WHITE PAPER

NVIDIA® Tesla® GPU Computing

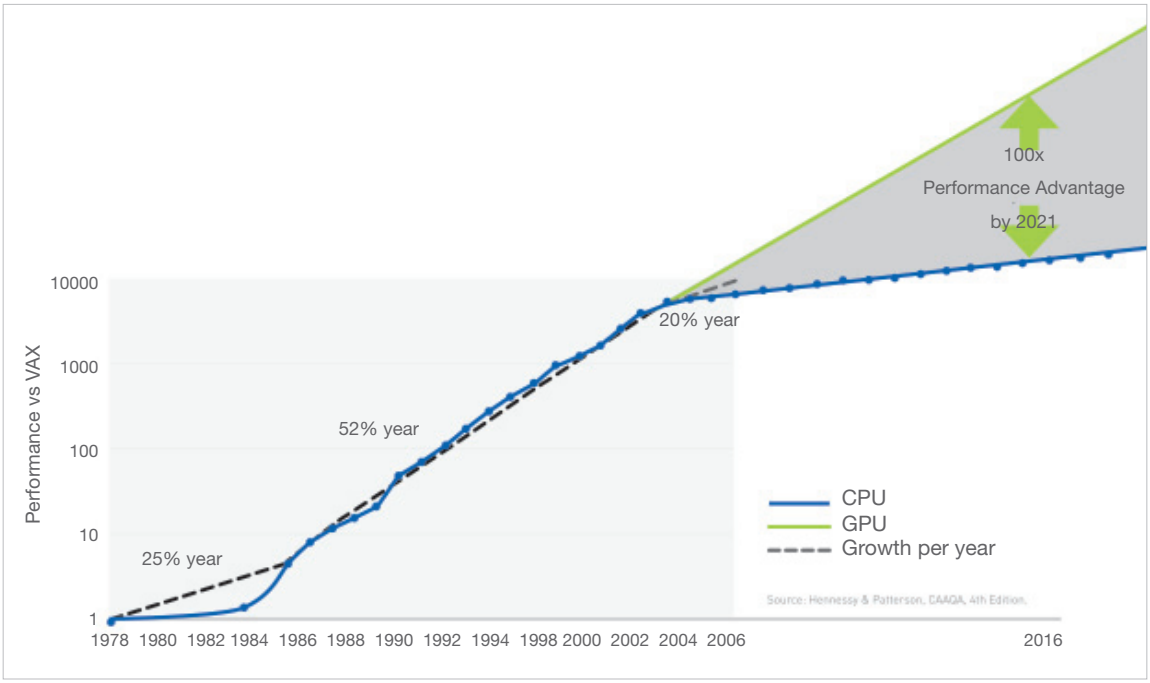
Revolutionizing High-Performance Computing

TABLE OF CONTENTS

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.

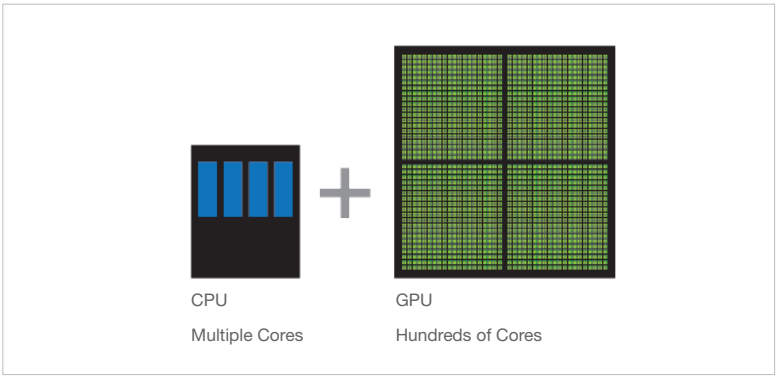


The use of GPUs for computation is a dramatic shift in HPC. GPUs deliver performance increases of 10x to 100x to solve problems in minutes instead of hours, outpacing the performance of traditional computing with x86-based CPUs alone. In addition, GPUs also deliver greater performance per watt of power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

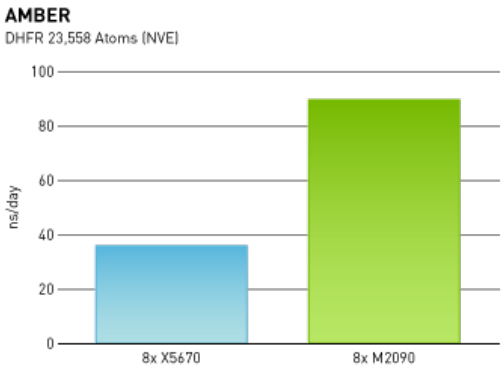
As parallel processors, GPUs excel at tackling large amounts of similar data because the problem can be split into hundreds or thousands of pieces and calculated simultaneously. As sequential processors, CPUs are not designed for this type of computation, but they are adept at more serial-based tasks 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.

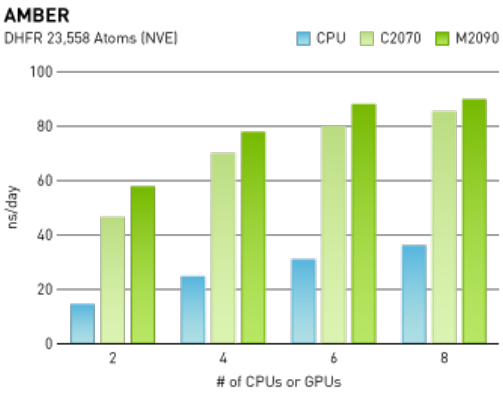


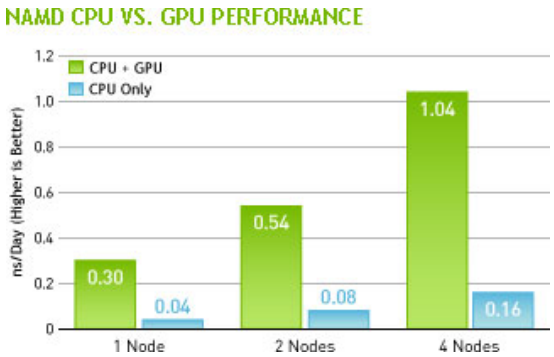
3.0 Product Architecture

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.

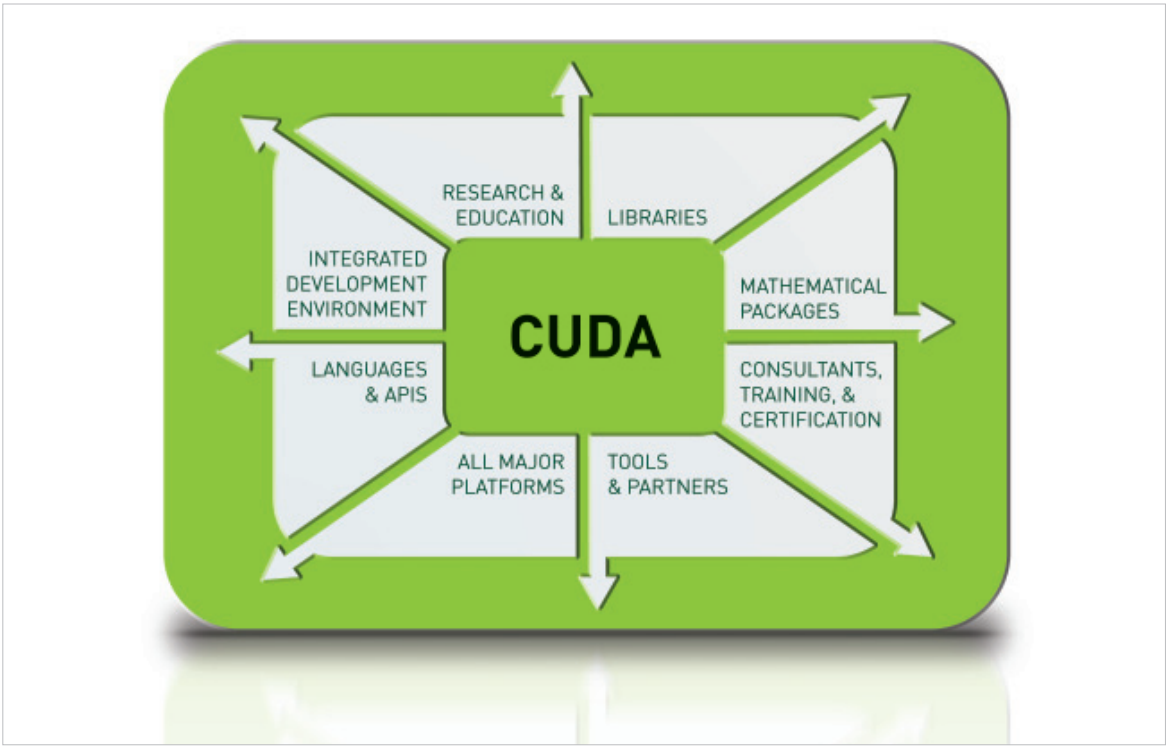




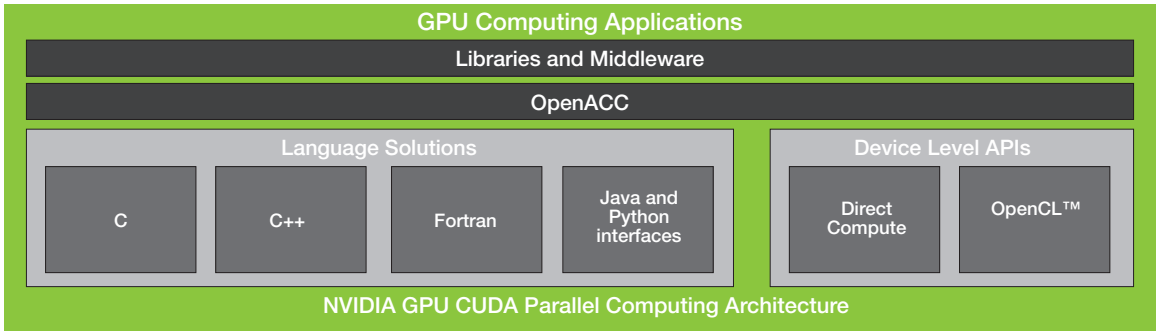


4.0 CUDA® Parallel Computing Architecture

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.



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.

To help customers get their GPU computer efforts up to speed quickly, SGI is now offering a factory built, preconfigured GPU solution of software and hardware starting at less than $200K, with NVIDIA's latest GPU, the Tesla K10 GPU Accelerator. This is all you need for a powerful, 75 teraflops peak performance deployment in a single rack. This SGI GPU solution is called the K10 GPU Starter Kit, and is ideal for numerous high performance computing applications including seismic processing; signal, image and video processing; and, radio astronomy applications.

K10 Starter Kit Configuration:

- One SGI® Rackable™ C1104G-RP5 1U server as head-node with two NVIDIA® Tesla® K10 GPU Accelerator
- Nine SGI Rackable C1104G-RP5 compute nodes with two Tesla K10 GPU Accelerator in each
- Mellanox® MIS5030Q-1BFC, 36 port QDR InfiniBand switch
- Twenty-four port GigE switch
- Rack, cables, and PDU, all ready for production
- Pre-installed software platform consisting of RHEL 6.2,SGI Performance Suite, PBS Pro and SGI Management Center

For more information please visit **sgi.com/products/gpu/**

Scale-Up and Scale-Out Solutions

SGI high-performance computing server solutions allow you to scale up and scale out, and even manage a HPC environment having both system capabilities via the same management software. All these server solutions can be augmented with GPUs for compute acceleration as well.

SGI Rackable™: As your needs grow beyond the K10 Starter Kit, Rackable rack mount servers offer additional server, GPU, software and networking configurations, scaling out to 100s of nodes with factory test and integration standard with each system.

SGI 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 UV scale-up platform will satisfy their needs. The 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 installed on 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.

SGI ICE™: For customers who want to manage large scale-out HPC environments that include GPUs, the SGI ICE 8400 and ICE X platforms offer 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 ICE platform.

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
Rackable™ C1103-TY12	1U	2S Intel® Xeon® 5600	12	NVIDIA® Tesla® M2075, M2090, NextIO vCore™ Express S2090
Rackable™ C1104-2TY12	1U	2 x 2S Intel® Xeon® 5600	12 per node; 2 nodes	NextIO vCore™ Express S2090
Rackable™ C2112-4TY14	2U	4 x 2S Intel® Xeon® 5600	12 per node; 4 nodes	NextIO vCore™ Express S2090
Rackable™ C3108-TY11	3U	2 x 2S Intel® Xeon® 5600	18	NVIDIA® C2075, NextIO vCore™ Express S2090
Rackable™ C1103-G15	1U	2 x 2S AMD Opteron™ 6200	16	NVIDIA® Tesla® M2075, M2090, NextIO vCore™ Express S2090
Rackable™ C1104G-RP5	1U	2 x 2S Intel® Xeon® E5-2600	8	NVIDIA® M2075, M2090, K10
Rackable™ C2110G-RP5	2U	2 x 2S Intel® Xeon® E5-2600	8	NVIDIA® M2075, M2090, K10
SGI® UV™ 10/100/1000	3U or 18U enclosure	Up to 256S Intel® Xeon® E7	Up to 16 TB memory	NextIO vCore™ Express S2090
SGI® ICE™ 8400		1 x 2S Intel® Xeon® 5600 or AMD Opteron™ 6200 blade, 1000s of blades	12 per blade	Rackable™ C3108 as a GPU service node
SGI ICE™ X		1 x 2S Intel® Xeon® E5-2600 blade, 1000s of blades	IP113: 8 or 16 IP115: 16	C1104G-RP5 or C2110G-RP5 as service nodes

Global Sales and Support: sgi.com/global