NVIDIA

NVIDIA® TESLA® K40
WORLD'S FASTEST ACCELERATORS

PART NUMBER:
TCSK40M-PB

TESLA K40 GPU COMPUTING ACCELERATOR

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.

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 usercontrolled performance boost.

The innovative design of the Kepler compute architecture includes:

>> **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.

TESLA K40 Module - PRODUCT SPECIFICATIONS

CUDA PARALLEL PROCESSING CORES	**2880**
FRAME BUFFER MEMORY	**12 GB GDDR5**
PEAK DOUBLE PRECISION FLOATING POINT PERFORMANCE	**1.43 Tflops**
PEAK SINGLE PRECISION FLOATING POINT PERFORMANCE	**4.29 Tflops**
INTERFACE	**384-bit**
MEMORY BANDWIDTH	**288 GB/s**
DISPLAY CONNECTORS	**None**
MAX POWER CONSUMPTION	**235 W**
PROCESSOR CORE CLOCK	**745 MHz**
POWER CONNECTORS	**1 × 6-pin PCI Express power connectors** **1 × 8-pin PCI Express power connectors**
GRAPHICS BUS	**PCI Express 3.0 x16**
FORM FACTOR	**110 mm (H) × 265 mm (L) - Dual Slot, Full-Height**
THERMAL SOLUTION	**Passive**

PNY Technologies Europe
sales@pny.eu | T: +33 (0)5 56 13 75 75

Tesla GPU Computing Accelerator Common Features

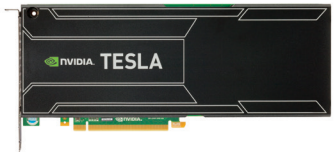
ECC MEMORY ERROR PROTECTION	Meets a critical requirement for computing accuracy and reliability in datacenters and supercomputing centers. Both external and internal memories are ECC protected in Tesla K40.
SYSTEM MONITORING FEATURES	Integrates the GPU subsystem with the host system's monitoring and 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.
TESLA GPUBoost	End-user can convert power headroom to higher clocks and achieve even greater acceleration for various HPC workloads on Tesla K40.

Software and Drivers

Software applications page	www.nvidia.com/teslaapps
Tesla GPU computing accelerators are supported for both Linux and Windows.	
Drivers	www.pny.eu/drivers
Learn more about Tesla data center management tools at	www.nvidia.com/softwarefortesla

Technical Specifications

	TESLA K40	**TESLA K20X**	**TESLA K20**	**TESLA K10[4]**
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	2880	2688	2496	2 x 1536
Memory size per board (GDDR5)	12GB	6GB	5GB	8GB
Memory bandwidth for board (ECC off)[2]	288 GB/s	250 GB/s	208 GB/s	320 GB/s
Architecture features	SMX, Dynamic Parallelism, Hyper-Q			SMX
System	Servers & workstations	Servers	Servers & workstations	Servers



1. Based on DGEMM performance: Tesla M2090 = 410 gigaflops, Tesla K40 > 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
4. Tesla K10 specifications are shown as aggregate of two GPUs.

PNY Technologies Europe
sales@pny.eu | T: +33 (0)5 56 13 75 75