Choosing the right NVIDIA GPU for your workload

ERIK BOHNHORST

TECHNICAL MARKETING @ NVIDIA

EMAIL: EBOHNHORST@NVIDIA.COM

TWITTER:@ERIKBOH

NVIDIA vGPU PRODUCTS

NVIDIA GRID Virtual Applications

For organizations deploying XenApp or other RDSH solutions. Designed to deliver Windows applications at full performance.

NVIDIA GRID Virtual PC

For users who want a virtual desktop but need great user experience leveraging PC Windows applications, browsers and high definition video.

NVIDIA Quadro Virtual Data Center Workstation

For users who want to be able to use remote professional graphics applications with full performance on any device, anywhere.

NVIDIA DATA CENTER GPUs

	V100	RTX 8000	RTX 6000	P40	T4	M10	P6
GPUs / Board (Architecture)	1 (Volta)	1 (Turing)	1 (Turing)	1 (Pascal)	1 (Turing)	4 (Maxwell)	1 (Pascal)
CUDA Cores	5,120	4,608	4,608	3,840	2,560	2,560 (640 per GPU)	2,048
Tensor Cores	640	576	576	---	320	---	---
RT Cores	---	72	72	---	40	---	---
Memory Size	32 GB/16 GB HBM2	48 GB GDDR6	24 GB GDDR6	24 GB GDDR5	16 GB GDDR6	32 GB GDDR5 (8 GB per GPU)	16 GB GDDR5
vGPU Profiles	1 GB, 2 GB, 4 GB, 8 GB, 16 GB, 32 GB	1 GB, 2 GB, 3 GB, 4 GB, 6 GB, 8 GB, 12 GB, 16 GB, 24 GB, 48 GB	1 GB, 2 GB, 3 GB, 4 GB, 6 GB, 8 GB, 12 GB, 24 GB	1 GB, 2 GB, 3 GB, 4 GB, 6 GB, 8 GB, 12 GB, 24 GB	1 GB, 2 GB, 4 GB, 8 GB, 16 GB	0.5 GB, 1 GB, 2 GB, 4 GB, 8 GB	1 GB, 2 GB, 4 GB, 8 GB, 16 GB
Form Factor	PCIe 3.0 Dual Slot & SXM2 (rack servers)	PCIe 3.0 Dual Slot	PCIe 3.0 Dual Slot	PCIe 3.0 Dual Slot (rack servers)	PCIe 3.0 Single Slot (rack servers)	PCIe 3.0 Dual Slot (rack servers)	MXM (blade servers)
Power	250W/300W	295W	295W	250W	70W	225W	90W
Thermal	passive	active	active	passive	passive	passive	bare board
vGPU Software Editions	GRID vPC/vApps, Quadro vDWS	Quadro vDWS	Quadro vDWS	GRID vPC/vApps, Quadro vDWS	GRID vPC/vApps, Quadro vDWS	GRID vPC/vApps, Quadro vDWS	GRID vPC/vApps, Quadro vDWS

PERFORMANCE Optimized	DENSITY Optimized	BLADE Optimized

NVIDIA T4 FOR VIRTUALIZATION

The New Generation of Computer Graphics on a Quadro Virtual Data Center Workstation

- **Virtual Quadro Workstation for the Professional Designer & Data Scientist**:
  - Up to 2X graphics performance versus M60
  - Real-time, interactive rendering
  - NGC support; run deep learning inferencing workloads 25x faster than CPU on a virtual machine

- **Virtual PCs for the Knowledge Worker:**
  - Support for VP9 decode and H.265 encode and decode for improved CPU offload

QUADRO vDWS POSITIONING

Deep learning, rendering,
and GPGPU compute applications

Largest CAD models, CAE,
Photorealistic rendering,
Seismic exploration, GPGPU compute

Large/complex CAD models,
Seismic exploration, complex
DCC effects, 3D Medical Imaging Recon

Large/complex CAD models,
Advanced DCC, Medical Imaging

Medium size/complexity CAD models,
Basic DCC, Medical Imaging, PLM

Small/simple CAD
models, video, Entry
PLM

High-End Quadro vDWS

NVIDIA V100

NVIDIA P40/RTX6000/RTX8000

NVIDIA T4

Entry – Mid Range Quadro vDWS

Office, Sketchup | PACS/Diagnostics | Schlumberger, Halliburton, DeltaGen, Catia Live Rendering

AutoCAD, Revit, Inventor | Ansys, Abaqus, Simulia

Solidworks, Siemens NX, Creo, Catia, ArcGIS Pro

Adobe CC Photoshop, Illustrator | Adobe CC Premiere Pro, After Effects, Autodesk Maya, 3ds Max, Mari, Nuke

RECOMMENDED NVIDIA GPU OPTIONS

Different workflows require different GPUs

Quadro vDWS:

NVIDIA T4 GPUs with **Quadro vDWS** for entry to mid end users provides the most flexible and cost effective solution

NVIDIA P40/V100/RTX6000/RTX8000 GPUs with **Quadro vDWS** provides graphics acceleration for few ultra high end users

GRID vPC:

NVIDIA M10/T4 GPUs with **GRID vPC** enhances user experience while being the most cost effective solution

Quadro vDWS

Intel Xeon Gold 6254
+
NVIDIA T4*
+
Quadro vDWS

GRID vPC

Intel Xeon Gold 6248 or AMD EPYC 7501
+
NVIDIA M10/T4**
+
GRID vPC

GRID vApps

Intel Xeon Gold 6248
+
NVIDIA M10/T4**
+
GRID vApps

** Tested with NVIDIA's Internal VDI Benchmarking tool using SPECviewperf 12.1 on a Dell PowerEdge R740 with 2x Intel Xeon Gold 6154 CPUs*
*** Tested with NVIDIA's Internal VDI Benchmarking tool using the Knowledge Worker workload and comparing End-User Latency and Remoted Frames*

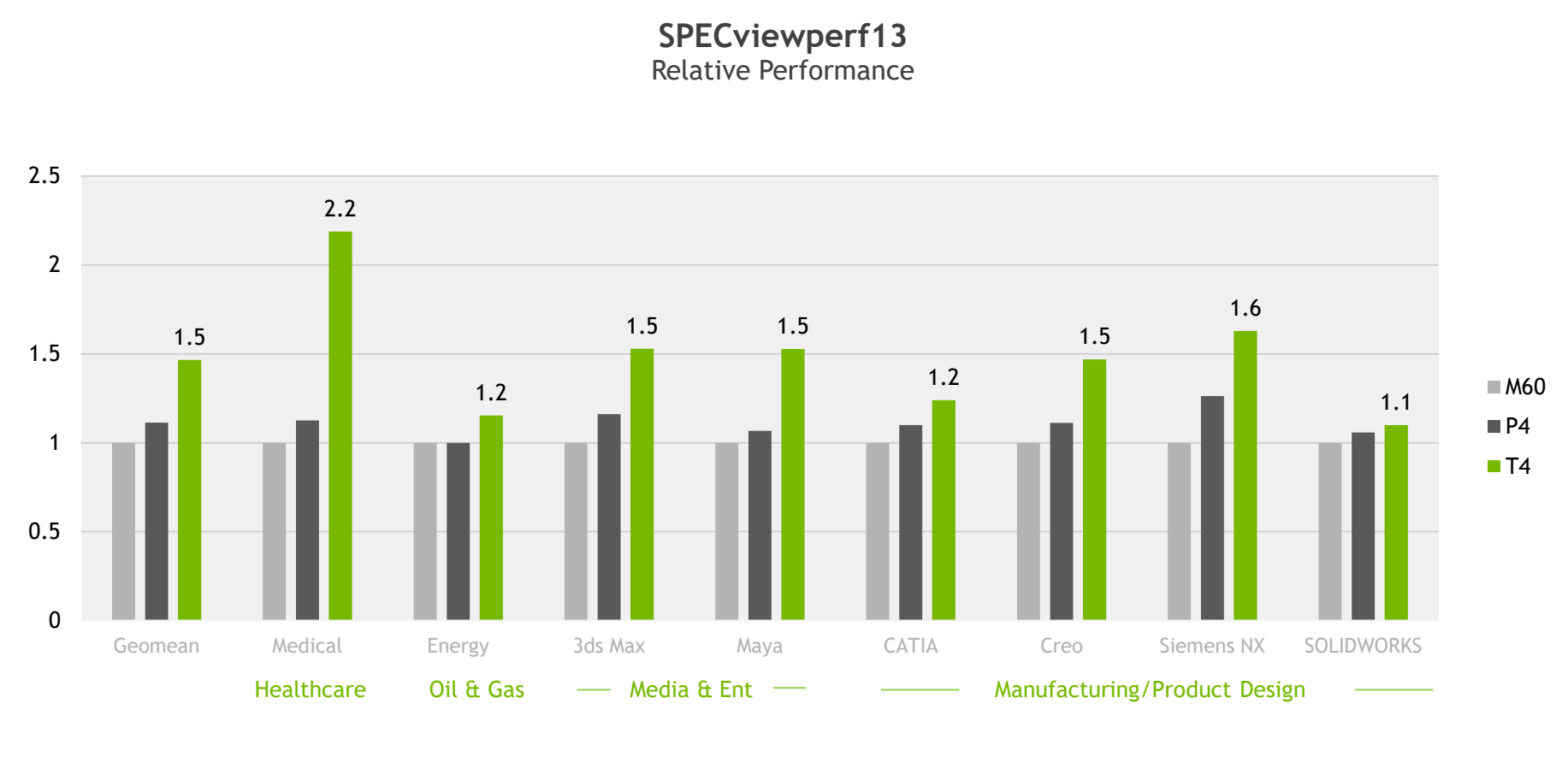
HIGHEST GRAPHICS PERFORMANCE ON A VIRTUAL WORKSTATION

Work Faster with Larger Models

Up to 2X performance compared to M60

2X framebuffer compared to P4 to support larger models

Professional Performance
- ✓ Healthcare
- ✓ Oil & Gas
- ✓ Media & Entertainment
- ✓ Manufacturing



SPECviewperf 13 results tested on a server with Intel Xeon Gold 6154 (18C, 3.0 GHz), Quadro vDWS with T4-16Q, VMware ESXi 6.7, host/guest driver 410.87/412.10, VM config, Windows 10, 8 vCPU, 16GB memory.

Run RTX Applications on a Virtual Workstation

Quadro vDWS with RTX-Capable NVIDIA T4

Run applications built on the **RTX platform**, the most powerful rendering platform, on any device, anywhere

Real-time ray tracing performance

Accelerate **batch rendering** for faster time-to-market

AI-enhanced denoising speeds creative workflows

Photorealistic design with accurate shadows, reflections & refractions

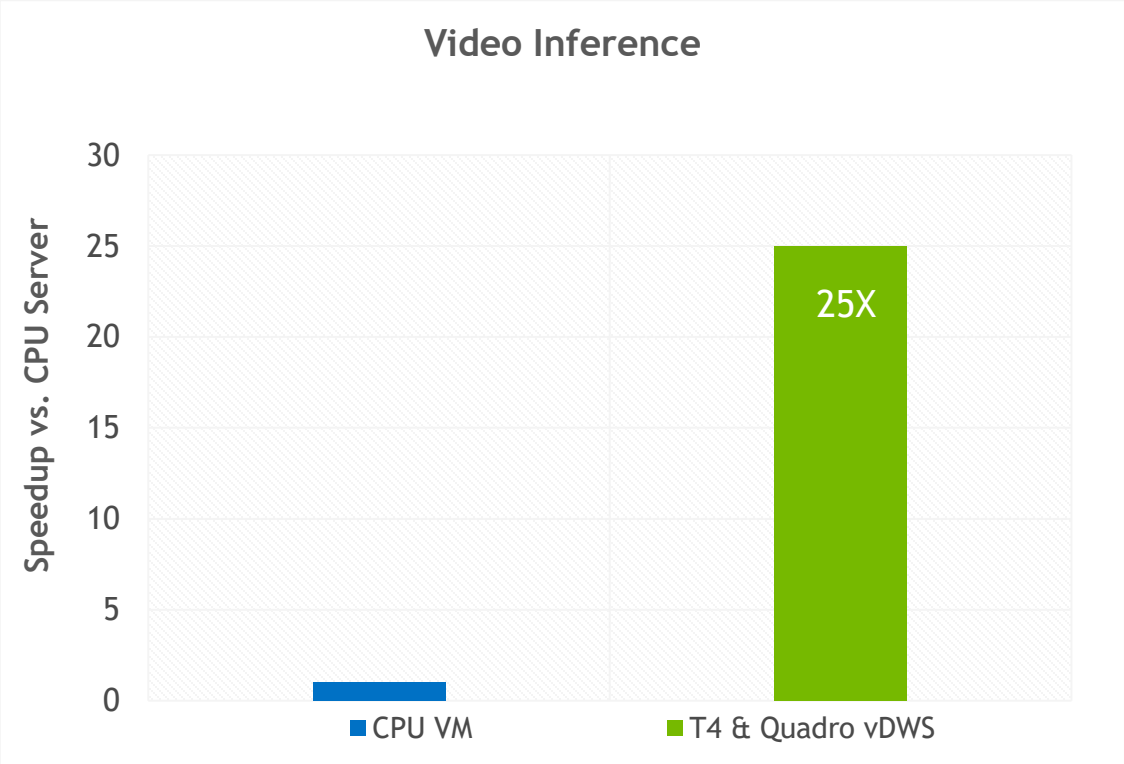
NVIDIA T4 WITH QUADRO vDWS

Real-Time Inference Performance

Quadro Virtual Workstation for deep learning inferencing workloads

Support for NVIDIA GPU Cloud (NGC)

Ideal for deep learning labs and classrooms



Speedup: 25x faster

ResNet-50 (7ms latency limit)

Tested on a server with Intel Xeon Gold 6154 (18C, 3.0 GHz), Quadro vDWS with T4-16Q, VMware ESXi 6.7, host/guest driver 410.87/412.10, VM config, Ubuntu 16.04, 8 vCPU, 32GB memory. 25X performance improvement over CPU VM.

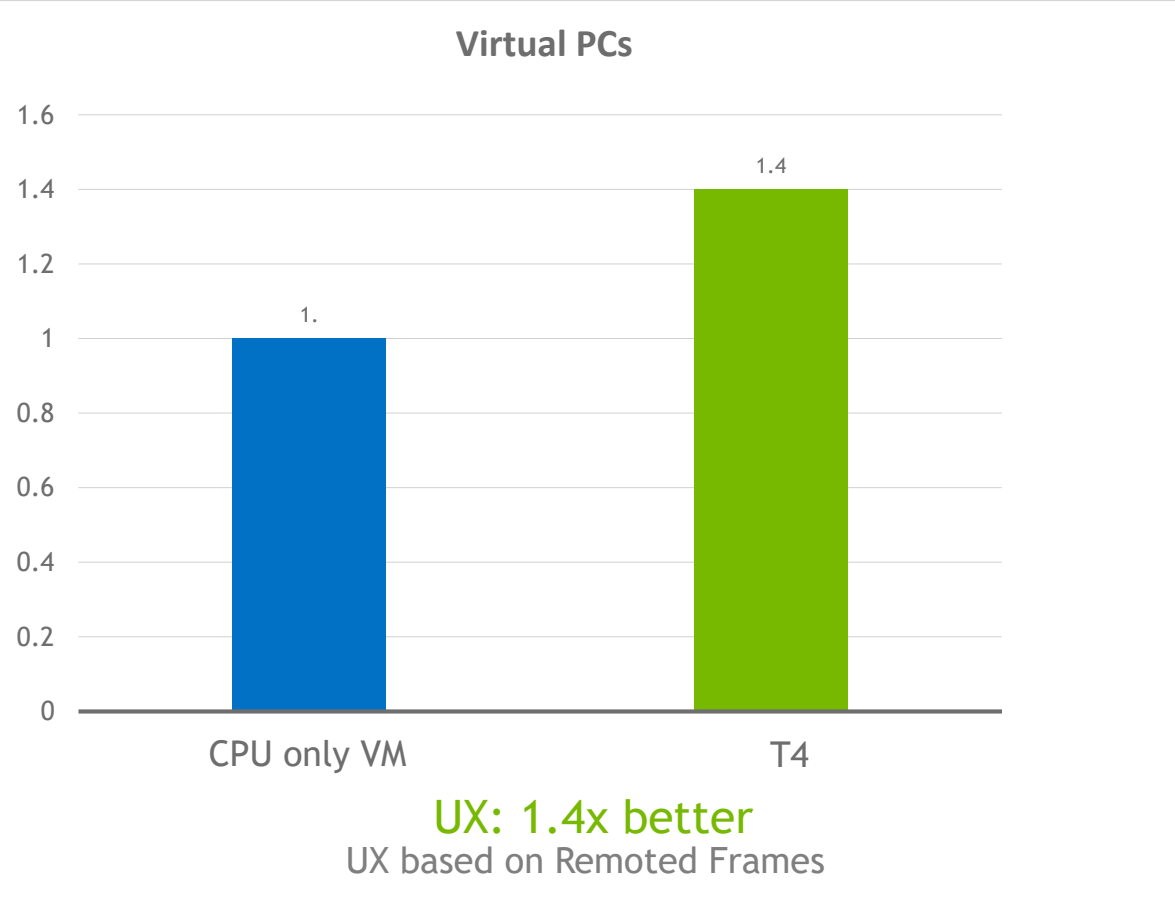
NVIDIA T4 FOR VIRTUAL PCs

Optimize Data Center Utilization with Mixed Workloads

T4 vs. CPU only: Adding NVIDIA GPUs results in 1.4X better user experience versus CPU only VMs**

T4 vs. M10: provides same user density with lower power consumption*

- Same user experience & performance**
- Support for VP9 decode
- Support for H.265 (HEVC) 4:4:4 encode and decode
- Support for >1TB system memory



- Two NVIDIA T4 GPUs support the same user density as a single M10 and fit in the same 2 slot PCIe form factor.

** NVIDIA internal benchmark running Microsoft PowerPoint, Word, Excel, Chrome, PDF viewing and video playback.

SELECTING THE RIGHT GPU

NVIDIA Quadro Virtual Data Center Workstation

Use Case: Entry to Midrange Quadro Workstations

Workloads: CAD, CAE, Digital Content Creation, Rendering, Inferencing, Training

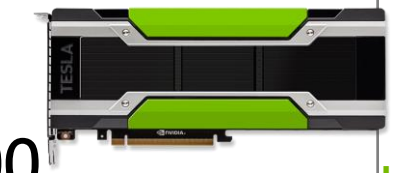
NVIDIA T4

My end users work with larger models or applications

Use Case: High-end Quadro Workstations

Workloads: Large, Complex CAD models, Seismic Exploration, Complex Digital Content Creation, Effects, 3D Medical Imaging

NVIDIA P40
RTX6000/8000



My end users use CAE applications, or are experimenting with DL/AI

Use Case: Ultra High-end Quadro Workstations

Workloads: Largest CAD models, CAE, Seismic Exploration, GPGPU compute, Deep Learning, Immersive Visualization

NVIDIA V100

Smaller Profiles, More Users

Increasing workflow/model complexity

Decreasing user density per server

Larger Profiles, Fewer Users

SELECTING THE RIGHT GPU

NVIDIA GRID vPC/vApps

	2 x NVIDIA T4	1 x NVIDIA M10
Density	32 users	32 users
Form Factor	PCIe 3.0 single slot	PCIe 3.0 dual slot
Power	140W (70W per GPU)	225W
Cores Available	CUDA, Tensor, RT	CUDA
CODECs	VP9, H.265	H.264
System Memory Support	> 1TB	< 1TB
Use Case	Universal GPU for virtual workstations, knowledge workers, rendering, inferencing, training	Lowest TCO for knowledge workers

Resources

NVIDIA GPUs for Virtualization Line Card:
https://www.nvidia.com/content/dam/en-zz/Solutions/design-visualization/solutions/resources/documents1/tesla-gpu-linecard-virtualization-us-nvidia-669786-r7.pdf

Webinar: Introducing NVIDIA T4 for Virtual Workstations:
https://info.nvidia.com/vgpu-vmug-nvidia-T4-reg-page.html

Sizing Guides:

- NVIDIA Quadro Virtual Data Center Workstation Application Sizing Guide for Siemens NX:
- https://images.nvidia.com/content/vGPU/pdf/nvidia-quadro-vdws-application-guide-siemens-nx.pdf

- NVIDIA Quadro Virtual Data Center Workstation Application Sizing Guide for Dassault Systèmes CATIA:
- https://images.nvidia.com/content/vGPU/pdf/nvidia-quadro-vdws-application-guide-catia.pdf

- NVIDIA GRID: Deployment Best Practices for the Digital Workplace Sizing Guide:
- https://www.nvidia.com/object/grid-win10-guide.html