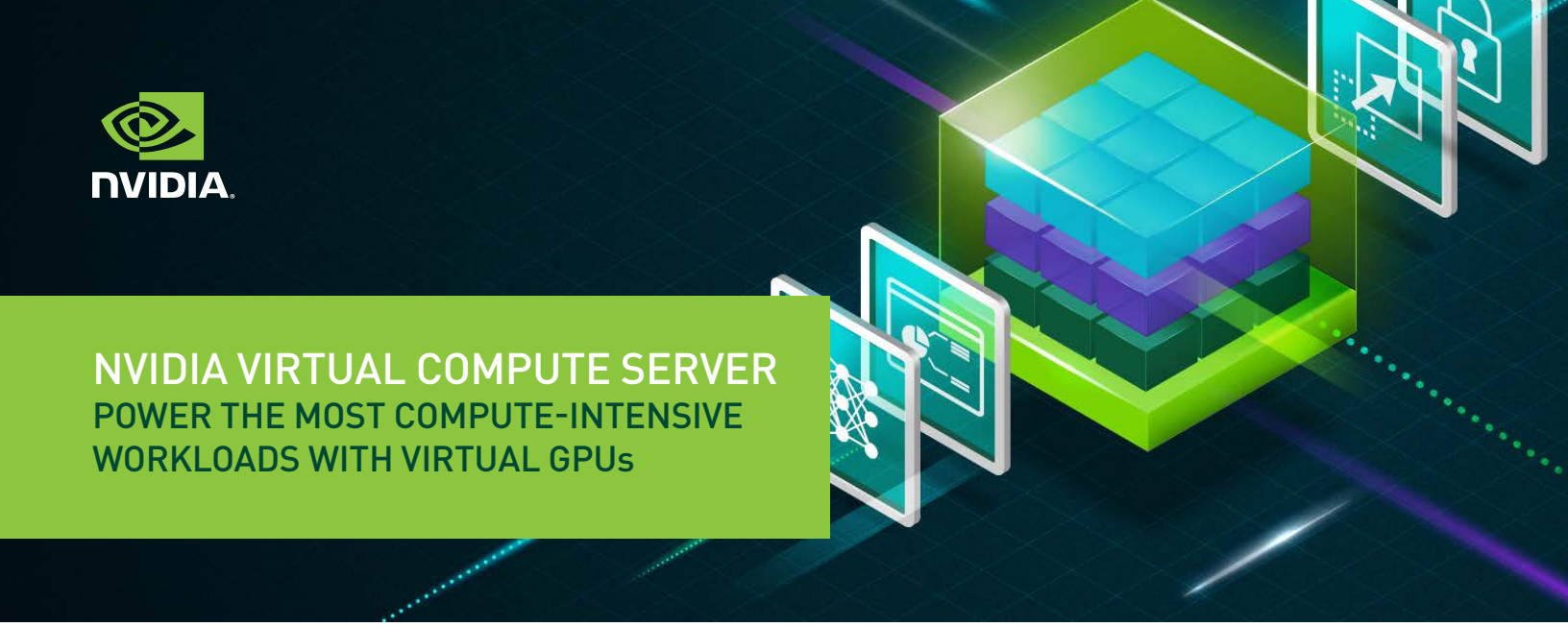

NVIDIA VIRTUAL COMPUTE SERVER

POWER THE MOST COMPUTE-INTENSIVE WORKLOADS WITH VIRTUAL GPUs

As the number of GPU servers grows across the data center, IT admins expect to manage them with standard server virtualization platforms from VMware, Red Hat, Nutanix, and Citrix. However, traditional data center architecture using hypervisor-based virtualization for server workloads—which accounts for 80-90% of midsize to large enterprises[1]—has been limited to CPU-only servers, with VDI as an exception. As a result, GPU servers running AI, deep learning, and high-performance computing (HPC) workloads are often isolated from the data center, limiting utilization, flexibility, and manageability.

Additionally, data centers frequently use server virtualization platforms to deploy containers. While this improves manageability and security, they still lack the ability to support GPUs with VM-integrated containers.

TRANSFORMING VIRTUALIZED COMPUTE

NVIDIA vComputeServer enables the benefits of hypervisor-based server virtualization for GPU-enabled servers. Data center admins are now able to run any compute-intensive workload that requires GPUs in a virtual machine (VM).

NVIDIA vComputeServer software virtualizes NVIDIA GPUs to accelerate compute-intensive workloads, including more than 600 GPU accelerated applications for AI, deep learning, data science, 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 compute-intensive workloads possible. And with support for all major hypervisor virtualization platforms, data center admins can use the same management tools for their GPU clusters as they do for the rest of their data center, maximizing GPU utilization and ensuring security.

LICENSED FOR COMPUTE

Unlike NVIDIA GRID vPC/vApps and Quadro vDWS, which are client compute products licensed per concurrent user (CCU), NVIDIA vComputeServer is a server compute product that is not tied to a user with a display. As such, NVIDIA vComputeServer is licensed per GPU as a 1-year subscription with NVIDIA enterprise support included. This allows multiple compute workloads in multiple VMs to be run on a single GPU without any added cost, maximizing utilization of resources and ROI.

OPTIMIZED FOR CONTAINERS WITH NGC SOFTWARE

NVIDIA vComputeServer 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.

FEATURES

- **GPU Performance** - Access the most powerful GPUs in a virtualized environment.
- **Management and Monitoring** - Streamline data center manageability by leveraging hypervisor-based management and monitoring tools to also manage GPU-enabled servers.
- **Live Migration** - Live migrate GPU-accelerated VMs without disruption, easing maintenance, availability, and upgrades.
- **Maximize Utilization** - Increase utilization and productivity with both GPU sharing (fractionalizing) and aggregation of multiple GPUs, including peer-to-peer over NVLink in a VM.
- **Security** - Enhance security using hypervisor-based security extending to GPU workloads.
- **Multi-Tenant** - Enable multi-tenant deployments to isolate workloads and securely support multiple users.

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

NVIDIA VCOMPUTESERVER FEATURES LIST

Configuration and Deployment	
GPU Sharing (fractional)	✓
GPU Aggregation (Multi-vGPU)	✓
Peer-to-Peer over NVLink	✓
ECC & Dynamic Page Retirement	✓
NVIDIA NGC Support	✓
Linux OS Support	✓
Windows OS Support	x
NVIDIA Compute Driver	✓
NVIDIA Graphics Driver	x
NVIDIA Quadro Driver	x
Quality-of-Service Scheduling	✓

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 for up to Three Years[2]	✓
NGC Support Services[3]	✓

- **Rapid Deployment** - Rapid deployment of GPU-optimized NGC containers for AI and HPC. Continuous performance improvement over time on the same host system with constant optimization of containerized software from NGC.
- **Reliability** - Error-correcting code (ECC) and dynamic page retirement prevents against data corruption, critical for HPC, scientific, and finance workloads.
- **Enterprise Software Support** - Fully supported with NVIDIA Enterprise and NVIDIA NGC Support Services.

VCOMPUTESERVER PROFILES

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

RECOMMENDED GPUS FOR VCOMPUTESERVER

	NVIDIA T4	NVIDIA V100 (SXM2)
RT Cores	48	-
Tensor Cores	320	640
CUDA® Cores	2,560	5,120
Memory	16 GB GDDR6	32 GB HBM2
FP 16/FP 32 (mixed precision)	64 TFLOPS	125 TFLOPS
FP 32 (single precision)	8.1 TFLOPS	15.7 TFLOPS
FP 64 (double precision)	-	7.8 TFLOPS
NVLink: Number of GPUs per VM	-	Up to 8
ECC and Page Retirement	✓	✓
Multi-GPU per VM	Up to 16	Up to 16

ADDITIONAL SUPPORTED GPUs

NVIDIA® Quadro RTX™ 6000, RTX 8000, NVIDIA P40, P100, and P6 for blade form factor.

[1] Data from Gartner's "Market Guide for Server Virtualization", April 2019
[2] Available with an active Support, Updates, and Maintenance (SUMs) contract.
[3] Not included with vComputeServer license, but available separately through NVIDIA NGC Support Service partners.
[4] 24C profile available with Quadro RTX 6000 and RTX 8000.
[5] 32C profile available with NVIDIA V100.
[6] 48C profile supported with Quadro RTX 8000.

NVIDIA