NVIDIA® VIRTUAL COMPUTE SERVER (vCS)
POWER THE MOST COMPUTE-INTENSIVE WORKLOADS WITH VIRTUAL GPUs

TRANSFORMING VIRTUALIZED COMPUTE

As the number of servers grow across the data center, IT admins expect to manage them with standard server virtualization platforms from VMware, Red Hat, Nutanix, and Citrix. According to Gartner, "hypervisor-based server virtualization is now mature, with 80% to 90% of server workloads running in a virtual machine (VM) for most midsize to large enterprises."[1] However, this traditional data center infrastructure using hypervisor-based virtualization has been limited to CPU-only servers, with VDI as an exception. As a result, GPU accelerated servers running AI, deep learning, data science, and high-performance computing (HPC) workloads are often isolated in other servers in the data center, limiting utilization, flexibility, and manageability.

NVIDIA® Virtual Compute Server (vCS) enables the benefits of hypervisor-based server virtualization for GPU-accelerated servers. Data center admins are now able to power any compute-intensive workload with GPUs in a virtual machine (VM).

vCS software virtualizes NVIDIA GPUs to accelerate large workloads, including more than 600 GPU accelerated applications for AI, deep learning, and HPC. With GPU sharing, multiple VMs can be powered by a single GPU, maximizing utilization and affordability, or a single VM can be powered by multiple virtual GPUs, making even the most intensive workloads possible. And with support for nearly all major hypervisor virtualization platforms, data center admins can use the same management tools for their GPU-accelerated servers as they do for the rest of their data center.

LICENSED FOR COMPUTE

Unlike NVIDIA Virtual PC/Virtual Applications (vPC/vApps) and RTX Virtual Workstation (vWS), vCS is not tied to a user with a display. It's licensed per GPU as a 1-year subscription with NVIDIA enterprise support included. This allows a number of compute workloads in multiple VMs to be run on a single GPU, maximizing utilization of resources and ROI.

OPTIMIZED FOR CONTAINERS WITH NGC SOFTWARE

vCS supports NVIDIA NGC GPU-optimized software for deep learning, machine learning, and HPC. NGC software includes containers for the top AI and data science software, tuned, tested, and optimized by NVIDIA, as well as fully-tested containers for HPC applications and data analytics.

NGC also offers pre-trained models for a variety of common AI tasks that are optimized for NVIDIA Tensor Core GPUs and includes instructions and scripts for creating deep learning models with sample performance and accuracy metrics. This allows data scientists, developers, and researchers to reduce deployment times and project complexity so that they can focus on building solutions, gathering insights, and delivering business value.

FEATURES

- **GPU Performance** - Access the most powerful GPUs in a virtualized environment.
- **Management and Monitoring** - Streamline data center manageability by leveraging hypervisor-based tools.
- **Live Migration** - Live migrate GPUaccelerated VMs without disruption, easing maintenance and upgrades.
- **Maximize Utilization** - Increase utilization and productivity with both GPU sharing and aggregation of multiple GPUs.
- **Security** - Extend the benefits of server virtualization to GPU workloads.
- **Multi-Tenant** - Isolate workloads and securely support multiple users.
- **Rapid Deployment** - Leverage GPU optimized NGC containers for AI, data science, and HPC using GPU Operator.
- **Reliability** - Prevent against data corruption with error-correcting code (ECC) and dynamic page retirement.
- **Enterprise Software Support** - Enterprise Software Support - Get support with NVIDIA Enterprise and NVIDIA NGC Support Services.
- **Low Latency Data Transfer** - Direct access to GPU memory with GPUDirect RDMA.
- **Multi-Instance GPU (MIG)** - Enable heterogenous operating systems and vCS profiles.

NVIDIA vCS FEATURES LIST

Configuration and Deployment	
GPU Sharing (fractional)	✓
GPU Aggregation (Multi-vGPU)	✓
Peer-to-Peer over NVLink	✓
ECC & Dynamic Page Retirement	✓
Linux OS Support	✓
Windows OS Support	X
NVIDIA Compute Driver	✓
NVIDIA Graphics Driver	X
NVIDIA RTX Enterprise Driver	X
Quality-of-Service Scheduling	✓
GPUDirect RDMA	✓
Multi-Instance GPU (MIG)	✓
GPU Operator	✓

Data Center Management	
Host-, Guest-, and Application-Level Monitoring	✓
Live Migration	✓

Support	
NVIDIA Direct Enterprise-Level Technical Support	✓
Maintenance Releases, Defect Resolutions, and Security Patches[2]	✓
NGC Support Services[3]	✓

vCS PROFILES

Maximum Frame Buffer Supported	80GB
Minimum Frame Buffer Supported	4GB
Maximum Multi-Tenancy	20:1[9]
Available Profiles	4C, 5C[8], 6C, 8C, 10C[8], 12C, 16C, 20C[8], 24C[4], 32C[5], 40C[8], 48C[6], 80C[8],

RECOMMENDED GPUs FOR vCS

	NVIDIA A100	NVIDIA A30[10]	NVIDIA V100	NVIDIA A40	NVIDIA A10[10]	NVIDIA RTX™ 8000/6000	NVIDIA T4
GPU/Boards Architecture	1 (Ampere)	1 (Ampere)	1 (Volta)	1 (Ampere)	1 (Ampere)	1 (Turing)	1 (Turing)
RT Cores	-	-	-	✓	✓	✓	✓
Memory Size	40 GB/80 GB HBM2	24 GB HBM2	32 GB/16 GB HBM2	48 GB GDDR6	24 GB GDDR6	48 GB/24 GB GDDR6	16 GB GDDR6
NVLink	✓	-	✓	-	-	✓	-
Form Factor	PCIe 4.0 Dual Slot/SXM4	PCIe Gen4 Dual Slot	PCIe 3.0 Dual Slot/SXM2	PCIe 4.0 Dual Slot	PCIe Gen4 Single Slot	PCIe Gen4 Dual Slot	PCIe 3.0 Single Slot
Power	250W/400W	165W	250W/300W	300W	150W	250W	70W

ADDITIONAL SUPPORTED GPUs

NVIDIA® P40, P100, and P6 for blade form factor.

[1] Gartner. Market Guide for Server Virtualization. April 24, 2019. ID G00350674.

[2] Available with an active Support, Updates, and Maintenance (SUMs) contract.

[3] Not included with vCS license, but available separately through NVIDIA NGC Support Service partners.

[4] 24C profile available with NVIDIA A40, and NVIDIA RTX 6000 and RTX 8000.

[5] 32C profile available with NVIDIA V100.

[6] 48C profile supported with RTX 8000.

[7] Number of multi-GPUs supported may vary by hypervisor.

[8] 5C, 10C, 20C, 40C profiles are available with NVIDIA A100 and 80C profile available with NVIDIA A100 80GB.

[9] 20:1 max multi-tenancy is available with NVIDIA A100 80GB.

[10] Availble in future vGPU software release