REVOLUTIONISING HIGH PERFORMANCE COMPUTING
NVIDIA® TESLA®
TABLE OF CONTENTS

GPUs are revolutionising computing 4
NVIDIA Tesla Kepler architecture 6
CUDA 5.0 and developer tools 8
Accelerate your code with OpenACC Directives 10
GPU acceleration for science and research 12
GPUs accelerate MATLAB, CST, MSC Nastran 14
GPUs accelerate SIMULIA Abaqus, automotive CFD 16
GPUs accelerate ANSYS Mechanical, ANSYS Fluent 18
GPUs for life science and material science applications 20
GPUs accelerate molecular dynamics and weather modelling 22
GPU computing case studies: Oil and Gas 24
Case studies: numerical analytics, HPC 26
Kepler vs. Fermi performance comparison 28
GPU computing solutions: Tesla Fermi, Tesla K10 30
GPU computing solutions: Tesla K20, K20X 32
PNY company background, useful links 34
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.

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.

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 modelling 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.

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.

One challenge in supercomputers today is power consumption. Combining GPUs and CPUs in a single system requires less power than CPUs alone and is a responsible move toward lowering our carbon footprint. Titan will provide unprecedented computing power for research in energy, climate change, materials and other disciplines to enable scientific leadership.”

**TITAN — THE WORLD’S MOST POWERFUL SUPERCOMPUTER**

ORNL’s Titan a supercomputer capable of churning through more than 20,000 trillion calculations each second — or 20 petaflops — by employing a family of processors called graphic processing units is proudly leading the Top500 list. Titan is 10 times more powerful than ORNL’s last world-leading system, Jaguar, while overcoming power and space limitations inherent in the previous generation of high performance computers.

Titan, which is supported by the Department of Energy, provides unprecedented computing power for research in energy, climate change, efficient engines, materials and other disciplines and pave the way for a wide range of achievements in science and technology.

Because they handle hundreds of calculations simultaneously, GPUs can go through many more than CPUs in a given time. By relying on its 299,008 CPU cores to guide simulations and allowing its new NVIDIA GPUs to do the heavy lifting, Titan will enable researchers to run scientific calculations with far greater speed and accuracy. Supported by the DOE, Titan will provide unprecedented computing power for research in energy, climate change, efficient engines, materials and other disciplines and pave the way for a wide range of achievements in science and technology.

www.ornl.gov/info/ornlreview

**#1 IN ENERGY EFFICIENCY: EURORA AND THE K20 GPU**

The “Eurora” supercomputer reached 3,150 megaflops per watt of sustained (Linpack) performance — a mark 26 percent higher than the top system on the most recent Green500 list of the world’s most efficient supercomputers. Parring NVIDIA Tesla K20 GPUs with Eurotech’s Aurora Hot Water Cooling technology, the Eurora system is more efficient and compact than conventional air-cooled solutions.

**TITAN PERFORMANCE**

<table>
<thead>
<tr>
<th>Memory per node</th>
<th>32 GB + 6 GB</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Opteron cores</td>
<td>299,008</td>
</tr>
<tr>
<td># of NVIDIA K20X*</td>
<td>18,688*</td>
</tr>
<tr>
<td>Total System Memory</td>
<td>710 TB</td>
</tr>
<tr>
<td>Total System Peak Performance</td>
<td>27 Petaflops</td>
</tr>
<tr>
<td>Sustained Performance</td>
<td>17,599 Petaflops/s</td>
</tr>
</tbody>
</table>

* 90% of performance from NVIDIA GPUs.
FASTEST, MOST EFFICIENT HPC ARCHITECTURE

With the launch of the Fermi GPU in 2009, NVIDIA ushered in a new era in the high performance computing (HPC) industry based on a hybrid computing model where CPUs and GPUs work together to solve computationally-intensive workloads. And in just a couple of years, NVIDIA Fermi GPUs powers some of the fastest supercomputers in the world as well as tens of thousands of research clusters globally. Now, with the new Tesla Kepler, NVIDIA raises the bar for the HPC industry, yet again.

Comprised of 7.1 billion transistors, the Tesla Kepler is an engineering marvel created to address the most daunting challenges in HPC. Kepler is designed from the ground up to maximise performance computing (HPC) industry, yet again.

NVIDIA® Tesla® Kepler — World’s fastest and most power efficient GPU accelerator.

DYNAMIC PARALLELISM — CREATING WORK ON-THE-FLY

One of the overarching goals in designing the Kepler GK110 architecture was to make it easier for developers to more easily take advantage of the immense parallel processing capability of the GPU.

To this end, the new Dynamic Parallelism feature enables the Tesla® Kepler to dynamically spawn new threads by adapting to the data without going back to the host CPU. This effectively allows more of a program to be run directly on the GPU, as kernels now have the ability to independently launch additional workloads as needed.

Any kernel can launch another kernel and can create the necessary streams, events, and dependencies needed to process additional work without the need for host CPU interaction. This simplified programming model is easier to create, optimize, and maintain. It also creates a programmer-friendly environment by maintaining the same syntax for GPU launched workloads as traditional CPU kernel launches.

Dynamic Parallelism broadens what applications can now accomplish with GPUs in various disciplines. Applications can launch small and medium sized parallel workloads dynamically where it was too expensive to do so previously.

HYPER-Q — MAXIMIZING THE GPU RESOURCES

Hyper-Q enables multiple CPU cores to launch work on a single GPU simultaneously, thereby dramatically increasing GPU utilization and slashing CPU idle times. This feature increases the total number of connections between the host and the Tesla® Kepler by allowing 32 simultaneous, hardware managed connections, compared to the single connection available with Fermi. Hyper-Q is a flexible solution that allows connections for both CUDA streams and Message Passing Interface (MPI) processes, or even threads from within a process. Existing applications that were previously limited by false dependencies can see up to a 32x performance increase without changing any existing code.

Hyper-Q offers significant benefits for use in MPI-based parallel computer systems. Legacy MPI-based algorithms were often created to run on multi-core CPU-based systems. Because the workload that could be efficiently handled by CPU-based systems is generally smaller than that available using GPUs, the amount of work passed in each MPI process is generally insufficient to fully occupy the GPU processor.

While it has always been possible to issue multiple MPI processes to concurrently run on the GPU, these processes could become bottlenecked by false dependencies, forcing the GPU to operate below peak efficiency. Hyper-Q removes false dependency bottlenecks and dramatically increases speed at which MPI processes can be moved from the system CPU(s) to the GPU for processing.

Hyper-Q promises to be a performance boost for MPI applications.

CONCLUSION

Tesla® Kepler is engineered to deliver ground-breaking performance with superior power efficiency while making GPUs easier than ever to use. SMX, Dynamic Parallelism, and Hyper-Q are three important innovations in the Tesla® Kepler to bring these benefits to reality for our customers. For further details on additional architectural features, please refer to the Kepler GK110 whitepaper at www.nvidia.co.uk/kepler

Figure 3: Without Dynamic Parallelism, the CPU launches every kernel onto the GPU. With the new feature, Tesla® Kepler can now launch nested kernels, eliminating the need to communicate with the CPU.

Figure 4: Hyper-Q allows all streams to run concurrently using a separate work queue. In the Fermi model, concurrency was limited due to intra-stream dependencies caused by the single hardware work queue.

TRY THE TESLA GPU TEST DRIVE TODAY!

A ready to use solution for science and research. www.nvidia.eu/cluster
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 “Kepler”, is the most advanced GPU computing architecture ever built. With over seven billion transistors, Kepler is making GPU and CPU co-processing pervasive by addressing the full spectrum of computing applications. With support for C++, GPUs based on the Kepler 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. www.nvidia.com/cuda

CUDA 5.0 AN DEVELOPER TOOLS

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

DEVELOPER ECOSYSTEM

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.

GPU-ACELERATED LIBRARIES

Take advantage of the massively parallel computing power of the GPU by using the GPU-accelerated versions of your existing libraries. Some examples include:

NVIDIA Math Libraries
A collection of GPU-accelerated libraries—including FFT, BLAS, sparse matrix operations, RNG, performance primitives for image/signal processing, and the Thrust C++ library of high-performance templated algorithms— that all deliver significant speedups when compared to CPU-only libraries.

These highly optimized libraries are free of charge in the NVIDIA® CUDA® Toolkit available at www.nvidia.com/gpuctk.

EM Photonics CUDA
A GPU-accelerated linear algebra (LA) library that dramatically improves the performance of sophisticated mathematics.
ACCELERATE YOUR CODE EASILY WITH OPENACC DIRECTIVES

I have written micromagnetic codes (written in Fortran 90) to study the properties of two and three dimensional magnetic systems. The directives approach enabled me to port my existing code with ease to perform my computations on the GPU which resulted in a significant speedup (more than 20 times) of the computation.”

Professor M. Amin Kayali, University of Houston

ACCELERATE YOUR CODE EASILY WITH OPENACC DIRECTIVES

Accelerate your code with directives and tap into the thousands of computing cores in GPUs. With directives, you simply insert compiler hints into your code and the compiler will automatically map compute-intensive portions of your code to the GPU.

By starting with a free, 30-day trial of PGI directives today, you are working on the technology that is the foundation of the OpenACC directives standard.

OpenACC is:
• Open: Future-proof your codes with this open standard
• Simple: Easy, high-level, compiler driven approach to parallel computing
• Portable: Ideal for accelerating legacy Fortran or C codes

OpenACC is supported by industry parallel computing tool leaders: Cray, CAPS, and The Portland Group (PGI).

FOR A FREE 30 DAY TRIAL OF THE DIRECTIVES REGISTER NOW!
www.nvidia.eu/openacc

DIRECTIVE-BASED SOLUTIONS

Directives allow you to quickly add GPU acceleration to the most performance critical sections of your application while maintaining portability. Directive-based solutions for C and Fortran available today include:

PGI ACCELERATOR™ COMPILERS

Similar to OpenMP, PGI Accelerator directives can be added to existing C99 or Fortran applications to achieve GPU acceleration on Linux, Mac OS X, and Windows.

OpenACC DIRECTIVES

The OpenACC Application Program Interface describes a collection of compiler directives to specify loops and regions of code in standard C, C++ and Fortran to be offloaded from a host CPU to an attached accelerator, providing portability across operating systems, host CPUs and accelerators.

The directives and programming model defined in this document allow programmers to create high-level host-accelerator programs without the need to explicitly initialize the accelerator, manage data or program transfers between the host and accelerator, or initiate accelerator startup and shutdown.

DIRECTIVE-BASED SOLUTIONS

Directed-based compiler for hybrid computing based on C and FORTRAN directives, HmPP offers a high level abstraction for hybrid programming that fully leverages the computing power of stream processors without the complexity associated with GPU programming. HmPP compiler integrates powerful data-parallel back ends for NVIDIA CUDA and OpenCL that drastically reduce development time. The HmPP runtime ensures application deployment on multi-GPU systems.

www.caps-entreprise.com

HINT

Add a simple compiler directive to provide a hint for the compiler to automatically parallelize your for loop and execute it on the GPU.

```
main() {
  double pi = 0.0f; long i;
  #pragma acc region
  for (i=0; i<N; i++)
  {
    double t= (double)((i+0.5)/N);
    pi +=4.0/(1.0+t*t);
  }
  printf("pi=%f\n",pi/N);
}
```

```
for (i=0; i<N ; i++)
{
  double t= (double)((i+0.5)/N);
  pi +=4.0/(1.0+t*t);
}
```

The PGI compiler is now showing us just how powerful it is. On the software we are writing, it’s at least 60 times faster on the NVIDIA card. We are very pleased and excited about the future uses. It’s like owning a personal supercomputer.”

Dr. Kerry Black, University of Melbourne

```
main() {
  double pi = 0.0f; long i;
  #pragma acc region
  for (i=0; i<N ; i++)
  {
    double t= (double)((i+0.5)/N);
    pi +=4.0/(1.0+t*t);
  }
  printf("pi=%f\n",pi/N);
}
```

```
for (i=0; i<N ; i++)
{
  double t= (double)((i+0.5)/N);
  pi +=4.0/(1.0+t*t);
}
```
NVIDIA’s Tesla GPU card includes almost everything scientists ever wanted.”

CT magazine (Germany)
MATLAB® ACCELERATIONS ON TESLA® GPUs
MATLAB PERFORMANCE WITH TESLA

NVIDIA and MathWorks have collaborated to deliver the power of GPU computing for MATLAB users. Available with the latest release of MATLAB, NVIDIA GPU acceleration enables faster results for users of the Parallel Computing Toolbox and MATLAB Distributed Computing Server. GPU support is also available in several domain-specific MATLAB toolboxes (including Signal Processing, Communications System, Phased Array, and Neural Network toolboxes). MATLAB supports NVIDIA® CUDA™ — enabled and Neural Network toolboxes). MATLAB communications System, Phased array, toolboxes (including Signal Processing, in several domain-specific MATLAB and MATLAB distributed computing users of the Parallel computing Toolbox acceleration enables faster results for the latest release of MATLAB, Nvidia GPU for MATLAB users. available with the alpha release of MATLAB® accelerations on TESLA® GPUs.

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

MATLAB PERFORMANCE WITH TESLA

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.

RECOMMENDED TESLA & QUADRO CONFIGURATIONS

High-End Workstation
- Two Tesla K20 GPUs
- Quadro 6000
- Two CPUs
- 12 GB system memory

Mid-Range Workstation
- Tesla K20 GPUs
- Quadro 2000
- One CPU
- 6 GB system memory

Entry Workstation
- TESLA K20 GPUs
- Quadro 600
- One CPU
- 6 GB system memory

MSC NASTRAN 2013 (ALPHA)
5X PERFORMANCE BOOST WITH SINGLE GPU OVER SINGLE CORE,
>1.5X WITH 2 GPUs OVER 8 CORE

NVIDIA® and MSC Software have collaborated to deliver the power of GPU computing for Nastran customers. Available with the alpha release of MSC Nastran 2013, NVIDIA® GPU acceleration enables faster results for more efficient computation and job turnaround times, delivering more license utilization for the same investment.

Recent performance studies conducted together with MSC engineers demonstrated that the Tesla K20X GPU acceleration of Intel Xeon (Sandy Bridge) CPUs resulted in speed-ups in the range of 5X+ with a single GPU over a serial run and in the range of 1.5-2.0X with 2 GPUs over 8 cores DMP run of SOL108 (NVH) models.

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.
The flow simulation software Lbultra works particularly fast on graphics processing units (GPUs). As a plug-in prototype, it is tightly integrated into the high-end 3D visualization software RTT DeltaGen by RTT. As a consequence, flow simulation computing can be proceeded directly within RTT DeltaGen. Due to this coupling system, the designer can do a direct flow simulation of their latest design draft — enabling the designer to do a parallel check of aerodynamic features of the vehicle’s design draft.

First of all, a certain scenario is selected, such as analyzing a spoiler or an outside mirror. Next, various simulation parameters and boundary conditions such as flow rates or resolution levels are set, which also influence the calculation time and result’s accuracy.

After determining the overall simulation, the geometry of the design is handed over to Lbultra and the simulation is started. While the simulation is running, data is being visualized in real-time in RTT DeltaGen. In addition, there is an opportunity of exporting simulation results. These results may then, for instance, be further analyzed in more detail by experts using highly-capable fluid mechanics specialist programs and tools.

**IMPROVED OPTIMUM PERFORMANCE THROUGH EARLY INVOLVEMENT OF AERODYNAMICS INTO DESIGN**

**FAST FLOW SIMULATIONS DIRECTLY WITHIN RTT DELTAGEN**

**Lbultra Plug-In for RTT Deltagen**

**Lbultra Developed by Fluidyna**

---

**SIMULIA Abaqus/Standard**

**Reduce Engineering Simulation Times**

NVIDIA and SIMULIA, the Dassault Systems brand for Realistic Simulation, have collaborated to deliver the power of GPU computing for Abaqus customers. Available in the Abaqus 6.11 release, NVIDIA GPU acceleration enables faster results for more efficient computation and job turnaround times, delivering more license utilization for the same investment.

**Recommended Tesla Configurations**

- **Workstation**
  - Tesla K20c
  - 4x-94GB memory
  - Dual CPUs

- **Server**
  - Tesla K20m/K20X
  - 96-128GB memory
  - Dual CPUs

**Recommended Tesla Configurations**

- Westmere (2x Xeon X5670, 2.93 GHz CPUs, 96GB memory), 2x Tesla M2090, Linux RHEL 6.2
- Sandy Bridge (2x E5-2670, 2.6GHz CPUs, 128GB memory), 2x Tesla K20X, Linux RHEL 6.2

**Test Drive Your SIMULIA Abaqus/Standard Jobs with GPUs for Free!**

www.accelerateabaqusongpu.com

---

**Simulia Abaqus/Standard**

**Reduce Engineering Simulation Times**

**NVIDIA and SIMULIA, the Dassault Systems brand for Realistic Simulation, have collaborated to deliver the power of GPU computing for Abaqus customers. Available in the Abaqus 6.11 release, NVIDIA GPU acceleration enables faster results for more efficient computation and job turnaround times, delivering more license utilization for the same investment.**

**Recommended Tesla Configurations**

- **Workstation**
  - Tesla K20c
  - 4x-94GB memory
  - Dual CPUs

- **Server**
  - Tesla K20m/K20X
  - 96-128GB memory
  - Dual CPUs

**Test Drive Your SIMULIA Abaqus/Standard Jobs with GPUs for Free!**

www.accelerateabaqusongpu.com

---

**Simulia Abaqus/Standard**

**Reduce Engineering Simulation Times**

**NVIDIA and SIMULIA, the Dassault Systems brand for Realistic Simulation, have collaborated to deliver the power of GPU computing for Abaqus customers. Available in the Abaqus 6.11 release, NVIDIA GPU acceleration enables faster results for more efficient computation and job turnaround times, delivering more license utilization for the same investment.**

**Recommended Tesla Configurations**

- **Workstation**
  - Tesla K20c
  - 4x-94GB memory
  - Dual CPUs

- **Server**
  - Tesla K20m/K20X
  - 96-128GB memory
  - Dual CPUs

**Test Drive Your SIMULIA Abaqus/Standard Jobs with GPUs for Free!**

www.accelerateabaqusongpu.com
**REVOLUTIONISING HIGH PERFORMANCE COMPUTING | NVIDIA® TESLA®**

**Speed up simulations with NVIDIA GPU**

**ANSYS MECHANICAL**

**SPEED UP simulations WITH NVIDIA GPU**

With ANSYS® Mechanical™ 14.5 and NVIDIA® Professional GPUs, you can:
- Improve product quality with 5x more design simulations
- Accelerate time-to-market by reducing engineering cycles
- Develop high fidelity models with practical solution times

How much more could you accomplish if simulation times could be reduced from one day to just a few hours? As an engineer, you depend on ANSYS Mechanical to design high quality products efficiently. To get the most out of ANSYS Mechanical 14.5, simply upgrade your Quadro GPU or add a Tesla GPU to your workstation, or configure a server with Tesla GPUs, and instantly unlock the highest levels of ANSYS simulation performance.

**ACTIVATE A GPU WITH ANSYS HPC PACK**

To unlock the GPU feature in ANSYS Mechanical 14.5, you must have an ANSYS HPC Pack license, the same scheme also required for going parallel for greater than 2 CPU cores. For academic license users, the GPU capability is included with the base ANSYS Academic license that provides access to ANSYS Mechanical and no additional Academic HPC licenses are required.

**RECOMMENDED CONFIGURATIONS**

<table>
<thead>
<tr>
<th>Workstation</th>
<th>CPU only</th>
<th>CPU + GPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quadro M4000 GPU</td>
<td>312</td>
<td>163</td>
</tr>
<tr>
<td>Tesla K20c GPU</td>
<td>294</td>
<td>213</td>
</tr>
<tr>
<td>2 x 8-core x86 CPUs (16 cores total)</td>
<td>213</td>
<td></td>
</tr>
<tr>
<td>128 – 256GB DDR3 Memory</td>
<td></td>
<td></td>
</tr>
<tr>
<td>InfBand interconnect</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**ANSYS FLUENT**

**SPEED UP simulations WITH NVIDIA GPU**

NVIDIA has made investments to develop multigrid schemes specific for GPUs and offer solver-level libraries that can plug into an existing CFD code as a linear solver. The result is high quality implementations of a modern Algebraic Multigrid (AMG) method for solving linear scalar and coupled systems of equations. AMG is optimal for elliptic-type or coupled elliptic-dominant partial differential equations (PDEs) discretized over irregular grids. AMG is the linear solver standard in the commercial CFD software market.

The initial implementation of ANSYS Fluent 14.5 is designed to run on a single GPU and in the future on a single node with multiple GPUs, or across multiple nodes with GPUs on a cluster. The multi-GPU interconnect will support GPU-parallel applications via either OpenMP or MPI. It will also take advantage of GPUDirect from NVIDIA for distributed performance.

The ANSYS Fluent 14.5 release currently supports GPU acceleration of the coupled incompressible Navier-Stokes solver. And just as with the CPU-only AMG solver, there are solver settings that can be custom configured in order to achieve the optimal performance for a particular model and/or flow condition.

**Note** as in the study, the optimal solver settings for CPU-only and CPU+GPU are most often different.

For each model, GPUs accelerate the solver time by 2.4x over the 6-core CPU-only results. Those results display the current generation of CPU and GPU architectures, and in both cases these latest architectures provide improved performance. The ANSYS Fluent solver is designed in such a way that as either CPU or GPU architectures improve, so will overall ANSYS Fluent performance. It was demonstrated that substantial CSM and CFD performance gains can be achieved by using the latest GPU technology NVIDIA Tesla Kepler-series, with GPU acceleration of x86-based CPUs. Based on these trends, we can expect that GPU acceleration will be a meaningful HPC trend in the future of scientific computing and engineering modelling.

---

By Jeff Beisheim, Senior Software Developer, ANSYS, Inc.
The Nvidia Tesla for Life and Material Sciences Applications enable biophysicists and computational chemists to push the boundaries of their research. With NVIDIA Tesla, scientists can turn a standard PC into a “computational laboratory” capable of running complex biochemistry codes, in fields such as drug discovery and DNA sequencing, more than 10-20 times faster.

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, GAMESS, HOOMD, LAMMPS, NAMD, TeraChem (Quantum Chemistry), VMD
- Bio Informatics
  - CUDA-BLASTP, CUDA-EC, CUDA-MEME, CUDA-SW++ (Smith-Waterman), GPU-HMMER, MUMmerGPU

For more information, visit: www.nvidia.com/object/gpu-applications

**AMBER GPU PERFORMANCE**

- CPU only
- CPU + GPU

**NAMD V2.9 GPU PERFORMANCE**

- CPU only
- CPU + GPU

**AMBER AND NAMD 5X FASTER. TAKE A FREE AND EASY TEST DRIVE TODAY!**

www.nvidia.com/cluster

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

**AMBER**

AMBER is a leading molecular dynamics software that simulates the force fields for molecular dynamics of biomolecules. Since the release of Amber10 in April 2010, the code has been GPU-enabled to accelerate PMEDMD for both explicit solvent PME and implicit solvent GB simulations, resulting in raw performance of more than 58 ns/day for the JAC benchmark on 2 NVIDIA Kepler K10 GPU boards.

**NAMD**

The Team at University of Illinois at Urbana-Champaign (UUC) has been enabling CUDA-acceleration on NAMD since 2007. They have performed scaling experiments on the NCSA Tesla-based Lincoln cluster and demonstrated that 4 Tesla GPUs can outperform a cluster with 16 quad core CPUs.

**RECOMMENDED HARDWARE CONFIGURATION**

**Workstation**

- 1-2x Tesla K10, K20, K20X
- 2x 6-core CPUs
- 16 GB of RAM

**Cluster single node Configuration**

- 2x Tesla K10, K20, K20X
- 2x 6 core CPU / node
- Infiniband QDR / node
- 24 GB of RAM per node
LAMMPS

LAMMPS is a classical molecular dynamics code that models an ensemble of particles in a liquid, solid, or gaseous state. It can model atoms, or more generically, as a parallel particle simulator at the atomic, meso, or continuum scale; i.e., polymeric, biological, metallic, granular, and coarse-grained systems using a variety of force fields and boundary conditions. Over a dozen potential models have been accelerated on the GPU and more are coming.

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 or Single Node
- 1-2x Tesla K10, K20, K20X (6GB)
- 2x 6-core SandyBridge
- 32 GB System Memory

NVIDIA GPUs ACCELERATE WEATHER SIMULATION

ASUCA (WEATHER MODELLING)
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.ecmwf.int/newsevents/meetings/workshops/2012/high_performance_computing_15th/Presentations/pdf/Govett.pdf
http://www.ecmwf.int/newsevents/meetings/workshops/2012/high_performance_computing_15th/Presentations/pdf/Henderson.pdf

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.

GROmacS GPU PERFORMANCE

GROmacS 4.6 Final Release Waters (192K atoms)

NIM PERFORMANCE

Cpus vs. GPs vs. 225km, 96-level

ENERGY EFFICIENT SUPERCOMPUTING

Model: HRRR — cONUS, Res: 3KM, freq: hourly

Chip Performance (Gflops)
OIL & GAS: ACCELERATING EXPLORATION REVOLUTIONISING SEISMIC IMAGING WITH NVIDIA GPUs

The oil and gas industry is steadily increasing its use of more computationally-intensive seismic imaging algorithms, ranging from Kirchhoff time and depth migration through to Reverse Time Migration and Full Waveform Inversion, to model subsurface structures and identify hydrocarbon traps. These applications are seeing significant acceleration using GPUs, and Tesla GPUs have been deployed in production to accelerate seismic imaging at international oil companies and major seismic contractors worldwide.

ACCELERATING SEISMIC VOLUME RENDERING & EXTRACTION WITH NVIDIA GPGPUs

Using GPU technology, leading interpretation and modeling software enables the interpreter to blend and render multiple seismic volumes interactively. Detecting anomalies, delineating structural and stratigraphic features, isolating areas of interest and then extracting geobodies can then be included directly in the 3D model, bridging the gap between volume-based interpretation and modeling software. Combining seismic attributes and volume-based interpretation is the fastest way to interrogate and analyze data for complex structures and strategic features. Interpretation software using GPU technology improves productivity for sharper, faster results. Complex volume attributes are generated on the fly and parameters are fine-tuned using new interactive controls. Geological bodies are isolated and extracted from the seismic and can be directly sampled into the model, so interpreters are able to isolate the reservoir bodies more quickly and intuitively.

ACCELERATED SEISMIC MIGRATION PROCESSING

Accelware Corporation’s seismic-data acceleration solution provides a significant performance increase, to reduce processing times and deliver faster results in order to make drilling decisions. The software solution uses the massively-parallel processing capabilities of graphics-processing-unit (GPU) accelerators to increase the throughput efficiency of data centers. This technique is designed for 2D and 3D land and marine data for which a typical processing job can run from several days to months. Accelerating the computation by this method can shorten turnaround times and increase confidence while handling larger data sets of increasing complexity.

WHAT THE EXPERTS ARE SAYING

“Paradigm actively seeks to take advantage of new hardware technologies, such as NVIDIA Maximus technology, to improve our customers’ performance and efficiency. By significantly accelerating the calculation of seismic trace attributes, interpreters can evaluate new workflows and practices to gain more from their seismic data investment.”

Laura Evins
Product manager of seismic attributes at Paradigm

“Geological Expression is all about maximising the amount of geological information that you can rapidly extract from 3D seismic. We’ve been working across multiple Quadro and Tesla GPUs in our workstations to do this for some time. Now, with NVIDIA Maximus certification for GeoTeric, and the additional supporting technology that comes with it, it is much easier for our clients to procure the Geological Expression platforms that are going to take their interpretation workflows to new levels of productivity.”

Steve Purves
Director of Technology at IFA

INCREASED VOLUME INSIGHT

Combining seismic attributes and volume-based interpretation is the fastest way to interrogate and analyze data for complex structures and strategic features. Interpretation software using GPU technology improves productivity for sharper, faster results. Complex volume attributes are generated on the fly and parameters are fine-tuned using new interactive controls. Geological bodies are isolated and extracted from the seismic and can be directly sampled into the model, so interpreters are able to isolate the reservoir bodies more quickly and intuitively.

THE CHALLENGE

Every energy company has a goal to make the best of their large drilling budgets. Drilling wells in the right place, and with accurate knowledge of the associated risks, really impacts the bottom line. When improved accuracy leads to more pay zone and less uncertainty about the production process, the ROI improves significantly.

Paradigm software aims at giving customers as much information as possible out of an existing seismic dataset, where some contractors may be advising the customer to shoot new data for better resolution. Tangible evidence for this is the ability to drill more productive feet in a well (GeoSteering, accurate models, fine-tuned depth conversion, better characterization leading to drilling in sweet spots).

HSE is a critical topic, where nothing can be left out. Avoiding overpressure zones, identifying fracture areas that could affect drilling all these factors should be taken into consideration, carefully processed and visualized with maximum detail. Working efficiently to reduce project turnaround without compromising quality has a definite impact on customer performance. It also implies that the data volumes are growing and with the users’ need to see more detail comes increasingly compute intensive data processing.

THE IMPACT

GPU rendering delivers 10x to 40x acceleration, increasing the massive data crunching efficiency by an order of magnitude. Users get a much more realistic rendering in 3D, taking voxel visualization to the next level. Porting the key processes to NVIDIA GPU architecture has brought about a new rendering process with significantly improved spatial perception and real-time computation of key attributes. Further integration of the CUDA – enabled code to the interpretation suite has produced a notable improvement in display quality, making it easier to perceive depth and volume.

As a result VoxelGeo remains the reference voxel product and the product of choice for industry leaders.
Revolutionising high performance computing | nvidia ® tesla ®

CASE STUDIES:

Numerical Analytics, HPC

GPU COMPUTING IN FINANCE — CASE STUDIES

Bloomberg: GPUs increase accuracy and reduce processing time for bond pricing

Bloomberg implemented an NVIDIA Tesla GPU computing solution in their datacenter. By porting their application to run on the NVIDIA CUDA parallel processing architecture, Bloomberg received dramatic improvements across the board. As Bloomberg customers make crucial 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.

Bloomberg: GPUs increase accuracy and reduce processing time for bond pricing

48 GPUs
144K
$14K/year
$1.2 Million/year

2000 CPUs
28x Lower Cost
31K/year
$4 Million

4x2 Lower Space

NVIDIA TESLA GPUs USED BY J.P. MORGAN TO RUN RISK CALCULATIONS IN MINUTES, NOT HOURS

The Challenge
Risk management is a huge and increasingly costly focus for the financial services industry. A cornerstone of J.P. Morgan’s cost-reduction plan involves accelerating its risk library. It is imperative to reduce the total cost of ownership of J.P. Morgan’s risk-management platform and create a leap forward in the speed with which client requests can be serviced.

The Solution
J.P. Morgan’s Equity Derivatives Group added NVIDIA® Tesla M2070 GPUs to its data centers. More than half the equity derivative-focused risk computations run by the bank have been moved to running on hybrid GPU/CPU-based systems. NVIDIA Tesla GPUs were deployed in multiple data centers across the bank’s global offices. J.P. Morgan was able to seamlessly share the GPUs between tens of global applications.

The Impact
Utilizing GPUs has accelerated application performance by 40X and delivered over 80 percent savings, enabling greener data centers that deliver higher performance for the same power. For J.P. Morgan, this is game-changing technology, enabling the bank to calculate risk across a range of products in a matter of minutes rather than overnight. Tesla GPUs give J.P. Morgan a significant market advantage.

The solution
J.P. Morgan’s Equity Derivatives Group added NVIDIA® Tesla M2070 GPUs to its data centers. More than half the equity derivative-focused risk computations run by the bank have been moved to running on hybrid GPU/CPU-based systems. NVIDIA Tesla GPUs were deployed in multiple data centers across the bank’s global offices. J.P. Morgan was able to seamlessly share the GPUs between tens of global applications.

The impact
Utilizing GPUs has accelerated application performance by 40X and delivered over 80 percent savings, enabling greener data centers that deliver higher performance for the same power. For J.P. Morgan, this is game-changing technology, enabling the bank to calculate risk across a range of products in a matter of minutes rather than overnight. Tesla GPUs give J.P. Morgan a significant market advantage.

GPUs ACCELERATE HPC IN EUROPE

NVIDIA KEPLER BASED HYBRID SYSTEM AT CSCS

It features 272 nodes, each one equipped with 16-core AMD Opteron CPU, 32 GB DDR3 memory and one NVIDIA Tesla K20X GPU with 6 GB of GDDR5 memory for a total of 4352 cores and 272 GPUs.

On Nov 14th, the new “Green500” list was published, it ranks the CSCS Cray XK7 Todi as the fourth most energy-efficient system in the world.

The Cray XK7 (Tod) is the first hybrid supercomputing system with high scalability at CSCS based on NVIDIA Kepler GPUs, designed to run data massively parallel computationally intensive applications.

ITALY’S TOP CLUSTER LEAPFROGS CURRENT GREEN500 #1 SYSTEM

italy’s Eurora supercomputer — which uses NVIDIA® Tesla® GPU accelerators based on NVIDIA Kepler™, the world’s fastest and most efficient high-performance computing (HPC) architecture — has set a new record for data centres energy efficiency.

Pairing NVIDIA Tesla K20 GPUs with Eurotech’s Aurora Hot Water Cooling technology, the Eurora system is more efficient and compact than conventional air-cooled solutions. HPC systems based on the Eurora hardware architecture, including the Eurotech Aurora Tigon, enable data centres to potentially reduce energy bills by up to 50 percent and reduce total cost of ownership by 30-50 percent.

CINECA
Battelle-Teesside
C-DAC, Greenest CPU System
ITCS Brainbox—Greenest Brain System

NVIDIA KEPLER BASED HYBRID SYSTEM AT CSCS

NVIDIA KEPLER BASED HYBRID SYSTEM AT CSCS

The Cray XK7 (Tod) is the first hybrid supercomputing system with high scalability at CSCS based on NVIDIA Kepler GPUs, designed to run data massively parallel computationally intensive applications.

NVIDIA KEPLER BASED HYBRID SYSTEM AT CSCS

The Cray XK7 (Tod) is the first hybrid supercomputing system with high scalability at CSCS based on NVIDIA Kepler GPUs, designed to run data massively parallel computationally intensive applications.

CINECA
Battelle-Teesside
C-DAC, Greenest CPU System
ITCS Brainbox—Greenest Brain System

NVIDIA KEPLER BASED HYBRID SYSTEM AT CSCS

The Cray XK7 (Tod) is the first hybrid supercomputing system with high scalability at CSCS based on NVIDIA Kepler GPUs, designed to run data massively parallel computationally intensive applications.

CINECA
Battelle-Teesside
C-DAC, Greenest CPU System
ITCS Brainbox—Greenest Brain System
As the demand for high performance parallel computing increases across many areas of science, medicine, engineering, and finance, NVIDIA continues to innovate and meet that demand with extraordinarily powerful GPU computing architectures. NVIDIA's existing Fermi GPUs have already redefined and accelerated High Performance Computing (HPC) capabilities in areas such as seismic processing, biochemistry simulations, weather and climate modeling, signal processing, computational finance, computer aided engineering, computational fluid dynamics, and data analysis. NVIDIA's new Kepler GK110 GPU raises the parallel computing bar considerably and will help solve the world's most difficult computing problems. By offering much higher processing power than the prior GPU generation and by providing new methods to optimize and increase parallel workload execution on the GPU, Kepler GK110 simplifies creation of parallel programs and will further revolutionize high performance computing.

Comprising 7.1 billion transistors, Kepler GK110 is not only the fastest, but also the most architecturally complex microprocessor ever built. Adding many new innovative features focused on compute performance, GK110 was designed to be a parallel processing powerhouse for Tesla® and the HPC market.

Kepler GK110 provides over 1 Tflop of double precision throughput with greater than 80% DGEMM efficiency versus 60-65% on the prior Fermi architecture.

We designed Kepler with an eye towards three things: performance, efficiency and accessibility. It represents an important milestone in GPU-accelerated computing and should foster the next wave of breakthroughs in computational research.

Jonah Albun, Senior VP of GPU Engineering and principal architect of Kepler at NVIDIA

Fermi was a major step forward in computing, Kepler will be equally disruptive, establishing GPUs broadly into technical computing, due to their ease of use, broad applicability and efficiency.”

Bill Dally, chief scientist and Senior VP of research at NVIDIA
NVIDIA® TESLA® FERMI
TESLA C-CLASS AND TESLA MODULES

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.

<table>
<thead>
<tr>
<th>FEATURE</th>
<th>TESLA C2075*</th>
<th>TESLA M2090</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Cores</td>
<td>448</td>
<td>512</td>
</tr>
<tr>
<td>Caches</td>
<td>64 KB L1 cache + Shared Memory / 32 cores, 768 KB L2 cache</td>
<td>64 KB L1 cache + Shared Memory / 32 cores, 768 KB L2 cache</td>
</tr>
<tr>
<td>Peak double precision for floating point</td>
<td>515 GFlops</td>
<td>665 GFlops</td>
</tr>
<tr>
<td>Floating point point performance</td>
<td>1038 GFlops</td>
<td>1331 GFlops</td>
</tr>
<tr>
<td>GPU Memory</td>
<td>6 GB GDDR5 ECC / 5.25 GB with ECC on</td>
<td>6 GB GDDR5 ECC / 5.25 GB with ECC on</td>
</tr>
<tr>
<td>Memory Bandwidth</td>
<td>144 GB/s ECC off / 115 GB/s ECC on</td>
<td>177.6 GB/s / 120 GB/s / 160 GB/s</td>
</tr>
<tr>
<td>Positioning</td>
<td>Best price/performance solutions for double precision codes and when ECC memory required</td>
<td>Best price/performance solutions for double precision codes and when ECC memory required</td>
</tr>
<tr>
<td>Power Consumption</td>
<td>200W</td>
<td>225W</td>
</tr>
</tbody>
</table>

*Extended ECC reporting and Dynamic power scaling capability available.

TECHNICAL SPEC COMPARISON: M2090 VS TESLA K10

<table>
<thead>
<tr>
<th>PRODUCT NAME</th>
<th>TESLA M2090</th>
<th>TESLA K10</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPU Architecture</td>
<td>Fermi</td>
<td>Kepler GK104</td>
</tr>
<tr>
<td># of GPUs</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Single Precision Flops</td>
<td>1.3 TF / 0.66 TF</td>
<td>4.58 TF / 0.99 TF</td>
</tr>
<tr>
<td>Double Precision Flops</td>
<td>0.66 TF / 0.190 TF</td>
<td>2.92 TF / 0.095 TF</td>
</tr>
<tr>
<td># CUDA Cores</td>
<td>512</td>
<td>3072 / 1536</td>
</tr>
<tr>
<td>Memory size</td>
<td>6 GB</td>
<td>8 GB / 4 GB</td>
</tr>
<tr>
<td>Memory BW (ECC off)</td>
<td>177.6 GB/s</td>
<td>120 GB/s / 160 GB/s</td>
</tr>
<tr>
<td>PCI-EXPRESS</td>
<td>Gen 2 (Gen 2 compatible)</td>
<td>Gen 3 (Gen 2 compatible)</td>
</tr>
<tr>
<td>Board Power</td>
<td>225 watts</td>
<td>225 watts</td>
</tr>
</tbody>
</table>

K10 FOR OIL & GAS
- 1.8X simulations per day for more accurate models
- Less risk and higher confidence
- 2X GPUs within the same footprint

K10 FOR DEFENSE
- 1.9X calculations per day for more accurate models
- Faster analysis and confident decisions
- 2X GPUs within the same footprint

K10 FOR LIFE SCIENCES
- 2.2X simulations for key MD applications
- Bigger experiments on smaller departmental clusters
- 2X GPUs within the same footprint

SEISMIC PROCESSING

INTERGER/LOGICAL OPERATIONS

www.nvidia.eu/tesla

Optimized for single precision applications, the Tesla K10 includes two ultra-efficient GK104 Kepler GPUs to deliver high throughput. It delivers up to 2x the performance for single precision applications compared to the previous generation Tesla M2090 GPU in the same power envelope. With an aggregate performance of 4.58 teraflop peak single precision and 320 gigabytes per second memory bandwidth for both GPUs put together, the Tesla K10 is optimized for computations in seismic, signal/image processing, and video analytics.
NVIDIA® TESLA® KEPLER
TESLA K20 AND K20X GPU ACCELERATORS MODULES

Designed to be the performance leader in double precision applications and the broader supercomputing market, the Tesla K20 and K20X GPU Accelerators deliver 10x the performance of a single CPU. Tesla K20 and K20X both feature a single GK110 Kepler GPU that includes the Dynamic Parallelism and Hyper-Q features. With more than one teraflop peak double precision performance, these GPU accelerators are ideal for the most aggressive high-performance computing workloads including climate and weather modeling, CFD, CAE, computational physics, biochemistry simulations, and computational finance.

www.nvidia.eu/tesla

TECHNICAL SPEC

<table>
<thead>
<tr>
<th>FEATURE</th>
<th>TESLA K20</th>
<th>TESLA K20X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak double precision floating point performance</td>
<td>1.17 Tflops</td>
<td>1.31 Tflops</td>
</tr>
<tr>
<td>Peak single precision floating point performance</td>
<td>3.52 Tflops</td>
<td>3.95 Tflops</td>
</tr>
<tr>
<td>Number of GPUs</td>
<td>1 x GK110</td>
<td></td>
</tr>
<tr>
<td>CUDA cores</td>
<td>2496</td>
<td>2688</td>
</tr>
<tr>
<td>Memory size per board (GDDR5)</td>
<td>5 GB</td>
<td>6 GB</td>
</tr>
<tr>
<td>Memory bandwidth for board (ECC off)</td>
<td>208 GB/sec</td>
<td>250 GB/sec</td>
</tr>
<tr>
<td>GPU Computing Applications</td>
<td>Seismic processing, CFD, CAE, Financial computing, Computational chemistry and Physics, Data analytics, Satellite imaging, Weather modelling</td>
<td></td>
</tr>
<tr>
<td>Architecture Features</td>
<td>SMX, Dynamic Parallelism, Hyper-Q</td>
<td></td>
</tr>
<tr>
<td>System</td>
<td>Servers and Workstations</td>
<td>Servers only</td>
</tr>
</tbody>
</table>

PERFORMANCE ON LEADING SCIENTIFIC APPLICATIONS

The NVIDIA® Tesla® K20 graphics processing unit (GPU) active accelerator board is a PCI Express, dual-slot full height (4.376 inches by 10.5 inches) form factor computing module comprising of a single GK110 GPU. The Tesla K20 active accelerator is designed for workstations and offers a total of 5 GB of GDDR5 on-board memory and supports PCI Express Gen2.

www.nvidia.eu/tesla

TESLA PERSONAL SUPERCOMPUTING GPUs

NVIDIA® Tesla® GPU Accelerators turn standard PCs and workstations into personal supercomputers. Powered by NVIDIA CUDA® — the world’s most pervasive parallel-computing model Tesla GPU Accelerators for workstations deliver cluster level performance right at your desk.

Tesla workstation products are supported under:

- Windows (Windows XP-c2075 only), Windows vista, Windows 7 — 32-bit (c2075 only) and 64-bit
- Linux 32-bit and 64-bit
- Fedora 12
- RHEL 5.4 Desktop
- Ubuntu 9.10 Desktop
- RHEL 4.8 Desktop (64-bit only)
- RHEL 6
- OpenSUSE 11.2
- SLED 11

www.nvidia.eu/tesla

TECHNICAL SPEC COMPARISON: TESLA K20 VS TESLA C2075

<table>
<thead>
<tr>
<th>FEATURE</th>
<th>TESLA K20</th>
<th>TESLA C2075</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak double precision floating point performance</td>
<td>1.17 Tflops</td>
<td>515 Gflops</td>
</tr>
<tr>
<td>Peak single precision floating point performance</td>
<td>3.52 Tflops</td>
<td>1030 Gflops</td>
</tr>
<tr>
<td>CUDA cores</td>
<td>2496</td>
<td>448</td>
</tr>
<tr>
<td>Memory size per board (GDDR5)</td>
<td>5 GB</td>
<td>6 GB</td>
</tr>
<tr>
<td>Memory bandwidth for board (ECC off)</td>
<td>208 GB/sec</td>
<td>250 GB/sec</td>
</tr>
<tr>
<td>GPU Computing Applications</td>
<td>Reservoir simulation, CAE (structural analysis), Molecular dynamics, Numerical analytics, Computational visualisation (ray tracing)</td>
<td></td>
</tr>
</tbody>
</table>

www.nvidia.eu/tesla

TRY THE TESLA GPU TEST DRIVE TODAY!

A ready to use solution for science and research.

www.nvidia.eu/cluster
PNy Technologies has over 25 years of excellence in the manufacturing of a full spectrum of high-quality products for everything in and around the computer. PNY Technologies has a long history of providing designers, engineers and developers with cutting-edge NVIDIA® Quadro™ and Tesla® Solutions. PNY understands the needs of its professional clients offering professional technical support and a constant commitment to customer satisfaction.

Based on the revolutionary architecture of massively parallel computing NVIDIA® CUDA™, the NVIDIA® Tesla® solutions by PNY are designed for high performance computing (HPC) and offer a wide range of development tools. In 2012 PNY Technologies enhances its presence and offer in the HPC market becoming the European distributor of Tyan™ servers based on NVIDIA® Tesla® processors.

PNY Professional Solutions are offered in cooperation with qualified distributors, specialty retailers, computer retailers, and system integrators. To find a qualified PNY Technologies partner visit: www.pny.eu/wheretobuy.php

WHERE TO BUY
Offering pre- and post- sales assistance, three year standard warranty, professional technical support, and an unwavering commitment to customer satisfaction, our partners and customers experience firsthand why PNY is considered a market leader in the professional industry.

Based on the revolutionary architecture of massively parallel computing NVIDIA® CUDA™, the NVIDIA® Tesla® solutions by PNY are designed for high performance computing (HPC) and offer a wide range of development tools.

PNY Technologies offers a 3 year manufacturer warranty on all Tesla based systems in accordance to PNY Technologies Terms of Guarantee.

SUPPORT
PNY offers individual support as well as comprehensive online support. Our support websites provide FAQs, the latest information and technical data sheets to ensure you receive the best performance from your PNY product.

SUPPORT CONTACT INFORMATION
Business hours: 9:00 a.m. - 5:00 p.m.
Germany, Austria, Switzerland, Russia, Eastern Europe
Email: vertrieb@pny.eu
Hotline Support: +49 (0)2405/40848-40

United Kingdom, Denmark, Sweden, Finland, Norway
Email: kvadrouk@pny.eu
Hotline Presales: +49 (0)2405/40848-55

France, Belgium, Netherlands, Luxembourg
Email: sales@pny.eu
Hotline Presales: +33 (0)1055613-7575

NVIDIA TESLA AND CUDA LINKS

NEWS AND SUCCESS STORIES
NVIDIA GPU Computing on Twitter
twitter.com/gpucomputing
CUDA weekly newsletter
www.nvidia.com/object/cuda_week_in_review_newsletter
News and articles
www.nvidia.co.uk/page/tesla-articles
Tesla Video on YouTube
www.youtube.com/nvidiatalesa
Success stories
www.nvidia.co.uk/page/tesla_testimonials

PNY Twitter
twitter.com/PNYProukUK (English)
twitter.com/PNYProuDE (German)
twitter.com/PNYProuFR (French)

DEVELOPMENT TOOLS
CUDA Zone
www.nvidia.co.uk/cuda
CUDA in Action
www.nvidia.co.uk/object/cuda_in_action
CUDA Books
www.nvidia.co.uk/object/cuda_books
CUDA Training & Consulting
www.nvidia.co.uk/page/cuda_consultants.html
Order CUDA course
www.parallel-compute.com

OpenACC Directives
www.nvidia.eu/openacc
Software Development Tools
www.nvidia.co.uk/object/tesla_software

HARDWARE
NVIDIA High Performance Computing
www.nvidia.eu/tesla
Personal Supercomputers with Tesla
www.nvidia.co.uk/psc

Tesla Data Center Solutions
www.nvidia.co.uk/page/preconfigured_clusters
Tesla Products Technical Descriptions
www.nvidia.co.uk/page/tesla_product_literature

SOFTWARE
CUDA Zone
www.nvidia.co.uk/cuda
CUDA in Action
www.nvidia.co.uk/object/cuda_in_action
CUDA Books
www.nvidia.co.uk/object/cuda_books
CUDA Training & Consulting
www.nvidia.co.uk/page/cuda_consultants.html
Order CUDA course
www.parallel-compute.com

OpenACC Directives
www.nvidia.eu/openacc
Software Development Tools
www.nvidia.co.uk/object/tesla_software

WHERE TO BUY
Offering pre- and post- sales assistance, three year standard warranty, professional technical support, and an unwavering commitment to customer satisfaction, our partners and customers experience firsthand why PNY is considered a market leader in the professional industry.

Based on the revolutionary architecture of massively parallel computing NVIDIA® CUDA™, the NVIDIA® Tesla® solutions by PNY are designed for high performance computing (HPC) and offer a wide range of development tools.

PNY Professional Solutions are offered in cooperation with qualified distributors, specialty retailers, computer retailers, and system integrators. To find a qualified PNY Technologies partner visit: www.pny.eu/wheretobuy.php

Germany, Austria, Switzerland, Russia, Eastern Europe
Email: vertrieb@pny.eu
Hotline Support: +49 (0)2405/40848-40

United Kingdom, Denmark, Sweden, Finland, Norway
Email: kvadrouk@pny.eu
Hotline Presales: +49 (0)2405/40848-55

France, Belgium, Netherlands, Luxembourg
Email: sales@pny.eu
Hotline Presales: +33 (0)1055613-7575

WARRANTY
PNY Technologies offers a 3 year manufacturer warranty on all Tesla based systems in accordance to PNY Technologies Terms of Guarantee.

SUPPORT
PNY offers individual support as well as comprehensive online support. Our support websites provide FAQs, the latest information and technical data sheets to ensure you receive the best performance from your PNY product.

SUPPORT CONTACT INFORMATION
Business hours: 9:00 a.m. - 5:00 p.m.
Germany, Austria, Switzerland, Russia, Eastern Europe
Email: vertrieb@pny.eu
Hotline Support: +49 (0)2405/40848-40

United Kingdom, Denmark, Sweden, Finland, Norway
Email: kvadrouk@pny.eu
Hotline Presales: +49 (0)2405/40848-55

France, Belgium, Netherlands, Luxembourg
Email: sales@pny.eu
Hotline Presales: +33 (0)1055613-7575

NVIDIA TESLA AND CUDA LINKS

NEWS AND SUCCESS STORIES
NVIDIA GPU Computing on Twitter
twitter.com/gpucomputing
CUDA weekly newsletter
www.nvidia.com/object/cuda_week_in_review_newsletter
News and articles
www.nvidia.co.uk/page/tesla-articles
Tesla Video on YouTube
www.youtube.com/nvidiatalesa
Success stories
www.nvidia.co.uk/page/tesla_testimonials

PNY Twitter
twitter.com/PNYProukUK (English)
twitter.com/PNYProuDE (German)
twitter.com/PNYProuFR (French)

DEVELOPMENT TOOLS
CUDA Zone
www.nvidia.co.uk/cuda
CUDA in Action
www.nvidia.co.uk/object/cuda_in_action
CUDA Books
www.nvidia.co.uk/object/cuda_books
CUDA Training & Consulting
www.nvidia.co.uk/page/cuda_consultants.html
Order CUDA course
www.parallel-compute.com

OpenACC Directives
www.nvidia.eu/openacc
Software Development Tools
www.nvidia.co.uk/object/tesla_software

HARDWARE
NVIDIA High Performance Computing
www.nvidia.eu/tesla
Personal Supercomputers with Tesla
www.nvidia.co.uk/psc

Tesla Data Center Solutions
www.nvidia.co.uk/page/preconfigured_clusters
Tesla Products Technical Descriptions
www.nvidia.co.uk/page/tesla_product_literature

SOFTWARE
CUDA Zone
www.nvidia.co.uk/cuda
CUDA in Action
www.nvidia.co.uk/object/cuda_in_action
CUDA Books
www.nvidia.co.uk/object/cuda_books
CUDA Training & Consulting
www.nvidia.co.uk/page/cuda_consultants.html
Order CUDA course
www.parallel-compute.com

OpenACC Directives
www.nvidia.eu/openacc
Software Development Tools
www.nvidia.co.uk/object/tesla_software