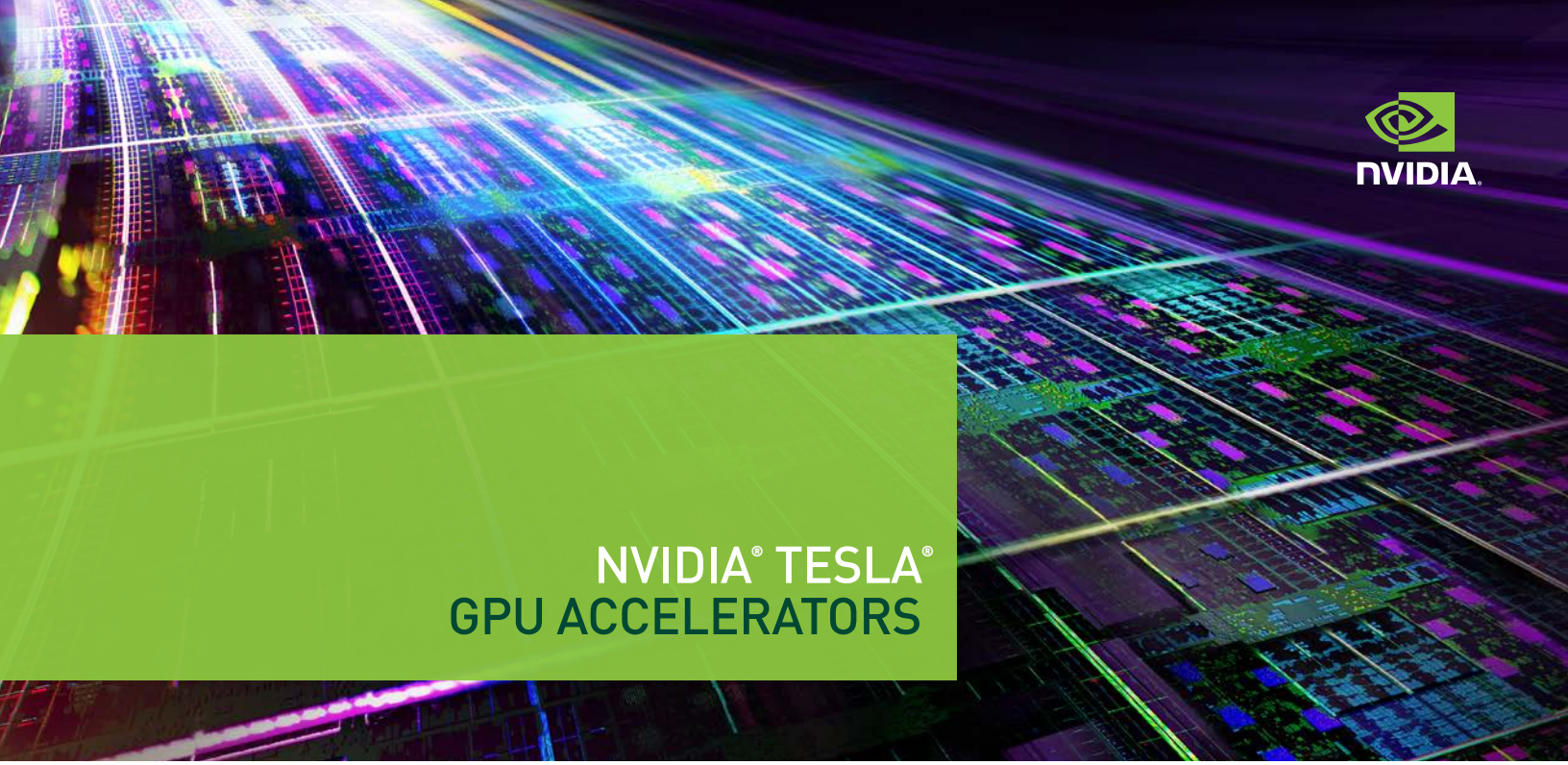

NVIDIA® TESLA® GPU ACCELERATORS

World's fastest accelerators

Solve your most demanding High-Performance Computing (HPC) challenges with NVIDIA Tesla family of GPUs. They're built on the NVIDIA Kepler™ compute architecture and powered by NVIDIA CUDA®, the world's most pervasive parallel computing model. This makes them ideal for delivering record acceleration and more efficient compute performance for big data applications in fields including seismic processing; computational biology and chemistry; weather and climate modeling; image, video and signal processing; computational finance, computational physics; CAE and CFD; and data analytics.

The innovative Kepler compute architecture design includes the following powerful technology features:

SMX (streaming multiprocessor)

Delivers up to 3x more performance per watt than the SM in last-generation NVIDIA Fermi GPUs[1].

Dynamic Parallelism

Enables GPU threads to automatically spawn new threads. By adapting to the data without going back to the CPU, this greatly simplifies parallel programming.

Hyper-Q

Allows multiple CPU cores to simultaneously use the CUDA cores on a single Kepler GPU. This dramatically increases GPU utilization and slashes CPU idle times.

The Tesla Kepler-based family of GPUs includes:

Tesla K40 GPU Accelerator

Equipped with 12 GB of memory, the Tesla K40 GPU accelerator is ideal for the most demanding HPC and big data problem sets. It outperforms CPUs by up to 10x[2] and includes a Tesla GPUBoost[3] feature that enables power headroom to be converted into user-controlled performance boost.

Tesla K20 and K20X GPU Accelerators

Designed for double-precision applications across the broader supercomputing market, the Tesla K20X delivers over 1.31 TFlops peak double-precision performance while the Tesla K20 delivers 1.17 Tflops.

Tesla K10 GPU Accelerator

Optimized for single-precision applications, the Tesla K10 combines two ultra-efficient Kepler GPUs to provide high throughput for computations in seismic, signal image processing, and video analytics. The K10 GPU delivers up to 2x the performance of the previous-generation Tesla M2090 GPU for single-precision applications.

1. Based on DGEMM performance: Tesla M2090 = 410 gigaflops, Tesla K20 > 1000 gigaflops | 2. Based on SPECFEM3D performance comparison between single E5-2687W @ 3.20GHz vs single Tesla K40 | 3. For details on GPUBoost refer to the K40 Board spec on **www.nvidia.com/object/tesla_product_literature.html**

TECHNICAL SPECIFICATIONS

	TESLA K40	TESLA K20X	TESLA K20	TESLA K10[1]
Peak double-precision floating point performance (board)	1.43 Tflops	1.31 Tflops	1.17 Tflops	0.19 Tflops
Peak single-precision floating point performance (board)	4.29 Tflops	3.95 Tflops	3.52 Tflops	4.58 Tflops
Number of GPUs	1 x GK110B	1 x GK110		2 x GK104s
Number of CUDA cores	2,880	2,688	2,496	2 x 1,536
Memory size per board (GDDR5)	12 GB	6 GB	5 GB	8 GB
Memory bandwidth for board (ECC off)[2]	288 Gbytes/sec	250 Gbytes/sec	208 Gbytes/sec	320 Gbytes/sec
Architecture features	SMX, Dynamic Parallelism, Hyper-Q			SMX
System	Servers and workstations	Servers	Servers and workstations	Servers

FEATURES AND BENEFITS

Memory error protection	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 K40, K20X, and K20.
System monitory features	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.
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
Tesla GPUBoost	End-user can convert power headroom to higher clocks and achieve even greater acceleration for various HPC workloads on Tesla K40.
Flexible programming environment with broad support of programming language 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: **www.nvidia.com/teslaapps**

> Tesla GPU computing accelerators are supported for both Linux (64-bit) and Windows (64-bit).

> Drivers - NVIDIA recommends users get their drivers for Tesla server products from their system OEM to ensure the driver is qualified by the OEM on their system. The latest drivers can be downloaded from **www.nvidia.com/drivers**

> Learn more about Tesla data center management tools at **www.nvidia.com/softwarefortesla**

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

1. Tesla K10 specifications are shown as aggregate of two GPUs. | 2. With ECC on, 6.25% of the GPU memory is used for ECC bits. So, for example, 6 GB total memory yields 5.625 GB of user available memory with ECC on.

NVIDIA