Shattering the 1U Server Performance Record

Supermicro and NVIDIA® recently announced a new class of servers that combines massively parallel GPUs with multi-core CPUs in a single server system. This unique configuration delivers performance at least an order of magnitude better than traditional quad-core CPU-based servers. This breakthrough technology immediately provides users with the ability to implement tasks that were traditionally addressed only with massive supercomputers or that were simply unsolvable. Supermicro calls this new server class, the latest in its rich history of technology innovations as shown in Figure 1, the GPU SuperServer.

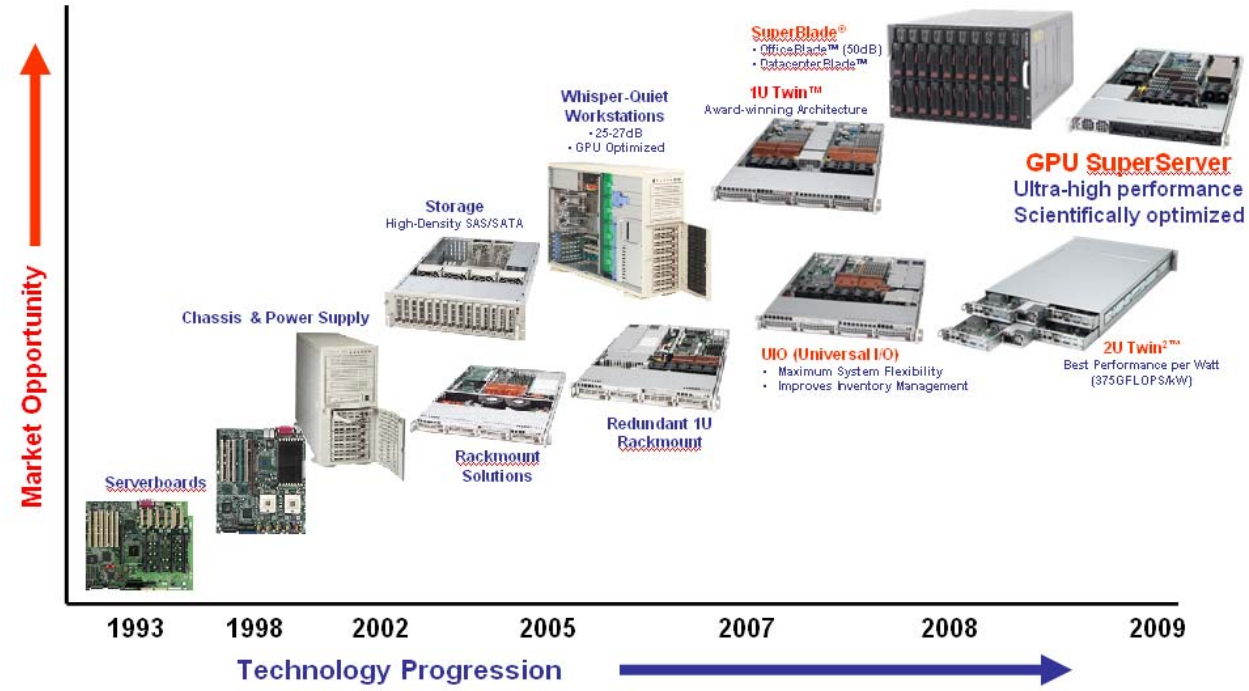


Figure 1: Supermicro Product and Market Opportunity Growth

First to Market

Supermicro demonstrated its first GPU SuperServer, the SuperServer 6016GT-GF series, at the Computex 2009 Show in Taiwan in early June. This server series features dual Intel® Xeon® (Nehalem) processors and two Gen2 PCI-Express x16 interfaces to support two NVIDIA Tesla™ M1060 GPU Processors. With two double-width GPUs in PCI-E 2.0 x16 lanes, this 1U server delivers truly non-blocking GPU performance, or up to 2 Teraflops of processing power, which makes it the highest performing 1U server on the planet.

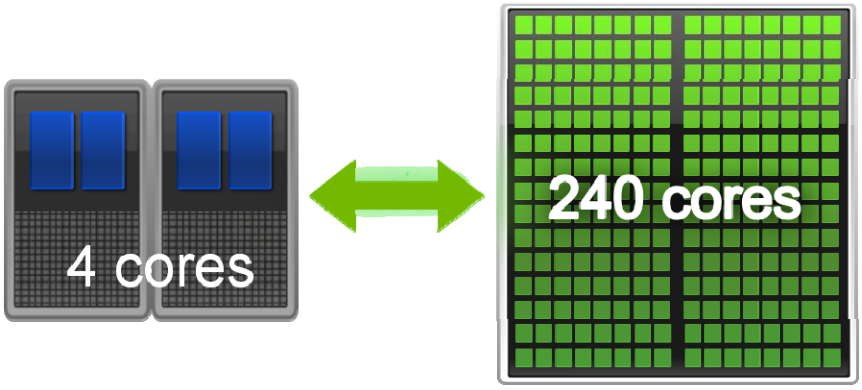
The SuperServer 6016GT 1U series is the first of an entire line of GPU-optimized systems Supermicro has created to meet the requirements of the high-performance computing market. By the end of June, Supermicro will launch a 4U/Tower system, the SW7046A-GRF, which supports four double-width GPUs. These platforms feature Supermicro's new Gold Level[1] (93% efficiency) power subsystems to deliver breakthrough performance-per-watt.

SS6016GT-TF-TM2 SW7046A-GRF

Figure 2: Supermicro GPU SuperServers

What is GPU Computing?

GPU computing is the use of a GPU (graphics processing unit) to perform general purpose scientific and engineering computing. The model for GPU computing is to use a CPU and GPU together in a heterogeneous computing model (Figure 3). The sequential part of the application runs on the CPU and the computationally-intensive part runs on the GPU. From the user's perspective, the application just runs faster because it is using the high performance of the GPU to boost overall system performance.



CPU GPU

Figure 3: Heterogeneous Computing Model

The application developer has to modify their application to take the compute-intensive kernels and map them to the GPU. The rest of the application remains on the CPU. Mapping a function to the GPU involves rewriting the function to expose its parallelisms and adding "C" keywords to move data to and from the GPU.

[1] www.80Plus.org

GPU computing is enabled by the massively parallel architecture of NVIDIA GPUs called the CUDA™ parallel computing architecture. The CUDA architecture consists of hundreds of processor cores that operate together to crunch through the data set in the application.

The NVIDIA® Tesla™ 10-series GPU is built on the second generation CUDA architecture with features optimized for scientific applications such as IEEE standard double precision floating point hardware support, local data caches in the form of shared memory dispersed throughout the GPU, and coalesced memory accesses.

Learn more about vertical applications that take advantage of the GPU at:
http://www.nvidia.com/object/vertical_solutions.html

NVIDIA provides a complete range of software tools to program the GPU, including a C compiler toolchain, libraries, and other tools. These are listed at:
http://www.nvidia.com/object/tesla_software.html

History of GPU Computing

Graphics chips started as fixed function graphics pipelines. Over the years, these graphics chips became increasingly programmable, which led NVIDIA to introduce the first GPU or Graphics Processing Unit. In the 1999-2000 timeframe, computer scientists in particular, along with researchers in fields such as medical imaging and electromagnetics started using GPUs for running general purpose computational applications. They found the excellent floating point performance in GPUs led to a huge performance boost for a range of scientific applications. This was the advent of the movement called **GPGPU** or General Purpose computing on GPUs.

The problem was that GPGPU required using graphics programming languages like OpenGL and Cg to program the GPU. Developers had to make their scientific applications look like graphics applications and map them into problems that drew triangles and polygons. This limited the accessibility of tremendous performance of GPUs for science.

NVIDIA realized the potential to bring this performance to the larger scientific community and decided to invest in modifying the GPU to make it fully programmable for scientific applications and added support for high-level languages like C and C++. This led to the CUDA general purpose parallel computing architecture for the GPU.

CUDA Parallel Computing Architecture and Programming Model

The CUDA parallel computing architecture is accompanied by the CUDA parallel programming model that provides a set of abstractions that enable expressing fine-grained and coarse-grained data and task parallelism. The programmer can choose to express the parallelism in high-level languages such as C, C++, Fortran or driver APIs such as OpenCL™ and DirectX™ Compute.

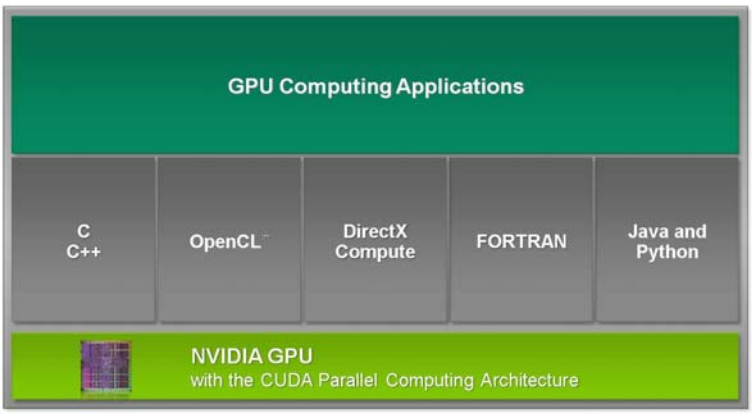


Figure 4: CUDA Programming Model

The first language support NVIDIA provided was for the C language. A set of C for CUDA software development tools enable the GPU to be programmed using C with a minimal set of keywords or extensions. Support for Fortran, OpenCL, and others will follow soon.

The CUDA parallel programming model guides programmers to partition the problem into coarse sub-problems that can be solved independently in parallel. Fine grain parallelism in the sub-problems is then expressed such that each sub-problem can be solved cooperatively in parallel.

The CUDA parallel computing architecture and the corresponding CUDA parallel computing model are now widely deployed with hundreds of applications and over one thousand published research papers. CUDA Zone lists many of these applications and papers.

Widest Variety of GPU-Optimized Servers

Supermicro has developed the widest variety of GPU SuperServers in the industry. These include the SS6016GT series of 1U SuperServers announced at Computex and the SW7046A series of 4U/Tower SuperWorkstations. These systems are identified in bold in Table 1 below together with the maximum number of double-width GPUs, and expansion slots that they can support simultaneously.

Supermicro's GPU SuperServer family also includes a growing number of systems that can be customized using Supermicro's unique Server Building Block Solutions®. These are unique combinations of Supermicro's servers, motherboards, and other components that are optimized for the customer's specific GPU application. These Server Building Block Solutions® are identified in Table 1 below by the number of dual-height GPUs and expansion slots that they can support for their customer application.

MB \ Chassis	SC743TQ-865B-SQ (4U/Tower)	SC745TQ-R1200B (4U/Tower)	SC747TQ-R1400B (4U/Tower)	SC818G-1400B (1U Rackmount)
X8DTG-QF			**SW7046A-GRF** 4 GPUs, 3 Expansion Slots	
X8DTG-DF				**SS6016GT-TF-TM2** Enterprise Level 2 integrated GPUs, 1 Expansion Slot (low-profile) **SS6016GT-TF-TC2** 2 integrated GPUs, 1 Expansion Slot (low-profile) **SS6016GT-TF** 2 GPUs, 1 Expansion Slot (low-profile) **SS6016XT-TF** 4 Expansion Slots + 1 Expansion Slot (low-profile)
X8DAH+		**SW7046A-HR+** 3 GPUs*, 1 Expansion Slot		
X8DTH-6/6F	2 GPUs*, 3 Expansion Slots	3 GPUs*, 1 Expansion Slot	3 GPUs*, 1 Expansion Slot	
X8DTH-i/F	2 GPUs*, 3 Expansion Slots	3 GPUs*, 1 Expansion Slot	3 GPUs*, 1 Expansion Slot	
X8DA3**	**SW7046A-3** 2 GPUs, 2 Expansion Slots	2 GPUs, 2 Expansion Slots	2 GPUs, 2 Expansion Slots	
X8DAi**	**SW7046A-T** 2 GPUs, 2 Expansion Slots	2 GPUs, 2 Expansion Slots	2 GPUs, 2 Expansion Slots	
X8DTT-F				1 GPU
X8DTT-IBXF				1 GPU
X8DTT-IBQF				1 GPU

* Some Gen2 PCI-E x8 in x16 slot ** requires video output card

Table 1: Supermicro GPU SuperServers (Bold) and Server Building Block Solutions®
(All Expansion Slots are Full-Height, Full-Length unless otherwise indicated)

Initially one 1U and three 4U/Tower chassis support GPU computing. A wide variety of Supermicro's next-generation motherboards supporting the Intel® Xeon® Processor 5500 Series (Nehalem) are also available as Building Blocks.

More information on these Supermicro chassis is available at: http://www.supermicro.com/products/chassis/
Additional motherboard information can be found at: http://www.supermicro.com/products/motherboard/Xeon1333/
Supermicro's GPU SuperServer webpage is: http://www.supermicro.com/GPU/

Advantages of Supermicro GPU SuperServers

Supermicro's GPU SuperServer family provides many feature advantages of interest to the user community.

Widest Selection

Supermicro's wide line of GPU systems, outlined in Table 1, can satisfy any GPU Supercomputing server application. The family comprises a wide variety of 1U and 4U/Tower systems, chassis and motherboards.

Highly Reliable Thermal Optimization

Supermicro's advanced thermal subsystem designs optimize the cooling elements of the server with the thermal characteristics of the GPU, CPU, motherboard, and other components. Each thermally controlled fan module is 1+1 paired in a counter rotating matched set, which maximizes the cooling effect. This design enhances system reliability and simplifies maintenance. In addition, by monitoring both CPU and GPU thermal readings though the I^2C bus, Supermicro's unique intelligent cooling control and auto fan-speed adjustment keeps the entire system in stable operation.

Direct GPU Connect Architecture

The Supermicro design architecture provides direct Gen2 PCI-E x16 non-blocking connectivity to each GPU to take advantage of the full bandwidth of the GPU. No extra cabling is required, since the GPUs are connected directly to the server motherboard via riser cards, thus improving reliability, airflow, and maintenance, as well as reducing costs.

Industry-Leading Power Efficiency

Supermicro's GPU SuperServer platforms feature Gold Level (93% efficiency) power subsystems, along with high-efficiency motherboard and thermal designs to deliver breakthrough performance-per-watt and increased system reliability.

Flexible Networking Connectivity

The SS6016GT GPU SuperServer series includes a single half-height PCI-E Gen2 slot for additional connectivity. The SW7046A-GRF 4U/Tower server with 4 double-width GPUs has 3 additional PCI-E expansion slots for add-on cards.

Advanced System Management

Supermicro's advanced IPMI capability allows the GPU system to be managed directly via the remote monitoring / controlling features. All the key elements of the server system- CPU, GPU and power supply- can be remotely monitored, along with Supermicro standard onboard IPMI functionalities.

Applications and Industries

This new line of highly parallel, multi-core, multi-GPU systems is an excellent choice for an extensive range of graphics and computationally intensive applications in a wide variety of industries. In general, these systems are expected to make the fastest teraflop clusters much more affordable and accessible for users throughout the world.

Applications requiring high arithmetic intensity, such as dense linear algebra, partial differential equations, n-body problems, and finite difference formulas, can be easily accommodated using these Supermicro SuperServers. High bandwidth problems such as sequencing (virus scanning, genomics), sorting, and database problems are also potential applications. Finally, visual computing problems such as graphics, image processing, tomography, and machine vision, being original applications for GPUs, are especially accessible.

Fields where these types of problems are common and can benefit from Supermicro's GPU SuperServers include medical, energy, telecommunications, finance, science, and engineering in a wide array of application areas such as:

- Computational Chemistry
- Fluid Dynamics
- Digital Content Creation
- Electronic Device Automation
- Financial Markets Modeling
- Game Physics
- Genomics
- Medical Imaging
- Oil & Gas Exploration
- Research & Scientific Computing
- Signal Processing
- Weather Simulation

Below in Figure 5 are some application examples from Supermicro's GPU partner NVIDIA®:

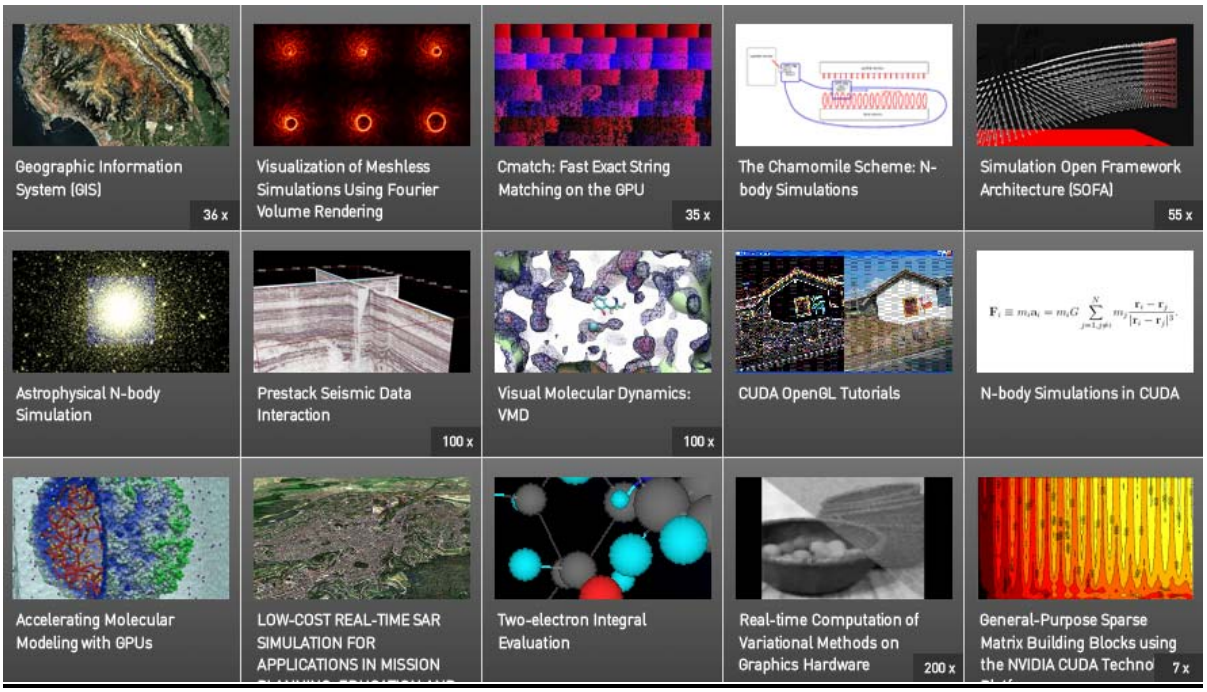


Figure 5: Example GPU Applications
http://www.nvidia.com/object/cuda_home.html#

Conclusion

Supermicro's GPU SuperServer family includes powerful 1U and 4U systems integrating up to four GPUs. These server systems offer many feature advantages applicable to a wide range of problems in many industries. Supermicro and GPU technology partner NVIDIA have aggressively positioned themselves to lead this exciting new business, beginning with the SS6016GT GPU SuperServer series, which shatters the 1U server performance record.

Additional Resources

Those wishing to learn more about this topic may find the following resources helpful places to start:

- Supermicro GPU: http://www.supermicro.com/GPU/
- YouTube video (Introducing the World's Fastest 1U Server): http://www.youtube.com/watch?v=KGoT7C8y5rQ
- GPGPU.org is a central resource for GPGPU news and information: http://gpgpu.org/
- Dr. Dobb's Journal discusses applications software for GPUs: http://www.ddj.com/hpc-high-performance-computing/206900471
- NVIDIA CUDA Zone: http://www.nvidia.com/object/cuda_home.html
- NVIDIA Tesla Vertical Applications: http://www.nvidia.com/object/vertical_solutions.html
- NVIDIA Tesla CUDA Software Development Tools: http://www.nvidia.com/object/Tesla_software.html
- Who's working on GPU apps: "GPGPU People" http://www.gpgpu.org/w/index.php/GPGPU_People