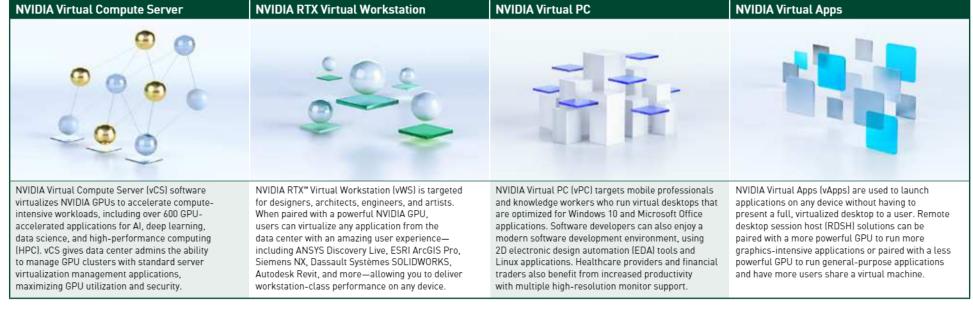


# NVIDIA GPUs FOR VIRTUALIZATION

NVIDIA virtual GPU (vGPU) software enables powerful GPU performance from the enterprise data center, as well as public and private clouds for any workload from virtual desktops (VDI) to data science and AI. Installed on a server with an NVIDIA GPU, the NVIDIA vGPU software creates virtual GPUs that can be shared between multiple virtual machines running on any device, anywhere. This enables the IT department standardized on software-defined and hyperconverged infrastructure to leverage the management and security benefits of virtualization –and get the performance of NVIDIA GPUs for modern graphics and compute workloads.

## Choose the Right Virtual GPU Software Edition for Your Use Case:



## NVIDIA GPUs Recommended for Virtualization (Ampere Architecture)

	A100	A30	A40	A10	A16
	[2]				
GPU Architecture	1 NVIDIA Ampere	1 NVIDIA Ampere	1 NVIDIA Ampere	1 NVIDIA Ampere	4 NVIDIA Ampere
RTX Technology	_	-	<b>√</b>	<b>√</b>	J
Guaranteed QoS (GPU Scheduler)	/	<b>J</b>	<b>/</b>	/	<b>J</b>
Live Migration	<b>J</b>	<b>√</b>	<b>√</b>	<b>√</b>	<b>√</b>
Multi-vGPU	J	<b>/</b>	<b>/</b>	/	J
Memory Size	40/80GB HBM2	24GB HBM2	48GB GDDR6	24GB GDDR6	64 GB GDDR6 (16 GB per GPU)
vGPU Profiles	4GB, 5GB, 8GB, 10GB, 16GB, 20GB, 40GB, 80GB	4 GB, 6GB, 8GB, 12GB, 24GB	1GB <sup>1</sup> , 2GB <sup>1</sup> , 3GB <sup>1</sup> , 4GB, 4GB, 8GB, 12GB, 16GB, 24GB, 48GB	1GB <sup>1</sup> , 2GB <sup>1</sup> , 3GB <sup>1</sup> , 4GB, 6GB, 8GB, 12GB, 24GB	1GB <sup>1</sup> , 2GB <sup>1</sup> , 4GB, 8GB, 16GB
Form Factor	SXM4, and PCIe 4.0 dual slot	PCIe 4.0 dual slot	PCIe 4.0 dual slot	PCIe 4.0 single slot	PCIe 4.0 dual slot
Power	400/250W	165W	300W	150W	250W
Thermal	passive	passive	passive	passive	passive
vGPU Software Support	vCS	vCS	vWS, vPC, vApps, vCS	vWS, vPC, vApps, vCS	vWS, vPC, vApps, vCS
Use Case	Al, data analytics, and HPC	Al inferencing, data analytics, and HPC	High-end 3D design and creative workflows with vWS; upgrade path for RTX 8000, RTX 6000	Mid-range 3D design and engineering workflows with vWS. High-density GPU acceleration for knowledge workers with vPC. Deep learning inferencing with vCS. Upgrade path for T4.	Knowledge worker virtual desktops using modern productivity apps and multimedia with NVIDIA vPC/ vApps. Upgrade path for M10.

The following NVIDIA GPUs are also supported for virtualization: NVIDIA V100S, RTX A6000, RTX A5000, RTX 8000 (active), RTX 6000 (active), P40, and P6. 

Not supported by vCS. Minimum profile size supported by vCS is 46B.

### WHAT MAKES NVIDIA VIRTUAL GPUS POWERFUL



#### **EXCEPTIONAL USER EXPERIENCE**

Ultimate user experience, with the ability to support both compute and graphics workloads.



#### BEST USER DENSITY

Industry's highest user-density solution with support for up to 64 virtual desktops per physical GPU. Lower TCO with more than ten vGPU profiles for the most flexibility to provision resources to match your users' needs.



#### CONTINUOUS INNOVATION

Regular cadence of new software releases to ensure you stay on top of the latest features and enhancements.



#### **PERFORMANCE**

Consistent near bare-metal performance, whether on premises or in the cloud.



#### **OPTIMAL MANAGEMENT AND MONITORING**

End-to-end management and monitoring for realtime insight into GPU performance. Broad partner integrations so you can use the tools you know and love.



#### **BROADEST ECOSYSTEM SUPPORT**

Support for all major hypervisors. Most extensive portfolio of professional apps certifications with NVIDIA RTX Enterprise Drivers.

## Explore NVIDIA GPUs From Previous Generations

Explore NVIDIA OF 05 From Frevious Generations								
	V100	RTX 8000	RTX 6000	T4	M10			
		The second						
		2	-	0)				
GPU Architecture	1 NVIDIA Volta™	1 NVIDIA Turing™	1 NVIDIA Turing	1 NVIDIA Turing	4 NVIDIA Maxwell™			
RTX Technology	-	<b>/</b>	<b>/</b>	J	-			
Guaranteed QoS (GPU Scheduler)	/	<b>√</b>	/	<b>V</b>	_			
Live Migration	J	<b>√</b>	<b>J</b>	J	J			
Multi-vGPU	J	<b>/</b>	<b>J</b>	J	J			
Memory Size	32/16GB HBM2	48GB GDDR6	24GB GDDR6	16GB GDDR6	32GB GDDR5 (8GB per GPU)			
vGPU Profiles	1GB¹, 2GB¹, 4GB, 8GB, 16GB, 32GB	1GB <sup>1</sup> , 2GB <sup>1</sup> , 3GB <sup>1</sup> , 4GB, 6GB, 8GB, 12GB, 16GB, 24GB, 48GB	1GB¹, 2GB¹, 3GB¹, 4GB, 6GB, 8GB, 12GB, 24GB	1GB¹, 2GB¹, 4GB, 8GB, 16GB	0.5GB, 1GB, 2GB, 4GB, 8GB			
Form Factor	PCIe 3.0 dual slot and SXM2	PCIe 3.0 dual slot	PCIe 3.0 dual slot	PCIe 3.0 single slot	PCIe 3.0 dual slot			
Power	250W/300W (SXM2)	250W	250W	70W	225W			
Thermal	passive	passive	passive	passive	passive			
vGPU Software Support	vWS, vPC, vApps, vCS	vWS, vPC, vApps, vCS	vWS, vPC, vApps, vCS	vWS, vPC, vApps, vCS	vWS, vPC, vApps			
Use Case	Ultra-high-end rendering, simulation, 3D design with vWS; compute- intensive AI, deep learning, HPC workloads with vCS	High-end rendering, 3D design, and creative workflows with vWS; AI and data science workloads with vCS	Mid-range to high-end rendering, 3D design, and creative workflows with vWS	Entry-level 3D design and engineering workflows with vWS; high-density, low-power GPU acceleration for knowledge workers with vPC software	Knowledge workers using modern productivity apps and Windows 10 with NVIDIA vPC/vApps			

The following NVIDIA GPUs are also supported for virtualization: NVIDIA V100S, RTX A6000, RTX A5000, RTX 8000 (active), RTX 6000 (active), P40, and P6.  $^1$  Not supported by vCS. Minimum profile size supported by vCS is 4GB.

