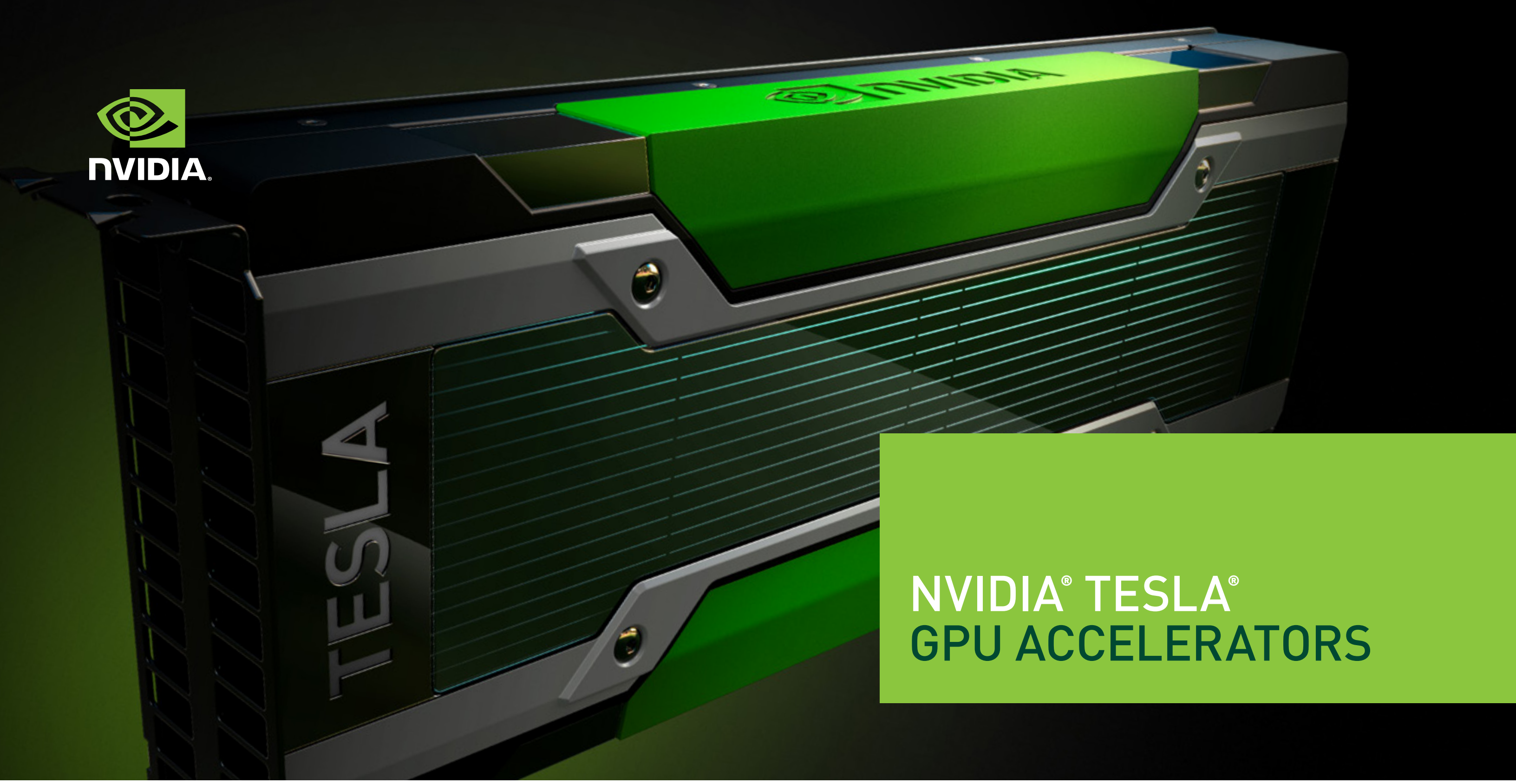


NVIDIA® TESLA® GPU ACCELERATORS

The world's fastest accelerators

Accelerate your most demanding high-performance data analytics and scientific computing applications with the NVIDIA Tesla Accelerated-Computing Platform.

Tesla GPU Accelerators are built on the NVIDIA Kepler™ compute architecture and powered by CUDA®, the world's most pervasive parallel-computing model. This makes them ideal for delivering record acceleration and compute performance efficiency for applications in fields including:

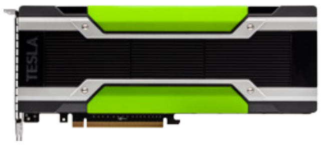
- Machine Learning and Data Analytics
- Seismic Processing
- Computational Biology and Chemistry
- Weather and Climate Modeling
- Image, Video, and Signal Processing
- Computational Finance/Physics
- CAE and CFD

Tesla Accelerated Computing Platform

The Kepler-based Tesla family of GPUs is part of the innovative Tesla Accelerated Computing Platform. As the leading platform for accelerating data analytics and scientific computing, it combines the world's fastest GPU accelerators, the widely used CUDA parallel computing model, and a comprehensive ecosystem of software developers, software vendors, and datacenter system OEMs.

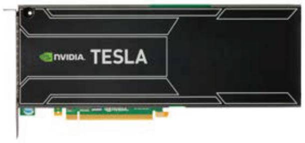
The Tesla family of GPU Accelerators includes:

Tesla K80 GPU Accelerator

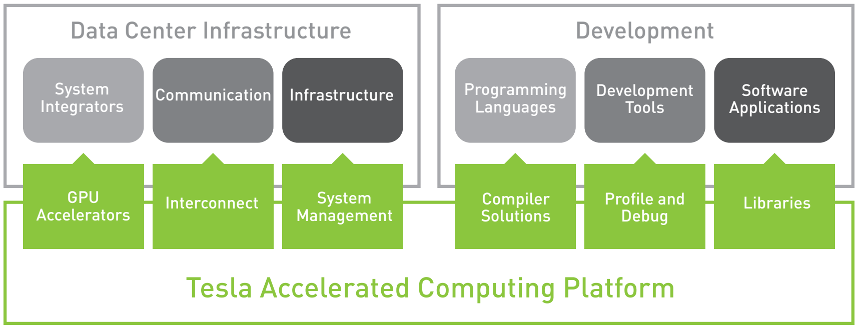


This accelerator is designed for the most demanding computational tasks, combining 24 GB of memory with blazing-fast memory bandwidth and leading compute performance for single and double precision workloads. Equipped with the latest NVIDIA GPU Boost™ technology, the Tesla K80 intelligently monitors GPU usage to maximize throughput[1] and outperforms CPUs by up to 10x[2].

Tesla K40 GPU Accelerator



This is a flexible solution for applications in high-performance computing and data analysis. The Tesla K40 comes equipped with 12 GB of memory, delivers 1.43 TFlops of double precision performance, and includes GPU Boost, enabling power headroom to be converted in a user-controlled performance increase[1].



Data Center Infrastructure

System Integrators | Communication | Infrastructure

GPU Accelerators | Interconnect | System Management

Development

Programming Languages | Development Tools | Software Applications

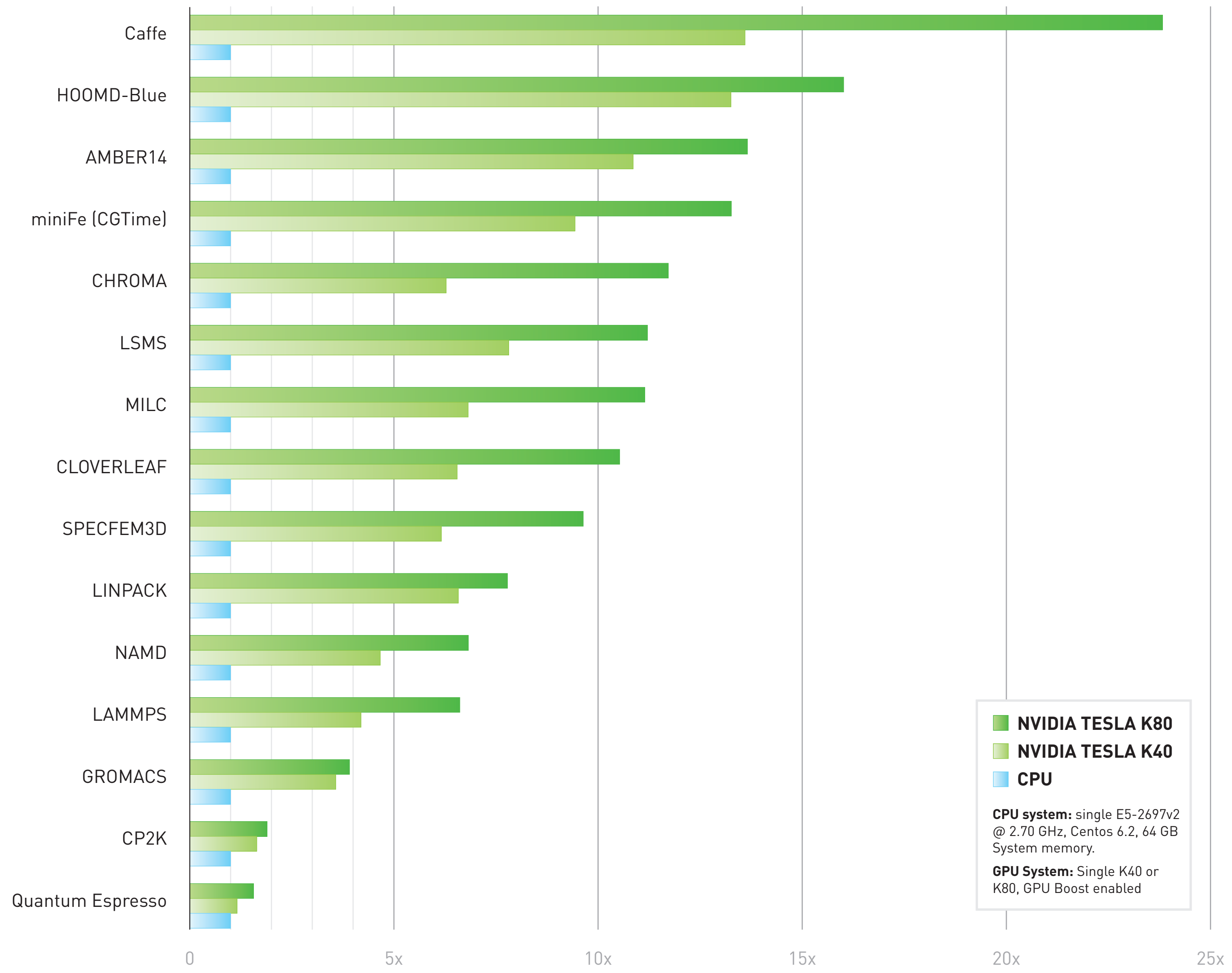
Compiler Solutions | Profile and Debug | Libraries

Tesla Accelerated Computing Platform

[1] For details on GPU Boost, refer to the GPU Boost Application Note on www.nvidia.com/object/tesla_product_literature.html

[2] Based on AMBER14 performance comparison between single E5-2697v2 @ 2.70 GHz vs single Tesla K80

TESLA GPU ACCELERATOR PERFORMANCE



CPU system: single E5-2697v2 @ 2.70 GHz, Centos 6.2, 64 GB System memory.

GPU System: Single K40 or K80, GPU Boost enabled

TECHNICAL SPECIFICATIONS

	Tesla K40	Tesla K80[1]
Peak double-precision floating point performance (board)	1.43 Tflops	1.87 Tflops
Peak single-precision floating point performance (board)	4.29 Tflops	5.6 Tflops
GPU	1 x GK110B	2 x GK210
CUDA cores	2,880	4,992
Memory size per board (GDDR5)	12 GB	24 GB
Memory bandwidth for board (ECC off)[2]	288 Gbytes/sec	480 Gbytes/sec
Architecture features	SMX, Dynamic Parallelism, Hyper-Q	
System	Servers and workstations	Servers

[1] Tesla K80 specifications are shown as aggregate of two GPUs.

[2] With ECC on, 6.25% of the GPU memory is used for ECC bits. For example, 6 GB total memory yields 5.625 GB of user available memory with ECC on.

FEATURES	Tesla K40	Tesla K80
Dynamic Parallelism Enables GPU threads to automatically spawn new threads. By adapting to the data without going back to the GPU, this greatly simplifies parallel programming.	✓	✓
Hyper-Q Allows multiple CPU cores to simultaneously use the CUDA cores on a single or multiple Kepler-based GPUs. This dramatically increases GPU utilization, simplifies programming, and slashes CPU idle times.	✓	✓
System Monitoring 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	✓	✓
Memory Error Protection Meets a critical requirement for computing accuracy and reliability in data centers and supercomputing centers. Both external and internal memories are ECC protected in the Tesla K80 and K40.	✓	✓
Asynchronous Transfer with Dual DMA Engines Turbocharges system performance by transferring data over the PCIe bus while the computing cores are crunching other data	✓	✓
GPU Boost Enables the end-user to convert power headroom to higher clocks and achieve even greater acceleration for various HPC workloads	✓	✓
Dynamically scales GPU clocks for maximum application performance and improved energy efficiency		✓
Flexible Programming Environment with Broad Support of Programming Language and APIs Offers the freedom to choose OpenACC, CUDA toolkits for C, C++, or Fortran to express application parallelism and take advantage of the innovative Kepler architecture	✓	✓
2x Shared Memory and 2x Register File Increases effective throughput and bandwidth with 2x shared memory and 2x register file compared to the K40		✓
Zero-power Idle Increases data center energy efficiency by powering down idle GPUs when running legacy non-accelerated workloads		✓

SOFTWARE AND DRIVERS

- Software applications page: **www.nvidia.com/teslaapps**
- 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**
- Tesla GPU computing accelerators are supported for both Linux (64-bit) and Windows (64-bit).
- 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**