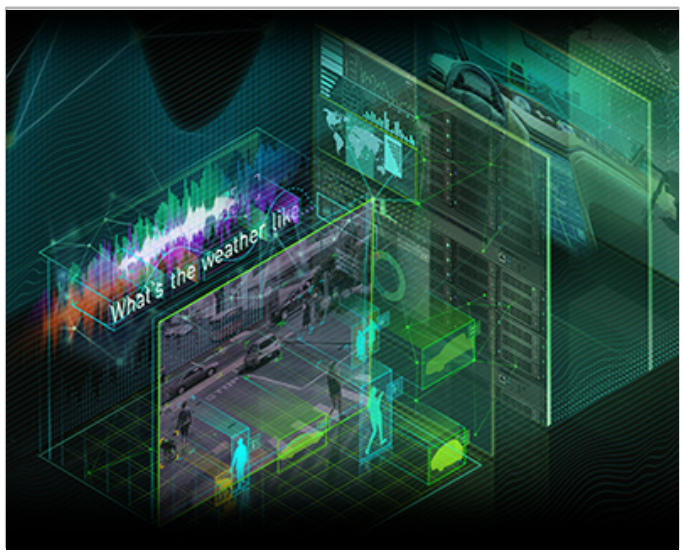
VIRTUAL GPU POWERS AI AND DEEP LEARNING VIRTUAL LABS

Emily Apsey, Performance Engineering, NVIDIA

John Meza - Performance Engineering Team Lead, Esri

Konstantin Cvetanov - Sr. Solution Architect, NVIDIA

AI AND DEEP LEARNING IN UNIVERSITIES

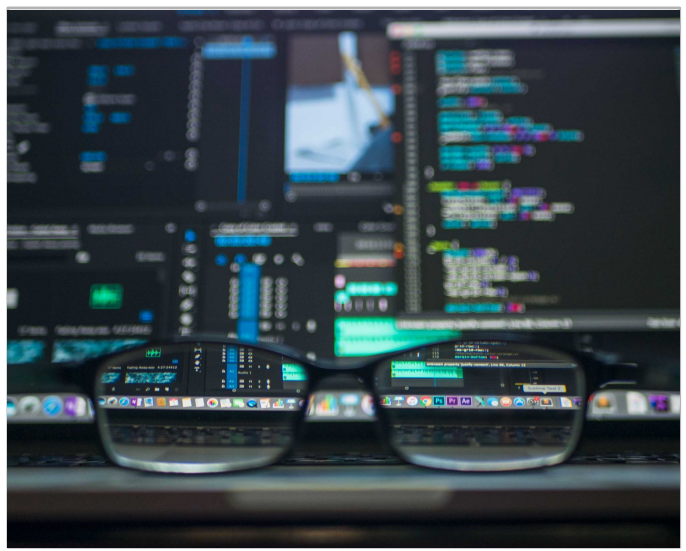


Increasing Demand for AI/DL Classrooms/Labs

Implementation is constrained by cost & availability of physical classroom labs

Need for a Highly Secure, Flexible, Accessible solution

Students require access to labs 24x7 to work on projects and assignments



Robust, Data Scientist Workstations are Expensive

Expensive workstations are not always needed due to smaller data sizes

EDUCATION ECOSYSTEM

General Education Software

Blackboard
ellucian.
Google Classroom™
Office
DRAGON NaturallySpeaking

Engineering

Research, Labs, Academic Disciplines, and Publications

SOLIDWORKS
ptc
cadence
SYNOPSYS
AUTODESK
NX SIEMENS
ANSYS

Architecture & Design

Labs, Theory, Technology and Design, Visual and Studio Art, Media, Design, and Gaming

AUTODESK
Bentley
ANSYS
ALLPLAN A NEMETSCHEK COMPANY
Ai
Ps
Id

Arts & Sciences

Research Centers, Labs, Earth Sciences and Geospatial Institutes

ESRI
Ai
Ps
Id

Medical

Collaborative Research, Statistical Analysis, Business Tools, and Labs

Epic
Cerner
McKESSON

Research

Training Deep Neural Networks, Debugging and Running Experiments

TensorFlow
Caffe
Caffe2
PYTORCH
Keras

Servers

DELL
CISCO
Hewlett Packard Enterprise
Lenovo

Clients

DELL
hp
Lenovo
NVIDIA

WHY CHOOSE NVIDIA T4 FOR VIRTUALIZATION?

DRIVING NEW WORKFLOWS

Empowering the Modern Digital Workplace

Photorealistic Rendering
Increasingly Complex Designs

Data Science
Increase in AI/DL & Inference

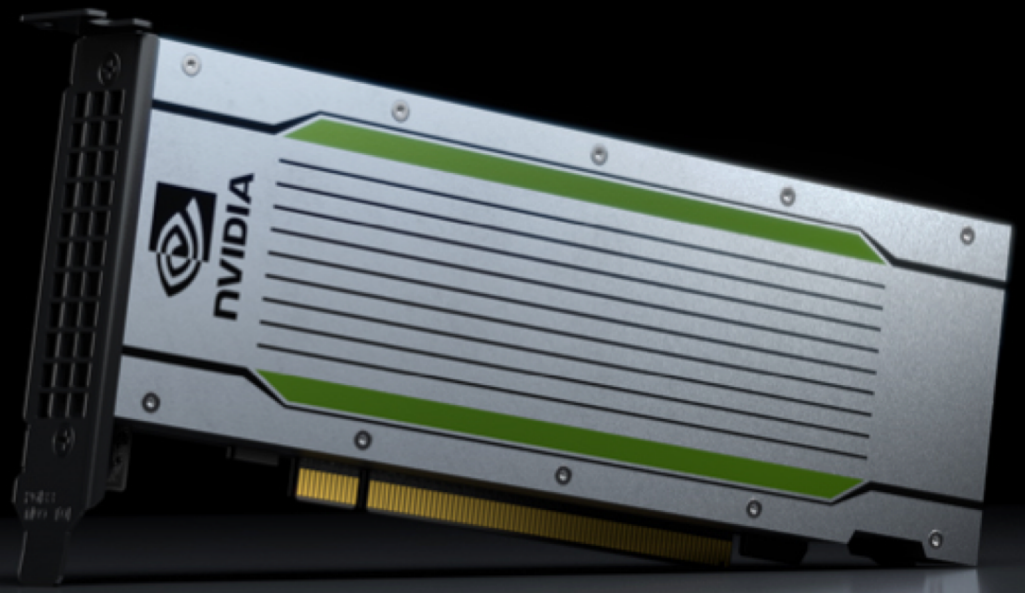
Digital Workplace
Windows 10 & Productivity Apps

ANNOUNCING 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
  - 5 Giga Rays per second for 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



RTX PERFORMANCE IN A QUADRO VIRTUAL WORKSTATION

Support for up to 5 Giga Rays/Sec

Media & Entertainment
Real-time Rendering

Manufacturing
Simulation, modeling, design

Architecture
Rendering, design

NVIDIA T4 KEY SPECIFICATIONS

GPU Architecture	**NVIDIA Turing**
NVIDIA CUDA® Cores	**2,560**
NVIDIA Turing™ Tensor Cores	**320**
RT Cores	**40**
Giga Rays/second	**5**
Memory Size	**16 GB GDDR6**
Memory BW	**Up to 320 GB/s**
vGPU Profiles	**1 GB, 2 GB, 4 GB, 8 GB, 16 GB**
Form Factor	**PCIe 3.0 single slot (half height & length)**
Power	**70W**
Thermal	**Passive**

NVIDIA DATA CENTER GPUs
Recommended for Virtualization

	V100	P40	T4	M10	P6
GPUs / Board (Architecture)	1 (Volta)	1 (Pascal)	1 (Turing)	4 (Maxwell)	1 (Pascal)
CUDA Cores	5,120	3,840	2,560	2,560 (640 per GPU)	2,048
Tensor Cores	640	---	320	---	---
RT Cores	---	---	40	---	---
Memory Size	32 GB/16 GB HBM2	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, 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 (rack servers)	PCIe 3.0 Single Slot (rack servers)	PCIe 3.0 Dual Slot (rack servers)	MXM (blade servers)
Power	250W/300W	250W	70W	225W	90W
Thermal	passive	passive	passive	passive	bare board
	PERFORMANCE Optimized		**DENSITY** Optimized		**BLADE** Optimized

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

NVIDIA P40

NVIDIA V100

High-End Quadro vDWS

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

NVIDIA

NVIDIA T4 PERFORMANCE FOR VIRTUALIZATION WORKLOADS

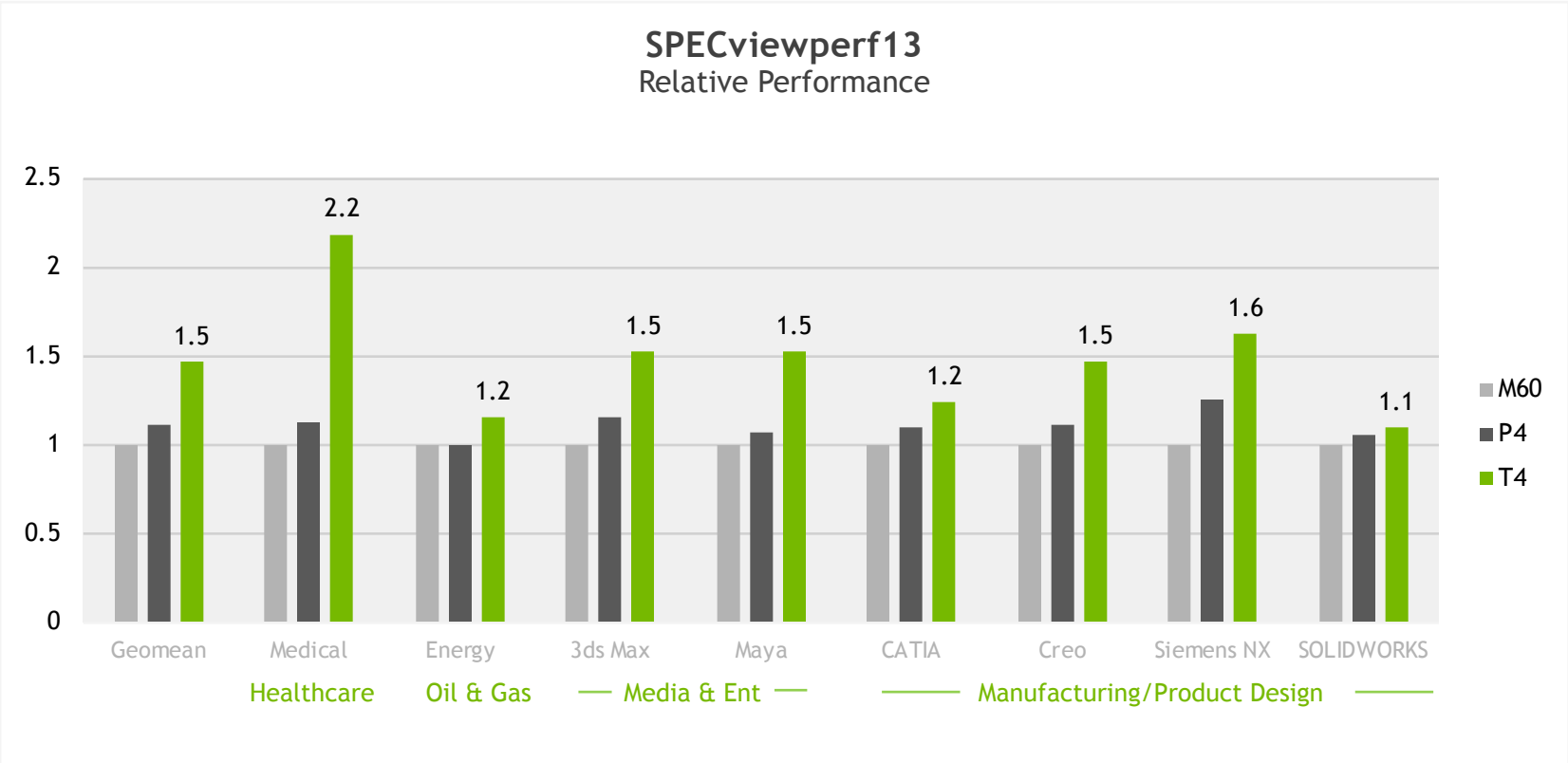
HIGHEST GRAPHICS PERFORMANCE ON A VIRTUAL WORKSTATION

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.

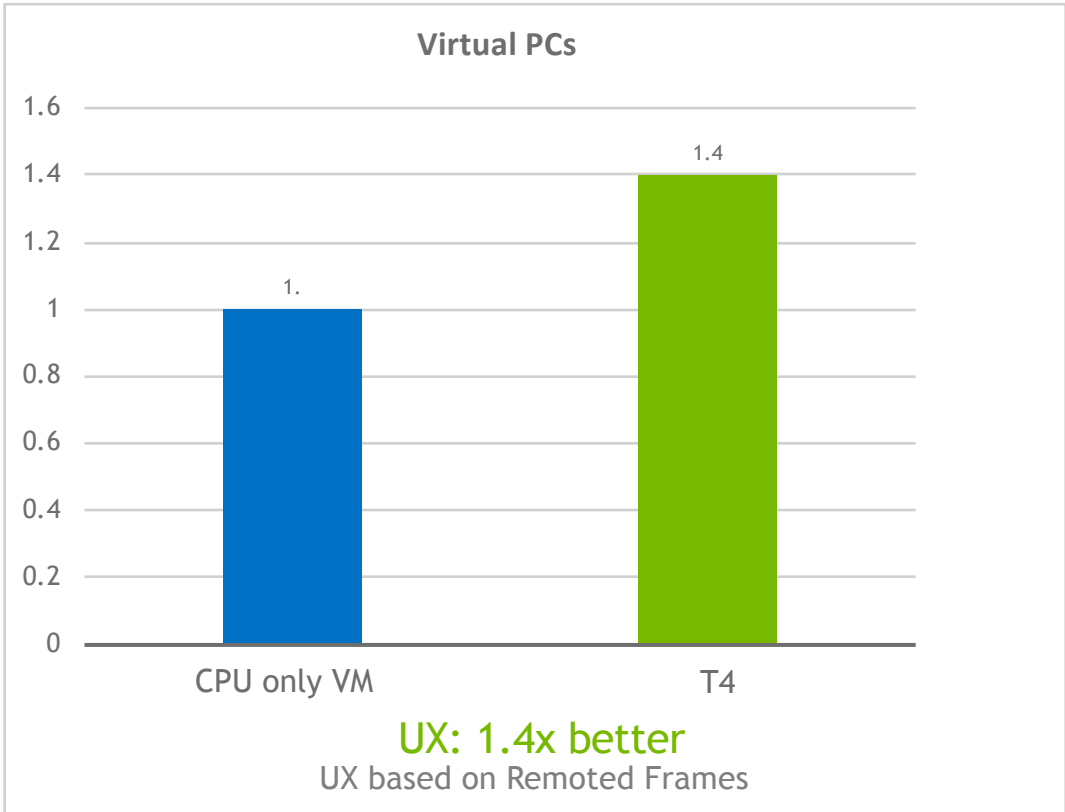
NVIDIA T4 FOR VIRTUAL PCs

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.

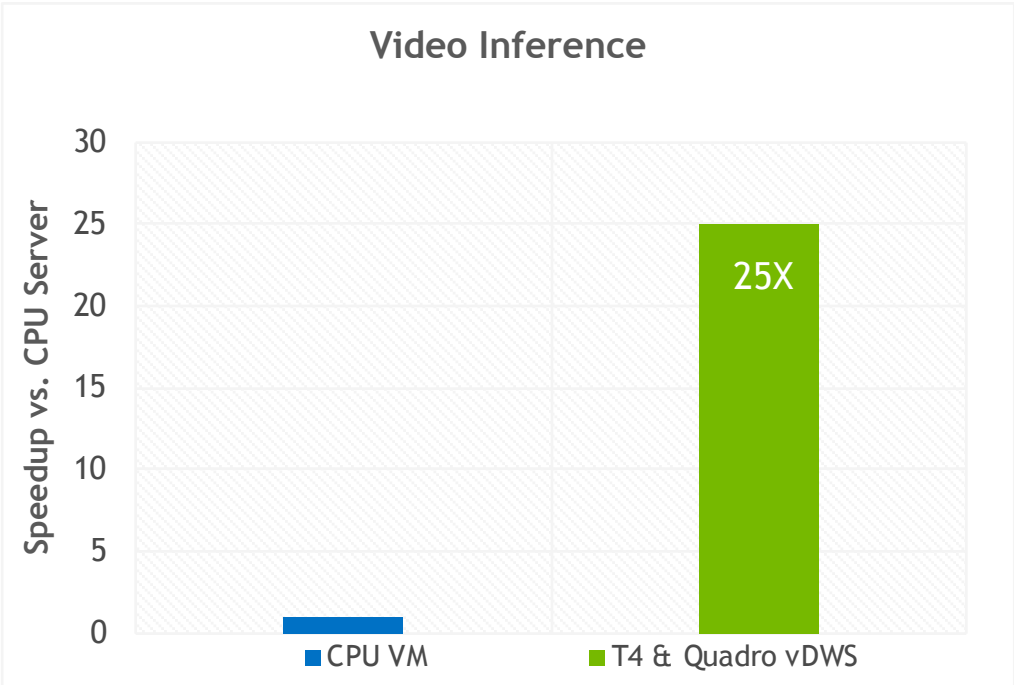
NVIDIA T4 WITH QUADRO vDWS

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.

T4 BENCHMARK ANALYSIS

TensorRT Inferencing

	Avg Images/s		
Tensor RT Inferencing - NGC 18.12	T4-16Q 410.49	T4 BM 410.79	T4 vGPU vs BM
Model:ResNet50 Python: 2 Precision: FP32 bs:1	309	337	-8%
Model:ResNet50 Python: 2 Precision: FP32 bs:2	394	427	-8%
Model:ResNet50 Python: 2 Precision: FP32 bs:4	444	484	-8%
Model:ResNet50 Python: 2 Precision: FP32 bs:8	482	521	-8%
Model:ResNet50 Python: 2 Precision: FP32 bs:16	505	561	-10%
Model:ResNet50 Python: 2 Precision: FP32 bs:32	495	553	-11%
Model:ResNet50 Python: 2 Precision: FP32 bs:64	518	575	-10%
Model:ResNet50 Python: 2 Precision: FP32 bs:128	515	580	-11%
Model:ResNet50 Python: 2 Precision: FP32 bs:7ms Target Latency	380	447	-15%

vGPU 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.

T4 BENCHMARK ANALYSIS

MXNet Training

MXNet - NGC 18.12	T4-16Q 410.49	T4 BM 410.79	% Diff T4 vGPU vs BM
Model: ResNet50 Precision: FP32 bs:32	130	137	-5%
Model: ResNet50 Precision: FP32 bs:48	131	DNA	DNA
Model: ResNet50 Precision: FP32 bs:128	131	139	-6%
Model: ResNet50 Precision: Mixed bs:32	378	DNA	DNA
Model: ResNet50 Precision: Mixed bs:48	390	DNA	DNA
Model: ResNet50 Precision: Mixed bs:64	396	431	-8%

vGPU 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.

MXNET TRAINING

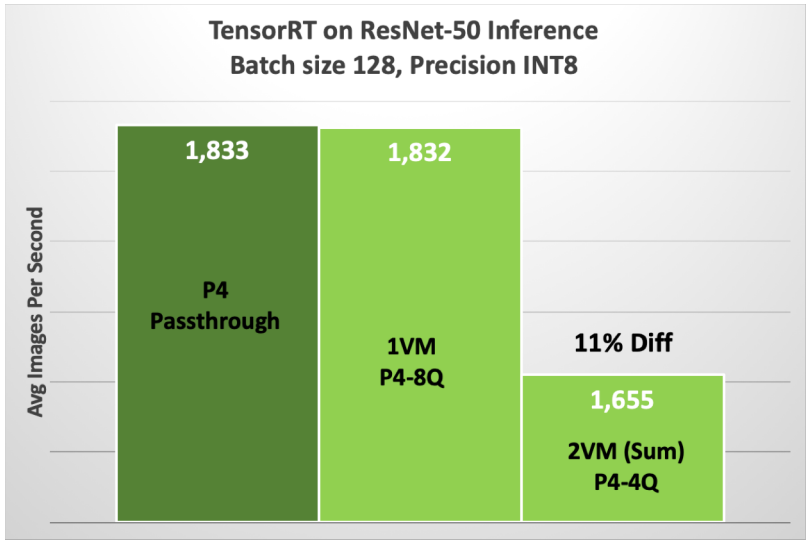
vGPU T4-16Q vs V100-16Q

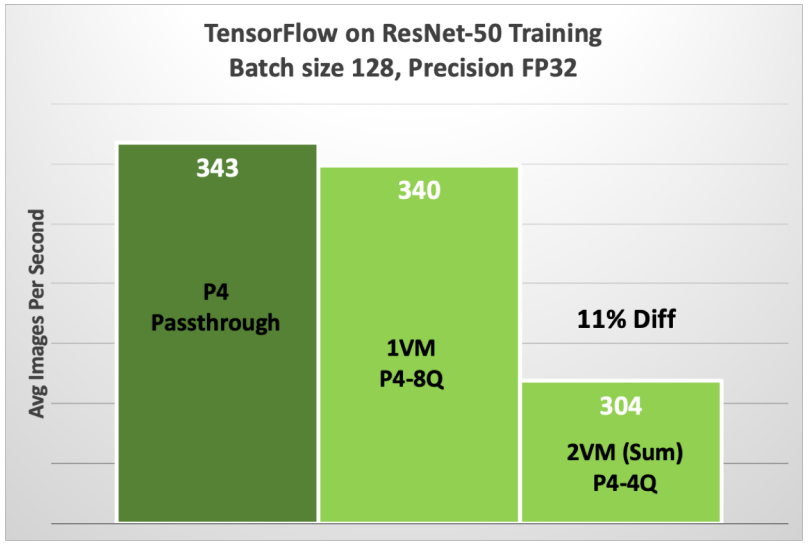
MXNet - NGC 18.12	T4-16Q 410.49	V100-16Q 410.49	% Diff T4 vs V100
Model: ResNet50 Precision: FP32 bs:32	130	363	-62%
Model: ResNet50 Precision: FP32 bs:48	131	390	-66%
Model: ResNet50 Precision: FP32 bs:128	131	402	-67%
Model: ResNet50 Precision: Mixed bs:32	378	985	-62%
Model: ResNet50 Precision: Mixed bs:48	390	1117	-65%
Model: ResNet50 Precision: Mixed bs:64	396	1152	-66%

vGPU 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.

FRACTIONAL PROFILE ANALYSIS

Inferencing and Training





[1] Represents the performance of 2x 4Q VMs (sum across the VMs) against a single 8Q VM and P4 PT

Equal or Fixed Share Scheduler recommended, Best Effort will result in significantly lower performance.

Performance dependent on dataset used (not all tests can run successfully on smaller profile sizes).

Tests ran on Intel Xeon Gold 6140 CPU @ 2.3 GHz, Esxi 6.7, VM Config – Ubuntu 16.04.3, 8vCPU, 16GB, Host/Guest driver: 410.91,410.92, based upon NGC 19.01

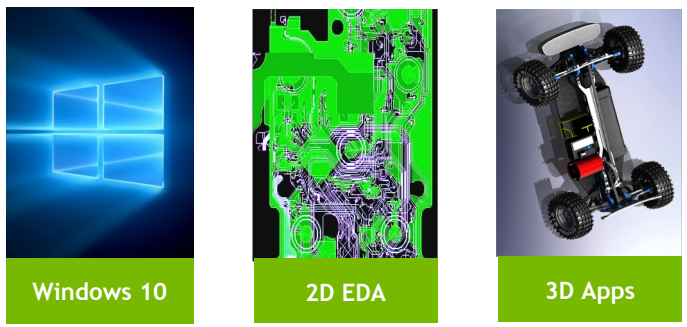
AI LABS IN EDUCATION

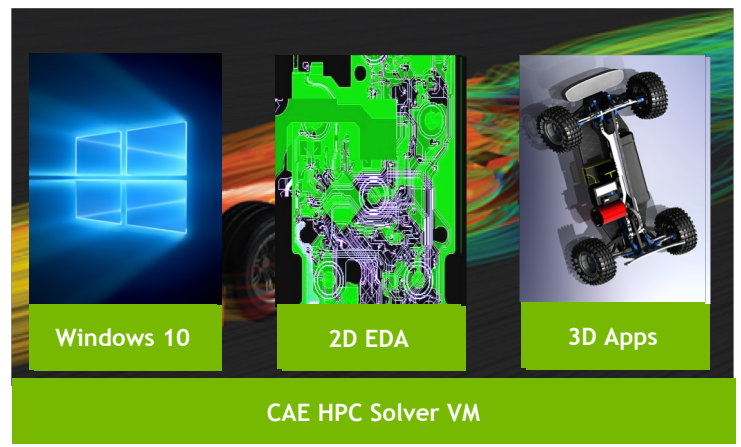
COMMON CHARACTERISTICS

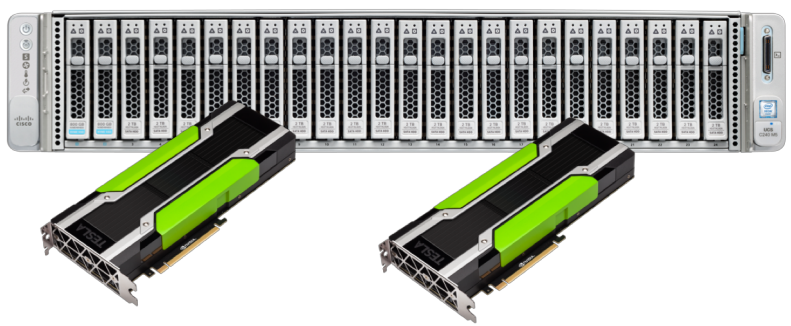
- Universities typically have an HPC cluster that students have ssh (secure shell) lab access to, over the duration of a course like deep learning.
- An HPC cluster would typically have a few servers providing 1:1 bare metal access.
- Students submit jobs into the cluster while debugging code on their laptop.
- Tensorflow is still the dominant framework but Pytorch and Keras are quickly catching up. Caffe and Torch are still used for legacy reasons and they are on a down trend.
- At least 8GB GPU memory is required for training (more is better, loading batches of high resolution images takes a lot of GPU memory).

MIXED WORKLOADS WITH NVIDIA vGPU

Increase productivity & utilization, decrease costs



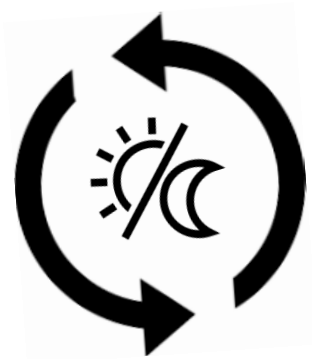




vSphere vMotion with vGPU

End-to-end GPU insights with vROPS

DESKTOP VIRTUALIZATION TECHNOLOGY SUPPLEMENTS HPC CLUSTERS



vGPU Technology Enables Ultimate Flexibility

Debug and run neural network training and inferencing from anywhere using NVIDIA Quadro vDWS for Education License and Tesla Data Center GPUs. Run multiple training and graphics jobs concurrently

HPC Cluster Runs Compute and Data Intensive Workloads

Apply deep learning frameworks to models debugged during the day:

- Finance: modeling and risk exposure
- Life Science: genomics
- Engineering: data analysis, training/inferencing

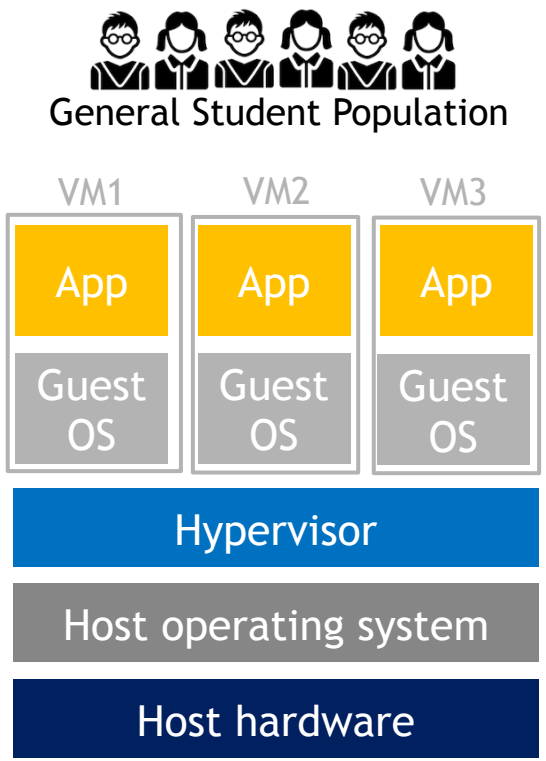
BENEFITS OF vGPU FOR AI

	BENEFIT	TESLA V100 GPU + QUADRO vDWS
Workflow Acceleration	Best Professional Application Performance	Highly tuned drivers for professional applications used for manufacturing design, architecture, energy, medical industries and many others
	Advanced Professional Features	ECC*, double precision FP64 support, multi-GPU support, Tensor cores
	Scalable Performance	Up to 32GB memory with V100
Infrastructure Agility	Right-Sizing Resource Allocation	- De-provision users and modify profile sizes
	Run Mixed Workloads	- Run VDI, HPC and compute workloads on the same infrastructure by using live migration and repurposing the hosts
	Support Other Virtual Labs	-Run learning labs on common infrastructure for other departments (AEC, Engineering) with support for 3D professional applications
IT Management	Ease of IT Management	- Configuration, monitoring and diagnostic tools, including vGPU Live migration - Local and remote access
	Global Support / Warranty	- Enterprise level technical support - Warrantied by NVIDIA
	Extended Product Availability	- Bulk availability - Full product lifecycle management
Enterprise Class Reliability	Mission Critical Drivers	- Long life, stable Drivers - Enterprise level verification by OEMS's comprehensive test suites - Enterprise level release management
	Certified for OEM Workstations	Extensive joint qualifications with major workstation OEMs for enterprise deployment
Security	Enterprise Grade Security	- Ability to sandbox users/ container isolation
Mobility	Virtual Deep Learning Labs	- Deep learning labs can be conducted anywhere
	Anywhere access to applications	- Training can be done on any device, anywhere

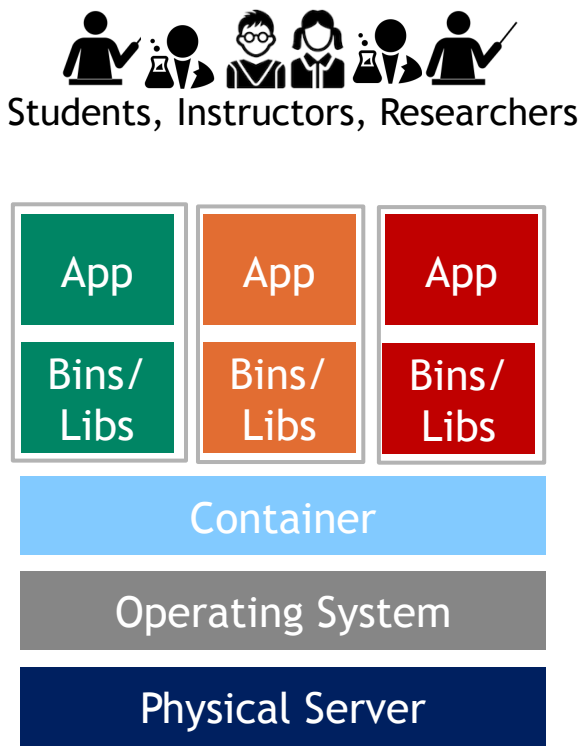
*TBS

GPU VIRTUALIZATION WITH NGC

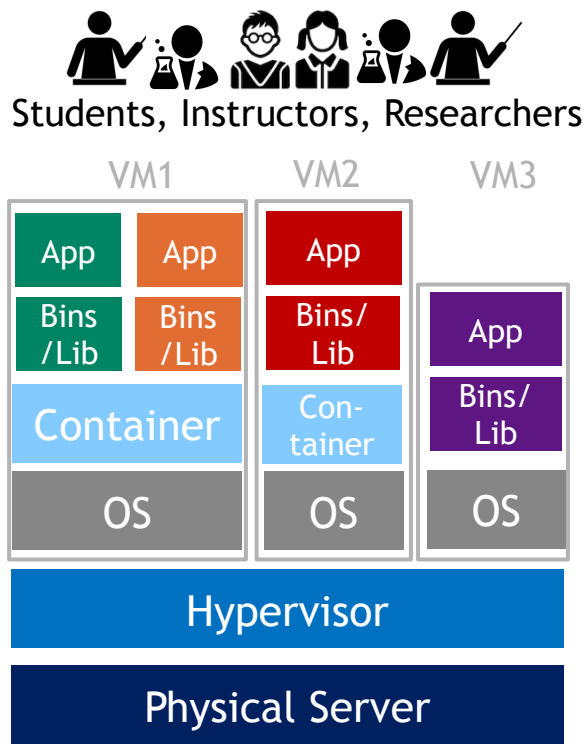
Virtual Machines



Containers



GPU Virtualization + NGC

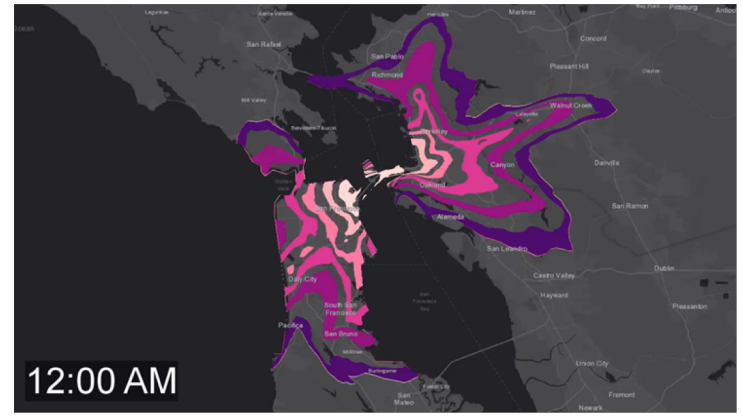


CUSTOMER EXAMPLE – NANJING UNIVERSITY

vForum Breakout Session: VMware云平台加速机器学习 (VMware Cloud Platform Accelerated Machine Learning)

- Using vGPU+NGC(DOCK) solution with 80 licenses
- Adopted vGPU to provide students mobility to study and practice anywhere
- Leveraged NGC to simplify the installation process and avoid DIY risks
- Working with VMware to develop a feature on VMware vRealize Automation to enable students/teachers to apply the GPU/vGPU resource for themselves
- Defined different vGPU profiles depending on user:
    - vGPU with 4/6GB FB (For undergraduate AI teaching)
    - vGPU with 8/12GB FB (For grad students AI teaching/DL beginner)
    - vGPU with 24GB FB (For teachers doing research)
- Uses Caffe/Tensorflow/Pytorch DL framework, with DL model(Lenet/Alexnet) and Dataset(Mnist/Cifar10) for teaching

ESRI DEEP LEARNING IN HIGHER ED WORKSHOP



First One Day workshop at Esri Developer Summit

Delivered completely from the cloud

Each attendee had an entire V100

Deep Learning integration into ArcGIS platform

Next –

Potential Esri Education Service class

ESRI DEEP LEARNING IN HIGHER ED

Universities

- Clemson – Center for Geospatial Technologies
  - Parallel Processing UAV Imagery
- University of California Riverside
- University of California San Diego
  - San Diego Supercomputer Center, Spatial Information Systems Laboratory
- Johns Hopkins
- University of Michigan

GEORGIA STATE UNIVERSITY

- Desire to implement GPU cluster solution for teaching HPC/ML/DL courses to 1200 students
- Existing HPC cluster will continue to operate
- Each Jypiter Notebook connects to a SLURM node on a cluster
- Potential implementation details:
  - CentOS 7 KVM
  - 4-8 GPUs per system (T4 and/or V100)
  - Multiple VMs per GPU today, and multi-GPU with NVLINK in the future
  - Lightweight simulations for molecular dynamics research
- Working on a National Science Foundation proposal for MRI grant

GETTING STARTED

Deploy Virtual labs for AI/DL with a 30% discount on V100 GPUs for educational institutions

NVIDIA Quadro Virtual Data Center Workstation for Education

Get up to 75 percent discount on NVIDIA Quadro® Virtual Data Center Workstation commercial list price, with a single SKU optimized for educational institutions providing all NVIDIA virtual GPU features.

$99 Perpetual License List Price

NVIDIA VIRTUAL GPU RESOURCES

Virtual GPU Test Drive
https://www.nvidia.com/tryvgpu

NVIDIA Virtual GPU Website
www.nvidia.com/virtualgpu

NVIDIA Virtual GPU YouTube Channel
http://tinyurl.com/gridvideos

Questions? Ask on our Forums
https://gridforums.nvidia.com

NVIDIA Virtual GPU on LinkedIn
http://linkd.in/QG4A6u

Follow us on Twitter
@NVIDIAVirt


NVIDIA