NVIDIA Accelerators for HPE ProLiant Servers

Hewlett Packard Enterprise supports, on select HPE ProLiant servers, computational accelerator modules based on NVIDIA® Tesla™, NVIDIA® GRID™, and NVIDIA® Quadro™ Graphical Processing Unit (GPU) technology.

Based on NVIDIA's CUDA™ architecture, the NVIDIA accelerators enable seamless integration of GPU computing with HPE ProLiant servers for high-performance computing, large data center graphics and virtual desktop deployments. These accelerators deliver all of the standard benefits of GPU computing while enabling maximum reliability and tight integration with system monitoring and management tools such as HPE Insight Cluster Management Utility.

The NVIDIA Tesla GPUs are general purpose accelerators which excel at boosting performance of structured numerical algorithms. These GPUs are powered by CUDA® and include technologies like Dynamic Parallelism and Hyper-Q to boost performance as well as power efficiency. Applications which benefit from accelerators include deep learning, seismic processing, biochemistry simulations, weather and climate modeling, image, video and signal processing, computational finance, computational physics, CAE, CFD, and data analytics.

The NVIDIA GRID GPUs are optimized for virtual desktop infrastructures (VDI). The Grid K1 adaptor has 4 GPUs on a single PCIe card, and supports large numbers of users with standard desktop applications. Also, together with NVIDIA GRID 2.0 software (to be purchased separately), VDI supported adapters can be used as GRID GPUs for highest-end virtual desktop applications.
See **https://www.hpe.com/us/en/product-catalog/storage/complete-storage-solution/pip.hpe-complete-nvidia-grid-solutions.1009949096.html** for details on required NVIDIA GRID software licenses.

The HPE GPU Ecosystem includes HPE Cluster Platform specification and qualification, HPE-supported GPU-aware cluster software, and also third-party GPU-aware cluster software for NVIDIA Tesla, Quadro and GRID Modules on HPE ProLiant Servers. In particular, the HPE Insight Cluster Management Utility (CMU) will monitor and display GPU health sensors such as temperature. Insight CMU will also install and provision the GPU drivers and the CUDA software. Insight CMU is integrated with popular schedulers such as Adaptive Moab, Altair PBS Professional, and Univa Grid Engine – all of which have the capability of scheduling jobs based on GPU requirements.

Standard Features

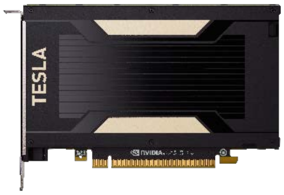
NVIDIA Accelerators

HPE NVIDIA Tesla V100 FHHL16GB Module	Q9U36A
HPE NVIDIA Tesla V100 PCIe 32GB Computational Accelerator	Q9U37A
HPE NVIDIA Tesla V100 SXM2 32GB Computational Accelerator	Q1K34A
HPE NVIDIA Tesla V100 SXM2 16GB Computational Accelerator	Q2N66A
HPE NVIDIA Tesla V100 PCIe 16GB Computational Accelerator	Q2N68A
HPE NVIDIA Tesla P100 SXM2 16GB Computational Accelerator	Q0C71A
HPE NVIDIA Tesla P100 PCIe 16GB Module	Q0E21A
HPE NVIDIA Tesla P100 PCIe 12GB Module	Q2S42A
HPE NVIDIA Tesla M10 Quad GPU Module	Q0J62A
HPE NVIDIA Tesla P40 24GB Computational Accelerator	Q0V80A
HPE NVIDIA Tesla P4 8GB Computational Accelerator	Q0V79A
HPE NVIDIA Quadro GV100 Graphics Accelerator	Q2N67A
HPE NVIDIA Quadro P2000 Graphics Accelerator	Q0V77A
HPE NVIDIA Quadro P4000 Graphics Accelerator	Q0V78A
HPE NVIDIA Quadro P6000 Graphics Accelerator	Q0V76A

NOTE: Please see the HPE ProLiant server Quickspecs for configuration rules including requirements for enablement kits.

Standard Features

Standard Features

HPE NVIDIA Tesla V100 FHHL 16GB Computational Accelerator	
NVIDIA® Tesla® V100 is the world's most advanced data center GPU ever built to accelerate AI, HPC, and Graphics. Powered by the latest GPU architecture, NVIDIA Volta™, Tesla V100 offers the performance of 100 CPUs in a single GPU—enabling data scientists, researchers, and engineers to tackle challenges that were once impossible. The Full Height Half Length form factor offers the same performance and functionality as the PCIe full length card with significantly less power draw 150W vs 250W.	
Form Factor	
Performance	7TF DP | 14TF SP
Memory Size	16GB HBM2 Stacked Memory
Memory Bandwidth	900 GB/s with CoWoS HBM2 Stacked Memory
Cores	5120 CUDA | 640 Tensor
GPU Peer to Peer	PCIe Gen3
Power	150W
Supported Servers	DL380 Gen10
Supported Operating Systems	RHEL7.4 SLES12 SP3 UBUNUTU 16.04 Windows Server 2016 VMware 6.5 U1 Citrix 7.4
	NOTE: NVIDIA GPUs are supported only on 64-bit versions
Product Positioning	Ultimate performance for Deep Leaning, highly versatile for all workloads

HPE NVIDIA Tesla V100 PCIE 32GB Computational Accelerator HPE NVIDIA Tesla V100 SXM2 32GB Computational Accelerator		
NVIDIA® Tesla® V100 now offers a 32GB high bandwidth memory configuration. Providing 2X the memory capacity improves deep learning training performance for next-generation AI models like language translations and ResNet 1K models by over 50%, by training more data in parallel with these larger models. Supporting larger AI data models also improves AI developer productivity, allowing developers to deliver more AI breakthroughs in less time. This higher memory configuration allows HPC applications to run larger simulations more efficiently than ever before.		
Form Factor		
	SXM2	**PCIe**

Standard Features

Performance	7.8TF DP, 15.7TF SP, 125TF FP16	7TF DP, 14TF SP, 112TF FP16	
Memory Size	32GB HBM2	32GB HBM2	
Memory Bandwidth	900 GB/s	900 GB/s	
GPU Peer to Peer	NVLINK	PCIe Gen3	
Power	300W	250W	
Supported Servers	Xl270d Gen10	Xl270d Gen10	
Supported Operating Systems	RHEL7.4 SLES12 SP3 UBUNUTU 16.04 Windows Server 2016		
	NOTE: NVIDIA GPUs are supported only on 64-bit versions		
Product Positioning	**Application**	**Current Product**	**For New Deployment**
	Deep Learning Training	P100, P40, V100 16GB	V100 32GB
	Memory Capacity-Bound HPC (e.g. Seismic, Graphs, CFD, Physics, Climate, GIS, Finance)	K80, P100, V100 16GB	V100 32GB
	Memory Capacity-Bound HPC (e.g. Seismic, Graphs, CFD, Physics, Climate, GIS, Finance)	K80, P100, V100 16GB	V100 32GB

HPE NVIDIA Quadro GV100 Graphics Accelerators

The NVIDIA® Quadro® GV100 reinvents the workstation GPU to meet the demands of AI-enhanced design and visualization workflows. It's powered by NVIDIA Volta, delivering extreme memory capacity, scalability, and performance that designers, architects, and scientists need to create, build, and solve the impossible.

Form Factor	
Performance	7.4TF DP | 14.8TF SP
Memory Size	32GB HBM2 Stacked Memory
Memory Bandwidth	870GB/s
Cores	5120 CUDA | 640 Tensor
GPU Peer to Peer	PCIe Gen3
Power	250W
Supported Servers	DL380 Gen10
Supported Operating Systems	RHEL7.4 SLES12 SP2 Windows Server 2016
NOTE: NVIDIA GPUs are supported only on 64-bit versions	
Product Positioning	Supercharge rendering with AI; Bring Optimal designs to market faster; Accelerate AI training/inferencing with tensor cores

Additional Options

Performance of the Tesla V100 PCIe 16GB Computational Accelerator

- The Tesla V100 for PCIe module is built for HPC and Deep Learning.
- 5120 CUDA cores
- NVIDIA GPU Boost™ enables opportunistic clock frequency bursts provided no thermal or power limits are hit
- 15 Tflops (Boost) of single-precision performance
- 7.5 Tflops (Boost) of double-precision performance
- 16GB CoWoS HBM2 at 900 GB/s
- Power consumption: 250W
- The NVIDIA Parallel DataCache™ accelerates algorithms where data addresses are not known beforehand. This includes a configurable L1 cache per Streaming Multiprocessor block and a unified L2 cache for all of the processor cores.
- Asynchronous transfer turbo charges system performance by transferring data over the PCIe bus while the computing cores are crunching other data. Even applications with heavy data-transfer requirements, such as seismic processing, can maximize the computing efficiency by transferring data to local memory before it is needed.
- Hyper-Q feature that enables multiple CPU cores to simultaneously utilize the CUDA cores on a single GPU.
- The high speed PCIe Gen 3.0 data transfer maximizes bandwidth between the HPE ProLiant server and the Tesla processors.

Performance of the Quadro, P2000, P4000, and P6000 24GB Graphic Accelerators

- P2000 has 1024 cores, P4000 has 1792 cores, and P6000 has 3840 cores.
- P6000 24GB has 24 GB GDDR5 memory
- Power consumption: 75W,105W, 250W
- Support OpenGL 4.3, Shader Model 5.0, DirectX 11
- Dedicated H.264 encode engine that's independent of 3D/compute pipeline and delivers real-time performance for transcoding, video editing, and other encoding applications.
- Provides the ability to texture from and render to 16K x 16K surfaces. This is beneficial for applications that demand the highest resolution and quality image processing.
- NVIDIA SMX delivers more processing performance and efficiency through a new, innovative streaming multiprocessor design that allows a greater percentage of space to be applied to processing cores versus control logic, enabling greater model complexity.
- Hyper-Q feature that enables multiple CPU cores to simultaneously utilize the CUDA cores on a single GPU.
- The high speed PCIe Gen 3.0 data transfer, available on the, P4000 (8GB), and P6000 (24GB) maximizes bandwidth between the HPE ProLiant server and the Tesla processors.

Performance of the Tesla P100 SXM2 16GB Computational Accelerators

- The Tesla P100 is built for HPC and Deep Learning.
- 3584 CUDA cores
- NVIDIA GPU Boost™ enables opportunistic clock frequency bursts provided no thermal or power limits are hit
- 10.6 Tflops (Boost) of single-precision performance
- 5.3 Tflops (Boost) of double-precision performance
- 21.2 Tflops (Boost) of Half-precision performance
- 16GB CoWoS HBM2 at 732 GB/s
- Power consumption: 300W
- The NVIDIA Parallel DataCache™ accelerates algorithms where data addresses are not known beforehand. This includes a configurable L1 cache per Streaming Multiprocessor block and a unified L2 cache for all of the processor cores.
- Tesla P100 also features NVIDIA NVLink™ technology that enables superior strong-scaling performance for HPC and hyperscale applications. Up to eight Tesla. P100 GPUs interconnected in a single node can deliver the performance of racks of commodity CPU servers.
- Hyper-Q feature that enables multiple CPU cores to simultaneously utilize the CUDA cores on a single GPU.

Additional Options

- VDI Enabled. See **https://www.hpe.com/us/en/product-catalog/storage/complete-storage-solution/pip.hpe-complete-nvidia-grid-solutions.1009949096.html** for details on required NVIDIA GRID software licenses.

Performance of the Tesla P100 PCIe 12GB Module

- The Tesla P100 for PCIe is built for HPC and Deep Learning.
- 3584 CUDA cores
- NVIDIA GPU Boost™ enables opportunistic clock frequency bursts provided no thermal or power limits are hit
- 9.3 Tflops (Boost) of single-precision performance
- 4.7 Tflops (Boost) of double-precision performance
- 12GB CoWoS HBM2 at 549 GB/s
- Power consumption: 300W
- The NVIDIA Parallel DataCache™ accelerates algorithms where data addresses are not known beforehand. This includes a configurable L1 cache per Streaming Multiprocessor block and a unified L2 cache for all of the processor cores.
- Asynchronous transfer turbo charges system performance by transferring data over the PCIe bus while the computing cores are crunching other data. Even applications with heavy data-transfer requirements, such as seismic processing, can maximize the computing efficiency by transferring data to local memory before it is needed.
- Hyper-Q feature that enables multiple CPU cores to simultaneously utilize the CUDA cores on a single GPU.
- The high speed PCIe Gen 3.0 data transfer maximizes bandwidth between the HPE ProLiant server and the Tesla processors.

Performance of the Tesla P40 24GB Computational Accelerator

- The Tesla P40 is for single-precision, especially deep-learning, applications
- 3840 CUDA cores
- GPU Boost enables opportunistic clock frequency bursts provided no thermal or power limits are hit
- 12 Tflops of single-precision peak performance
- Power consumption: 250W
- 24 GB of GDDR5 memory optimizes performance and reduces data transfers by keeping large data sets in local memory
- The NVIDIA Parallel DataCache™ accelerates algorithms where data addresses are not known beforehand. This includes a configurable L1 cache per Streaming Multiprocessor block and a unified L2 cache for all of the processor cores.
- Asynchronous transfer turbo charges system performance by transferring data over the PCIe bus while the computing cores are crunching other data. Even applications with heavy data-transfer requirements, such as seismic processing, can maximize the computing efficiency by transferring data to local memory before it is needed.
- Dynamic Parallelism capability that enables GPU threads to automatically spawn new threads.
- Hyper-Q feature that enables multiple CPU cores to simultaneously utilize the CUDA cores on a single GPU.
- The high speed PCIe Gen 3.0 data transfer maximizes bandwidth between the HPE ProLiant server and the Tesla processors.
- VDI Enabled. See **https://www.hpe.com/us/en/product-catalog/storage/complete-storage-solution/pip.hpe-complete-nvidia-grid-solutions.1009949096.html** for details on required NVIDIA GRID software licenses.

Performance of the Tesla P4 8GB Computational Accelerator

- The Tesla M4 is for single-precisions, especially deep learning and inference.
- 2560 CUDA cores
- NVIDIA GPU Boost™ enables opportunistic clock frequency bursts provided no thermal or power limits are hit
- 5.5 Tflops (Boost) of single-precision performance
- Total 8 GB of GDDR5 memory optimizes performance and reduces data transfers by keeping large data sets in local memory
- Power consumption: 75W
- The NVIDIA Parallel DataCache™ accelerates algorithms where data addresses are not known beforehand. This includes a configurable L1 cache per Streaming Multiprocessor block and a unified L2 cache for all of the processor cores.

Additional Options

- Asynchronous transfer turbo charges system performance by transferring data over the PCIe bus while the computing cores are crunching other data. Even applications with heavy data-transfer requirements, such as seismic processing, can maximize the computing efficiency by transferring data to local memory before it is needed.
- Hyper-Q feature that enables multiple CPU cores to simultaneously utilize the CUDA cores on a single GPU.
- The high speed PCIe Gen 3.0 data transfer maximizes bandwidth between the HPE ProLiant server and the Tesla processors.
- VDI Enabled. See **https://www.hpe.com/us/en/product-catalog/storage/complete-storage-solution/pip.hpe-complete-nvidia-grid-solutions.1009949096.html** for details on required NVIDIA GRID software licenses.

Performance of the Tesla M10 Quad Module

- The Tesla M60 and Tesla M60 RAF are for GRID computing only. To enable NVIDIA GRID, customer must purchase GRID licensing from an authorized NVIDIA distributor.
- 2560 CUDA cores (640 per GPU)
- GPU Boost enables opportunistic clock frequency bursts provided no thermal or power limits are hit
- Total 32 GB of GDDR5 memory optimizes performance and reduces data transfers
- Power consumption: 225W
- The NVIDIA Parallel DataCache™ accelerates algorithms such as physics solvers, ray-tracing, and sparse matrix multiplication where data addresses are not known beforehand. This includes a configurable L1 cache per Streaming Multiprocessor block and a unified L2 cache for all of the processor cores.
- Asynchronous transfer turbo charges system performance by transferring data over the PCIe bus while the computing cores are crunching other data. Even applications with heavy data-transfer requirements, such as seismic processing, can maximize the computing efficiency by transferring data to local memory before it is needed.
- Dynamic Parallelism capability that enables GPU threads to automatically spawn new threads.
- Hyper-Q feature that enables multiple CPU cores to simultaneously utilize the CUDA cores on a single GPU.
- The high speed PCIe Gen 3.0 data transfer maximizes bandwidth between the HPE ProLiant server and the Tesla processors.
- VDI Enabled. See **https://www.hpe.com/us/en/product-catalog/storage/complete-storage-solution/pip.hpe-complete-nvidia-grid-solutions.1009949096.html** for details on required NVIDIA GRID software licenses.

Performance of the Tesla P100 PCIe 16GB Module

- The Tesla P100 for PCIe is built for HPC and Deep Learning.
- 3584 CUDA cores
- NVIDIA GPU Boost™ enables opportunistic clock frequency bursts provided no thermal or power limits are hit
- 9.3 Tflops (Boost) of single-precision performance
- 4.7 Tflops (Boost) of double-precision performance
- 16GB CoWoS HBM2 at 720 GB/s
- Power consumption: 250W
- The NVIDIA Parallel DataCache™ accelerates algorithms where data addresses are not known beforehand. This includes a configurable L1 cache per Streaming Multiprocessor block and a unified L2 cache for all of the processor cores.
- Asynchronous transfer turbo charges system performance by transferring data over the PCIe bus while the computing cores are crunching other data. Even applications with heavy data-transfer requirements, such as seismic processing, can maximize the computing efficiency by transferring data to local memory before it is needed.
- Hyper-Q feature that enables multiple CPU cores to simultaneously utilize the CUDA cores on a single GPU.
- The high speed PCIe Gen 3.0 data transfer maximizes bandwidth between the HPE ProLiant server and the Tesla processors.

Programming and Management Ecosystem

- The CUDA programming environment has broad support of programming languages and APIs. Choose C, C++, OpenCL, DirectCompute, or Fortran to express application parallelism and take advantage of the innovative Tesla architectures.

Additional Options

The CUDA software, as well as the GPU drivers, can be automatically installed on HPE ProLiant servers, by HPE Insight Cluster Management Utility.

- Exclusive mode" enables application-exclusive access to a particular GPU. CUDA environment variables enable cluster management software to limit the Tesla and GRID GPUs an application can use.
- With HPE ProLiant servers, application programmers can control the mapping between processes running on individual cores, and the GPUs with which those processes communicate. By judicious mappings, the GPU bandwidth, and thus overall performance, can be optimized. The technique is described in a white paper available to Hewlett Packard Enterprise customers at: **https://www.hpe.com/us/en/servers/density-optimized.html** A heuristic version of this affinity-mapping has also been implemented by Hewlett Packard Enterprise as an option to the micron command as used for example with HPE-MPI, available as part of HPE HPC Linux Value Pack.

GPU control is available through the nvidia-smi tool which lets you control compute-mode (e.g. exclusive), enable/disable/report ECC and check/reset double-bit error count. IPMI and iLO gather data such as GPU temperature. HPE Cluster Management Utility has incorporated these sensors into its monitoring features so that cluster-wide GPU data can be presented in real time, can be stored for historical analysis and can be easily used to set up management alerts.

Service and Support

Service and Support

If this is a qualified option, it is covered under the HPE Support Service(s) applied to the HPE ProLiant Server. Please check HPE ProLiant Server documentation for more details on the services for this particular option.

Warranty and Support Services

Warranty and Support Services will extend to include HPE options configured with your server or storage device. The price of support service is not impacted by configuration details. HPE sourced options that are compatible with your product will be covered under your server support at the same level of coverage allowing you to upgrade freely. Installation for HPE options is available as needed. To keep support costs low for everyone, some high value options will require additional support. Additional support is only required on select high value workload accelerators, fibre switches, InfiniBand and UPS options 12KVA and over. Coverage of the UPS battery is not included under TS support services; standard warranty terms and conditions apply.

Protect your business beyond warranty with HPE Support Services

HPE Technology Services delivers confidence, reduces risk and helps customers realize agility and stability.

Connect to HPE to help prevent problems and solve issues faster. HPE Support Services enable you to choose the right service level, length of coverage and response time as you purchase your new server, giving you full entitlement to the support you need for your IT and business.

Protect your product, beyond warranty.

Parts and materials

Hewlett Packard Enterprise will provide HPE-supported replacement parts and materials necessary to maintain the covered hardware product in operating condition, including parts and materials for available and recommended engineering improvements.

Parts and components that have reached their maximum supported lifetime and/or the maximum usage limitations as set forth in the manufacturer's operating manual, product quick-specs, or the technical product data sheet will not be provided, repaired, or replaced as part of these services.

The defective media retention service feature option applies only to Disk or eligible SSD/Flash Drives replaced by Hewlett Packard Enterprise due to malfunction.

For more information

Visit the Hewlett Packard Enterprise Service and Support **website**.

Summary of Changes

Date	Version History	Action	Description of Change
4-Jun-2018	Version 24	Updated	Changes made throughout document; adding V100 FHHL card
7-May-2018	Version 23	Updated	Add HPE NVIDIA Tesla V100 PCIE 32GB Computational Accelerator / HPE NVIDIA Tesla V100 SXM2 32GB Computational Accelerator
2-Apr-2018	Version 22	Changed	Updates throughout document; added GV100
4-Dec-2017	From version 20 to 21	Updated	New modules added
16-Oct-2017	From version 19 to 20	Updated	Update verbiage in the Standard
25-Sept-2017	From version 18 to 19	Updated	Update verbiage in the Standard Features Sections
11-Jul-2017	From version 17 to 18	Updated	New models were added and updates throughout the whole document
05-Jun-2017	From version 16 to 17	Updated	Remove obsolete info and SKUs and update the recent info
08-May-2017	From version 15 to 16	Updated	New models were added and updates throughout the whole document
28-Nov-2016	From version 14 to 15	Changed	Updates throughout the whole document
26-Sep-2016	From version 13 to 14	Updated	Added new SKUs to QS
6-Jun -2016	From version 12 to 13	Updated	Add The new information to the QS and remove obsolete SKU´s
31-Mar-2016	From version 11 to 12	Updated	Update all the sections and general info throughout the QuickSpecs
01-Dec-2015	From version 10 to 11	Updated	Update the Standard Features and the technical Specifications section
17-Aug-2015	From version 9 to 10	Changed	Update several Overview and technical specifications.
09-Feb-2015	From version 8 to 9	Changed	Update several Overview and technical specifications.
01-Dec-2014	From version 7 to 8	Revised	Revised wording and Technical Specifications
09-Sept-2014	From Version 6 to 7	Changed	Changes made throughout the QuickSpecs.
05-Jun-2014	From Version 5 to 6	Changed	High Performance Clusters and Thermal Solutions were revised
31-Mar-2014	From Version 4 to 5	Added	NVIDIA Tesla K40C 12 GB Computational Accelerator and NVIDIA Quadro K2000 PCIe Graphics Adapter were added
18-Feb-2014	From Version 3 to 4	Changed	Changes made throughout the QuickSpecs
09-Dec-2013	From Version 2 to 3	Added	NVIDIA Tesla K10 Rev B Dual GPU Module and NVIDIA Tesla K40 12 GB Module were added.
20-Sep-2013	From Version 1 to 2	Changed	Changes made in the following Sections Standard Features Optional Features Technical Specifications

Sign up for updates

Hewlett Packard
Enterprise