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

The following Tesla Modules are available from HP, for use in certain HP SL-series servers.

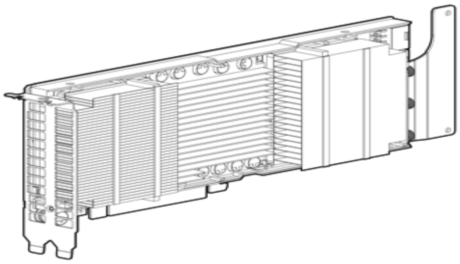
- NVIDIA Tesla K10 Dual GPU PCIe Module
- NVIDIA Tesla K10 Rev B Dual GPU Module
- NVIDIA Tesla K20 5 GB Module
- NVIDIA Tesla K20X 6 GB Module
- NVIDIA Tesla K40 12 GB Module
- NVIDIA GRID K2 PCIe GPU Kit

The NVIDIA GRID K2 PCIe GPU Kit can also be used in HP ProLiant WS460c workstation blades.

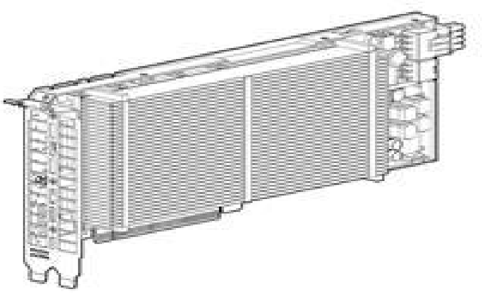
Based on NVIDIA's CUDA™ architecture, the Tesla and GRID Modules enable seamless integration of GPU computing with HP ProLiant servers for high-performance computing and large data center, scale-out deployments. These Tesla and GRID Modules deliver all of the standard benefits of GPU computing while enabling maximum reliability and tight integration with system monitoring and management tools such as HP Insight Cluster Management Utility. The GRID K2 uses the NVIDIA Kepler GPU architecture to provide NVIDIA Quadro® professional-class visualization for graphics intensive applications and for virtual desktop environments.

The GRID K2 uses the NVIDIA Kepler GPU architecture that combines Tesla's high performance computing - found in other Tesla Modules - and the NVIDIA Quadro® professional-class visualization in the same GPU. The GRID K2 is the ideal solution for customers who want to deploy high performance computing in addition to advanced and remote visualization in the same datacenter.

The HP GPU Ecosystem includes HP Cluster Platform specification and qualification, HP-supported GPU-aware cluster software, and also third-party GPU-aware cluster software for NVIDIA Tesla and GRID Modules on HP ProLiant Servers. In particular, the HP 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 IBM Platform LSF - all of which have the capability of scheduling jobs based on GPU requirements.



NVIDIA K10, K2



NVIDIA K20, K20X, K40

What's New

- Support for the NVIDIA Tesla K40 12 GB Module

Models

NVIDIA Passive Tesla Modules	NVIDIA GRID K2 Dual GPU PCIe Graphics Accelerator **NOTE:** 2-slot passively cooled GRID module with 8 GB memory. **NOTE:** Please see the HP ProLiant SL250s and SL270s Gen8 server QuickSpecs or HP ProLiant WS460c Generation 8 Workstation Blade QuickSpecs for configuration rules including requirements, if any, for enablement kits. http://h18004.www1.hp.com/products/quickspecs/14232_div/14232_div.html http://h18004.www1.hp.com/products/quickspecs/14405_div/14405_div.html http://h18004.www1.hp.com/products/quickspecs/13713_div/13713_div.html http://h18004.www1.hp.com/products/quickspecs/14409_div/14409_div.html	729851-B21
	NVIDIA Tesla K10 Dual GPU PCIe Computational Accelerator	B3M66A
	NVIDIA Tesla K10 Rev B Dual GPU PCIe Computational Accelerator **NOTE:** 2-slot passively cooled pair of Tesla GPUs each with 4 GB memory. **NOTE:** Please see the HP ProLiant SL250s and SL270s Gen 8 (SE) server QuickSpecs for configuration rules including requirements, if any, for enablement kits. http://h18004.www1.hp.com/products/quickspecs/14232_div/14232_div.html http://h18004.www1.hp.com/products/quickspecs/14405_div/14405_div.html	E5V47A
	NVIDIA Tesla K20 5 GB Computational Accelerator	C7S14A
	NVIDIA Tesla K20X 6 GB Computational Accelerator	C7S15A
	NVIDIA Tesla K40 12 GB Computational Accelerator **NOTE:** 2-slot passively cooled Tesla GPUs based on NVIDIA Kepler architecture. **NOTE:** Please see the HP ProLiant SL250s and SL270s Gen8 server QuickSpecs for configuration rules including requirements, if any, for enablement kits. http://h18004.www1.hp.com/products/quickspecs/14232_div/14232_div.html http://h18004.www1.hp.com/products/quickspecs/14405_div/14405_div.html	F1R08A

K2, K10, K20, K20X and K40 Modules

Performance of the K2 Module

- 3072 CUDA cores (1536 per GPU)
- GDDR5 memory optimizes performance and reduces data transfers by keeping large data sets in 8 GB of local memory, 4 GB attached directly to each GPU.
- The Kepler GPU includes a high-performance H.264 encoding engine capable of encoding simultaneous streams with superior quality. This provides a giant leap forward in cloud server efficiency by offloading the CPU from encoding functions and allowing the encode function to scale with the number of GPUs in a server.
- GRID boards enable GPU-capable virtualization solutions from Citrix, Microsoft, and VMware, delivering the flexibility to choose from a wide range of proven solutions.
- The high speed PCIe Gen 3.0 data transfer maximizes bandwidth between the HP ProLiant server and the GRID processors.

Performance of the K10 Module

- 3072 CUDA cores (1536 per GPU)
- 190 Gigaflops of double-precision peak performance (95 Gflops in each GPU)
- 4577 Gigaflops of single-precision peak performance (2288 Gigaflops in each GPU)
- GDDR5 memory optimizes performance and reduces data transfers by keeping large data sets in 8 GB of local memory, 4 GB attached directly to each GPU.
- 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.
- The high speed PCIe Gen 3.0 data transfer maximizes bandwidth between the HP ProLiant server and the Tesla processors.

Performance of the K20 Module

- 2496 CUDA cores
- 1.17 Tflops of double-precision peak performance
- 3.52 Tflops of single-precision peak performance
- GDDR5 memory optimizes performance and reduces data transfers by keeping large data sets in 5 GB of local memory that is attached to the GPU
- 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 2.0 data transfer maximizes bandwidth between the HP ProLiant server and the Tesla processors.

Performance of the K20X Module

- 1..32 Tflops of double-precision peak performance
- 3.95 Tflops of single-precision peak performance
- GDDR5 memory optimizes performance and reduces data transfers by keeping large data sets in 6 GB of local memory that is

attached to the GPU

- 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 2.0 data transfer maximizes bandwidth between the HP ProLiant server and the Tesla processors.

Performance of the K40 Module

- 2880 CUDA cores
- 1.43 Tflops of double-precision peak performance
- 4.29 Tflops of single-precision peak performance
- GDDR5 memory optimizes performance and reduces data transfers by keeping large data sets in 12 GB of local memory that is attached to the GPU
- 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 HP ProLiant server and the Tesla processors.

Reliability

- ECC Memory meets a critical requirement for computing accuracy and reliability for datacenters and supercomputing centers. It offers protection of data in memory to enhance data integrity and reliability for applications. For M2075, K20 and K20X register files, L1/L2 caches, shared memory, and DRAM all are ECC protected. For K2 and K10, only external DRAM is ECC protected. Double-bit errors are detected and can trigger alerts with the HP Cluster Management Utility.
- Passive heatsink design eliminates moving parts and cables reduces mean time between failures.

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. The CUDA software, as well as the GPU drivers, can be automatically installed on HP ProLiant servers, by HP 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 HP 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 HP customers at: www.hp.com/go/hpc. A heuristic version of this affinity-mapping has also been implemented by HP as an option to the mpirun command as used for example with HP-MPI, available as part of HP 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. HP 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.

Supported Operating Systems

NOTE: The NVIDIA Tesla and GRID modules are supported only on 64-bit versions of Linux and Windows operating systems. The supported operating systems are those below.

RHEL 5
RHEL 6
SLES 11
Windows Server 2008

Supported Servers and Workstation Blades

HP ProLiant SL250s (K2, K10, K20, K20X, K40)

NOTE: The ambient temperature for SL250s systems with between one and three NVIDIA GPUs, must be 30 degrees Celsius or less.

NOTE: Consult an HP Solution Architect for precise configuration rules.

HP ProLiant SL270s (K10, K20, K20X, K40)

NOTE: The ambient temperature for SL270s systems with between five and eight NVIDIA GPUs, must be 30 degrees Celsius or less. All other SL270s systems may be operated with ambient temperatures up to 35 degrees Celsius

NOTE: Consult an HP Solution Architect for precise configuration rules.

HP ProLiant WS460c Generation8 (K2 only)

HP Warranty

The NVIDIA Tesla or GRID GPU Module has a one year parts exchange warranty. For details on HP Qualified Options Limited Warranty visit:
http://h18004.www1.hp.com/products/servers/platforms/warranty/index.html

Optional Features

HP High Performance Clusters

HP Cluster Platforms

HP Cluster Platforms are specifically engineered, factory-integrated large-scale ProLiant clusters optimized for High Performance Computing, with a choice of servers, networks and software. Operating system options include specially priced offerings for Red Hat Enterprise Linux and SUSE Linux Enterprise Server, as well as Microsoft Windows HPC Server. A Cluster Platform Configurator simplifies ordering. http://www.hp.com/go/clusters

HP HPC Interconnects

High Performance Computing (HPC) interconnect technologies are available for this server as part of the HP Cluster Platform portfolio. These high-speed InfiniBand and Gigabit interconnects are fully supported by HP when integrated within an HP cluster. Flexible, validated solutions can be defined with the help of configuration tools.
http://www.hp.com/techservers/clusters/ucp/index.html

HP Insight Cluster Management Utility

HP Insight Cluster Management Utility (CMU) is an HP-licensed and HP-supported suite of tools that are used for lifecycle management of hyperscale clusters of Linux ProLiant systems. CMU includes software for the centralized provisioning, management and monitoring of nodes. CMU makes the administration of clusters user friendly, efficient, and effective.
http://www.hp.com/go/cmu

Third Party GPU Cluster and Development Software

More software for applications and development tools for general purpose GPU enabled systems are available every week. Examples of software available for various vendors are listed below.

PGI Accelerator: Fortran and C Compilers (directive-based generation of CUDA code, and additionally a CUDA Fortran compiler)

CAPS HMPP C and Fortran to CUDA C Compiler (directive-based generation of CUDA code)

TotalView Dynamic Source Code and Memory Debugging for C, C++ and FORTRAN HPC Applications

Allinea DDT Distributed Debugging Tool

Wolfram Mathematica mathematical analysis software

Altair PBS Professional workload scheduler

Platform LSF workload scheduler

Adaptive Computing Moab scheduler

Microsoft Windows HPC Server 2008

Optional Features

Service and Support

HP Technology Services for ProLiant Servers
Capitalizing on HP ProLiant server capabilities requires a service partner who understands your increasingly complex business technology environment. That's why it makes sense to team up with the people who know HP infrastructure hardware and software best - the experienced professionals at HP Services.

Protect your business beyond warranty with HP Care Pack Services
When you buy HP Options, it's also a good time to think about what level of service you may need. HP Care Pack services provide total care and support expertise with committed response choices designed to meet your IT and business need.

HP Foundation Care services offer scalable reactive support-packages for HP industry-standard servers and software. You can choose the type and level of service that is most suitable for your IT and business needs. HP Proactive Care delivers high levels of system availability through proactive service management and advanced technical response.

Recommended HP Care Pack Services for your HP product

Optimized Care

3-Year HP 6 hour Call to Repair Response, Proactive Care
Combined reactive and proactive support for hardware and software helping optimize your systems and delivering high levels of availability through proactive service management and advanced technical response. Hardware problem resolution to return the hardware in operating condition within 6 hours of the initial service request. A Technical Account Manager, as your single point of contact, will own your call or issue end to end until resolved.
http://h20195.www2.hp.com/v2/GetPDF.aspx/4AA3-8855EEE.pdf

HP Installation of ProLiant Add On Options Service
This easy-to-buy, easy-to-use HP Care Pack service helps ensure that your new HP hardware or software is installed smoothly, efficiently, and with minimal disruption of your IT and business

Standard Care

3-Year HP 24x7 4 hour response, Proactive Care Service
This service gives you combined reactive and proactive support including rapid access to our Advanced Solution Center to manage and prevent problems and a Technical Support Specialist with a broad level of technical knowledge that will engage with additional technical expertise as needed from HP's vast global resources.
http://h20195.www2.hp.com/v2/GetPDF.aspx/4AA3-8855EEE.pdf

HP Installation of ProLiant Add On Options Service
This easy-to-buy, easy-to-use HP Care Pack service helps ensure that your new HP hardware or software is installed smoothly, efficiently, and with minimal disruption of your IT and business

Optional Features

Related Services

HP Proactive Care Personalized Support - Environmental Option
The Personalized Support option provides an assigned Account Support Manager who can bring best practices from across the industry plus extra technical skills to your IT team. This option is only available as an add-on to HP Proactive Care Support.

HP Proactive Select Service
Provides a flexible way to purchase HP best-in-class consultancy and technical services. You can buy Proactive Select Service Credits when you purchase your hardware and then use the credits over the next 12 months. http://h20195.www2.hp.com/V2/GetPDF.aspx/4AA2-3842ENN.pdf

NOTE: Additional HP Care Pack services can be found at: http://hp.com/go/cpc

Insight Remote Support

HP Insight Remote Support provides 24 X 7 remote monitoring, proactive notifications, and problem resolution. This comes at no additional cost with your HP solution. Learn more about Insight Remote Support http://www.hp.com/go/insightremotesupport and Insight Online http://h18013.www1.hp.com/products/servers/management/insight-online/index.html

NOTE: Insight Remote Support is a prerequisite for HP Proactive Care.

HP Support Center

Personalized online support portal with access to information, tools and experts to support HP business products. Submit support cases online, chat with HP experts, access support resources or collaborate with peers. Learn more http://www.hp.com/go/hpsc

HP's Support Center Mobile App allows you to resolve issues yourself or quickly connect to an agent for live support. Now, you can get access to personalized IT support anywhere, anytime.

HP Insight Remote Support and HP Support Center are available at no additional cost with a HP warranty, HP Care Pack or HP contractual support agreement.

NOTE: HP Support Center Mobile App above is subject to local availability.

Parts and materials

HP will provide HP-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.

For more information

To learn more about HP Care Pack Services, please contact your HP sales representative or HP Authorized ServiceOne Channel Partner. Or visit: http://www.hp.com/services/proliant or www.hp.com/services/bladesystem

Related Options

HP High Performance Cluster Models	HP Insight Cluster Management Utility 1yr 24x7 Flexible License **NOTE:** This part number can be used to purchase one certificate for multiple licenses with a single activation key. Each license is for one node (server). Customer will receive a printed end user license agreement and license entitlement certificate via physical shipment. The license entitlement certificate must be redeemed online in order to obtain a license key. **NOTE:** For additional license kits please see the QuickSpecs at: http://h18004.www1.hp.com/products/quickspecs/12612_div/12612_div.html	QL803B
	HP Insight Cluster Management Utility 3yr 24x7 Flexible License **NOTE:** These part numbers can be used to purchase one certificate for multiple licenses and support with a single activation key. Each license is for one node (server). Customer will receive a printed end user license agreement and license entitlement certificate via physical shipment. The license entitlement certificate must be redeemed online in order to obtain a license key. Customer also will receive a support agreement.	BD476A
	HP Insight Cluster Management Utility Media **NOTE:** Order a minimum of one license per cluster to purchase media including software and documentation, which will be delivered to the customer, and also licenses CMU management. No license key is delivered or required NOTE: For additional license kits please see the QuickSpecs at: http://h18004.www1.hp.com/products/quickspecs/12612_div/12612_div.html	BD477A

Technical Specifications

Form Factor	10.7 in (27.2 cm) PCIe x16 form factor	
Number of Tesla GPUs	**Tesla K20, K20X, K40**	1 GPU
	Tesla K10, GRID K2	2 GPUs
Double Precision floating point performance (peak)	**Tesla K10**	190 Gflops (95 Gflops per GPU)
	Tesla K20	1.17 Tflops
	Tesla K20X	1.32 Tflops
	Tesla K40	1.43 Tflops
Single Precision floating point performance (peak)	**Tesla K10**	4.577 Tflops (2.288 Tflops per GPU)
	Tesla K20	3.52 Tflops
	Tesla K20X	3.95 Tflops
	Tesla K40	4.29 Tflops
Total Dedicated Memory	**Tesla K20X**	6 GB GDDR5
	Tesla K40	12 GB GDDR4
	Tesla K10, GRID K2	8GB GDDR5 (4 GB per GPU)
	Tesla K20	5GB GDDR5
Memory Bandwidth	**Tesla K10**	320 GB/sec (160 GB per GPU)
	Tesla K40	288 GB/sec
	Tesla K20	200 GB/sec
	Tesla K20X	250 GB/sec
Power Consumption	**Tesla K20, K20X**	225W TDP
	Tesla K20X, K40	235W TDP
	GRID K2	235W TDP
	Tesla K10	235W TDP
System Interface	**Tesla K20, K20X**	PCIe x16 Gen2
	Tesla K10, K40, GRID K2	PCIe x16 Gen3
Thermal Solution	Passive heatsink cooled by host system airflow	

Technical Specifications

Environment-friendly Products and Approach	**End-of-life Management and Recycling**	Hewlett-Packard offers end-of-life HP product return, trade-in, and recycling programs in many geographic areas. For trade-in information, please go to: http://www.hp.com/go/green. To recycle your product, please go to: http://www.hp.com/go/green or contact your nearest HP sales office. Products returned to HP will be recycled, recovered or disposed of in a responsible manner. The EU WEEE directive (2002/95/EC) requires manufacturers to provide treatment information for each product type for use by treatment facilities. This information (product disassembly instructions) is posted on the Hewlett Packard web site at: http://www.hp.com/go/green. These instructions may be used by recyclers and other WEEE treatment facilities as well as HP OEM customers who integrate and re-sell HP equipment.

The information contained herein is subject to change without notice.

The only warranties for HP products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. HP shall not be liable for technical or editorial errors or omissions contained herein.