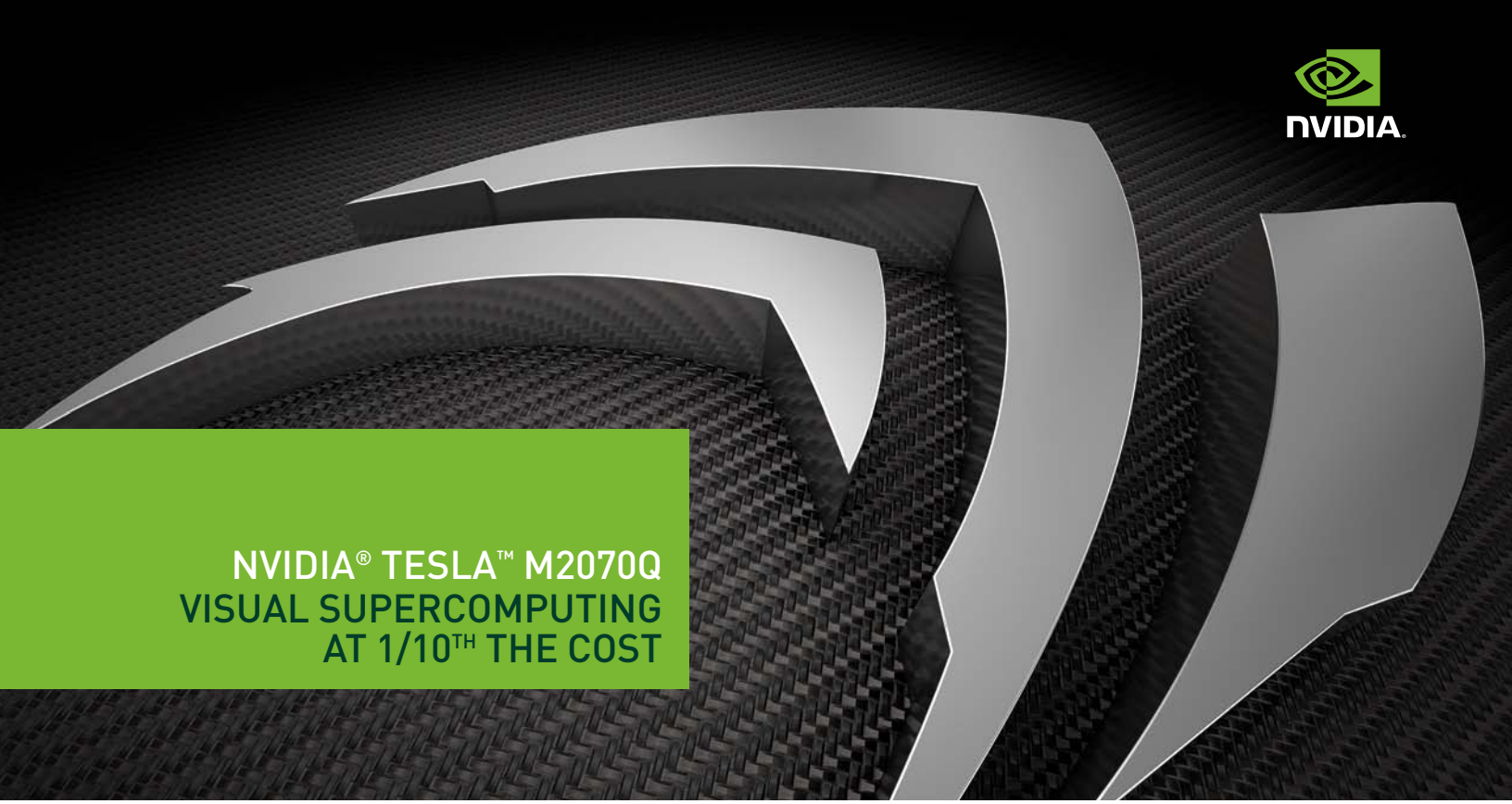

NVIDIA® TESLA™ M2070Q VISUAL SUPERCOMPUTING AT 1/10TH THE COST

Based on the next-generation CUDA™ architecture codenamed "Fermi", the Tesla™ M2070Q Computing and Visualization Module enables integration of GPU computing and visualization with host systems for high-performance computing and large data center, scale-out deployments.

The Tesla M2070Q is the first to deliver greater than 10X the double-precision horsepower of a quad-core x86 CPU and the first to deliver ECC memory. The Tesla M2070Q module delivers all of the standard benefits of GPU computing while enabling maximum reliability and tight integration with system monitoring and management tools. In addition, the "Q" means Quadro graphics features are also enabled, making the M2070Q a great solution for server based graphics delivering up to an unheard of 1.3 billion triangles per second, shattering previous 3D performance benchmarks.[1] This gives data center IT staff the ability to deploy GPUs for both compute and visualization in one convenient solution. With a wide variety of rack-mount and blade systems now supporting the Tesla M2070Q, and with remote monitoring, remote management, and remote graphics capabilities, IT managers can build a compute cluster and visualization cluster in one.

Compared to CPU-only systems, servers with Tesla M2070Q deliver supercomputing power at 1/10th the cost and 1/20th the power consumption while providing the highest compute density.

From medical imaging to structural analysis applications, data integrity and precision is assured, without sacrificing graphics performance. The Tesla M2070Q is even a great solution for enabling rich media VDI desktops using Microsoft's RemoteFX, giving multiple VDI users the power of hardware accelerated graphics from a single graphics device.

Tesla M2070Q is not only a graphics processor or computing device; it's an entire visual supercomputing platform, incorporating hardware and software that enables advanced capabilities. Giving researchers and engineers the power to compute and visualize their solutions in one simple hardware package, the M2070Q may well be the most versatile computing solution ever, capable of changing the way you work every day.

TECHNICAL SPECIFICATIONS

CUDA PARALLEL PROCESSING CORES
- 448

FORM FACTOR
- 9.75" PCIe x16 Dual Slot

OF TESLA GPUs
- 1

DOUBLE PRECISION FLOATING POINT PERFORMANCE (PEAK)
- 515 Gflops

SINGLE PRECISION FLOATING POINT PERFORMANCE (PEAK)
- 1.03 Tflops

TOTAL DEDICATED MEMORY[2]
- 6GB GDDR5

MEMORY SPEED
- 1.55 GHz

MEMORY INTERFACE
- 384-bit

MEMORY BANDWIDTH
- 148 GB/s

MAX POWER CONSUMPTION
- 225W TDP

SYSTEM INTERFACE
- PCIe x16 Gen2

ECC MEMORY
- Yes

FAST DOUBLE PRECISION
- Yes

DISPLAY CONNECTORS
- None, virtual remote displays only

THERMAL SOLUTION
- Passive heatsink cooled by host system airflow

SOFT WARE DEVELOPMENT TOOLS
- CUDA C/C++/Fortran, OpenCL,
- DirectCompute Toolkits.
- NVIDIA Parallel Nsight™ for Visual Studio

NVIDIA® TESLA™ M2070Q

Features	Benefits
448 CUDA CORES	Delivers up to 515 Gigaflops of double-precision peak performance in each GPU, enabling servers from leading OEMs to deliver a Teraflop or more of double-precision performance per 1 RU of space. Single precision peak performance is over one Teraflop per GPU.
6GB OF GDDR5 MEMORY PER GPU WITH ULTRA-FAST BANDWIDTH	Maximizes performance and reduces data transfers by keeping larger data sets and complex scenes in local memory that is attached directly to the GPU.
ECC MEMORY	Meets a critical requirement for computing accuracy and reliability in datacenters and super-computing centers. Offers protection of data in memory to enhance data integrity and reliability for applications. Register files, L1/L2 caches, shared memory, and DRAM all are ECC protected
NVIDIA® SCALABLE GEOMETRY ENGINE™	Dramatically improves geometry performance across a broad range of CAD, DCC and medical applications, enabling you to work interactively with models and scenes that are an order of magnitude more complex than ever before
SYSTEM MONITORING FEATURES	Integrates the GPU subsystem with the host system's monitoring and management capabilities. This means IT staff can manage all of the critical components of the computing system through a common management interface such as IPMI or OEM-proprietary tools.
DESIGNED FOR MAXIMUM RELIABILITY	Passive heatsink design aligns with server cooling airflows, and eliminates moving parts and cables.
NVIDIA PARALLEL DATA CACHE™	Accelerates algorithms such as physics solvers, ray-tracing, and sparse matrix multiplication where data addresses are not known beforehand. This includes a configurable L1 cache per Streaming Multiprocessor block and a unified L2 cache for all of the processor cores.
NVIDIA GIGA THREAD™ ENGINE	Maximizes the throughput by faster context switching that is 10X faster than previous architecture, concurrent kernel execution, and improved thread block scheduling.
DUAL COPY ENGINES	Enables the highest rates of parallel data processing and concurrent throughput between the GPU and host, accelerating techniques such as ray tracing, color grading and physical simulation.
ASYNCHRONOUS TRANSFER	Turbocharges system performance by transferring data over the PCIe bus while the computing cores are crunching other data. Even applications with heavy data-transfer requirements, such as seismic processing, can maximize the computing efficiency by transferring data to local memory before it is needed.
CUDA PROGRAMMING ENVIRONMENT WITH BROAD SUPPORT OF PROGRAMMING LANGUAGES AND APIs	Choose C, C++, OpenCL, DirectCompute, or Fortran to express application parallelism and take advantage of the innovative "Fermi" architecture.
HIGH SPEED , PCIE GEN 2.0 DATA TRANSFER	Maximizes bandwidth between the host system and the Tesla processors. Enables Tesla systems to work with virtually any PCIe-compliant host system with an open PCIe slot (x8 or x16).

DRIVERS AND DOWNLOADS

- Tesla M2070Q is supported under:
  - > Linux 32-bit and 64-bit
  - > Microsoft Windows Server 2003 and 2008
  - > Microsoft Windows 7 (64-bit and 32-bit)
  - > Microsoft Windows Vista (64-bit and 32-bit)
  - > Microsoft Windows XP (64-bit and 32-bit)
- Vertical Solutions and Software page: **www.nvidia.com/object/ vertical_solutions.htm**
- Software
  - > Drivers — NVIDIA recommends that users get drivers for M-series products from their System OEM to ensure that driver is qualified by the OEM on their system.
  - > Tools — Software development tools are available at: **www.nvidia.com/object/ tesla_software.html**

SUPPORT

- Hard ware Support
  For RMA requests, replacements and warranty issues regarding your NVIDIA based product, please contact the OEM that you purchased it from.
- Knowledgebase
  Our knowledgebase is available online 24x7x365 and contains answers to the most common questions and issues: **www.nvidia.custhelp.com/cgi-bin/ nvidia.cfg/php/enduser/std_alp.php**
- User Forums
  Discuss Tesla products, talk about CUDA development, and share interesting issues, tips and solutions with your fellow NVIDIA Tesla users on the CUDA discussion forums: **www.forums.nvidia.com**

To learn more about NVIDIA Tesla, go to **www.nvidia.com/tesla**
To learn more about NVIDIA Quadro, go to **www.nvidia.com/quadro**

[1] Raw throughput number calculated by graphics processing clusters, GPU clock rate, and triangle throughput.
[2] Note: With ECC on, a portion of the dedicated memory is used for ECC bits, so the available user memory is reduced by 12.5%. (e.g. 3 GB total memory yields 2.625 GB of user available memory.)

NVIDIA