



Shattering the 1U Server Performance Record

UPERMICRO 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, the GPU Supercomputing Server.

FIRST TO MARKET

Supermicro demonstrated its first GPU Supercomputing Servers, the SuperServer 6016TGF series, at the Computex 2009 Show in Taiwan in early June. This server series features dual Intel® Xeon® Processor 5500 series (Nehalem) and two PCI-Express 2.0 x16 interfaces to support multiple GPUs. With two double-width GPUs and double PCI-Express 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 fastest 1U server on the planet.

The SuperServer 6016TGF 1U series is the first of an entire line

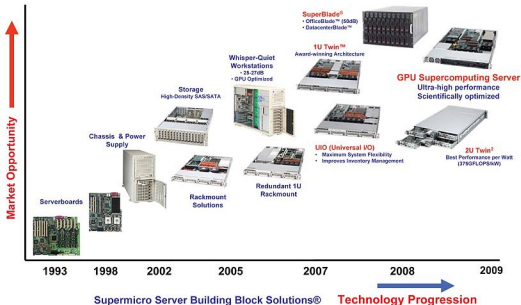


Figure 1: Supermicro Product and Market Opportunity Growth



Figure 2: Supermicro GPU Supercomputing Servers. Left to right: SS6016GT-TF-TM2, SW7046GT-TRF

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 SuperWorkstation 7046T-GRF, which supports four double-width GPUs. These platforms feature Supermicro's new Gold Level¹ (93% efficiency) power subsystems to deliver breakthrough performance-per-watt.

WHAT IS A GPU SUPERCOMPUTING SERVER?

To understand GPU Supercomputing Servers, we first need to understand the architecture of a GPU. GPUs (Graphics Processing Units) are high-performance multi-core processors capable of very high computation and data throughput. Once specially designed for computer graphics and difficult to program, today's GPUs are general-purpose parallel processors that support standard application programming interfaces (e.g. APIs like OpenGL) and industry-standard languages such as C. Applications that are run on GPUs often achieve substantial speed increases compared with optimized CPU-only implementations.

The model for GPU supercomputing is to use a CPU and GPU together in a heterogeneous (e.g. mixed) supercomputing model. 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 uses the high performance of the GPU to boost computing speed. This operation is sometimes called GPGPU (General-Purpose computation on Graphics Processing Units).

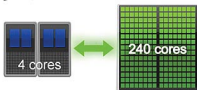


Figure 3: Heterogeneous Computing Model. Left to right: CPU, GPU

Such leaps in performance do not come easily. Application developers must modify their code by taking the compute-intensive portions and mapping them to the GPU. The rest of the application remains on the CPU. Mapping a function to the GPU involves rewriting the functions to expose their parallelisms and adding the proper "C" programming language instructions to move data to and from the GPU efficiently. New parallel processing architecture tools such as CUDA help to automate this process, saving programming resources and expanding the applications universe for CPU+GPU heterogeneous supercomputing.

A GPU Supercomputing Server integrates GPUs into the server architecture. The Supermicro SS6016GT-TF-TM2, for example, utilizes two NVIDIA Tesla M1060 double-width GPU cards, each of which has 240 processor cores generating up to 1 Teraflop, for a total computing capability of up to 2 Teraflops. This powerful system, with intelligent cooling control technology, is optimized for mission critical server cluster applications. The 4U/Tower Supermicro SW7046GT-TRF integrates four double-width GPUs for a total of up to 4 Teraflops.

WIDEST VARIETY OF GPU SUPERCOMPUTING SERVERS

Supermicro has developed the widest variety of GPU Supercomputing Servers in the industry. These include the SS6016GT series of four 1U servers announced at Computex and two 4U/Tower systems, the SW7046GT-TRF and SW7046A-HR+. 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.

Please review the Supermicro Supercomputing Server Solutions web page for more information:
<http://www.supermicro.com/GPU/>

Supermicro's GPU Supercomputing Server family also includes a growing number of systems that can be customized using Supermicro's unique Server Building Block Solutions*. These systems are identified in Table 1 by the number of GPUs and expansion slots that they can support. Initially four 1U and 4U/Tower chassis are supported and 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/>

¹ 80Plus.org

		Chassis			
		SC7437Q-865B-SQ (4U/TOWER)	SC7457Q-R1201B (4U/TOWER)	SC7477Q-R1406B (4U/TOWER)	SC816G-1401B (1U BACKMOUNT)
MotherBoard	X80TG-0F	SW7046T-GRF 4 GPUs, 3 Expansion Slots			
	X80TQ-0F	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) SS6016KT-TF: 4 Expansion slots + 1 Expansion slot (low profile)			
	X80AH+	SW7045A-HR+ 3 GPUs, 1 Expansion Slot			
	X80TH-0/6F	2 GPUs*, 3 Expansion Slot	3 GPUs*, 1 Expansion Slot	3 GPUs*, 1 Expansion Slot	
	X80TH-4F	2 GPUs*, 3 Expansion Slot	3 GPUs*, 1 Expansion Slot	3 GPUs*, 1 Expansion Slot	
	X80A3**	SW7046A-Q: 2 GPUs, 2 Expansion Slots	2 GPUs, 2 Expansion Slots	2 GPUs, 2 Expansion Slots	
	X80A1**	SW7046A-T: 2 GPUs, 2 Expansion Slots	2 GPUs, 2 Expansion Slots	2 GPUs, 2 Expansion Slots	
	X80TT-F	SS6016T-6TF: 1 GPU			
	X80TT-1BXF	SS6016T-6IBXF: 1 GPU			
	X80TT-1B0F	SS6016T-6IB0F: 1 GPU			

Table 1: Supermicro GPU Supercomputing Servers (Bold) and Server Building Block Solutions™

* Some Gen2 PCI-E x8 in x16 slot

** Requires video output card

ADVANTAGES OF SUPERMICRO GPU SUPERCOMPUTING SERVERS

Supermicro's GPU Supercomputing Server 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 line 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. The cooling system is designed to protect against any single point of fan failure, enhancing system reliability and simplifying maintenance. In addition, by monitoring both CPU and GPU thermal readings through the PC bus, Supermicro's unique intelligent cooling control and auto fan-speed adjustment keep the entire system operating smoothly.

Direct Connect Architecture

The Supermicro design architecture provides direct PCI-Express 2.0 x16 non-blocking connectivity to each GPU to take advantage of the GPU's full bandwidth. 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 Supercomputing server 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 Supercomputing Server series includes a single half-height PCI-E 2.0 slot for additional connectivity. The SW7046GT-TRF 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 remote monitor and controlling functions. All the key elements of the server system- CPU, GPU and power supply- are remotely monitored from Supermicro's extensive onboard IPMI management and control subsystem.

APPLICATIONS AND INDUSTRIES

This new line of highly parallel, many-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

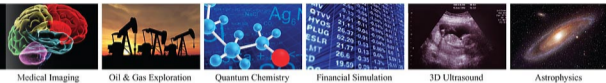


Figure 4: GPU Supercomputing Server Markets

dense linear algebra, partial differential equations, n-body problems, and finite difference formulas, can be easily accommodated using Supermicro Supercomputing Servers. High bandwidth problems such as sequencing (virus scanning, genomics), sorting, and database problems are also attractive 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 applications are common include medical imaging, oil and gas exploration, computational chemistry, telecommunications, financial simulation, weather forecasting, and astrophysics. In addition research and engineering projects and general scientific computing investigations can also benefit from Supermicro's GPU Supercomputing Servers.

S I G

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/>

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>

ATI Stream Technology Site:

<http://ati.amd.com/technology/streamcomputing/>

NVIDIA's CUDA Zone:

http://www.nvidia.com/object/cuda_home.html

Who's working on GPU apps: "GPGPU People"

http://www.gpgpu.org/w/index.php/GPGPU_People