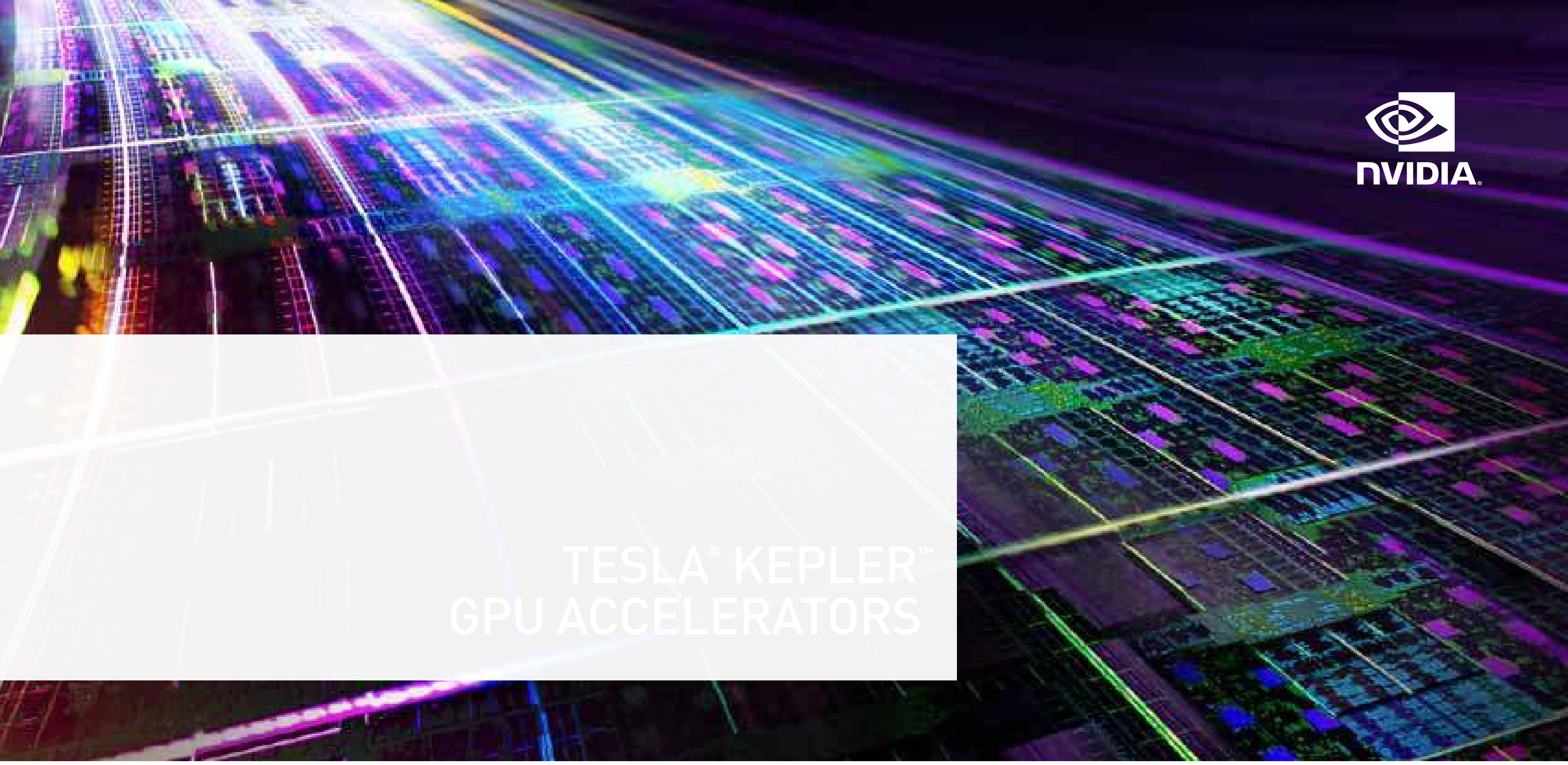

TESLA® KEPLER™ GPU ACCELERATORS

NVIDIA Tesla K-series GPU Accelerators are based on the NVIDIA Kepler™ compute architecture and powered by CUDA®, the world's most pervasive parallel computing model. They include innovative technologies like Dynamic Parallelism and Hyper-Q to boost performance as well as power efficiency and deliver record application speeds 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 Kepler compute architecture design includes:

SMX (streaming multiprocessor) design that delivers up to 3x more performance per watt compared to the SM in Fermi[1]. It also delivers one petaflop of computing in just ten 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. Plus it 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. This dramatically increases GPU utilization, slashes CPU idle times, and advances programmability—ideal for cluster applications that use MPI.

The Tesla K-series family of products includes:

Tesla K10 GPU Accelerator – Optimized for single precision applications, the Tesla K10 includes two ultra-efficient GK104 Kepler GPUs to deliver high throughput. It delivers up to 2x the performance for single precision applications compared to the previous generation Tesla M2090 GPU 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 and K20X GPU Accelerators – Designed to be the performance leader in double precision applications and the broader supercomputing market, the Tesla K20 and K20X GPU Accelerators deliver 10x the performance of a single CPU[2]. Tesla K20 and K20X both feature a single GK110 Kepler GPU that includes the Dynamic Parallelism and Hyper-Q features. With more than one teraflop peak double precision performance, these GPU accelerators are ideal for the most aggressive 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 = 410 gigaflops, Tesla K20 (expected) > 1000 gigaflops

[2] Based on WS-LSMS performance comparison between single E5-2687W @ 3.10GHz vs single Tesla K20X. Tesla K20X > 650 gigaflops

0.19 teraflops	1.17 teraflops	1.31 teraflops
4.58 teraflops	3.52 teraflops	3.95 teraflops
2 x GK104s	1 x GK110	
2 x 1536	2496	2688
8 GB	5 GB	6 GB
320 GBytes/sec	208 GBytes/sec	250 GBytes/sec
Seismic, image, signal processing, video analytics	CFD, CAE, financial computing, computational chemistry and physics, data analytics, satellite imaging, weather modeling	
SMX	SMX, Dynamic Parallelism, Hyper-Q	
Servers only	Servers and Workstations	Servers only

[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.

Meets a critical requirement for computing accuracy and reliability in data centers and supercomputing centers. External DRAM is ECC protected in Tesla K10. Both external and internal memories are ECC protected in Tesla K20 and K20X.
Integrates the GPU subsystem with the host system's monitoring and management capabilities such as IPMI or OEM-proprietary tools. IT staff can now manage the GPU processors in the computing system using widely used cluster/grid management solutions.
Accelerates algorithms such as physics solvers, ray-tracing, and sparse matrix multiplication where data addresses are not known beforehand.
Turbocharges system performance by transferring data over the PCIe bus while the computing cores are crunching other data.
Choose OpenACC, CUDA toolkits for C, C++, or Fortran to express application parallelism and take advantage of the innovative Kepler architecture.

> Software applications page:

> 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 the driver is qualified by the OEM on their system. The latest drivers can be downloaded from

> Learn more about Tesla data center management tools at

> Software development tools are available at

To learn more about NVIDIA Tesla, go to