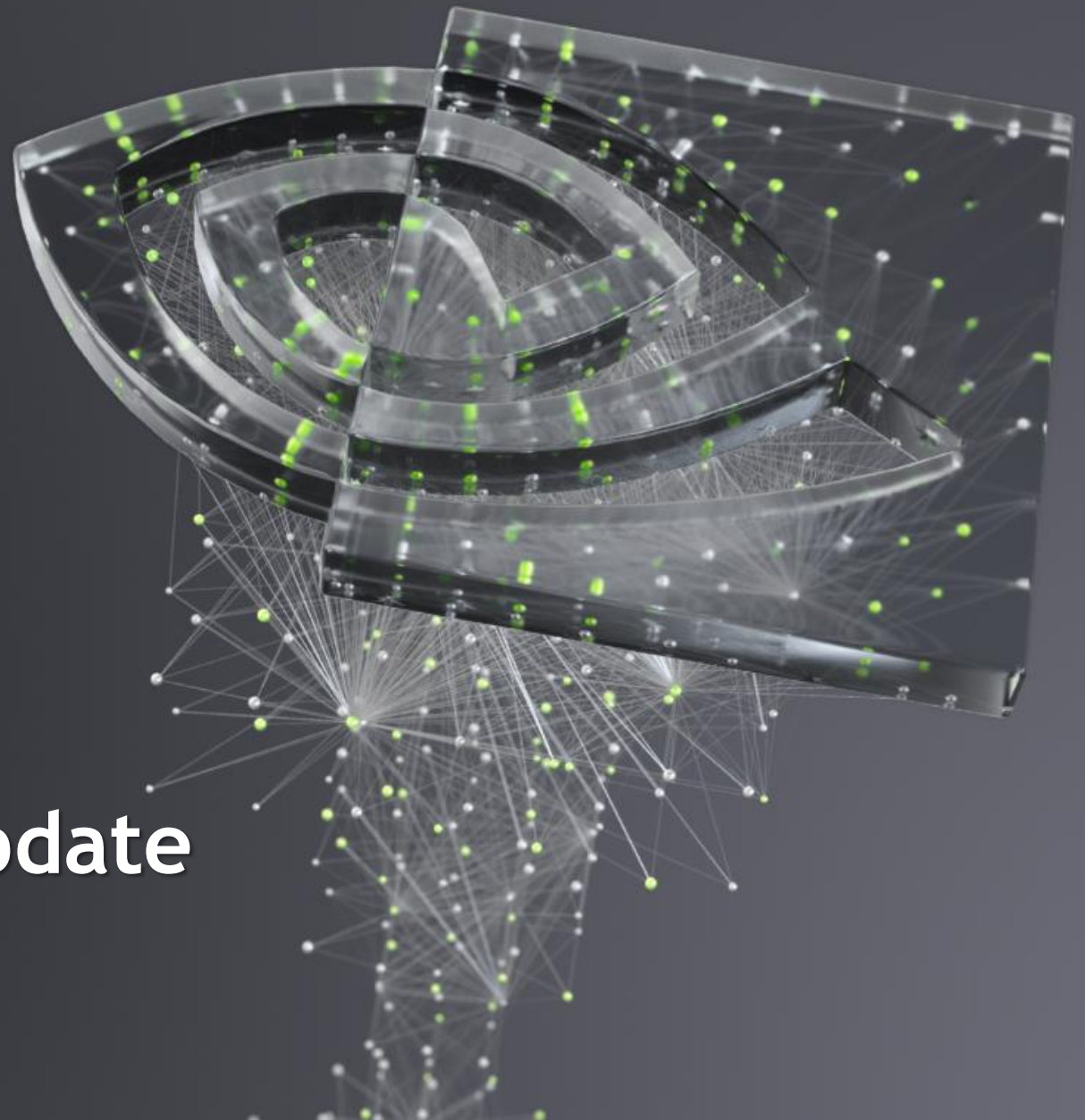


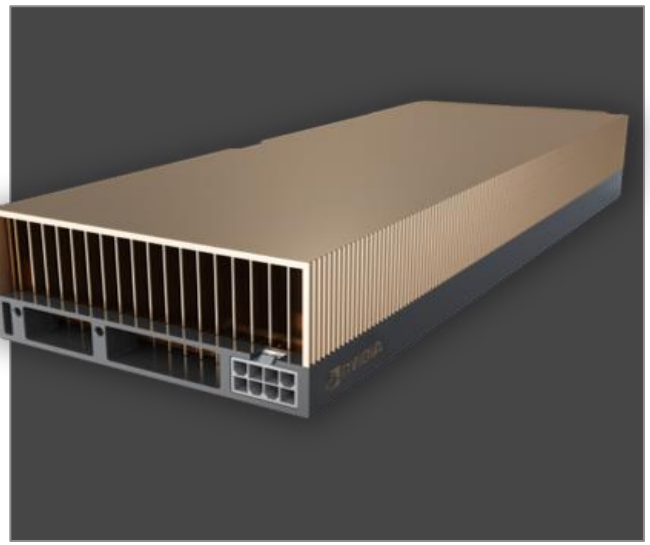
vGPU 12 product update

June 23, 2020

VIRTUAL GPU 2021 (12.0)

Next Generation Ampere Architecture Powers Virtualization

vGPU 12.0 GA



NVIDIA A40
High-Performance RTX Graphics
(vGPU 12.0)

NVIDIA A6000
the world's most powerful visual computing GPU for desktop workstations.

Virtual Compute Features
A100 80 GB, GPU Operator, CUDA Tools, UVM, GPUDirect Storage, NV Certified
(vGPU 12.1)

NVIDIA A40 GRAPHICS USE CASES

The World's Most Powerful Data Center GPU for Visual Computing

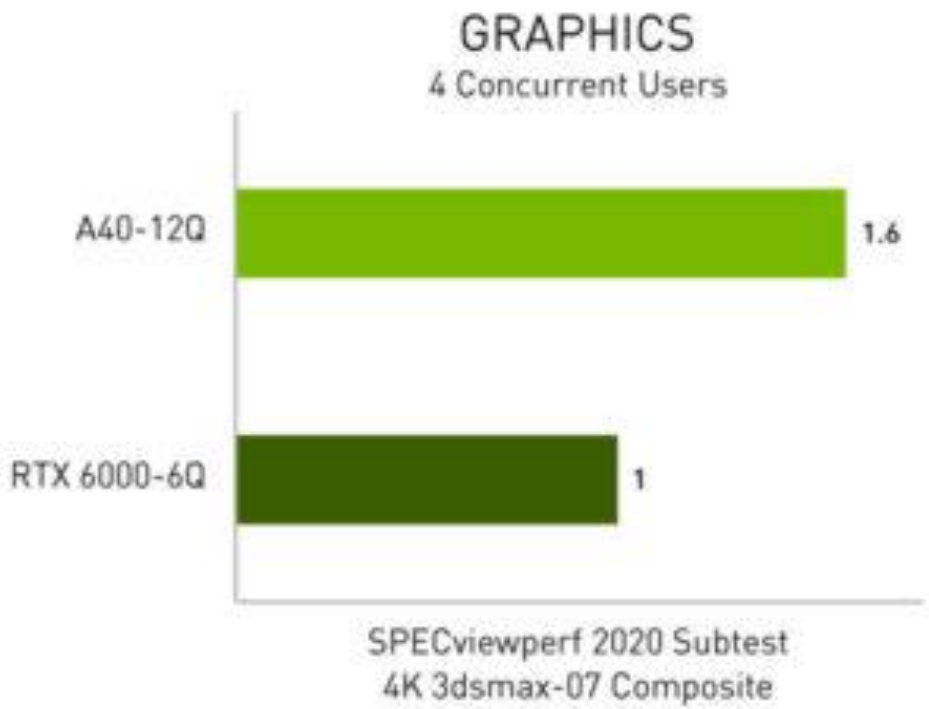
Batch Rendering & Simulation
Compute workloads

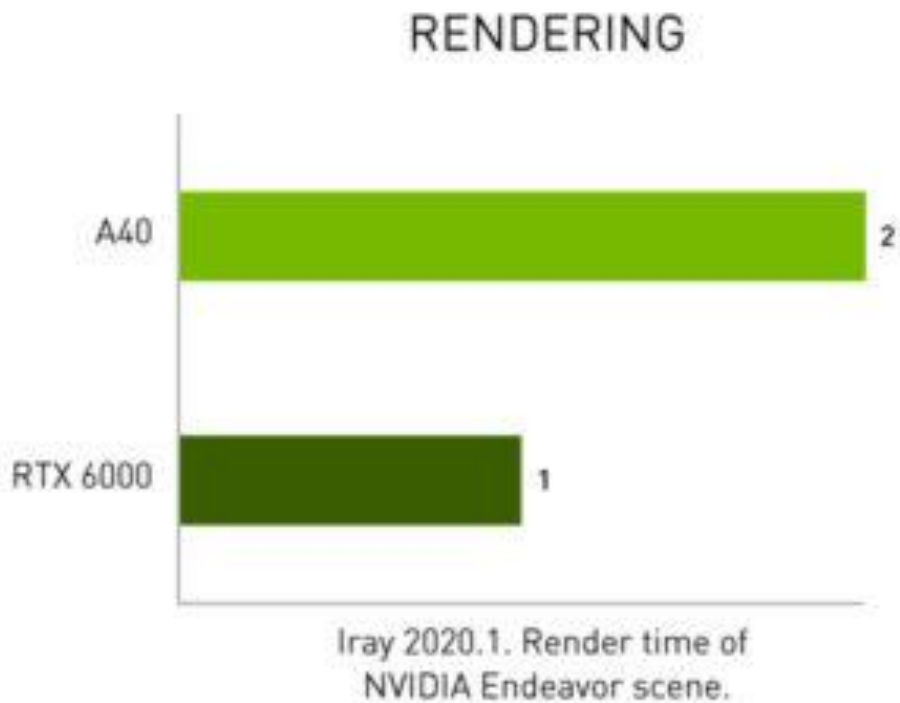
Interactive Graphics with vGPU
High-performance virtual workstations

Media & Entertainment
CAVES, Virtual Production, Location-based Entertainment

NVIDIA A40 VIRTUAL WORKSTATION PERFORMANCE

Up to 60% Better Graphics Performance & 2X Faster Rendering





NVIDIA DATA CENTER GPUS

NVIDIA T4

Versatile Data Center GPU for Mainstream Computing

- Multi-Purpose GPU for Enterprise Acceleration, Graphics, Inference
- Entry – Mid Range Quadro vDWS
- 16GB GPU Memory
- NVIDIA Turing Architecture Tensor Cores

NVIDIA A40

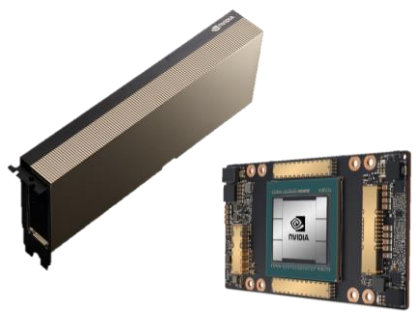
(RTX 6000/8000 Passive)

World's Most Powerful Data Center GPU for Visual Computing

- Fastest Graphics & Ray Tracing
- Largest models, Mid – High-end Quadro vDWS[1]
- 48GB GPU Memory
- 2-Way NVLink
- NVIDIA Ampere Architecture Tensor Cores, RT Cores

NVIDIA A100

World's Most Powerful Data Center GPU



- Fastest Compute
- FP64 Precision
- Virtualization with NVIDIA vCS
- 8-way NVLink and NVSwitch
- 40GB GPU Memory
- MIG for max utilization
- NVIDIA Ampere Architecture Tensor Cores

1. *NVIDIA A40 support coming in a future virtual GPU (vGPU) software release (early 2021)*

GEN-TO-GEN PRODUCT COMPARISON

	NVIDIA A40	RTX 6000 / 8000 Passive
GPU Architecture	NVIDIA Ampere Architecture	NVIDIA Turing Architecture
CUDA Cores	10752	4608
Tensor Cores	336 (3rd Gen)	576 (2nd Gen)
RT Cores	84 (2nd Gen)	72 (1st Gen)
Memory Size	48 GB GDDR6 w/ECC	24/48 GB GDDR6 w/ECC
Memory Bandwidth	696 GB/s	672 GB/s
NVLink	2-way (low profile)	2-way
Display Connectors	3 x DP 1.4	None
Quadro Sync	Yes	No
Max Power Consumption	300W	250W
Graphics Bus	PCI Express Gen 4 x 16	PCI Express Gen 3 x 16

GEN-TO-GEN PRODUCT COMPARISON

	RTX A6000	RTX 6000	P6000
GPU Architecture	Ampere	Turing	Pascal
CUDA Cores	10752	4608	3840
Tensor Cores	336 Ampere Arch Cores	576 Turing Arch Cores	-
RT Cores	84	72	-
Peak Single-Precision Performance	40.0 TFLOPS	16.3 TFLOPS	12 TFLOPS
Memory Size	48 GB GDDR6 w/ECC	24 GB GDDR6 w/ECC	24 GB GDDR5X w/ECC
Memory Bandwidth	768 GB/s	672 GB/s	432 GB/s
NVLink	2-way	2-way	-
Display Connectors	4x DP 1.4	4x DP 1.4 + 1x USB-C	4x DP 1.4
Max Power Consumption	300W	260W*	250W
Graphics Bus	PCI Express Gen 4 x 16	PCI Express Gen 3 x 16	PCI Express Gen 3 x 16

*RTX 6000: Total board power 295W, total graphics power 260W

DISPLAY MODE SELECTOR TOOL

Summary

- Goal: Simplify Customer Purchase based on Use Case
- Recommend: RTX A6000 for workstations, NVIDIA A40 for servers
- Default display modes are aligned with the recommended use cases
- Display Mode Selector Tool – provides flexibility for customers to change the default display mode **only for specific use cases**
- Display Mode Selector Tool can only be accessed with prior approval
- Improper use of the tool could result in undesired effects (ex. GPU can be bricked)

NVIDIA A40 & RTX A6000

Display Mode Options & Use Cases

Physical Display Ports Enabled With 256MB BAR1	Physical Display Ports Enabled With 8GB BAR1	Physical Display Ports Disabled
Use for standard workstation deployments with physically attached displays which can be optionally synchronized.	Use for broadcast, virtual production, and location-based entertainment deployments requiring optionally synchronized hi-res displays that are physically attached.	Use for running NVIDIA Virtual GPU or deployments where no physically attached displays are required, such as compute use cases.
RTX A6000 Default		A40 Default

DISPLAY MODE SUPPORTED FEATURES

	Virtual GPU	Pass Through	Bare Metal
Physical Display Ports Enabled with 256MB or 8GB BAR1	Not Supported	Supported with NVIDIA RTX Enterprise driver	Supported with NVIDIA RTX Enterprise driver
Physical Display Ports Disabled	Supported with NVIDIA vGPU software	Supported with vGPU software or NVIDIA Data Center driver	Supported with NVIDIA vGPU software or NVIDIA Data Center driver

Supported with the following NVIDIA vGPU software editions:

- NVIDIA RTX Virtual Workstation (vWS)
- NVIDIA Virtual PC (vPC) or Virtual Applications (vApps)
- NVIDIA Virtual Compute Server (vCS)

vGPU ECOSYSTEM EVOLVES

KVM Partners

NVIDIA's KVM partners include Red Hat RHEL/RHV, Nutanix AHV, and CSP partners, with the following new partnerships:

Red Hat Virtualization (RHV)

Red Hat Enterprise Linux with KVM

OpenStack Platform

SUSE Linux Enterprise Server (SLES) will support vGPU (coming soon)

SLES known for:

- Used in the majority of Fortune Global 100 companies*
- SLES already supports GPU for HPC, this will add support for vGPU

*Source: www.suse.com/company/about/

NVIDIA VIRTUAL GPU BRANDING REFRESH

Current Naming	New Naming (as of Jan 2021)
NVIDIA GRID Virtual PC (GRID vPC)	NVIDIA Virtual PC (vPC)
NVIDIA GRID Virtual Applications (GRID vApps)	NVIDIA Virtual Applications (vApps)
NVIDIA Quadro Virtual Data Center Workstation (Quadro vDWS) NVIDIA Quadro Virtual Workstation (Quadro vWS)	NVIDIA RTX Virtual Workstation (vWS)
NVIDIA Virtual Compute Server (vCS)	NVIDIA Virtual Compute Server (vCS)

SUMMARY OF NVIDIA VIRTUAL GPU BRANDING UPDATES

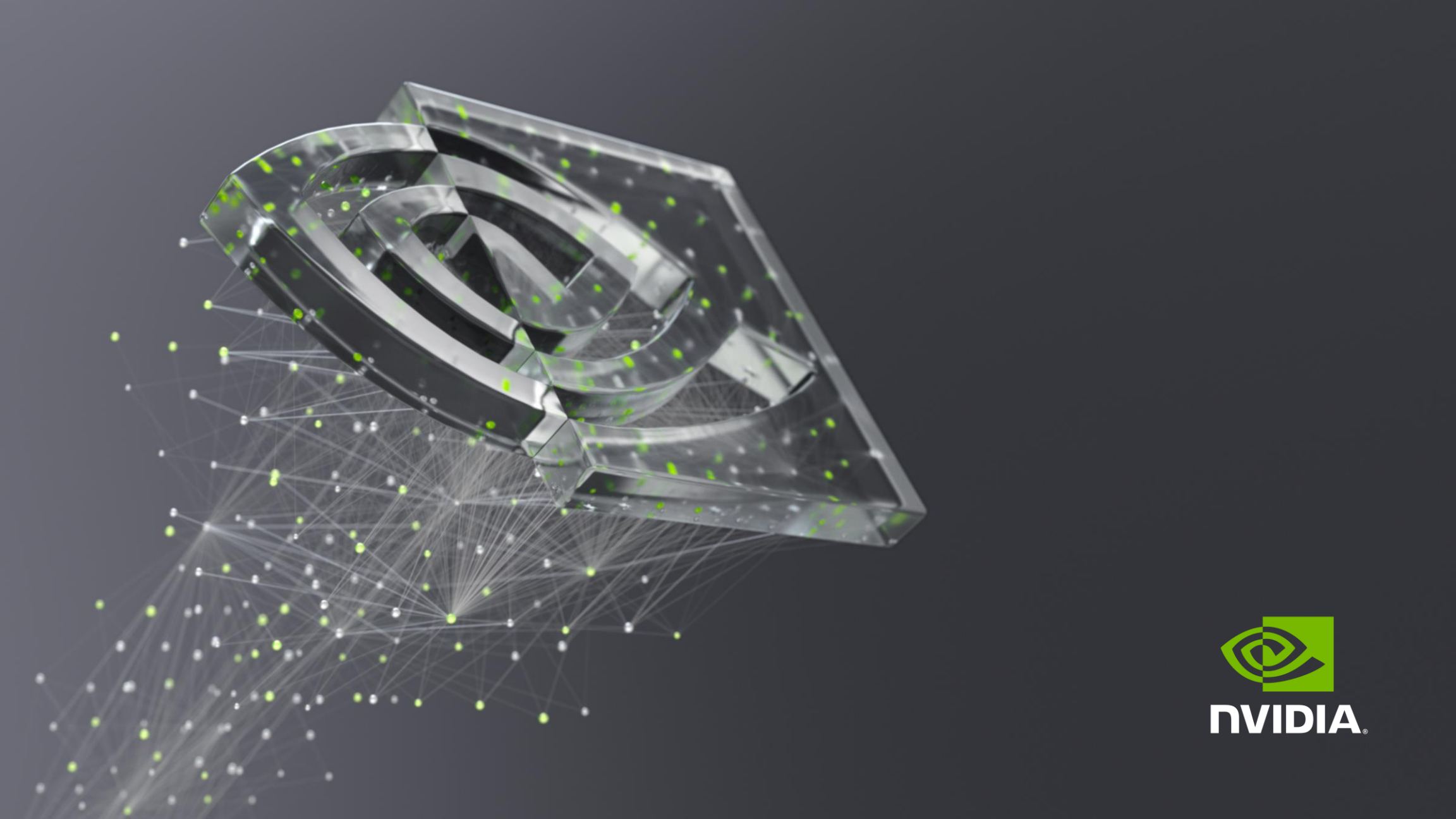
- Quadro branding goes away
  - NVIDIA Quadro Virtual Data Center Workstation and NVIDIA Quadro Virtual Workstation → NVIDIA RTX Virtual Workstation (vWS)
- GRID branding goes away
  - NVIDIA GRID Virtual PC → NVIDIA Virtual PC (vPC)
  - NVIDIA GRID Virtual Apps → NVIDIA Virtual Apps (vApps)
  - GRID T4-1B → NVIDIA vGPU T4-1B
  - GRID vGaming → NVIDIA vGaming
- NVIDIA Virtual Compute Server (vCS) → stays the same
- NVIDIA Q profile remains the same
- Drivers: NVIDIA RTX Enterprise Driver (vWS), Graphics Driver (vPC), Compute Driver (vCS)

VIRTUAL GPU 12

Who should use this release

- Customers looking to move to the Ampere hardware
- Suse enterprise users
- Admins looking to upgrade their nodes to the latest hypervisor and guest OS versions
- Customers looking to try new feature


Q&A


NVIDIA