NVIDIA® TESLA® KEPLER GPU COMPUTING ACCELERATORS

FASTEST, MOST EFFICIENT HPC ARCHITECTURE EVER BUILT

With teraflops of single and double precision performance, NVIDIA® Kepler GPU Computing Accelerators are the world's fastest and most efficient high performance computing (HPC) companion processors. Based on the Kepler compute architecture, which is 3 times higher performance per watt than the previous "Fermi" compute architecture[1], the Tesla Kepler GPU Computing Accelerators make hybrid computing dramatically easier, and applicable to a broader set of computing applications. NVIDIA Tesla GPUs deliver the best performance and power efficiency for seismic processing, biochemistry simulations, weather and climate modeling, image, video and signal processing, computational finance, computational physics, CAE, CFD, and data analytics.

The innovative design of the Kepler compute architecture includes:

- **SMX** (streaming multiprocessor) design that delivers up to 3x more performance per watt compared to the SM in Fermi. It also delivers 1 petaflop of computing in just 10 server racks.
- **Dynamic Parallelism** capability that enables GPU threads to automatically spawn new threads. By adapting to the data without going back to the CPU, it greatly simplifies parallel programming and enables GPU acceleration of a broader set of popular algorithms, like adaptive mesh refinement (AMR), fast multipole method (FMM), and multigrid methods.
- **Hyper-Q** feature that enables multiple CPU cores to simultaneously utilize the CUDA cores on a single Kepler GPU, dramatically increasing GPU utilization, slashing CPU idle times, and advancing programmability. Ideal for cluster applications that use MPI.

Tesla products based on two different Kepler GPUs:

Tesla K10 GPU Computing Accelerator – Optimized for single precision applications, the Tesla K10 is a throughput monster based on the ultra-efficient GK104 Kepler GPU. The accelerator board features two GK104 GPUs and delivers up to 2x the performance for single precision applications compared to the previous generation Fermi-based Tesla M2090 in the same power envelope. With an aggregate performance of 4.58 teraflop peak single precision and 320 gigabytes per second memory bandwidth for both GPUs put together, the Tesla K10 is optimized for computations in seismic, signal, image processing, and video analytics.

Tesla K20 GPU Computing Accelerator – Designed for double precision applications and the broader supercomputing market, the Tesla K20 delivers 3x the double precision performance compared to the previous generation Fermi-based Tesla M2090, in the same power envelope. Tesla K20 features a single GK110 Kepler GPU that includes the Dynamic Parallelism and Hyper-Q features. With more than one teraflop peak double precision performance, the Tesla K20 is ideal for a wide range of high performance computing workloads including climate and weather modeling, CFD, CAE, computational physics, biochemistry simulations, and computational finance.

[1]Based on DGEMM performance: Tesla M2090 (Fermi) = 330 gigaflops, Tesla K20 (expected) > 1000 gigaflops

TECHNICAL SPECIFICATIONS

	TESLA K10[a]	TESLA K20
Peak double precision floating point performance (board)	0.19 teraflops	To be announced
Peak single precision floating point performance (board)	4.58 teraflops	To be announced
Number of GPUs	2 x GK104s	1 x GK110
CUDA cores	2 x 1536	To be announced
Memory size per board (GDDR5)	8 GB	To be announced
Memory bandwidth for board (ECC off)[b]	320 GBytes/sec	To be announced
GPU Computing Applications	Seismic, Image, Signal Processing, Video analytics	CFD, CAE, Financial computing, Computational chemistry and Physics, Data analytics, Satellite imaging, Weather modeling
Architecture Features	SMX	SMX, Dynamic Parallelism, Hyper-Q
System	Servers only	Servers and Workstations.
Available	May 2012	Q4 2012

[a] Tesla K10 specifications are shown as aggregate of two GPUs.

[b] With ECC on, 12.5% of the GPU memory is used for ECC bits. So, for example, 6 GB total memory yields 5.25 GB of user available memory with ECC on.

TESLA GPU COMPUTING ACCELERATOR COMMON FEATURES

ECC Memory Error Protection	Meets a critical requirement for computing accuracy and reliability in datacenters and supercomputing centers. External DRAM is ECC protected in Tesla K10 and both external and internal memories are ECC protected in Tesla K20.
System Monitoring Features	Integrates the GPU subsystem with the host system's monitoring and management capabilities such as IPMI or OEM-proprietary tools. IT staff can thus manage the GPU processors in the computing system using widely used cluster/grid management solutions.
L1 and L2 caches	Accelerates algorithms such as physics solvers, ray-tracing, and sparse matrix multiplication where data addresses are not known beforehand
Asynchronous Transfer with dual DMA engines	Turbocharges system performance by transferring data over the PCIe bus while the computing cores are crunching other data.
Flexible programming environment with broad support of programming languages and APIs	Choose OpenACC, CUDA toolkits for C, C++, or Fortran to express application parallelism and take advantage of the innovative Kepler architecture.

SOFTWARE AND DRIVERS

- Software applications page: http://www.nvidia.com/object/vertical_solutions.html
- Tesla GPU computing accelerators are supported for both Linux and Windows. Server modules are only supported on 64-bit OSes and workstation / desktop modules are supported for 32-bit as well.
  - Drivers– NVIDIA recommends that users get drivers for Tesla server products from their system OEM to ensure that driver is qualified by the OEM on their system. Latest drivers can be downloaded from http://www.nvidia.com/drivers
  - Learn more about Tesla data center management tools at http://www.nvidia.com/object/software-for-tesla-products.html
- Software development tools are available at http://developer.nvidia.com/getting-started-parallel-computing

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

NVIDIA