance on your Desktop	The performance of a cluster in a desktop system. Four Tesla GPU computing processors deliver nearly 4 teraflops of performance.
Designed for Office Use	Plugs into a standard office power socket and quiet enough for use at your desk.
Massively Parallel Many Core GPU architecture	240 parallel processor cores per GPU that can execute thousands of concurrent threads.
High-Speed Memory per GPU	Dedicated compute memory enables larger datasets to be stored locally for each processor minimizing data movement around the system.
IEEE 754 Floating Point Precision (single -precision and double -precision)	Provides results that are consistent across platforms and meet industry standards.

	Tesla C2050	Tesla C2070
Architecture	Tesla 20-series GPU	
Number of Cores	448	
Caches	64 KB L1 cache + Shared Memory / 32 cores, 768 KB L2 cache	
FP Peak Performance	1,03 TFlops (single) 515 GFlops (double)	
FP Application Efficiency (Tesla C1060 reference)	1.5 - 2 (single) 3 - 4 (double)	
GPU Memory	3 GB GDDR5 ECC 2.625 GB with ECC on	6 GB GDDR5 ECC 5.25 GB with ECC on
Memory Bandwidth	144 GB/s ECC off 115 GB/s ECC on	
Video Output	DVI-I	
System I/O	PCIe x16 Gen2 (bi-directional async. transfer)	
Positioning	Best price/performance solutions for double precision codes and when ECC memory required	

	Tesla M2050	Tesla M2070	Tesla M2090
Architecture	**Tesla 20-series GPU**		
Number of Cores	448		512
Caches	64 KB L1 cache + Shared Memory / 32 cores, 768 KB L2 cache		
FP Peak Performance	1030 GFlops (single) 515 GFlops (double)		1331 GFlops (single) 665 GFlops (double)
FP Application Efficiency (Tesla C1060 reference)	1.5 - 2 (single) 3 - 4 (double)		
GPU Memory	3 GB GDDR5 ECC 2,625 GB with ECC on	6 GB GDDR5 ECC 5,25 GB with ECC on	6 GB GDDR5 ECC 5,25 GB with ECC on
Memory Bandwidth	144 GB/s ECC off 115 GB/s ECC on		
Positioning	Best price/performance solutions for double precision codes and when ECC memory required		

GPU COMPUTING SOLUTIONS: TESLA MODULES

TESLA M-CLASS GPU COMPUTING MODULES

DESIGNED FOR DATACENTER INTEGRATION

Tesla M-class GPU Computing Modules enable the seamless integration of GPU computing with host systems for high-performance computing and large data center, scale-out deployments.

SYSTEM MONITORING FEATURES

Tesla Modules deliver all the standard benefits of GPU computing while enabling maximum reliability and tight integration with system monitoring and management tools. This gives data centre IT staff much greater choice in how they deploy GPUs, offering a wide variety of rack-mount and blade systems plus the remote monitoring and remote management capabilities they need.

ENDORSED BY THE HPC OEMS

The Tesla Modules are only available in OEM systems that have been specifically designed and qualified along with NVIDIA engineers. They are designed for maximum reliability; their passive heatsink design eliminates moving parts and cables. Tesla Modules have been selected by the most important HPC vendors world-wide.

Tesla GPU's high performance makes them ideal for seismic processing, biochemistry simulations, weather and climate modeling, signal processing, computational finance, CAE, CFD, and data analystics.

The Tesla 20-series GPU Computing Processors are the first to deliver greater than 10X the double precision horsepower of a quad-core x86 CPU and the first GPUs to deliver ECC memory. Based on the Fermi architecture, these GPUs feature up to 665 gigaflops of double precision performance, 1 teraflops of single precision performance, ECC memory error protection, and L1 and L2 caches.

Based on the CUDA™ architecture codenamed "Fermi", the Tesla™ M-class GPU Computing Modules are the world's fastest parallel computing processors for high performance computing (HPC).

NVIDIA TESLA AND CUDA LINKS

NEWS AND SUCCESS STORI ES

NVIDIA GPU Computing on Twitter
http://twitter.com/gpucomputing

CUDA weekly newsletter
www.nvidia.com/object/cuda_week_in_review_newsletter.html

News and articles
www.nvidia.co.uk/page/tesla-articles.html

Tesla Video on YouTube
www.youtube.com/nvidiatesla

Success stories
www.nvidia.co.uk/page/tesla_testimonials.html

SOFTWARE

CUDA Zone
www.nvidia.co.uk/cuda

CUDA in Action
www.nvidia.co.uk/object/cuda_in_action.html

CUDA Books
www.nvidia.co.uk/object/cuda_books.html

CUDA Training & Consulting
www.nvidia.co.uk/page/cuda_consultants.html

Software Development Tools
www.nvidia.co.uk/object/tesla_software_uk.html

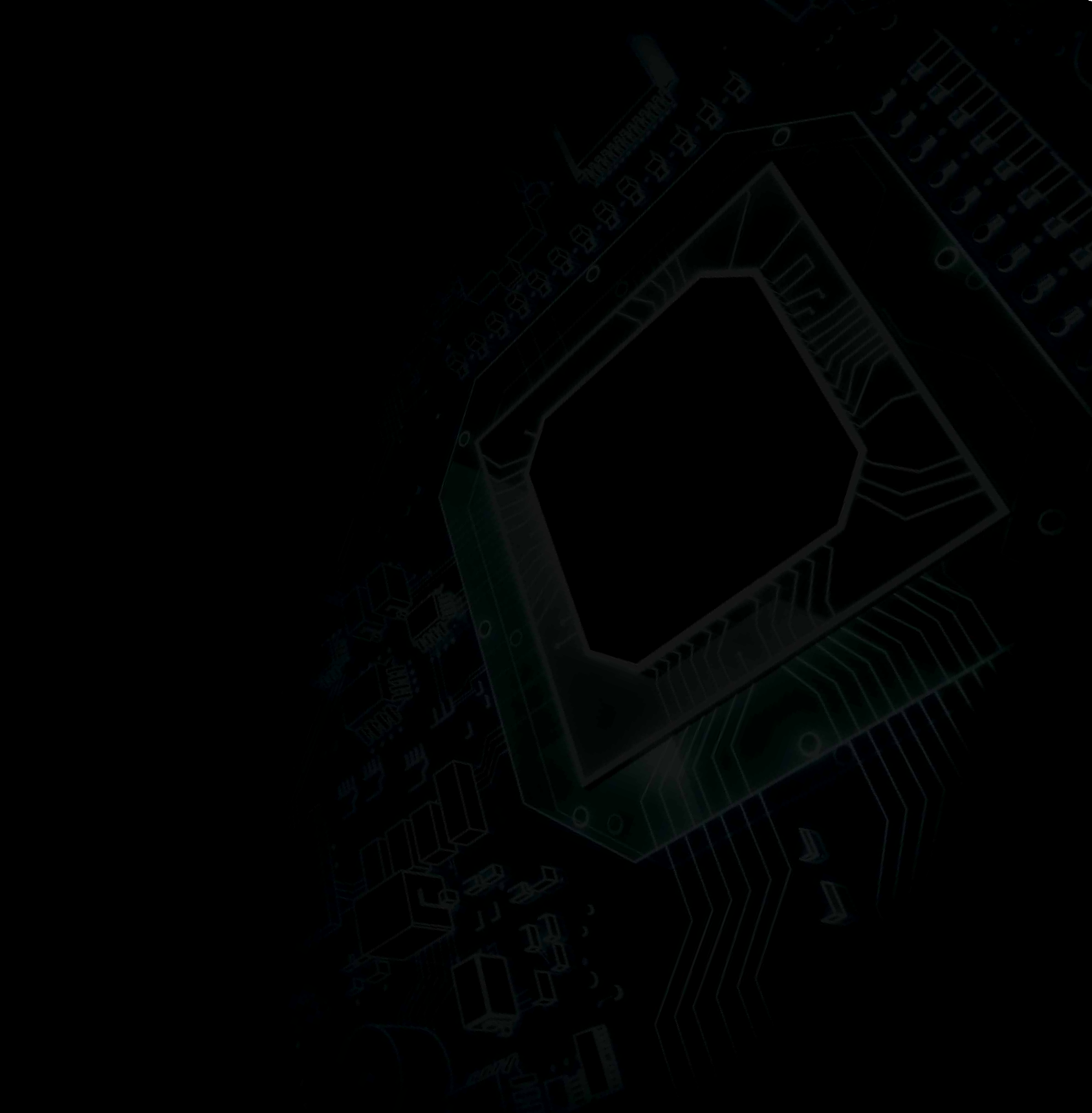
HARD WARE

NVIDIA High Performance Computing
www.nvidia.eu/tesla

Tesla for Personal Supercomputers
www.nvidia.co.uk/psc

Tesla Data Center Solutions
www.nvidia.co.uk/page/preconfigured_clusters.html

Tesla Products Technical Descriptions
www.nvidia.co.uk/page/tesla_product_literature.html



NVIDIA®