WHITE PAPER

NVIDIA® Tesla™ GPU Computing

Revolutionizing High-Performance Computing

TABLE OF CONTENTS

1.0 GPUs are Revolutionizing Computing 1

2.0 Why GPU Computing?........ 1

3.0 Parallel Acceleration 2

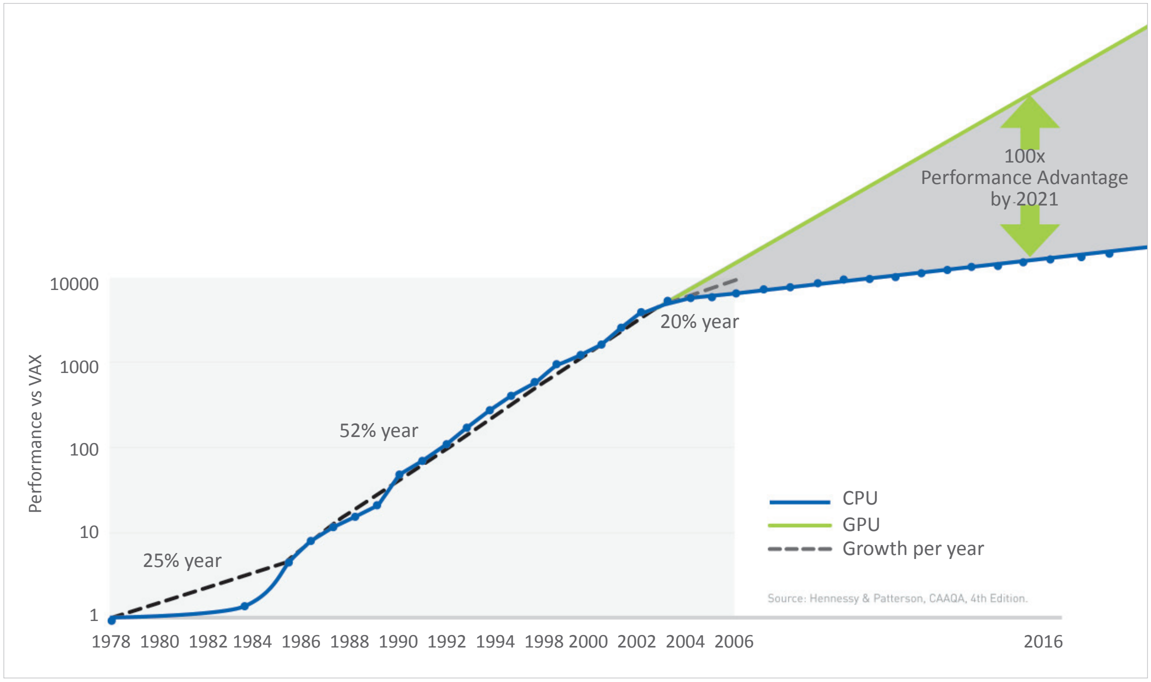
4.0 CUDA Parallel Computing Architecture........ 3

5.0 SGI® GPU Compute Solutions........ 4

1.0 GPUs are Revolutionizing Computing

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 Graphics Processing Unit (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.



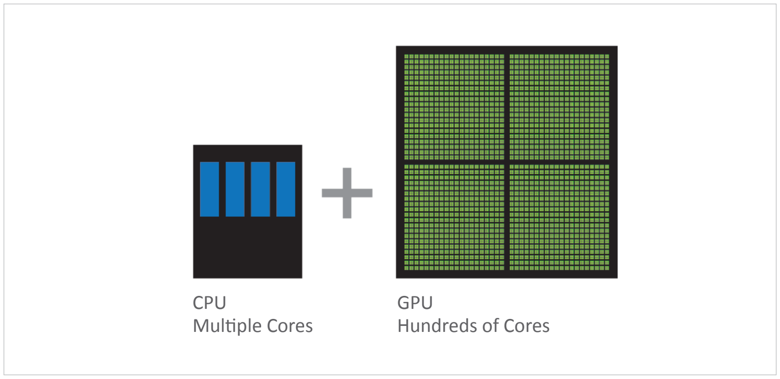
Source: Hennessy & Patterson, CAAQA, 4th Edition.

The use of GPUs for computation is a dramatic shift in HPC. GPUs deliver performance increases of 10x to 100x to solve problems in minutes instead of hours, outpacing the performance of traditional computing with x86-based CPUs alone. In addition, GPUs also deliver greater performance per watt of power consumed. From climate modeling to 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.

2.0 Why GPU Computing?

With the ever-increasing demand for more computing performance, the HPC industry is moving toward a hybrid computing model, where GPUs and CPUs work together to perform general purpose computing tasks.

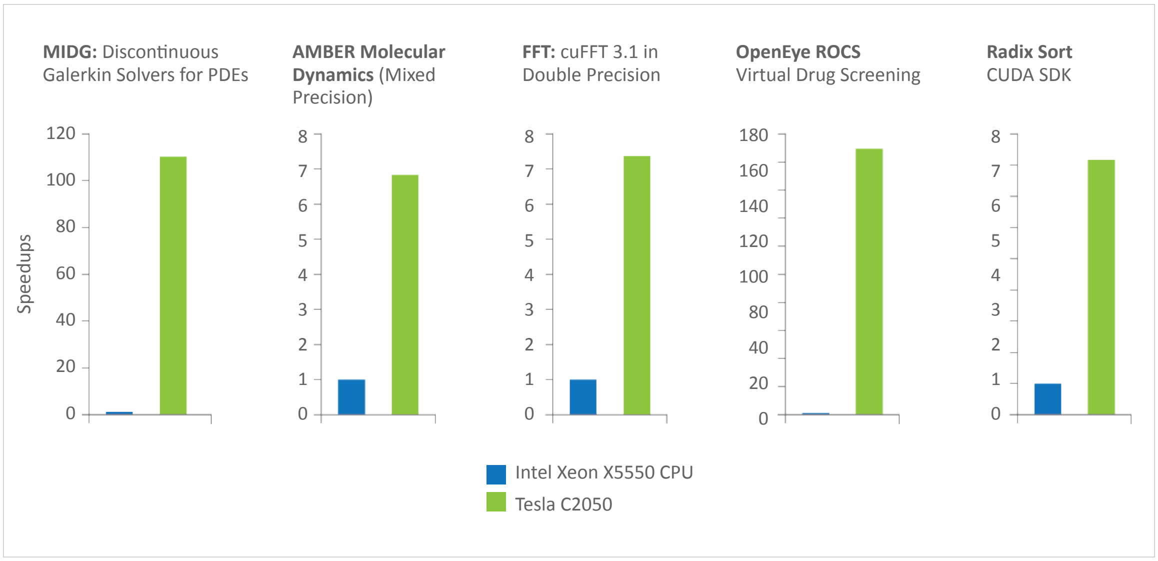
As parallel processors, GPUs excel at tackling large amounts of similar data because the problem can be split into hundreds or thousands of pieces and calculated simultaneously. As sequential processors, CPUs are not designed for this type of computation, but they are adept at more serial-based tasks such as running operating systems and organizing data. NVIDIA's GPU solutions outpace others as they apply the most relevant processor to the specific task in hand.



3.0 Parallel Acceleration

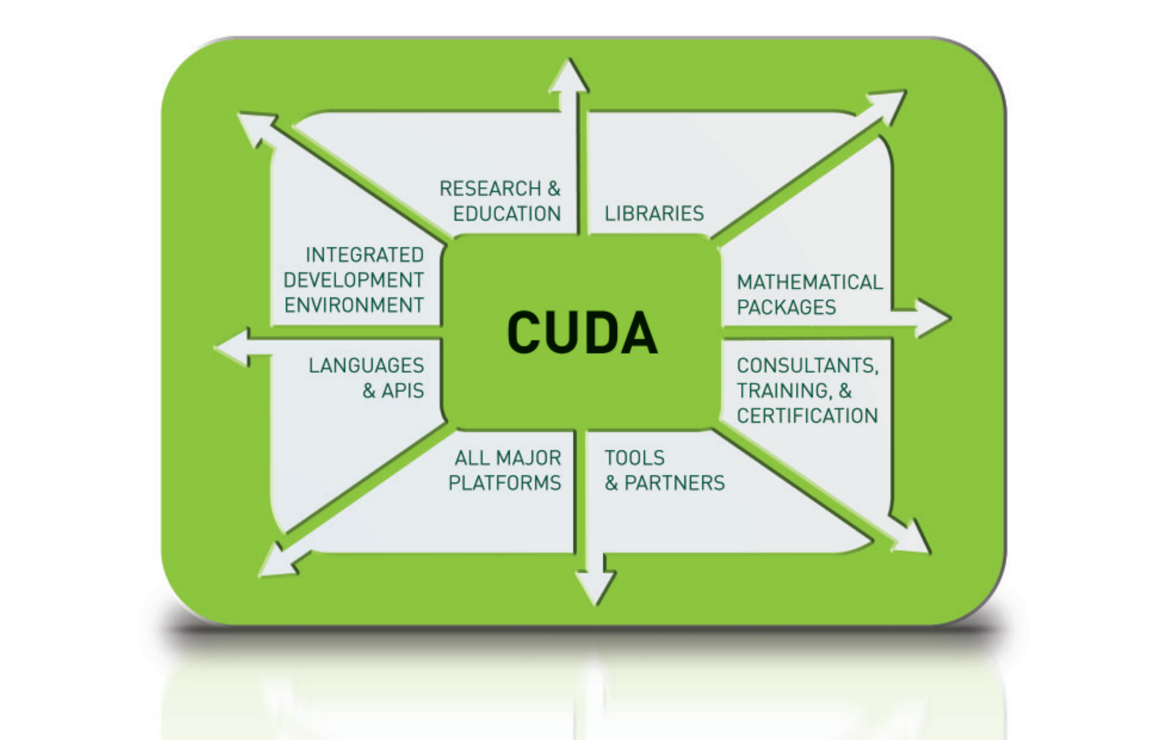
Multi-core programming with x86 CPUs is difficult and often results in marginal performance gains when going from 1 core to 4 cores to 16 cores. Beyond 4 cores, memory bandwidth becomes the bottleneck to further performance increases.

To harness the parallel computing power of GPUs, programmers can simply modify the performance critical portions of an application to take advantage of the hundreds of parallel cores in the GPU. The rest of the application remains the same, making the most efficient use of all cores in the system. Running a function on the GPU involves rewriting that function to expose its parallelism, then adding a few new function-calls to indicate which functions will run on the GPU or the CPU. With these modifications, the performance-critical portions of the application can now run significantly faster on the GPU.

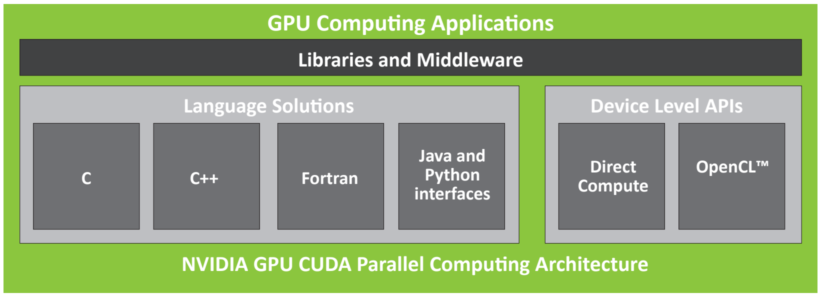


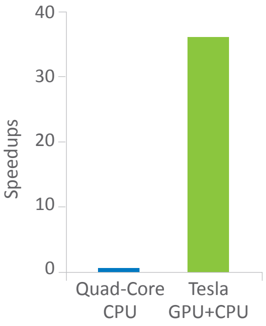
4.0 CUDA® Parallel Computing Architecture

CUDA® is NVIDIA's parallel computing architecture. 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 latest 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 proc essing 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.


GPU Computing Applications
Libraries and Middleware
Language Solutions
C
C++
Fortran
Java and Python interfaces
Device Level APIs
Direct Compute
OpenCL™
NVIDIA GPU CUDA Parallel Computing Architecture



5.0 SGI® GPU Compute Solutions

Accelerating Results with GPU Compute Solutions

SGI leads the industry in delivering application-specific acceleration, dating back to the Geometry Engine™ which accelerated graphics applications in the 1980s. SGI then co-developed the SGI Tensor Processing Unit (TPU), followed by RASC™ technology, FPGA's that were tightly-coupled to our shared memory architecture. With RASC technology, SGI created the world's largest single system image server with accelerators, to solve the most challenging life-sciences problems. With the backing of a team of application experts, SGI is in a unique position to help customers solve problems with GPU computing technology. SGI has services and support personnel ready to help customers port and debug specific applications.

Deskside High-performance Computing

SGI® Octane III: SGI Octane III offers the capabilities of a high-performance cluster with the portability and usability of a workstation. It is available as a single or dual-node graphics workstation with support for the fastest NVIDIA® Quadro® professional graphics and compute GPU cards. Octane III is optimized for use with multiple display solutions for advanced visual computing scenarios.

Workgroup to Enterprise

SGI® Rackable™ Servers: Leveraging the winning combination of the latest Intel® Xeon® Processor architecture, NVIDIA graphics and Tesla™ 10 and 20 -series solutions, these servers deliver top value and performance. Rackable servers also offer the Design-to-Order model for great flexibility in a variety of chassis and motherboard formats.

Supercomputer

Altix® UV: Customers trying to solve the world's toughest computational challenges independent of the typical limits of CPU, memory and I/O inherent in most twin-socket or even quad-socket designs will find that the SGI Altix UV platform will exceed their needs. The Altix UV platform brings GPUs to a new class of solutions in chemistry, homeland defense, fluid dynamics and biosciences. The Center for Remote Data Analysis and Visualization at the University of Tennessee recently installed Altix UV 1000 system with 128 CPUs, 4 TB of main memory and 8 NVIDIA GPUs to enhance the capabilities of the National Science Foundation (NSF) to 'see and understand' large volumes of data produced on the NSF's TeraGrid.

Altix® ICE: For customers who want to manage large scale-out HPC environments that include GPUs, the SGI Altix ICE 8400 platform offers the ability to integrate service nodes containing GPUs into dual-plane, high-bandwidth, low-latency InfiniBand networking topologies. With the assistance of the SGI Professional Services team, SGI has implemented some of the largest hybrid clusters in the world by combining NVIDIA GPUs with the Altix ICE platform.

Accelerating Customer Results

SGI has been working with the Irish Center for High-End Computing (ICHEC) on accelerating applications for GPUs. Founded in 2005, ICHEC was established as a national high performance computing (HPC) provider with offices in Dublin and Galway, Ireland. "Our focus has been placed firmly on developing new algorithms capable of exploiting the considerable power of GPUs. For example, recent results from our port of the DL_POLY MD application suggest that this technology indeed has the potential to make HPC accessible to a broader audience. We expect similar success with our ongoing port of Quantum Espresso," said James Slevin, director of ICHEC. "By partnering with SGI, ICHEC will provide business, scientific and academic customers with a high level of consultancy to support the development of effective computing solutions."

Services and Support

SGI has a team of GPU experts who have ported code to both CUDA and OpenCL and are available on-site to accelerate applications in a wide range of technical disciplines. SGI Professional Services is available to integrate hybrid clusters either at the factory, so it reaches your floor ready for immediate availability, or at your site.

SGI GPU Compute Solutions at a Glance

Solution	Vertical "U"	Sockets	DIMM Slots	NVIDIA GPU options
Octane III	Deskside	Up to 18S	12 per 'node', up to 9 nodes	NVIDIA Tesla C1060, C2050, C2070 (up to four GPUs, two per workstation)
Rackable C1103-TY12	1U	2S Intel Xeon 5600	12	NVIDIA Tesla, M2050, M2070, M2090, S2050, NextIO vCore Express S2070, S2090
Rackable 1104-2Y12	1U	2 x 2S Intel Xeon 5600	12 per node; 2 nodes	NVIDIA S2050, NextIO vCore Express S2070, S2090
Rackable C2112-4Y14	2U	4 x 2S Intel Xeon 5600	12 per node; 4 nodes	NVIDIA Tesla S2050 (up to two), NextIO vCore Express S2070, S2090
Rackable C3108-TY11	3U	2S Intel Xeon 5600	18	NVIDIA Tesla C1060, S1070, C2050, C2070, S2050, NextIO vCore Express S2070, S2090
Altix UV 10, Altix UV 100, or Altix UV 1000	3U enclosure or 18U enclosure	Up to 256S	Up to 16 TB memory	NVIDIA Tesla S2050 (up to eight GPUs in a single system image), NextIO vCore Express S2070, S2090
Altix ICE 8400		1 x 2S Intel Xeon 5600 or AMD Opteron 6100 up to 1000s	12 per blade	Rackable C3108 as GPU service node

Corporate Headquarters
46600 Landing Parkway
Fremont, CA 94538
tel 510.933.8300
fax 408.321.0293
www.sgi.com

Global Sales and Support
North America +1 800.800.7441
Latin America +55 11.5185.2860
Europe +44 118.927.8000
Asia Pacific +61 2.9448.1463