Overview

NVIDIA Accelerators for HP ProLiant Servers

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

The following NVIDIA accelerators are available from HP, for use in certain HP ProLiant DL-series, ML-series and SL-series servers.

- 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
- HP NVIDIA Tesla K80 Dual GPU Module
- NVIDIA Tesla K40C 12 GB Module
- NVIDIA GRID K1 PCIe GPU FIO Adapter
- NVIDIA GRID K2 PCIe GPU Kit
- NVIDIA GRID K2 RAF PCIe GPU Kit
- NVIDIA Quadro K2000 PCIe Graphics Adptr
- NVIDIA Quadro K4000 PCIe Graphics Adapter
- NVIDIA Quadro K5000 PCIe Graphics Adapter
- NVIDIA Quadro K6000 PCIe Graphics Adapter
- HP NVIDIA Quadro K2200 GPU Module
- HP NVIDIA Quadro K4200 GPU Module
- HP NVIDIA Quadro K5200 GPU Module
- HP NVIDIA GRID K1 Quad GPU Module

For the set of accelerators supported in a specific HP ProLiant server, see the QuickSpecs for that server. Some of these accelerators can also be used in HP ProLiant WS460c workstation blades (see QuickSpecs at http://h18000.www1.hp.com/products/quickspecs/14409_na/14409_na.pdf)

Based on NVIDIA's CUDA™ architecture, the NVIDIA accelerators enable seamless integration of GPU computing with HP 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 HP 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 seismic processing, biochemistry simulations, weather and climate modeling, image, video and signal processing, computational finance, computational physics, CAE, CFD, and data analytics. The NVIDIA Tesla K10 modules are optimized for single-precision algorithms such as those used in certain key seismic applications. The NVIDIA Tesla K20, K20X, K40(C), K80 modules are all general-purpose, optimized for both double-precision algorithms, with 5 GB, 6 GB, 12 GB and 24 GB respectively of onboard memory.

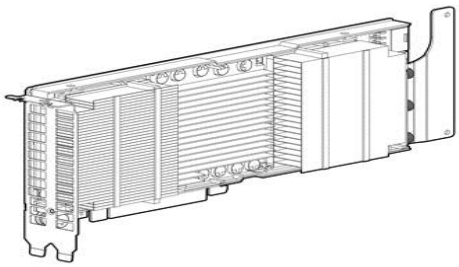
The NVIDIA Quadro GPUs offer outstanding graphics performance on a range of professional applications. The Quadro K2000, K4000, K2200 and K4200 adapters have 2 GB, 3 GB, 4 GB and 4 GB respectively of onboard memory and excel at remote visualization with multi-monitor capability. The K5000, K5200 and K6000 adapters, with 4 GB, 8 GB and 12 GB respectively of onboard memory, are the adapters of choice for large-scale and high-resolution 3D remote visualization.

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. The Grid K2 (RAF) adaptor has 2 GPUs which enable the NVIDIA Quadro® professional-class visualization features of the high-end Quadro cards and also virtual desktop applications all in the same datacenter.

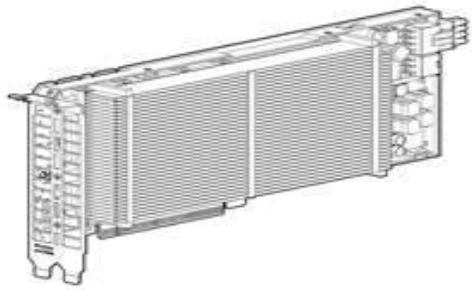
The HP GPU Ecosystem includes HP Cluster Platform specification and qualification, HP-supported GPU-aware cluster software, and also third-party GPU-aware cluster software for NVIDIA Tesla, Quadro and GRID Modules on HP ProLiant Servers. In particular,

Overview

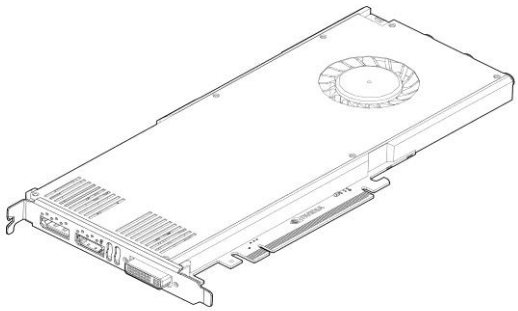
the HP Insight Cluster Management Utility (CMU) will monitor and display GPU health sensors such as temperature. Insight CMU will also install and provision the GPU drivers and the CUDA software. Insight CMU is integrated with popular schedulers such as Adaptive Moab, Altair PBS Professional, and IBM Platform LSF - all of which have the capability of scheduling jobs based on GPU requirements.



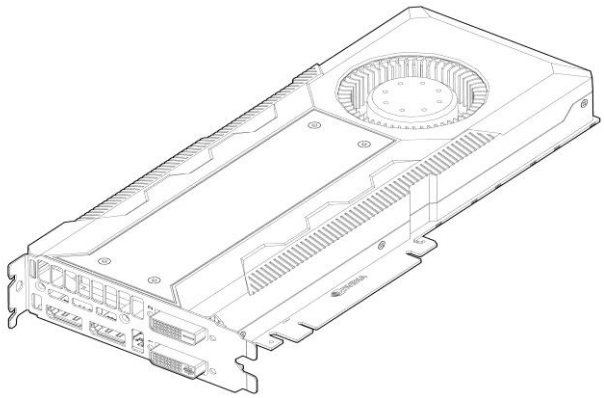
NVIDIA K10 (RAF), K2 (RAF)



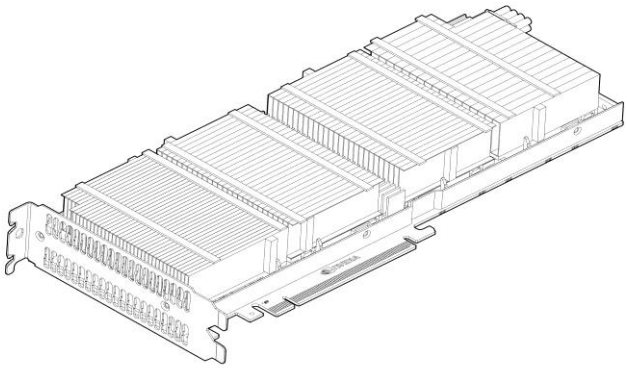
NVIDIA K20, K20X, K40



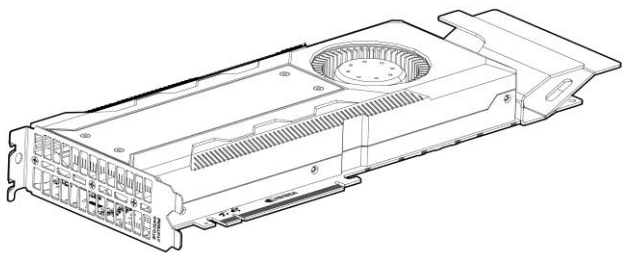
NVIDIA Quadro K4000, K4200



NVIDIA Quadro K5000, K5200, K6000

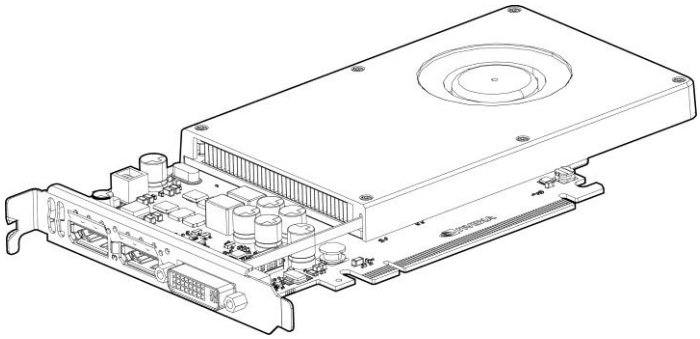


NVIDIA GRID K1

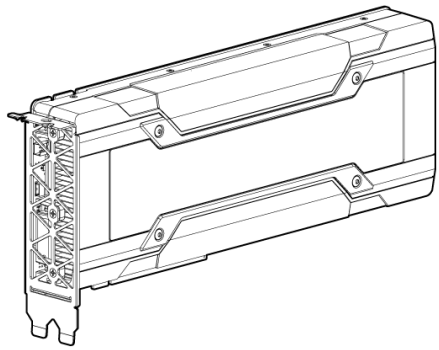


NVIDIA Tesla K40C

Overview



NVIDIA Quadro K2000, K2200



HP NVIDIA Tesla K80 Dual GPU Module

What's New

- Support for new servers and support for the NVIDIA Tesla K80 Dual GPU Module

Models

NVIDIA Accelerators	NVIDIA GRID K1 PCIe GPU FIO Adapter	730876-B21
	HP NVIDIA GRID K1 Quad GPU Module	J0G94A
	NVIDIA GRID K2 Dual GPU PCIe Graphics Accelerator	729851-B21
	NVIDIA GRID K2 Reverse Air Flow Dual GPU PCIe Graphics Accelerator	753958-B21
	NVIDIA Tesla K10 Rev B Dual GPU PCIe Computational Accelerator	E5V47A
	NVIDIA Tesla K20 5 GB Computational Accelerator	C7S14A
	NVIDIA Tesla K20X 6 GB Computational Accelerator	C7S15A
	NVIDIA Tesla K40 12 GB Computational Accelerator	F1R08A
	NVIDIA Tesla K40C 12 GB Computational Accelerator	753960-B21
	HP NVIDIA Tesla K80 Dual GPU Module	J0G95A
	NVIDIA Quadro K2000 PCIe Graphics Adapter	753959-B21
	HP NVIDIA Quadro K2200 GPU Module	J0G89A
	NVIDIA Quadro K4000 PCI-E Graphics Adapter	730870-B21
	HP NVIDIA Quadro K4200 GPU Module	J0G90A
	NVIDIA Quadro K5000 PCI-E Graphics Adapter	730872-B21
	HP NVIDIA Quadro K5200 GPU Module	J0G91A
	NVIDIA Quadro K6000 PCI-E Graphics Adapter	730874-B21

NOTE: Please see the HP ProLiant SL250s, SL270s, SL2500, DL360e, DL380e, DL380p, DL580, ML310e, ML350e or ML350p Gen8, DL80, DL120, DL180, DL360, DL380, ML150 or ML350 Gen9 server QuickSpecs or HP ProLiant WS460c Generation 8 Workstation Blade QuickSpecs for which accelerators are supported and for configuration rules including requirements, if any, for enablement kits.

NOTE: The Tesla K40 PCIe speed depends on configuration. When used as an option in the ProLiant SL250c server, the Tesla K40 operates at PCIe Gen 2. When used as an option in the ProLiant SL270c server, the Tesla K40 can be configured at delivery to operate at PCIe Gen 3 (see your HP Sales Representative) but ships by default at PCIe Gen 2.

NOTE: The Tesla K40C, Quadro K6000 PCIe and K2 speeds by default are PCIe Gen 3. However, on ProLiant DL580 servers, those cards run at PCIe Gen 2.

NOTE: The Tesla K80 speed by default is PCIe Gen 3. However, on ProLiant DL380 Gen9 servers, those cards run at PCIe Gen 2.

NOTE: The GRID K1 and K2 speed by default are PCIe Gen 3. However, on ProLiant DL380 Gen9 and DL380p Gen8 servers, those cards run at PCIe Gen 2.

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://h18000.www1.hp.com/products/quickspecs/14656_div/14656_div.html
http://h18000.www1.hp.com/products/quickspecs/14327_div/14327_div.html
http://h18004.www1.hp.com/products/quickspecs/14328_div/14328_div.html
http://h18000.www1.hp.com/products/quickspecs/14212_div/14212_div.html
http://h18000.www1.hp.com/products/quickspecs/14743_div/14743_div.html
http://h18000.www1.hp.com/products/quickspecs/14339_div/14339_div.html
http://h18000.www1.hp.com/products/quickspecs/14226_div/14226_div.html
http://h18004.www1.hp.com/products/quickspecs/14409_div/14409_div.html
http://h18000.www1.hp.com/products/quickspecs/14340_div/14340_div.html
http://h20195.www2.hp.com/v2/gethtml.aspx?docname=c04346247
http://h20195.www2.hp.com/v2/gethtml.aspx?docname=c04447843
http://h20195.www2.hp.com/v2/gethtml.aspx?docname=c04447832
http://h20195.www2.hp.com/v2/gethtml.aspx?docname=c04447806
http://h20195.www2.hp.com/v2/gethtml.aspx?docname=c04346227

Standard Features

NVIDIA Accelerators

Performance of the GRID K1 and K2 (RAF) Adapters

- K1 has 768 CUDA cores (192 per GPU), K2 has 3072 CUDA cores (1536 per GPU)
- GDDR5 memory optimizes performance and reduces data transfers by keeping large data sets in 4 GB of local memory attached to each GPU (16 GB total for K1, 8 GB total for K2).
- 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 Tesla 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 Tesla 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 Tesla K20X Module

- 1.32 Tflops of double-precision peak performance
- 3.95 Tflops of single-precision peak performance

Standard Features

- 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 Tesla K40 and K40c Modules

- 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.

Performance of the Tesla K80 Module

- 4992 CUDA cores (2496 per GPU)
- GPU Boost enables opportunistic clock frequency bursts provided no thermal or power limits are hit
- 1.87 Tflops (Base) / 2.7 Tflops (Boost) of double-precision peak performance (aggregate on 2 GPUs)
- 5.6 Tflops (Base) / 8.1 Tflops (Boost) of single-precision peak performance (aggregate on 2 GPUs)
- Total 24 GB of GDDR5 memory optimizes performance and reduces data transfers by keeping large data sets in 12 GB of local memory that is attached 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.
- 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.

Performance of the Quadro K2000, K2200, K4000, K4200, K5000, K5200 and K6000 Adapters

- K2000 has 384 CUDA cores, K2200 has 640 CUDA cores, K4000 has 768 CUDA cores, K4200 has 1344 CUDA cores, K5000

Standard Features

has 1436 CUDA cores, K5200 has 2304 CUDA cores and K6000 has 2880 CUDA cores.

- K2000 has 2 GB GDDR5 memory, K2200 has 4 GB GDDR5 memory, K4000 has 3 GB GDDR5 memory, K4200 has 4 GB GDDR5 memory, K5000 has 4 GB GDDR5 memory, K5200 has 8 GB GDDR5 memory and K6000 has 12 GB GDDR5 memory.
- Support OpenGL 4.3, Shader Model 5.0, DirectX 11 (K2200, K4200 and K5200 support OpenGL4.4)
- 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 K5200 and K6000, 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 K20, K20X and K40(C) 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, GRID and Quadro modules are supported only on 64-bit versions of Linux and Windows operating systems as well as on Virtual Machine client operating systems. The supported bare metal operating systems are those below. See server QuickSpecs for more details. RHEL 5 (not supported on K2200, K4200 and K5200) RHEL 6 SLES 11 Windows Server 2012 R2
HP Warranty	The NVIDIA Tesla, GRID or Quadro GPU Modules have 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. 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 Adaptive Computing Moab scheduler	
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.

Optional Features

	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
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.

Optional Features

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	**Tesla K10, K20, K20X, K40, K40C, K80, GRID K1, K2 (RAF), Quadro K5000, K5200, K6000**	10.5 in x 4.4 in PCIe x16 form factor
	Quadro K2000, K2200	8.0 in x 4.4 in PCIe x 16 form factor
	Quadro K4000, K4200	9.5 in x 4.4 in PCIe x 16 form factor
Number of GPUs	**Tesla K20, K20X, K40, K40C, Quadro K2000, K4000, K5000, K6000**	1 GPU
	Tesla K10, K80, GRID K2 (RAF)	2 GPUs
	GRID K1	4 GPUs
Double Precision floating point performance (peak) [Tesla only]	**Tesla K10**	190 Gflops (95 Gflops per GPU)
	Tesla K20	1.17 Tflops
	Tesla K20X	1.32 Tflops
	Tesla K40, K40c	1.43 Tflops
	Tesla K80	1.87 Tflops (base) / 2.7 Tflops (boost) (aggregate 2 GPUs)
Single Precision floating point performance (peak) [Tesla only]	**Tesla K10**	4.577 Tflops (2.288 Tflops per GPU)
	Tesla K20	3.52 Tflops
	Tesla K20X	3.95 Tflops
	Tesla K40, K40C	4.29 Tflops
	Tesla K80	5.6 Gflops (base) / 8.1 Tflops (boost) (aggregate 2 GPUs)
Total Dedicated Memory	**Tesla K20X**	6 GB GDDR5
	Tesla K40, K40C	12 GB GDDR5
	Tesla K10, GRID K2 (RAF)	8 GB GDDR5 (4 GB/GPU)
	Tesla K20	5 GB GDDR5
	Tesla K80	24 GB GDDR5 (12 GB/GPU)
	Quadro K2000	2 GB GDDR5
	Quadro K2200	4 GB GDDR5
	Quadro K4000	3 GB GDDR5
	Quadro K4200	4 GB GDDR5
	Quadro K5000	4 GB GDDR5
	Quadro K5200	8 GB GDDR5
	Quadro K6000	12 GB GDDR5
	GRID K1	16 GB GDDR5 (4 GB per GPU)
Memory Bandwidth [Tesla and Quadro only]	**Tesla K10**	320 GB/sec (160 GB/sec per GPU)
	Tesla K40, K40C,Quadro K6000	288 GB/sec
	Tesla K20	200 GB/sec
	Tesla K20X	250 GB/sec
	Tesla K80	480 GB/sec (240 GB/sec per GPU)
	Quadro K2000	64 GB/sec
	Quadro K2200	80 GB/sec
	Quadro K4000	134 GB/s
	Quadro K4200	173 GB/s
	Quadro K5000	173 GB/s
	Quadro K5200	192 GB/s

Technical Specifications

Number of slots	**Quadro K2000,K2200 K4000, K4200**	1
	Tesla K10 K20, K20X, K40, K40C, K80, GRID K1, K2 (RAF), Quadro K5000, K5200, K6000	2
Power Consumption	**Tesla K20, K20X, Quadro K6000**	225W TDP
	Tesla K20X, K40, K40C	235W TDP
	Tesla K10, GRID K1, K2 (RAF)	235W TDP
	Tesla K80	300W TDP
	Quadro K2000	51W TDP
	Quadro K2200	68W TDP
	Quadro K4000	80W TDP
	Quadro K4200	108W TDP
	Quadro K5000	122W TDP
	Quadro K5200	150W TDP
System Interface	**Tesla K20, K20X, K40*, Quadro K2000, K2200, K4000, K4200, K5000**	PCIe x16 Gen2
	Tesla K10, K40, K40C, K80**, GRID K1***, K2*** (RAF), Quadro K5200, K6000**	PCIe x16 Gen3
Thermal Solution	**Tesla K10, K20, K20X, K40, GRID K2 (RAF)**	Passive cooling by host system airflow
	Tesla K40C, Quadro K2000, K200, K4000, K4200, K5000, K5200, K6000****	Active cooling by on-board fan

* The Tesla K40 PCIe speed depends on configuration. When used as an option in the ProLiant SL250c server, the Tesla K40 operates at PCIe Gen2. When used as an option in the ProLiant SL270c server, the Tesla K40 operates at PCIe Gen3.
** The Tesla K40C and Quadro K6000 PCIe speed by default is PCIe Gen3. However, on ProLiant DL580 servers, those cards run at PCIe Gen2.
*** The GRID K1 and K2 RAF speed by default is PCIe Gen3. However, on ProLiant DL380 Gen9 servers, those cards run at PCIe Gen2. On the ProLiant DL580 servers, the GRID K2 RAF runs at PCIe Gen2
**** The Tesla K80 speed by default is PCIe Gen3. However, on ProLiant DL380 Gen9 servers, the Tesla K80 runs at PCIe Gen2

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.

Summary of Changes

Date	Version History	Action	Description of Change:
09-Feb-2014	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 Overview - Introduction Models Standard Features Optional Features Technical Specifications