HPE Hardware Accelerated Graphics for Desktop Virtualization

Contents

Technical white paper

Purpose of this document

- Overview of concepts and hardware-accelerated virtual desktop technologies.
- Overview of features and options for the HPE ProLiant WS460c Graphics Server Blade.

Abbreviations and naming conventions

Table 1. Abbreviation and naming conventions

Convention	Definition
Bare Metal OS	Operating system installed directly on the system, not virtualized
GPU	Graphical processing unit (graphics card). It is important to note that some graphics cards have more than one GPU processor
GPU compute	Synonymous to GPGPU. See GPGPU
GPGPU	General-purpose graphics processing unit—GPU technology that performs application computation traditionally handled by the CPU
GRID vGPU	NVIDIA® technology for sharing true virtual GPU (GRID vGPU) hardware acceleration between multiple users
HDX	Citrix® set of advanced desktop remoting technologies to deliver a high definition experience
HDX 3D Pro	Feature of XenDesktop® HDX protocol for delivering high-end 3D professional graphics
Hypervisor	Virtualization host platform (VMware vSphere® Hypervisor, Microsoft® Hyper-V, Citrix XenServer)
MxGPU	AMD Multiuser GPU (MxGPU) technology
PCoIP	Teradici remote desktop protocol used in VMware View®
RDP	Microsoft Remote Desktop Protocol
RemoteFX	Microsoft's set of advanced desktop remoting technologies
RFX	Microsoft RemoteFX
RGS	HP Remote Graphics software
SR-IOV	Single Root I/O Virtualization
VDI	Virtual desktop infrastructure
vGPU	NVIDIA GRID vGPU: hardware-virtualized GPU solution used on VMware vSphere® and Citrix XenServer
vSGA	VMware®-specific terminology: software-virtualized GPU (API capture model)
vDGA	VMware-specific terminology for GPU pass-through
VM	Virtual machine
vRAM	GPU video RAM

Concepts and technology

This section contains a conceptual overview of the technologies behind Hardware Accelerated Graphics for Desktop Virtualization. This is a high-level discussion on the differences between the technologies as well as how the major desktop virtualization providers implement these technologies into their products.

Bare Metal OS

This method is the classic workstation and PC blade remoting architecture (see figure 1). The client OS is installed directly on the blade or server hardware and no virtualization is used. End users connect to the workstation using remote protocols such as HP RGS, Microsoft RDP, and Citrix HDX 3D Pro from client hardware. This method is still used today for users that demand the power and the performance of dedicated hardware.

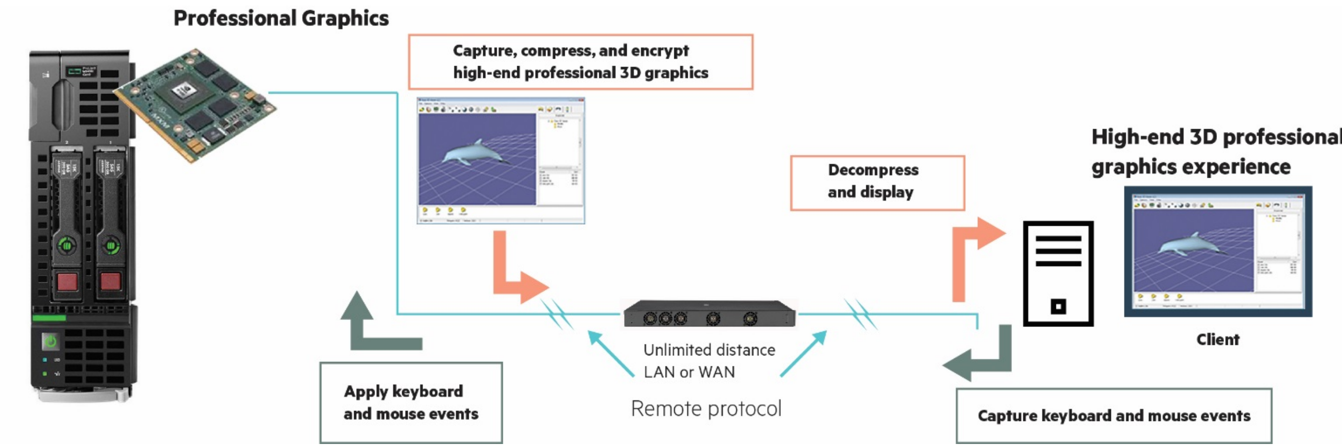


Figure 1. Bare Metal GPU model (example)

GPU pass-through

It is also referred generically as "direct-attached GPU," or vendor-specific "vDGA" (VMware) and "GPU pass-through" (Citrix). This method allows PCI GPU devices to be directly mapped to a virtual machine for dedicated one-to-one use by the VM (see figure 2). The virtual machine has full and direct access to GPU, including the native graphics driver, allowing for full workstation-class graphics and GPU computation functionality in a virtual machine. Typically intended for high-end 3D and GPU computational users, the GPU device is directly owned and managed by the VM operating systems just as in a desktop workstation. The GPU driver is loaded within the virtual machine.

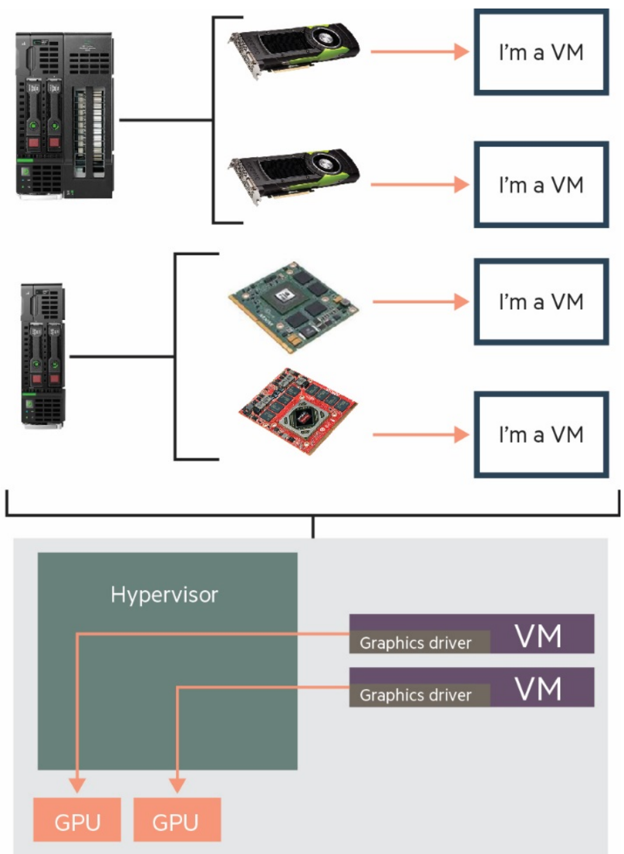


Figure 2. Pass-through GPU model

Enterprise hypervisors using this technology include the following:

- Microsoft Remote Desktop Session Host
- Citrix XenServer 6.0 and newer
- VMware vSphere 5.5 and newer

Advantages

- Up to six workstation-class VMs per host using the HPE WS460c Gen9 Graphics Server Blade with HPE Multi-GPU Carrier Card
- Support for all 3D technologies including DirectX, OpenGL, OpenCL, and NVIDIA CUDA via the native NVIDIA driver in the VM
- Best performing solution as the graphics driver resides in the VM. Virtual machines have full and direct access to a dedicated GPU (not shared)
- Can mix accelerated and non-accelerated VMs on the same host to maximize resource utilization

Disadvantages

- Higher cost of ownership per connection as it has a dedicated GPU per virtual machine
- Lower VM density per host when compared to other virtualized GPU solutions
- Live migration of VM with pass-through devices is not supported

Software-virtualized GPU

It is also referred to generically as “shared GPU,” “API intercept model,” or vendor-specific of “vSGA” (VMware), and “RemoteFX vGPU” with Microsoft Hyper-V RemoteFX. This method uses an API intercept model where the GPU is owned and managed by the hypervisor. All incoming graphics API requests from the VMs are intercepted via the API capture driver in the VM and redirected to and executed by the hypervisor and then sent back to VM (see figure 3). The VM does not have direct access to the GPU, and the GPU driver is loaded within the hypervisor. This solution is primarily a 3D offload solution to save CPU cycles and increase host performance, but is not a high-end 3D rendering solution.

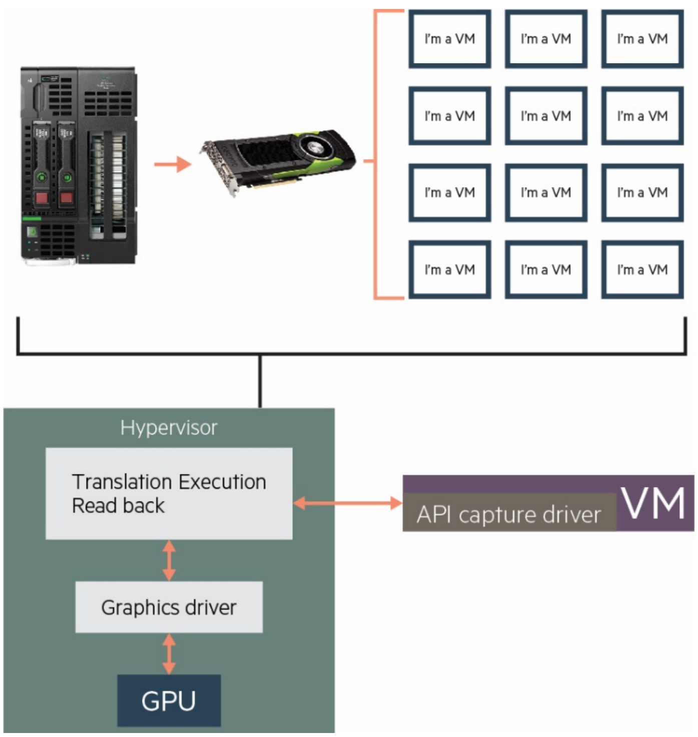


Figure 3. Software-virtualized GPU model

Enterprise hypervisors and servers using this technology include:

- Microsoft Hyper-V RemoteFX
- VMware vSGA

Advantages

- Scalability to over 50+ users per GPU depending on the workload.
- It can load balance between multiple installed cards as VMs start.
- Some solutions such as VMware vSGA can dynamically switch between GPU offload and full software rendering (CPU).
- Lower cost per user compared to other technologies, with high density of VMs per GPU.
- Allows each user to have power user performance with enhanced support for DirectX 3D and Windows® Aero.
- Live migration of VMs with vSGA supported.
- Up to 20 percent+ CPU utilization drop.

Disadvantages

- May exhibit unacceptable performance for mid- to high-end 3D knowledge or workstation user workloads.
- This solution is primarily a 3D offload solution to save CPU cycles and increase host performance, but is not a high-end 3D rendering solution.
- Potential application compatibility issues due to limitations of 3D the APIs supported:
  - Very limited OpenGL support
  - DirectX supported versions limited to DirectX 9 in some cases

Graphics-accelerated desktop sessions and application virtualization

A type of shared GPU virtualization: In this model, 3D applications are installed on the host system or VM and published as hosted, shared desktops or as a hosted, published application supporting large number of sessions per host. If the application or session is running on a host equipped with a supported 3D graphics card, each hosted application or session can utilize the graphics card for 3D rendering. Figure 4 shows the conceptual structure of accelerated desktop and application publishing using Citrix XenDesktop or XenApp.

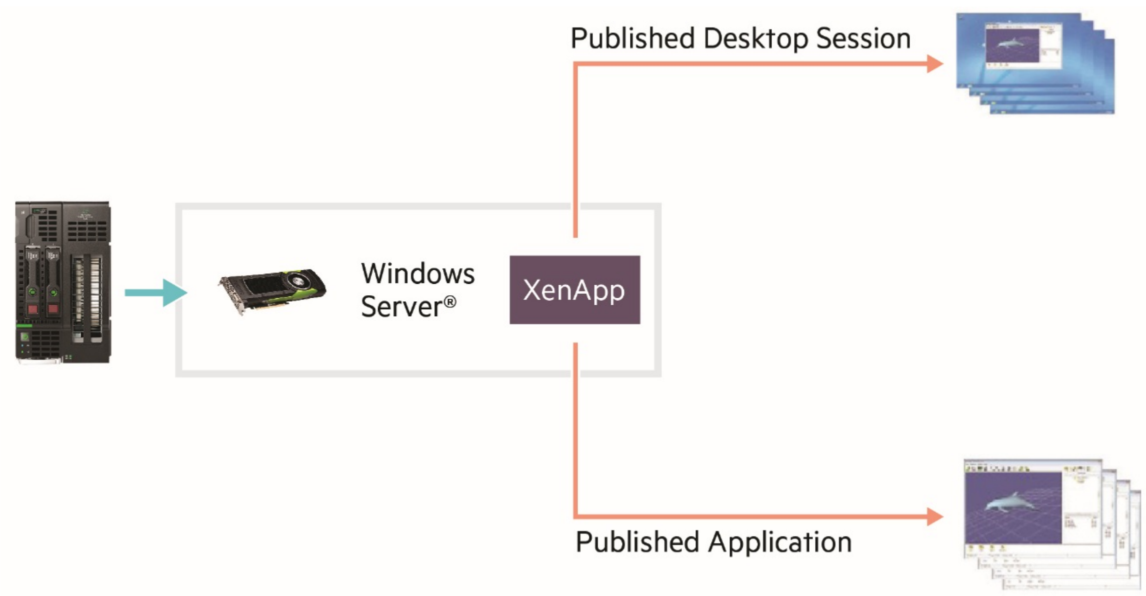


Figure 4. Graphics-accelerated session and application virtualization

Enterprise solutions that support 3D hosted applications and desktops:

- Citrix XenApp
- Citrix XenDesktop 7.x

Advantages

- GPU sharing with direct access to the video driver maximizing user density and cost.
- Ability to publish hosted applications with OpenGL and DirectX 3D support when a full desktop is not required.
- Lower cost of ownership when compared to other accelerated graphics for VDI technologies.
- Application virtualization is supported on both Bare Metal and GPU pass-through (for greater density). Both VMware vDGA and XenServer pass-through are supported.

Disadvantages

- Some 3D applications may not work or be certified as published application or multi-user published desktops.
- It may exhibit unacceptable performance for high-end 3D user.
- The GPU can become a performance bottleneck as many users draw on the resources of one card; it is possible that one user can consume all of the resources of the GPU.

Hardware-virtualized GPU

Hardware-virtualized GPU technology is conceptually a hybrid of software-virtualized GPU and pass-through GPU models. It offers the benefit of GPU sharing (similar to software-virtualized GPU model) as well as gives the performance, functionality, and features of native high-end graphics as it has direct access to GPU functionality (similar to pass-through model). Each virtual GPU has a set amount of video RAM (frame buffer) and the number of virtual GPUs per physical GPU is determined by total video memory on the physical GPU. For example, if we have a physical GPU that has 8 GB of vRAM, the vRAM can be divided into two virtual GPUs with 4 GB vRAM, or four virtual GPUs with 2 GB vRAM, etc. Depending on solution, the virtual GPUs are managed by either a software manager that is installed in the hypervisor (NVIDIA GRID vGPU) or managed by logic embedded on GPU hardware (AMD MxGPU). These solutions are discussed in more depth in next section.

Hardware virtualized GPU advantages

- Increased number of true hardware graphics-accelerated VMs per host, supporting up to 16 users per physical GPU.
- Support for all 3D technologies including DirectX 9/10/11, and OpenGL 4.4, OpenCL, CUDA via the native GPU driver in the VM.
- The virtual machine has full and direct access to the GPU, including the native graphics driver, for full workstation performance.

Hardware virtualized GPU advantages/considerations

- Hot VM migration not supported at this time (e.g., VMware vSphere® vMotion®, XenMotion).
- With NVIDIA vGPU, vRAM amount is guaranteed. The GPU resource, however, is shared among all VMs configured to use it, allowing any VM to have up to 100 percent of the GPU resource if no other VMs are using the resource, but VMs running heavy workload can take performance from other GPUs.
- With MxGPU, there is some overhead associated with MxGPU/SR-IOV and vRAM availability is slightly reduced. For example, if you create two VFs on an 8 GB physical MxGPU-enabled GPU, the VFs created will have slightly less than 4 GB of vRAM.

NVIDIA GRID vGPU

NVIDIA GRID vGPU is a software-defined solution where the physical GPU is broken into multiple virtual GPUs managed by the NVIDIA GRID vGPU Manager that resides in the hypervisor that resides in the hypervisor, see figure 5a. All vGPUs receive a dedicated amount of physical video memory and share the resource of GPU. The amount of video memory and displays supported is defined by GRID vGPU types or profiles. vGPU types are defined by: video memory size, maximum supported display heads, and maximum resolution supported. For NVIDIA GRID vGPU v3.1, the list in table 2 shows a few of the available predefined vGPU types/profiles for the NVIDIA Tesla M6. Once installed, vGPU types can easily be added by editing the VM setting and choosing the desired vGPU type.

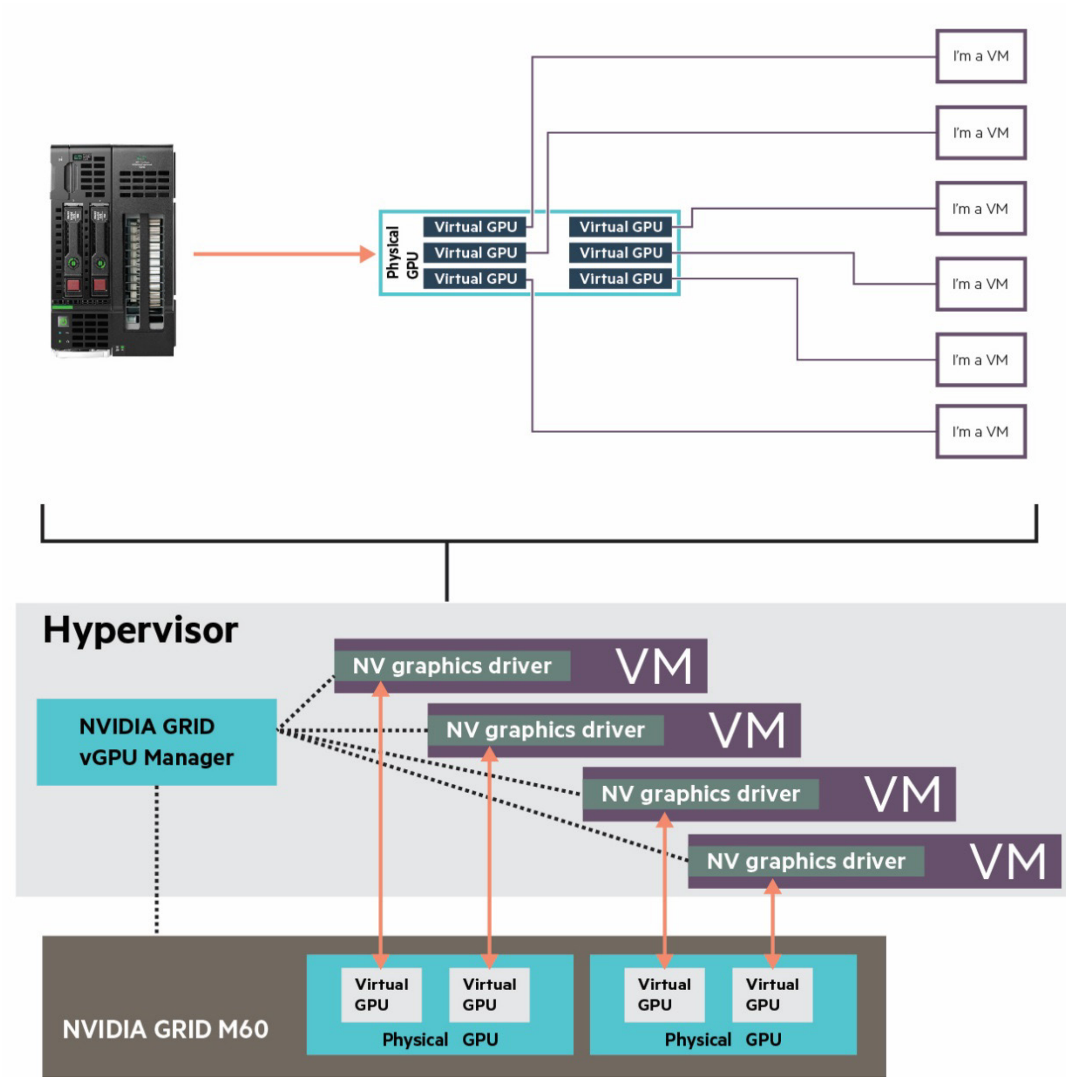


Figure 5a. NVIDIA GRID vGPU hardware-virtualized GPU model

The maximum number of vGPUs supported depends on the vGPU type selected. For example, if you have an NVIDIA Tesla M60 that has two physical GPUs on a single card with 8 GB of vRAM on each, then one GPU could have 2 to 4 GB vGPUs and the other could have 8 to 1 GB vGPUs. Only one vGPU type can run on a single physical GPU at a time. For example, if you only have a single NVIDIA Tesla M60 (one card with two physical GPUs), once you start a vGPU profile on a physical GPU (8Q for example) no other vGPU profile type can start on that physical GPU. You can however start a new vGPU profile type on another physical GPU. With NVIDIA vGPU, vRAM amount is guaranteed. The GPU resource is shared among all VMs configured to use it, allowing any VM to have up to 100 percent of the GPU resource if no other VMs are using the resource. In general, this works very well as most workloads are sporadic. However, if one or more heavy workloads are running, it may reduce the performance of other VMs running on the same GPU.

GRID vGPU key benefits

- Up to 16 vGPUs/users/VMs per GPU with direct access to high performance GPU.
- Support for provisioning technologies such as VMware-linked clone pools.
- Ease of configuration, vGPU types appear as VM configuration option.
- vGPU type changes without host system reboot.

Table 2. Example of NVIDIA GRID vGPU v3.1 M6 types

Type	Use case	Frame buffer	Display heads	Max. resolution	Use/GPU
M6-8Q	Designer	8192	4	4096x2160	1
M6-4Q	Designer	4096	4	4096x2160	2
M6-2Q	Designer	2048	4	4096x2160	4
M6-1B	Power user	1024	4	2560x1600	8
M6-8A	Virtual application	8192	1	1280x1024	1

Enterprise hypervisors using hardware-virtualized GPU technology support by HPE ProLiant WS460c include:

- NVIDIA vGPU
  - Citrix XenServer 6.2 and newer
  - Citrix XenServer 7.0 and newer
  - VMware vSphere 6.0 and newer

AMD MxGPU

AMD MxGPU hardware-virtualized GPU solution is based on the industry-standard Single Root I/O Virtualization (SR-IOV) specification. It has been used in network cards for some time. AMD MxGPU is the first truly all-hardware–virtualized GPU. The GPU can be used as a dedicated resource using GPU pass-through or it can be split into up to 16 virtual functions (VF), see figure 5a. Examples of the VF types that can be created are shown in table 3. Once the VF's are defined, the physical GPU presents itself as multiple small GPUs. For example, take a single physical MxGPU-enabled GPU and create four VFs. After the host is rebooted, the single MxGPU is presented to the host as four separate GPUs. Once configured, VF's are connected to the VMs in the same way as pass-through GPUs. The solution offers hardware partitioning of the GPU, creating virtual functions with hardware enforced memory isolation and deterministic performance. If a GPU is configured into four VFs, each VF/VM/user will get 25 percent on the GPU resource, no less, no more, even if other VMs are not using the resource. This allows for a quality of service that is both measurable and predicable. Because MxGPU technology exposes all GPU functionality to VF, there are no API limitations allowing support for DirectX, OpenGL, and OpenCL.

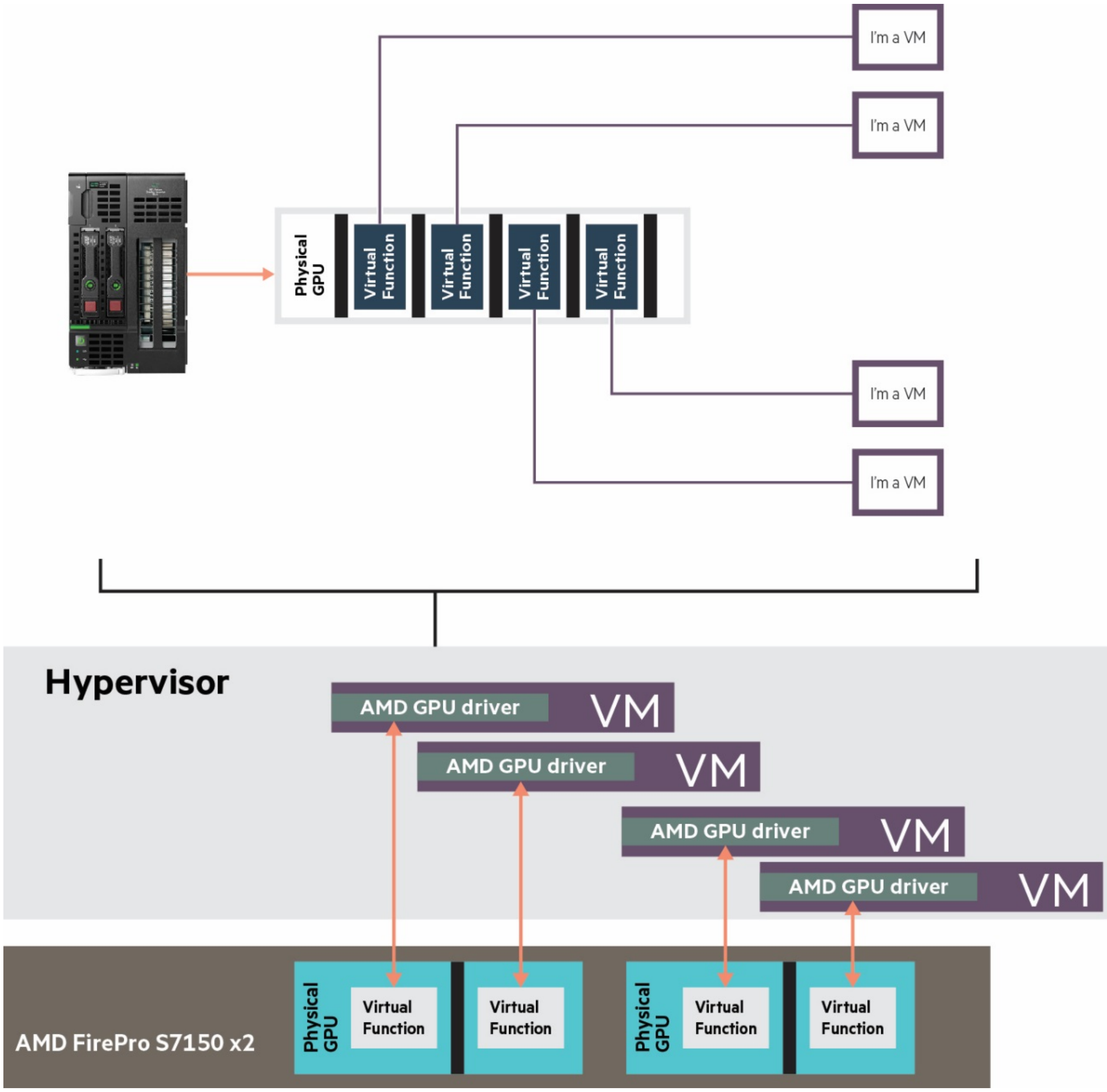


Figure 5b. AMD MxGPU hardware-virtualized GPU model

Table 3. Example of AMD FirePro S7100X MxGPU VF definitions

# of VFs per GPU	Frame buffer
2	3840
4	1920
8	960
16	480

MxGPU key benefits

- Predictable performance allows for a quality of service that is both measurable and predicable.
- True hardware-based GPU virtualization.
- Full API support including DirectX, OpenGL, and OpenCL.
- VF memory isolation provides security between VFs.

Enterprise hypervisors using hardware-virtualized GPU technology include:

- AMD MxGPU
  - VMware vSphere 6.0 and newer

Planning considerations for implementing hardware-accelerated desktop virtualization technologies

Determining the right GPU and platform for your use case

The following chart (figure 6a and 6b) shows a comparison of GPU acceleration on desktop virtualization technologies, GPU types, and what industry segment and use case they are best fitted to.

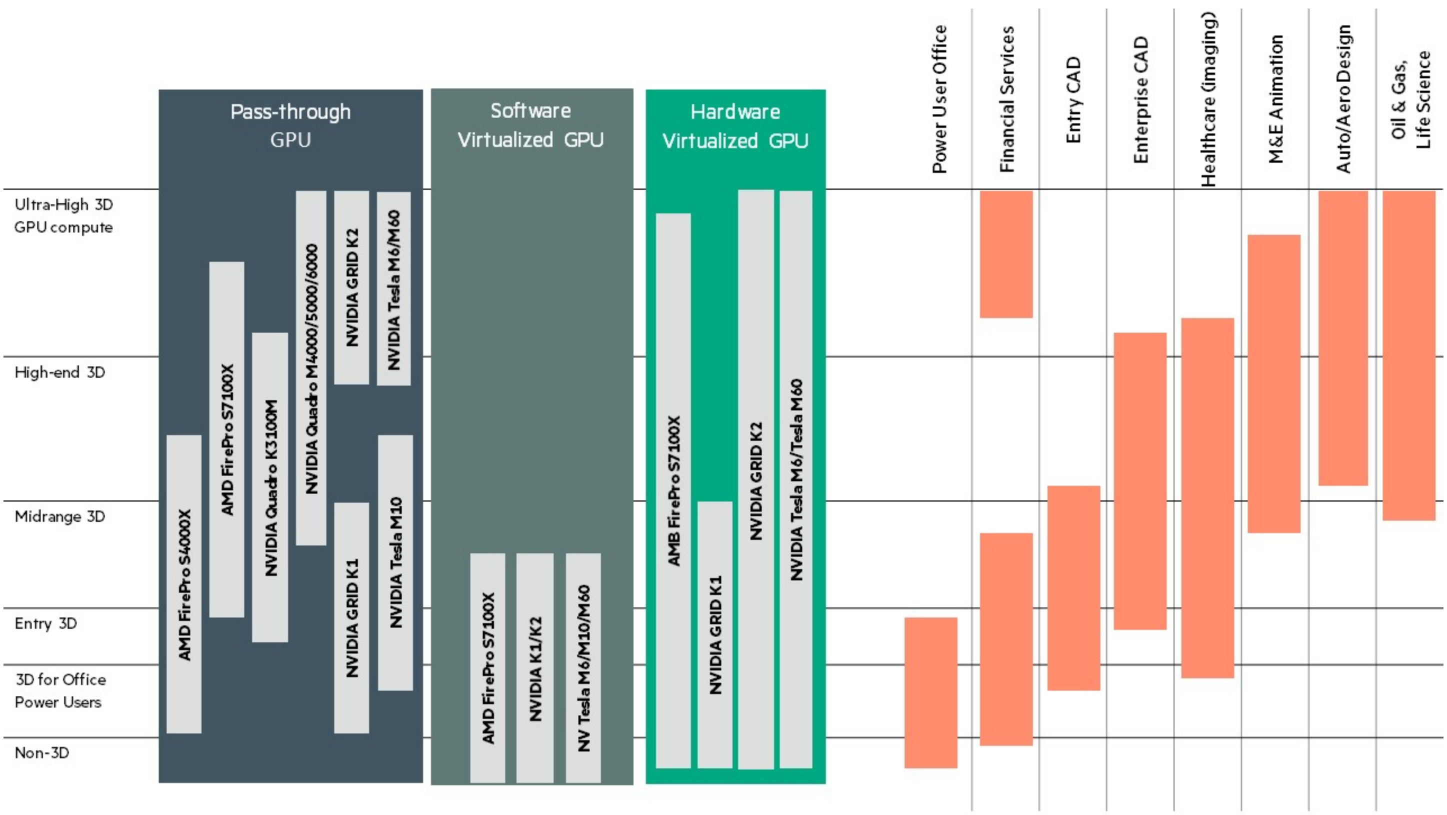


Figure 6a. Virtualized graphics technology use cases and segment positioning

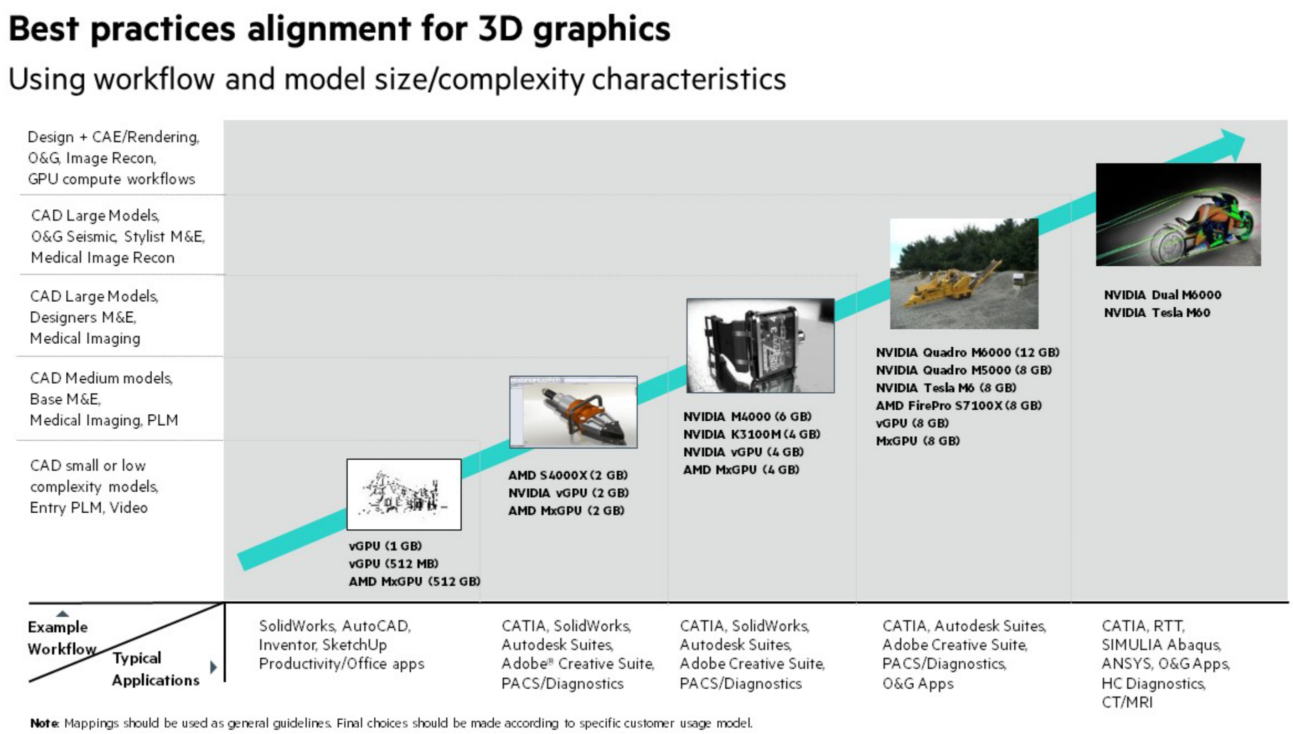


Figure 6b. 3D graphics application alignment

Virtualization solution feature comparison and considerations

Table 4. Solution feature comparison

	VMware vSGA	VMware vDGA with NVIDIA GPU	VMware vDGA with AMD GPU	Microsoft Server 2012 R2 for RemoteFX	Citrix XenServer GPU pass-through with NVIDIA	Citrix XenServer GPU pass-through with AMD	VMware or Citrix w/ NVIDIA vGPU	VMware AMD MxGPU
OpenCL	N/A	2.x[1]	2.x[1]	N/A	2.x[1]	2.x[1]	2.x[1]	2.x[1]
OpenGL	2.x	2.x, 3.x, 4.x	2.x, 3.x, 4.x	N/A	2.x, 3.x, 4.x[1]	2.x, 3.x, 4.x	2.x, 3.x, 4.x[1]	2.x, 3.x, 4.x
DirectX	9[2]	9, 10, 11, 12[2]	9, 10, 11[2]	9, 10, 11[2]	9, 10, 11, 12[2]	9, 10, 11[2]	9, 10, 11, 12[2]	9, 10, 11[2]
Displays supported	8[3]	8[3]	6[3]	8[3]	8[3]	6[3]	8[3]	6[3]

[1] Only supported using 8 GB profile (M60-8Q, M6-8Q).
[2] Contingent on actual GPU used.
[3] GPU and protocol support may vary.

XenServer considerations

- XenServer with NVIDIA vGPU
  - The NVIDIA vGPU Manager must be installed on the hypervisor to use vGPU features.
  - WS460c Gen9 Blades require XenServer 6.5 or newer.
  - XenMotion, Storage XenMotion, and VM suspend are not supported at this time.

- VMs can be migrated and started on any server with a compatible GPU configuration.
- CUDA and OpenCL are supported with NVIDIA GRID 3.1 and above.
- Support for linked clones.

- XenServer with AMD MxGPU
  - XenServer and XenDesktop do not support AMD cards at time of this writing.
- GPU pass-through
  - GPU pass-through technology is one-to-one, meaning each GPU is directly attached to a GPU and is not shared between VMs.
  - GPU pass-through technology allows for maximum guaranteed GPU performance as each VM has a dedicated GPU attached and it is not shared in any way.
  - GPU to VM density is determined by the number of GPUs. This model uses one GPU per VM, no sharing.
  - XenMotion, Storage XenMotion, or VM suspend are not supported at this time.
  - VMs can be migrated and started on any server with a compatible GPU configuration.

VMware considerations

- VMware with NVIDIA vGPU
  - The NVIDIA vGPU Manager must be installed on the hypervisor to use vGPU features.
  - VMware vSphere® ESXi™ or newer is required for latest vGPU features.
  - Live VM migration (vMotion) is not supported at this time. Standard migration is supported as long as both hosts have equivalent vGPU features configured.
  - CUDA and OpenCL are supported with NVIDIA GRID 3.1 and above.
  - Support for linked clones.
  - Variable performance, determined by number of VMs sharing the GPU and the type of workload the user are doing. VMs are sharing the resource of GPU. One VM running heavy workload may decrease performance on other VMs.
- VMware with AMD MxGPU
  - The AMD MxGPU driver must be installed on the hypervisor to use MxGPU features.
  - VMware vSphere ESXi 6.x or newer is required for vGPU features.
  - Live VM migration (vMotion) is not supported at this time. Standard migration is supported as long as both hosts have equivalent vGPU features configured.
  - CUDA and OpenCL are supported.
  - Deterministic predicable performance: The GPU is partitioned and each VF/VM/user gets a specific and consistent performance. If GPU has four virtual functions defined, each will get 25 percent of GPU resource, no more, no less. This can be good for deployment needing a guaranteed quality of service. However, one user cannot take advantage of extra GPU resource if it available.
- VMware vDGA
  - VMware vDGA pass-through GPU technology allows for maximum graphical performance (workstation grade) as each VM has a dedicated GPU attached.
  - GPU to VM density is determined by the number of GPUs; this model uses one GPU per VM, no sharing.
  - CUDA and OpenCL are supported.

- VMware vSGA
  - Live migration is supported on virtual machines that have 3D graphics enabled.
  - There are significant limitations in API support (DirectX/OpenGL/OpenCL) and overall performance may not be adequate for most workloads.

Hyper-V RemoteFX considerations

- The GPU has a dedicated amount of video RAM. Virtual machines consume a specific amount of video RAM based on the maximum number of monitors and resolution set for each virtual machine. This will dictate the maximum number of virtual machine per physical GPU.
- RemoteFX uses software-virtualized GPU API intercept technology, allowing multiple virtual machines to use the resources of that GPU. This model uses the GPU as an offload engine for the software 3D drivers in the VM. It is not directly rendering 3D content of VM, which means there is significant performance and API (DirectX, OpenGL) support limitations.
- Multiple physical graphics cards enhance performance and scalability; if multiples are installed, Hyper-V will load balance between cards as virtual machines start up.

The HPE ProLiant WS460c Gen9 Graphics Server Blade

The HPE ProLiant WS460c Graphics Server Blade portfolio (formerly WS460c "Workstation Blade") has been the cutting edge of remote workstation (i.e., centralized in a data center) computing and graphics virtualization since its inception. Rather than placing the workstation's computing power at the user's desk, the computing power (in the form of a server blade) is moved to the data center, where systems can be more easily, securely, and economically managed, shared, virtualized, and accessed from anywhere.

The WS460c Graphics Server Blade is ideal for Bare Metal or virtualized multi-tenancy high-end graphics users. Its features enable users to complete large 3D model visualizations with uncompromised workstation-class performance. New graphics virtualization technologies such as vGPU and MxGPU allow for a much broader range of workload types to take advantage of graphics virtualization and increased density in virtualized environments. The WS460c Graphics Server Blade portfolio of options allow for maximum flexibility from a single ultra-high-end 3D or GPU compute user, to a few mid- to high-end users, or many lower end users that demand less computational horse power but require full fidelity graphics-accelerated desktops.

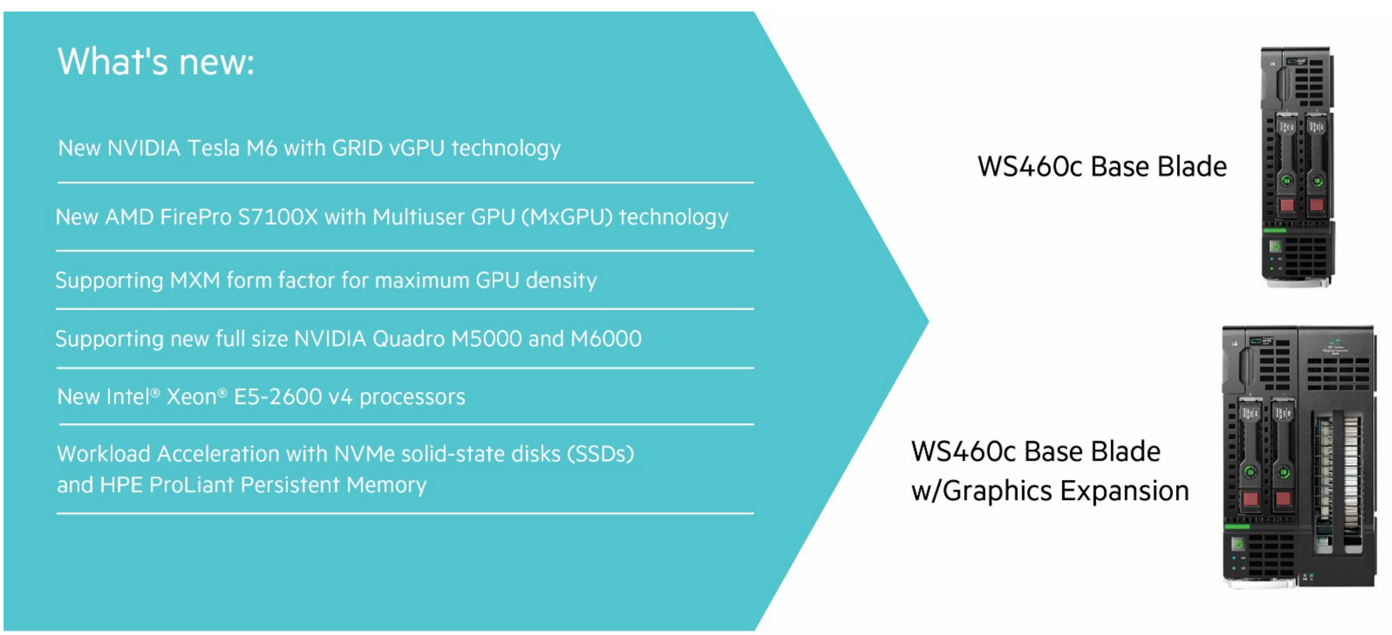


Figure 7a. HPE ProLiant WS460c Gen9 Graphics Server Blade

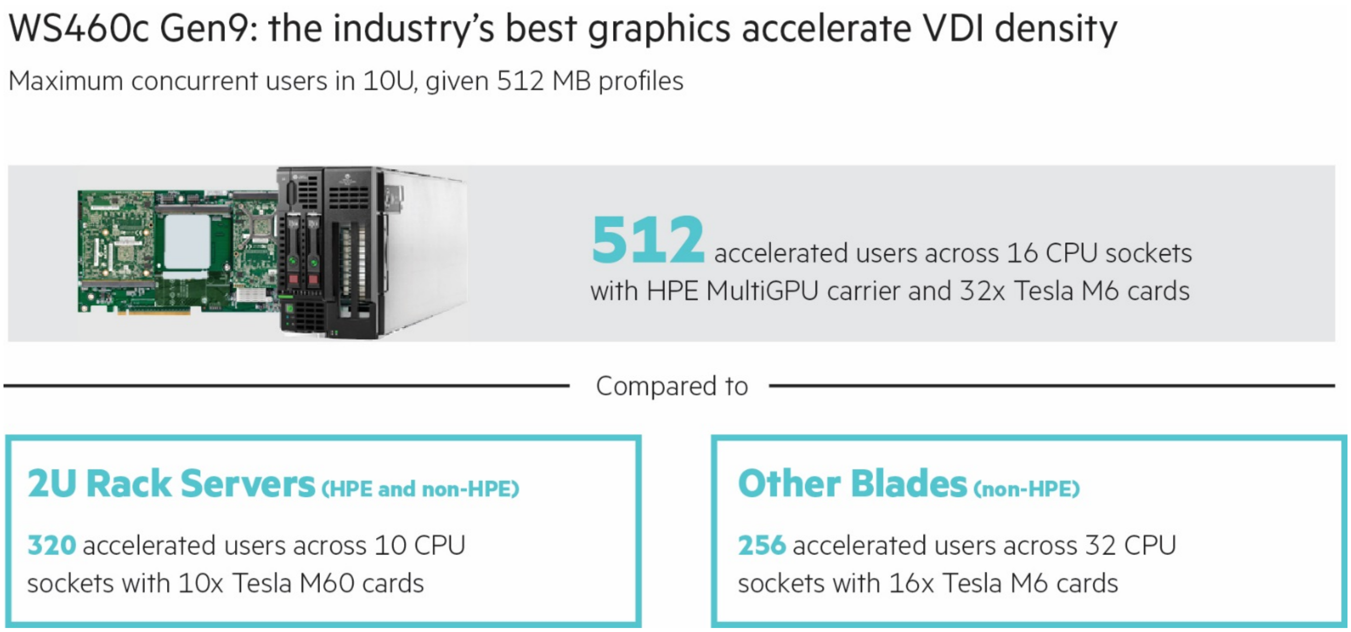


Figure 7b. HPE ProLiant WS460c Gen9 Graphics density comparison for competition and other platforms

Figure 8a. WS460c single-wide configuration options

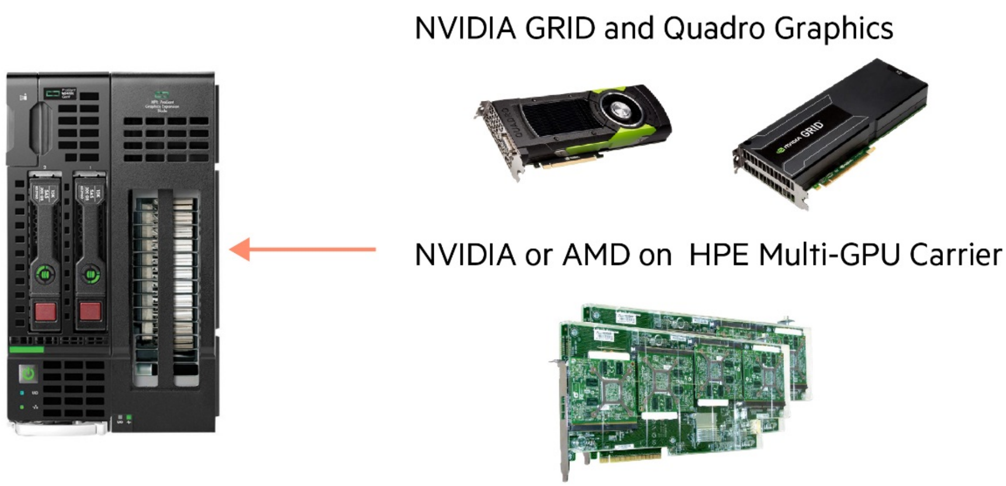


Figure 8b. WS460c Gen9 w/ expansion bay configuration options

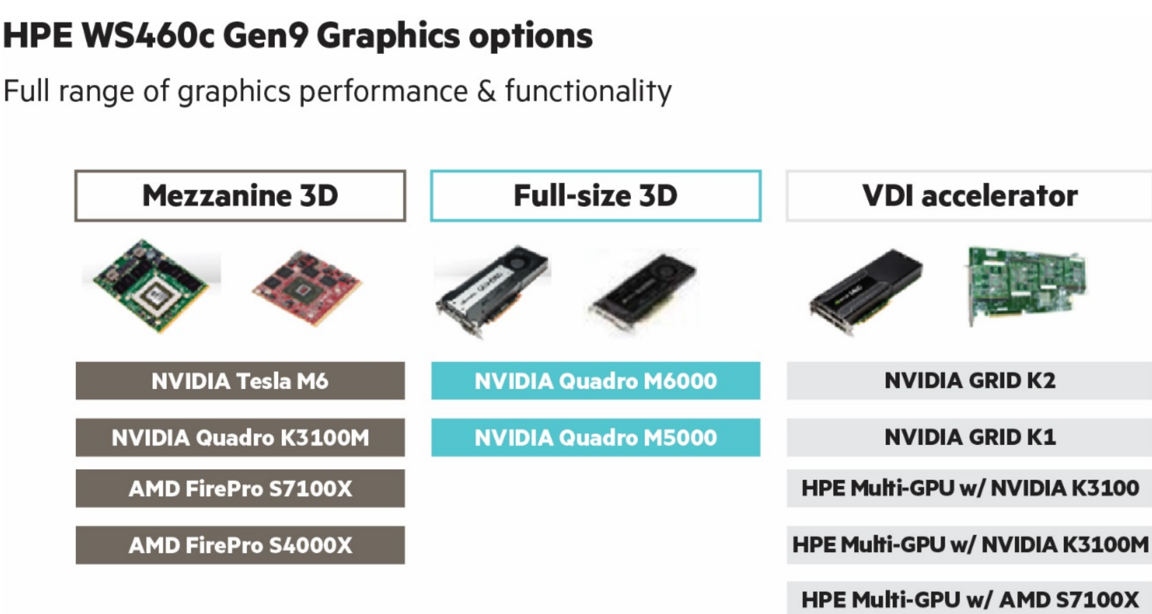


Figure 8c. WS460c Gen9 GPU support

The HPE Multi-GPU Carrier Card

The HPE Multi-GPU Carrier Card (see figure 9) for the WS460c Gen8 and Gen9 Graphics Server Blades is the industry's first MXM (small form factor GPU card) carrier card technology with four MXM slots. Providing high-density, high-end 3D graphics for GPU-accelerated desktop virtualization with support for up to six GPUs in an HPE ProLiant WS460c Gen9 Blade form factor to support all accelerated graphics for desktop virtualization technologies.

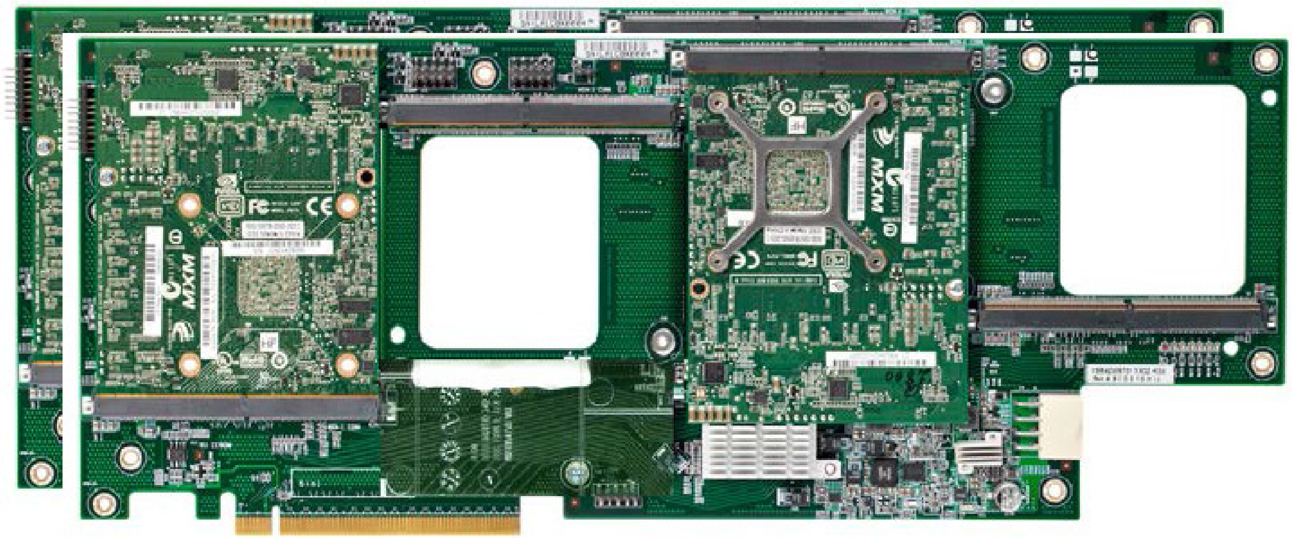


Figure 9. HPE Multi-GPU Carrier Card

HPE ProLiant WS460c Graphics Server Blade supported GPUs, specification, and configuration options

HPE Multi-GPU Carrier Card options and configurations

Table 5. HPE Multi-GPU Carrier Card options (up to two Multi-GPU cards)

	 **HPE Multi-GPU with 3 NVIDIA Quadro K3100M**	 **HPE Multi-GPU with 2 NVIDIA Tesla M6**	 **HPE Multi-GPU with 2 AMD FirePro S7100X**
Max. Multi-GPU cards supported in WS460c	2 (6 GPU)	2 (4 GPU)	2 (4 GPU)
GPU	3 NVIDIA Kepler GPU (K3100M) per Multi-GPU Card	2 NVIDIA Maxwell GPUs (K5000-class) per Multi-GPU Card	2 high-end AMD Tonga (FirePro S7100X) per Multi-GPU Card
CUDA/Stream cores per Multi-GPU Card	768 (192 per GPU)	3,072 (1536 per GPU)	4,096 (2048 per GPU)
Memory size	12 GB GDDR3 (4 GB per GPU)	16 GB GDDR5 (8 GB per GPU)	16 GB GDDR5 (8 GB per GPU)
Max. power	130 W	100 W per card	100 W per card
Form factor (Multi-GPU)	PCIe 3.0 Single Slot	PCIe 3.0 Single Slot	PCIe 3.0 Single Slot
PCIe	x16	x16	x16
PCIe generation	Gen3 (Gen2 compatible)	Gen3 (Gen2 compatible)	Gen3 (Gen2 compatible)
Cooling solution	Passive	Passive	Passive
Max. # of user	1 per GPU 3 per Multi-GPU Card	16 per GPU (vGPU) 32 per Multi-GPU Card	16 per GPU (MxGPU) 32 per Multi-GPU Card

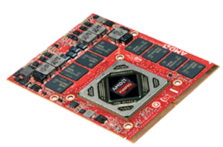
HPE ProLiant WS460c Graphics Server Blade supported NVIDIA GRID/Tesla cards

Table 6. GRID configuration options

	 **NVIDIA GRID K1**	 **NVIDIA GRID K2**	 **NVIDIA Tesla M6**
GPU	4 Kepler GPUs (K6000-class)	2 Kepler GPUs (K5000-class) (K5000-class)	Maxwell M5000-class
CUDA cores	768 (192 per GPU)	3072 (1,536 per GPU)	1536
Memory size	16 GB GDDR3 (4 GB per GPU)	8 GB GDDR5 (4 GB per GPU)	8 GB GDDR5
Max. power	130 W	225 W	100 W
Form factor	Dual slot ATX, 10.5"	Dual slot ATX, 10.5"	x16 MXM
Aux. power	6-pin connector	8-pin connector	N/A
PCIe	x16	x16	x16
PCIe generation	Gen3 (Gen2 compatible)	Gen3 (Gen2 compatible)	Gen3 (Gen2 compatible)
Cooling solution	Passive	Passive	Passive
# of users	4–64	2–64	1–16
OpenGL	4.3	4.3	4.4
Microsoft DirectX	9/10/11	9/10/11	9/10/11/12

HPE ProLiant WS460c Graphics Server Blade supported AMD MxGPU cards

Table 7. AMD MxGPU configuration options

	 **AMD FirePro S7100X**
Stream processors	2048
Memory size	8 GB GDDR5
Max. power	100 W
PCIe	x16
PCIe generation	Gen3
Cooling solution	Passive
# of users	1–16
OpenCL	2.0
OpenGL	4.2
Microsoft DirectX	11.1

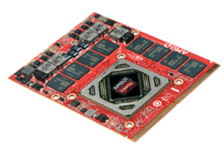
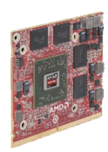
HPE ProLiant WS460c Graphics Server Blade supported Quadro full-size cards

Table 8. Quadro configuration options

					
	NVIDIA M6000	**NVIDIA M5000**	**NVIDIA K6000**	**NVIDIA K5000**	**NVIDIA K4000**
GPU	Maxwell	Maxwell	Kepler	Kepler	Kepler
CUDA cores	3072	2048	2880	1536	768
Memory size	12 GB GDDR5	8 GB GDDR5	12 GB GDDR5	4 GB GDDR5	3 GB GDDR5
Max. power	250 W	150 W	225 W	122 W	80 W
Form factor	Double-width	Double-width	Double-width	Double-width	Double-width
PCIe	x16	x16	x16	x16	x16
PCIe generation	Gen3	Gen3	Gen3	Gen2	Gen2
Cooling solution	Active	Active	Active	Active	Active
OpenGL	4.4	4.4	4.3	4.3	4.3
Microsoft DirectX	9/10/11/12	9/10/11/12	9/10/11	9/10/11	9/10/11

HPE ProLiant WS460c Graphics Server Blade supported MXM cards

Table 9. MXM configuration options

				
	NVIDIA Tesla M6	**AMD FirePro S7100X**	**AMD FirePro S4000X**	**NVIDIA Quadro K3100M**
CUDA/Stream cores	1536	2048	640	768
Memory size	8 GB GDDR5	8 GB GDDR5	2 GB	4 GB
Max. power	100 W	100 W	45 W	75 W
PCIe	x16 MXM	x16 MXM	x16 MXM	x16 MXM
PCIe generation	3	3	3	2
Cooling solution	Passive	Passive	Passive	Passive
OpenGL	4.4	4.4	4.3	4.3
OpenCL	2.0	2.0	2.0	N/A
Microsoft DirectX	9/10/11/12	9/10/11	9/10/11	9/10/11

WS460c Graphics Server Blade supported operating systems matrix

The following information complements the Server Blade documentation with supported operating systems and the hypervisor specific to the WS460c Graphics Server Blade being used for Hardware Accelerated Graphics for Desktop Virtualization.

Table 10. WS460c-supported operating systems

	Bare Metal Windows 7 32 bit	Bare Metal Windows 10 64 bit	Bare Metal Windows 7 64 bit	Bare Metal Windows 8.1 64 bit	VMware vSphere 6.x	Citrix XenServer 6.x	Citrix XenServer 7.x	Windows Server 2008 R2	Windows Server 2012 R2	Red Hat® Enterprise Linux® Workstation 6.x/7.x
WS460c Gen9 with Single MXM GPU	No	Coming soon	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
WS460c Gen9 with Graphics Expansion[4]	No	Coming soon	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
WS460c Gen9 with Multi-GPU Carrier	No	No	No	No	Yes	Yes	Yes	No	Yes	No
WS460c Gen8 MXM GPU	No	No	Yes	No	Yes	Yes	Yes	Yes	No	Yes
WS460c Gen8 with Graphics Expansion[4]	No	No	Yes	No	Yes	Yes	Yes	Yes	No	Yes
WS460c Gen8 with Multi-GPU Carrier	No	No	No	No	Yes	No	Yes	No	No	No

Table 11. WS460c Gen9 maximum supported graphics cards per platform and configuration

Hardware	Bare Metal	Microsoft 2008 R2	Microsoft 2012 R2	VMware vDGA	VMware vSGA	VMware w/ NVIDIA vGPU	VMware w/ AMD MxGPU	XenServer GPU pass-through	vGPU/ MxGPU
AMD FirePro S7100X	Max. 1	No	Max. 4[5,6]	Max. 4[5,6]	No	No	Max. 4[5,6]	No	No
NVIDIA Tesla M6	Max. 1	No	Max. 4[5,6]	Max. 4[5,6]	Max. 4[5,6]	Max. 4[5,6]	No	Max. 4[5,6]	Max. 4[5,6]
NV Quadro M5000	Max. 1[5]	No	Max. 1[5]	Max. 1[5]	No	No	No	Max. 2[5]	No
NV Quadro M6000	Max. 1[5]	No	Max. 1[5]	Max. 1[5]	No[5]	No	No	Max. 1[5]	No
NV Quadro K3100M	Max. 1	No	Max. 6[5,6]	Max. 6[5,6]	No	No	No	Max. 6[6]	No
NV GRID K1	No	No	Max. 1[5]	Max. 1[5]	Max. 1[5]	Max. 1[5]	No	Max. 1[5]	No
NV GRID K2	No	No	Max. 1[5]	Max. 1[5]	Max. 1[5]	Max. 1[5]	No	Max. 1[5]	No
NV Quadro K4000	Max. 2[5]	No	Max. 2[5]	Max. 2[5]	No	No	No	Max. 2[5]	No
NV Quadro K5000	Max. 1[5]	No	Max. 1[5]	Max. 1[5]	No[5]	No	No	Max. 1[5]	No
NV Quadro K6000	Max. 1[5]	No	Max. 1[5]	Max. 1[5]	No	No	No	Max. 1[5]	No
NV Tesla K40	No	No	No	No	No	No	No	No	No
NV Tesla K20	No	No	No	No	No	No	No	No	No
NV Quadro 4000	No	No	No	No	No	No	No	No	No
NV Quadro 5000	No	No	No	No	No	No	No	No	No
NV Quadro 6000	No	No	No	No	No	No	No	No	No
NV Quadro 3000M	No	No	No	No	No	No	No	No	No
NV Quadro 1000M	No	No	No	No	No	No	No	No	No
NV Quadro 500M	No	No	No	No	No	No	No	No	No
Teradici APEX MXM	No	No	No	No	No	No	No	No	No
NV Tesla M2070Q	No	No	No	No	No	No	No	No	No
AMD FirePro S4000X MXM	No	No	No	No	No	No	No	No	No

[4] HPE WS460c with Graphics Expansion allows for all full-size cards and HPE Multi-GPU Carrier Cards to be installed.
[5] Requires Graphics Expansion Blade.
[6] Requires HPE Multi-GPU Carrier Card.

Table 12. WS460c Gen8 maximum supported graphics cards per platform and configuration

Hardware	Bare Metal	Microsoft 2008 R2	Microsoft 2012 R2	VMware vDGA	VMware vSGA	VMware NVIDIA vGPU	VMware AMD MxGPU	XenServer GPU pass-through	XenServer NVIDIA vGPU
AMD FirePro S7100X	No	No	No	No	No	No	No	No	No
NVIDIA Tesla M6	No	No	No	No	No	No	No	No	No
NV Quadro K3100M	Max. 1	Max. 1	Max. 6[7, 8]	Max. 6[7, 8]	No	No	No	Max. 6[7, 8]	No
NV GRID K1	No	Max. 1[7]	Max. 1[7]	Max. 1[7]	Max. 1[7]	Max. 1[7]	No	Max. 1[7]	No
NV GRID K2	No	Max. 1[7]	Max. 1[7]	Max. 1[7]	Max. 1[7]	Max. 1[7]	No	Max. 1[7]	Max. 1[7]
NV Quadro K4000	Max. 2[7]	Max. 2[7]	Max. 2[7]	Max. 2[7]	No	No	No	Max. 1[7]	Max. 1[7]
NV Quadro K5000	Max. 1[7]	Max. 1[7]	Max. 1[7]	Max. 1[7]	No	No	No	Max. 1[7]	No
NV Quadro K6000	Max. 1[7]	Max. 1[7]	Max. 1[7]	Max. 1[7]	No	No	No	Max. 1[7]	No
NV Tesla K20	Max. 1[7]	No	No	Max. 1[7]	No	No	No	Max. 1[7]	No
NV Quadro 5000	Max. 1[7]	Max. 1[7]	Max. 1[7]	Max. 1[7]	No	No	No	Max. 1[7]	No
NV Quadro 6000	Max. 1[7]	Max. 1[7]	Max. 1[7]	Max. 1[7]	No	No	No	Max. 1[7]	No
NV Quadro 3000M	Max. 1	Max. 1	Max. 6[7, 8]	Max. 6[7, 8]	No	No	No	Max. 6[7, 8]	No
NV Quadro 1000M	Max. 2	Max. 2	Max. 8[7, 8]	Max. 8[7, 8]	No	No	No	Max. 8[7, 8]	No
NV Quadro 500M	Max. 2	No	No	No	No	No	No	No	No
Teradici APEX MXM	No	No	No	Max. 1	Max. 1	No	No	No	No
NV Tesla M2070Q	No	No	No	No	No	No	No	No	No
AMD FirePro S4000X MXM	Max. 2	No	No	No	No	No	No	No	No

[7] Requires Graphics Expansion Blade.
[8] Requires HPE Multi-GPU Carrier Card.

Learn more at
hpe.com/us/en/integrated-systems/bladesystem.html

Sign up for updates

Hewlett Packard Enterprise

4AA4-1701ENW, August 2016, Rev. 7