MAXIMIZING UTILIZATION OF NVIDIA VGPUS IN VMWARE VSPHERE FOR END-TO-END MACHINE LEARNING

Manvender Rawat, NVIDIA

Uday Kurkure, VMware

3/19/2019

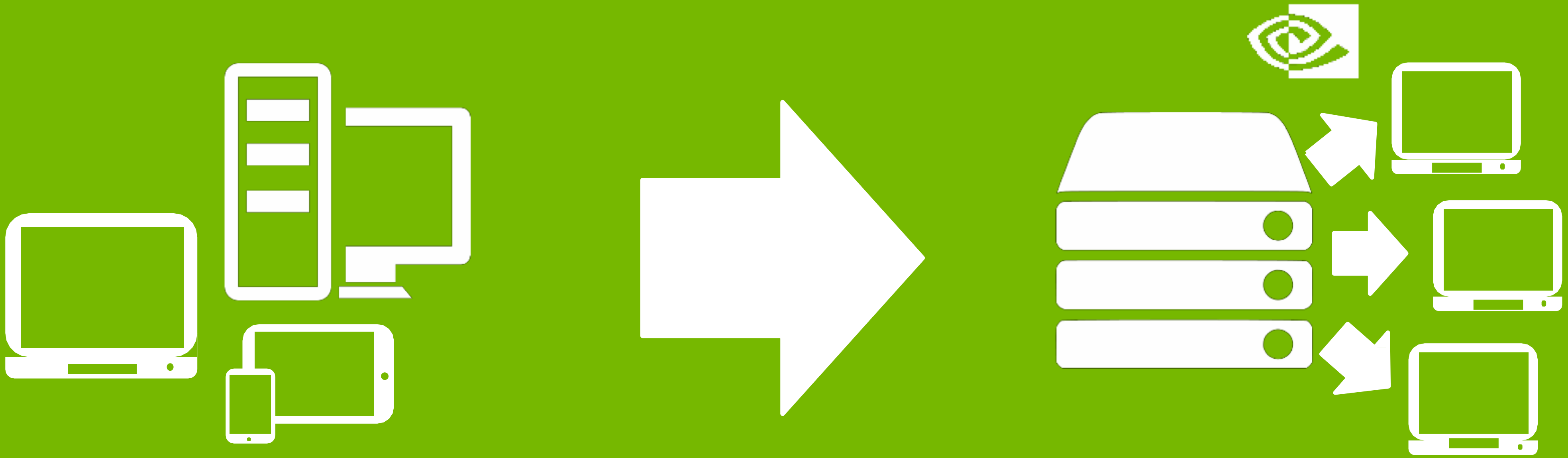
GPU VIRTUALIZATION FOR ANY WORKLOAD

NVIDIA delivers GPU virtualization for both graphics and compute workloads

WHAT IS NVIDIA VIRTUAL GPU TECHNOLOGY?

PERFORMANCE FROM THE DATA CENTER

NVIDIA Virtual GPU technology delivers graphics accelerated virtual desktops and applications

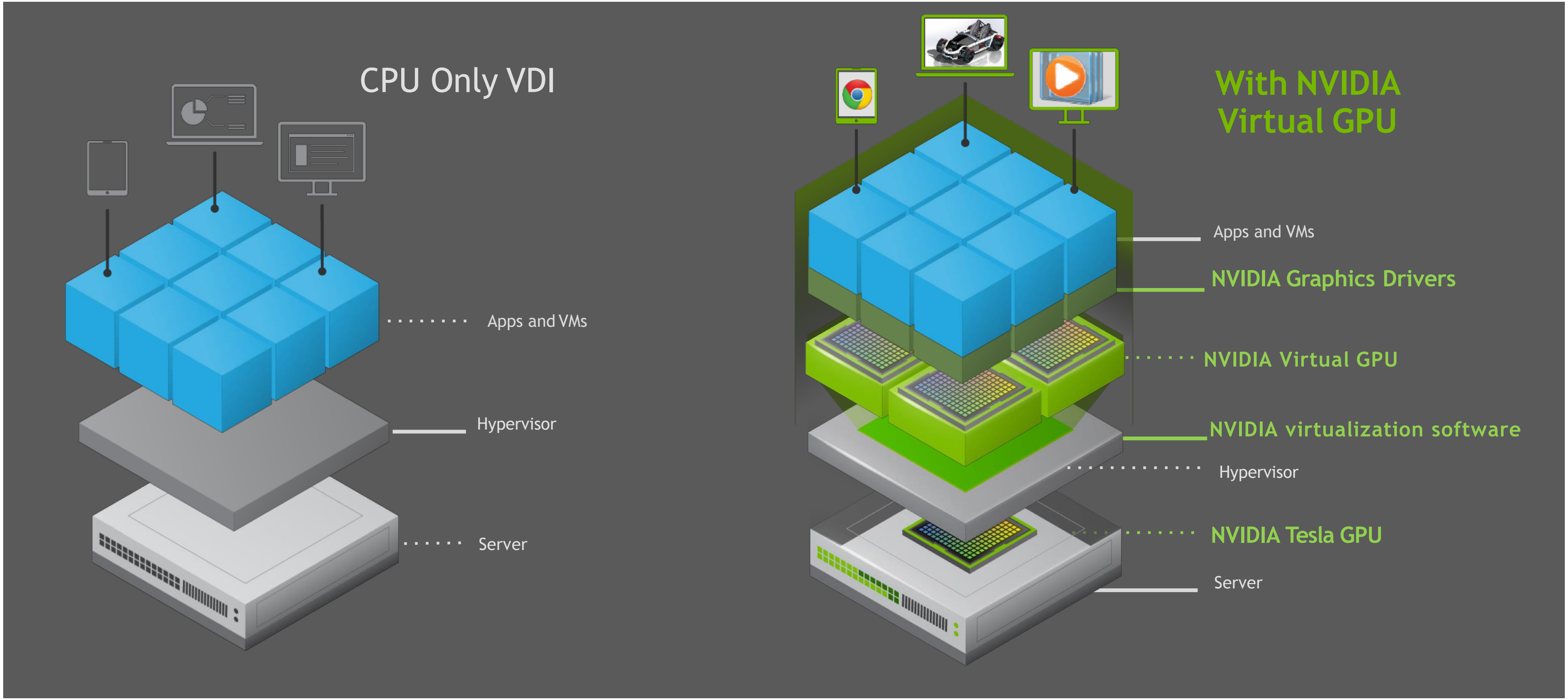


All devices have graphics

Virtual machines also need a GPU

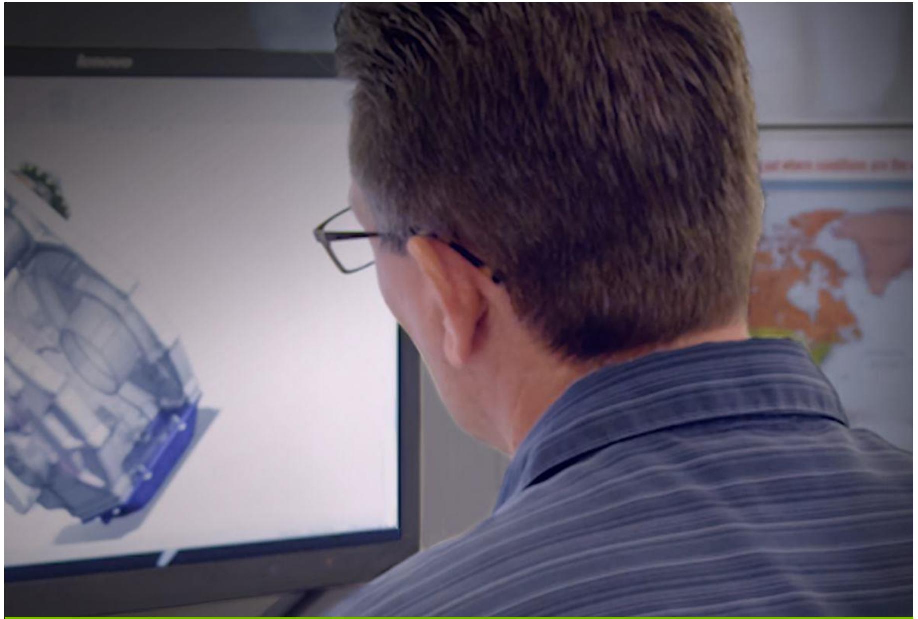
NVIDIA VIRTUAL GPU

GPU Accelerated Experience to every Virtual Desktop with NVIDIA HW and SW



SERVICING THE DIGITAL WORKPLACE

Knowledge/Business Worker



Creative/Technical Professional

ACCELERATE PRODUCTIVITY

For Every User, Any App

Knowledge Worker

Providing business users the highest level of experience for all their apps on any device

Office

Ai

Ps

Id

Creative & Technical Professional

Uncompromised experience for professional graphics users allowing them to design on the go

AUTODESK AUTOCAD

SOLIDWORKS

Petrel

CATIA

AUTODESK REVIT

ESRI ArcGIS

PLM Software SIEMENS NX

AUTODESK MAYA

3DEXCITE

Scalability

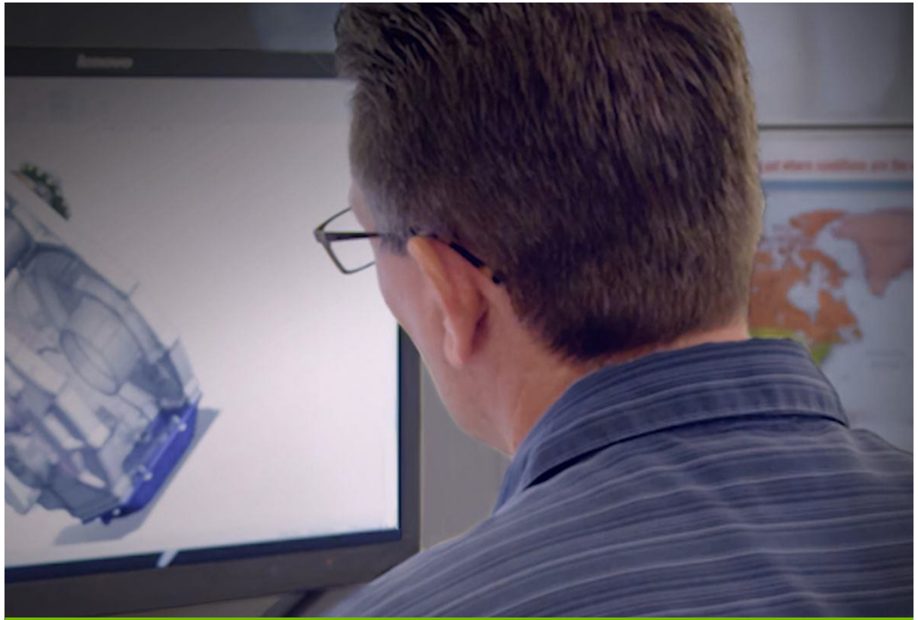
Performance

IMMERSIVE VIRTUAL WORKSPACES

Knowledge/Business Worker

NVIDIA Virtual PC

NVIDIA VirtualApps



Creative/Technical Professional

NVIDIA Quadro Virtual
Data Center Workstation

VIRTUAL GPU SOFTWARE FEATURES

Graphics Features and APIs	Quadro vDWS	vPC	vApps
NVENC	✓	✓	✓
OpenGL Extensions, including WebGL	✓	✓	✓
Quadro Performance Features and Optimizations	✓		
DirectX	✓	✓	✓
Vulkan support	✓		

Profiles	Quadro vDWS	vPC	vApps
Max Frame Buffer Supported	32 GB	2 GB	24 GB
Available Profiles	0Q, 1Q, 2Q, 3Q, 4Q, 6Q, 8Q, 12Q, 16Q, 24Q, 32Q	0B, 1B, 2B	24A, 16A, 12A, 8A, 6A, 4A, 3A, 2A, 1A

Advanced Professional Features	Quadro vDWS	vPC	vApps
ISV Certifications	✓		
CUDA/OpenCL	✓		✓

Data Center Management	Quadro vDWS	vPC	vApps
Host, Guest, and Application Level Monitoring	✓	✓	✓
Live Migration	✓	✓	✓
Multi-vGPU support	✓		

Display	Quadro vDWS	vPC	vApps
Maximum Hardware Rendered Display	Four 4K	Four QHD, Two 4K	One
Maximum Resolution	4096 x 2160	4096 x 2160	1280 x 1024

See the full details at www.nvidia.com/vGPU

Schedulers Compared

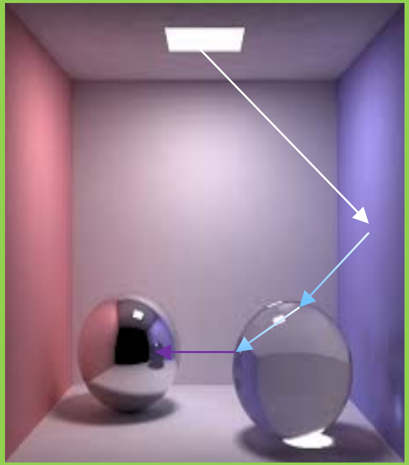
	Best Effort Scheduler	Equal Share / Fixed Share Scheduler
Deterministic GPU Cycles per VM	No	Yes
Deterministic QoS	No	Yes
Aware of vGPU that originated task	No	Yes
Potential of Noisy Neighbor Impact	Yes	No
FRL Required	Yes	No
HW Support	Maxwell, Pascal, Volta, Turing	Pascal, Volta, Turing

NVIDIA T4 FOR UNIVERSAL WORKLOADS

NVIDIA TURING: GRAPHICS REINVENTED

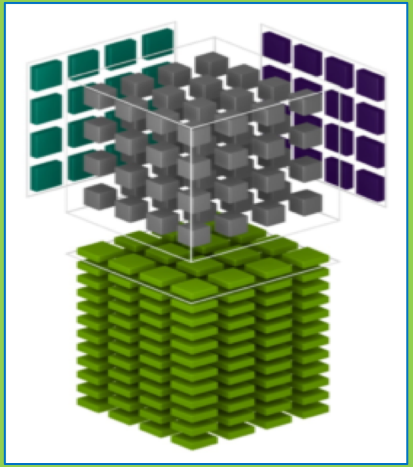
The Fusion of Ray Tracing, Deep Learning, Advanced Shading

ACCELERATED RAY TRACING



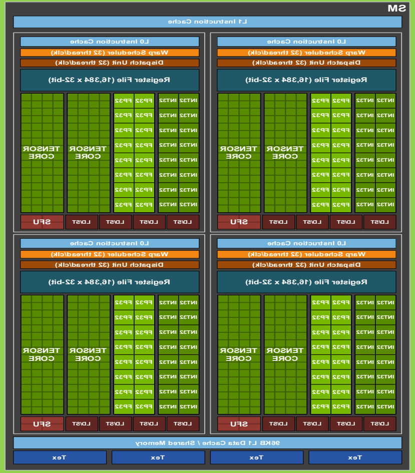
RT CORES

ENHANCED WITH DEEP LEARNING



TENSOR CORES

ADVANCED PROGRAMMABLE SHADING


SM
L1 Instruction Cache
L0 Instruction Cache
Warp Scheduler (32 thread/clk)
Dispatch Unit (32 thread/clk)
Register File (16,384 x 32-bit)
TENSOR CORE
FP32
INT32
SFU
LD/ST
96KB L1 Data Cache / Shared Memory
Tex

STREAMING MULTIPROCESSOR

NVIDIA T4 KEY SPECIFICATIONS

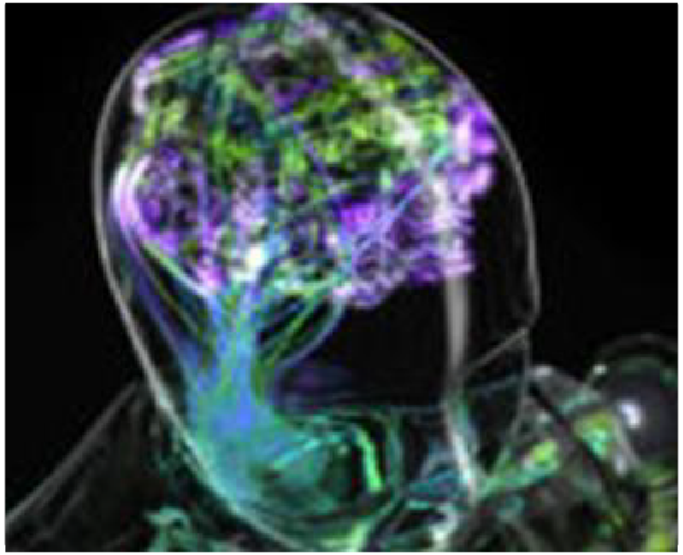
GPU Architecture	**NVIDIA Turing**
NVIDIA CUDA® Cores	**2,560**
NVIDIA Turing™ Tensor Cores	**320**
RT Cores	**40**
Giga Rays/second	**5**
Memory Size	**16 GB GDDR6**
Memory BW	**Up to 320 GB/s**
vGPU Profiles	**1 GB, 2 GB, 4 GB, 8 GB, 16 GB**
Form Factor	**PCIe 3.0 single slot (half height & length)**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

DRIVING NEW WORKFLOWS

Empowering the Modern Digital Workplace

Photorealistic Rendering

Increasingly Complex Designs



Data Science

Increase in AI/DL & Inference

Digital Workplace

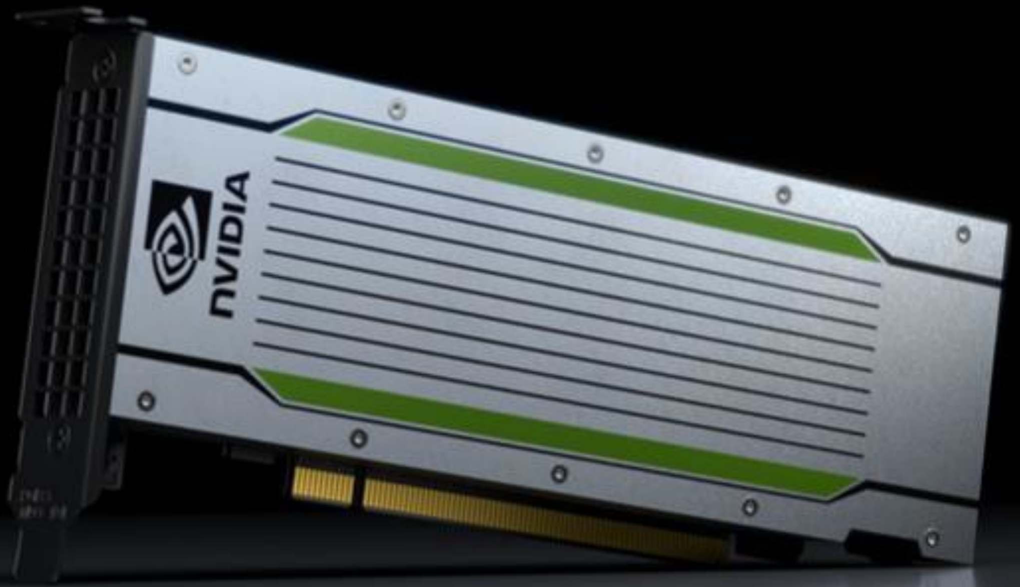
Windows 10 & Productivity Apps

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



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 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

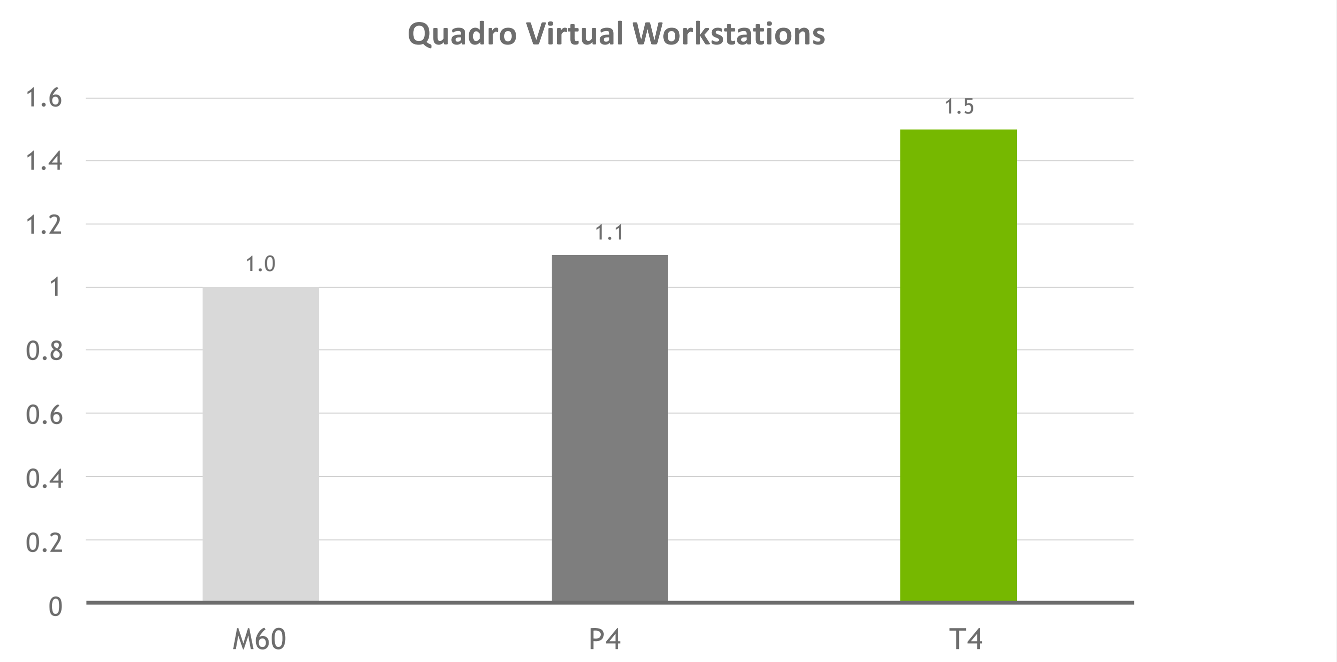
NVIDIA T4 PERFORMANCE FOR VIRTUALIZATION WORKLOADS

LATEST GENERATION QUADRO VIRTUAL WORKSTATION

Work Faster with Larger Models

Continued performance increases with latest generation GPUs

Added AI support and ray tracing support with Tensor and RT cores



3D Graphics: 1.5x performance

SPECviewperf13

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.

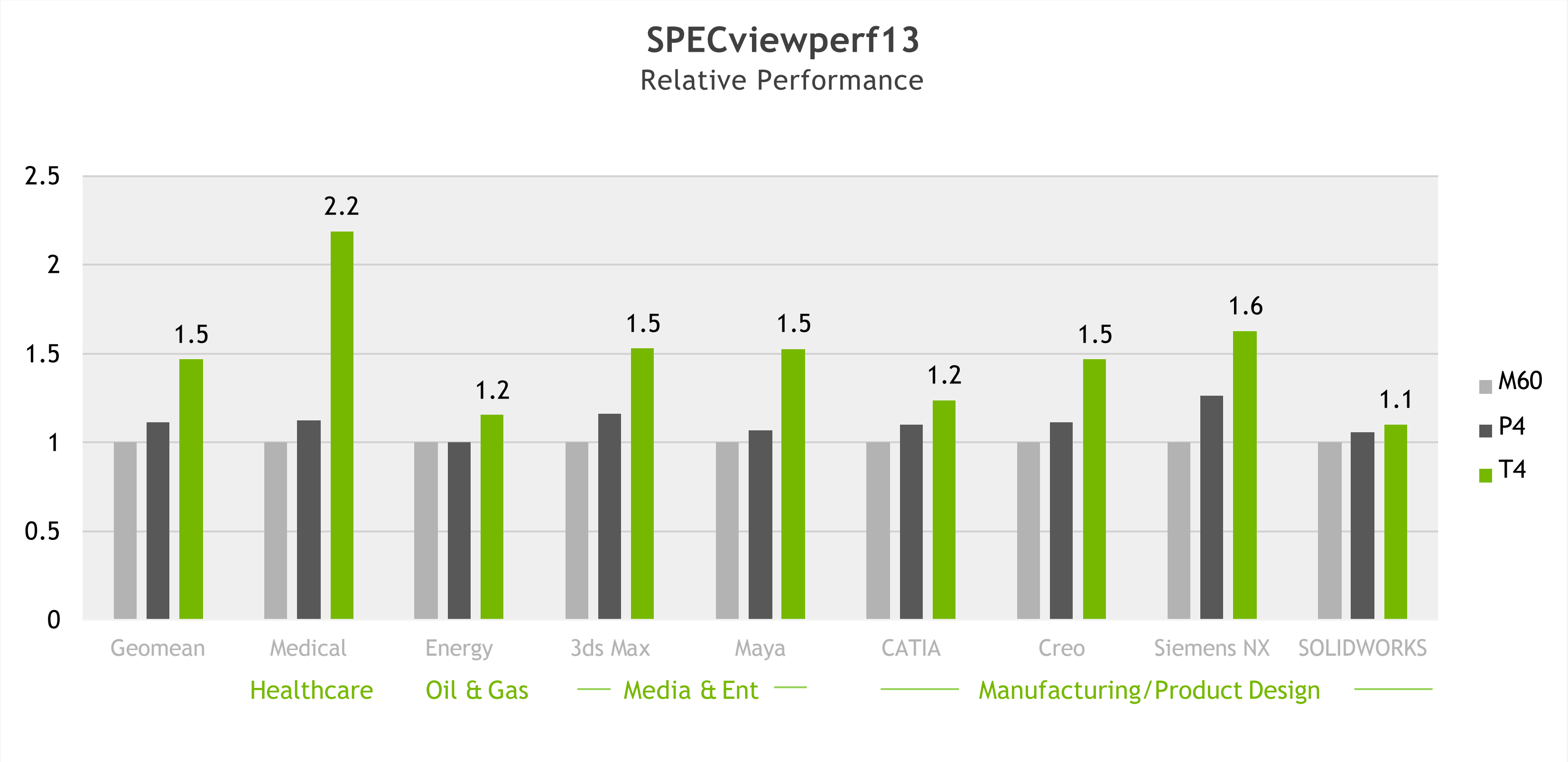
HIGHEST GRAPHICS PERFORMANCE ON A VIRTUAL WORKSTATION

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.

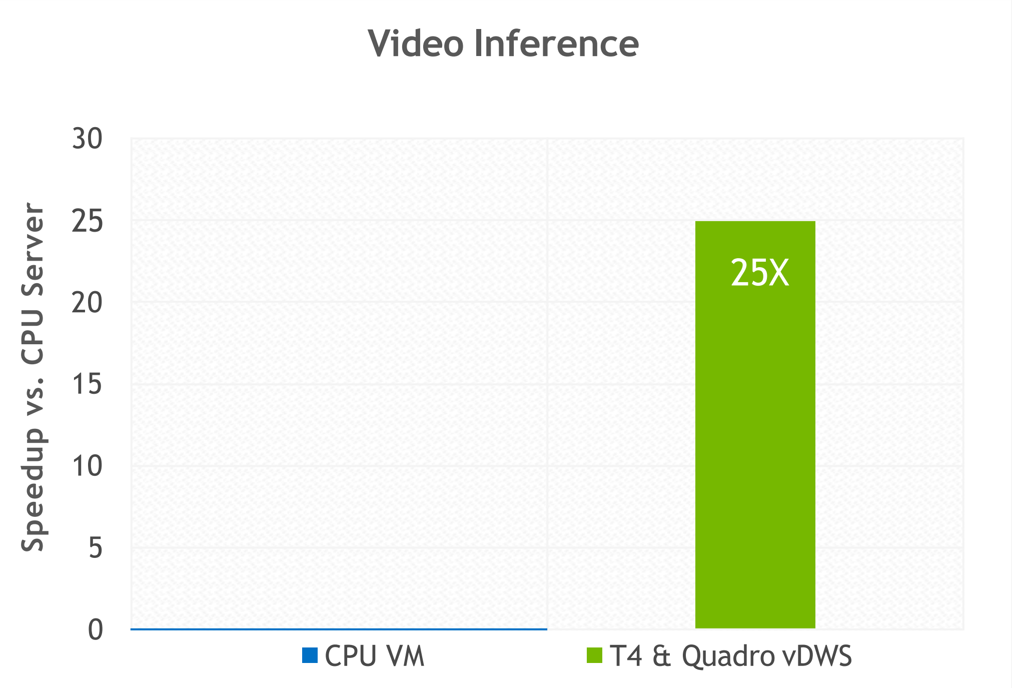
NVIDIA T4 WITH QUADRO vDWS

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**

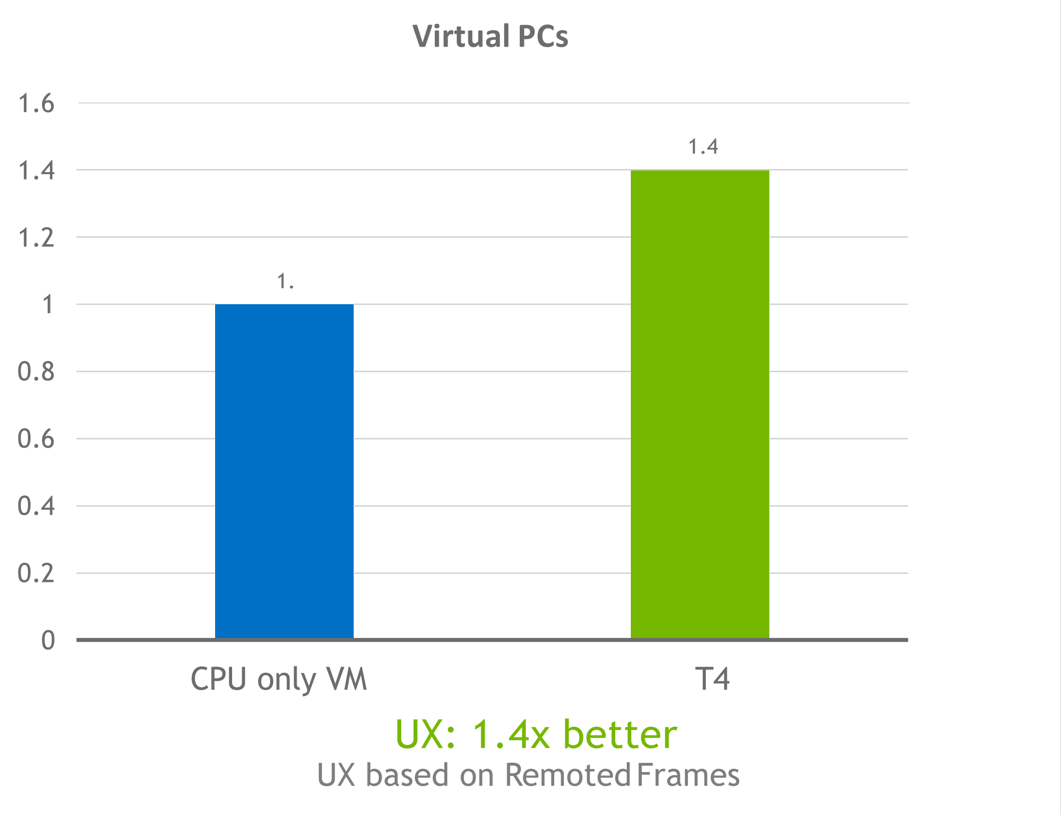
T4 vs. M10: provides same user density with lower power consumption*

Same user experience & performance**

Support for VP9 decode

Support for H.265 (HEVC) 4:4:4 encode and decode

Support for >1TB system memory

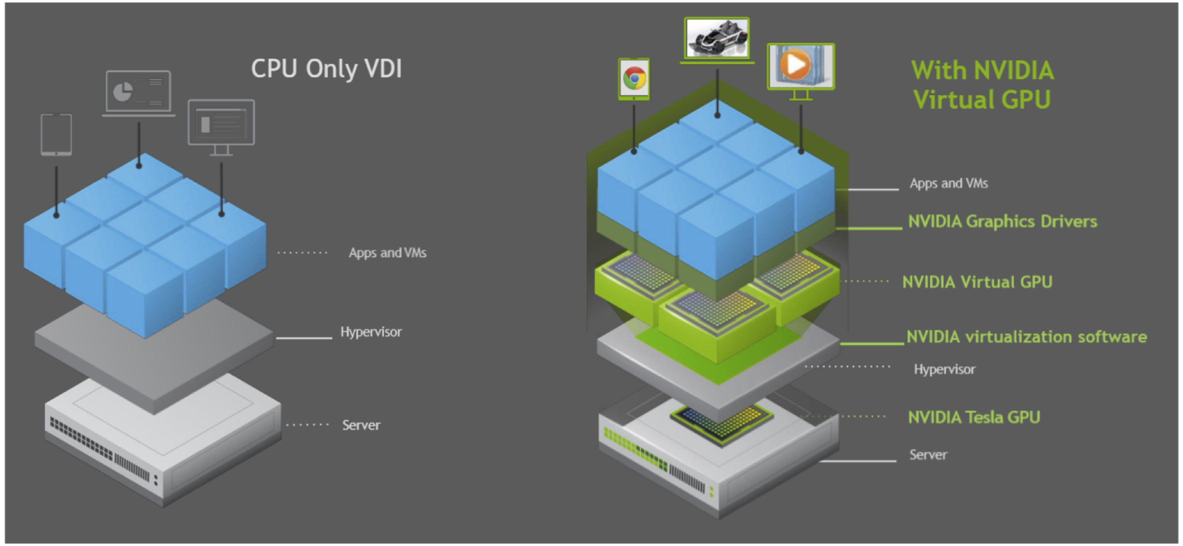


- Two NVIDIA T4 GPUs support the same user density as a single M10 and fit in the same 2 slot PCIe form factor.

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

NVIDIA VIRTUAL GPU

GPU Accelerated Experience to every Virtual Desktop with NVIDIA HW and SW



- NVIDIA QvDWS delivers best in class performance for compute and Graphics workload

- NVIDIA T4 is uniquely suited to deliver universal workloads

- The performance at scale and manageability for virtualized deployments made possible by vGPU architecture

NVIDIA VGPUS IN VMWARE VSPHERE FOR END-TO-END MACHINE LEARNING

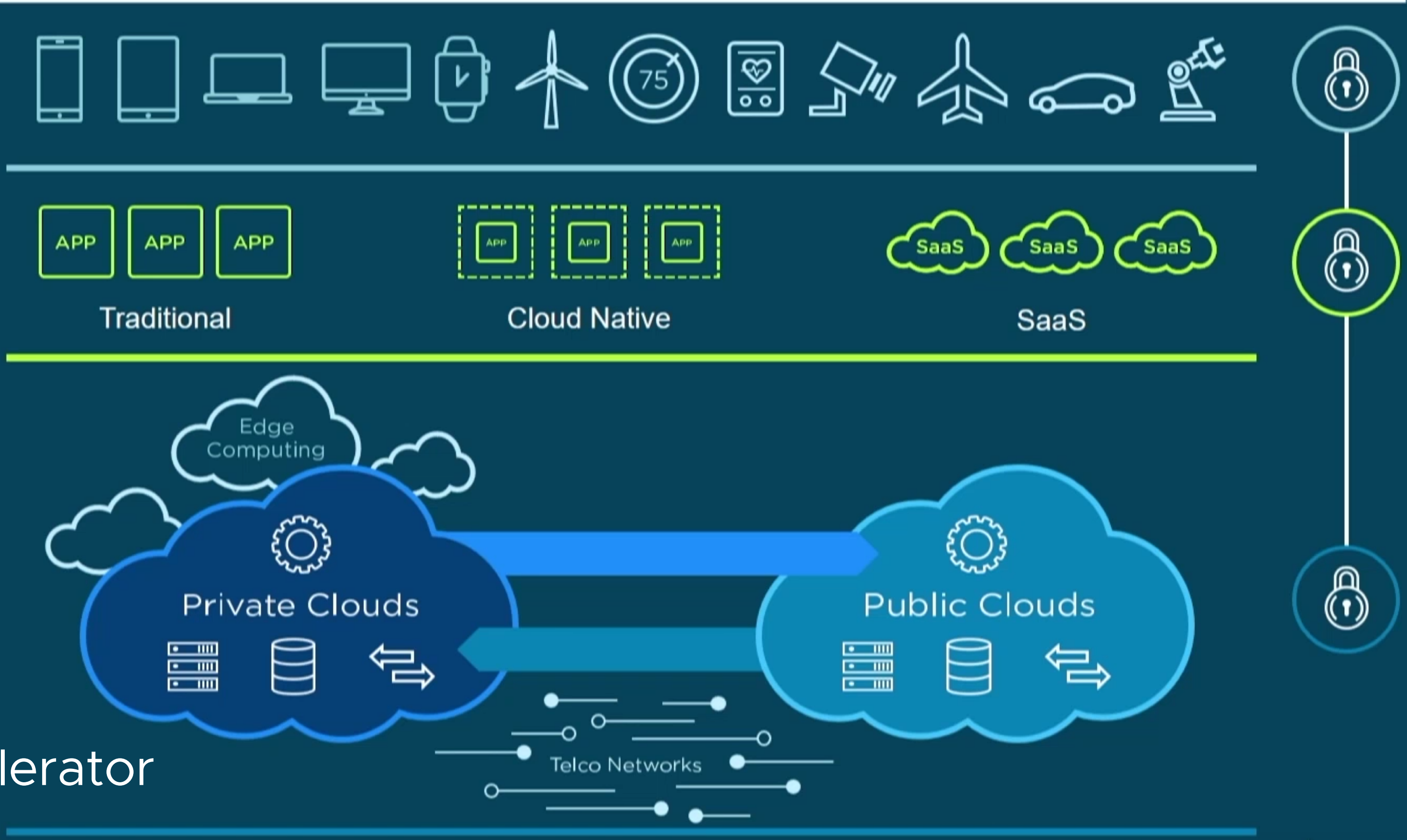
Our Vision

A Digital Foundation Built on VMware

Any Device

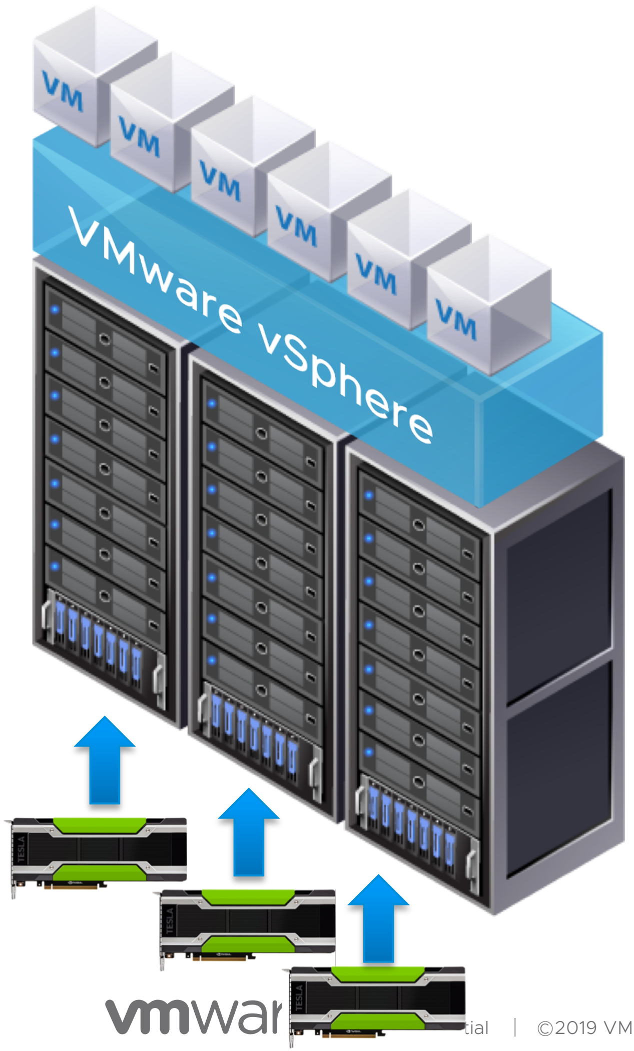
Any Application

Any Cloud



Any GPU/Accelerator

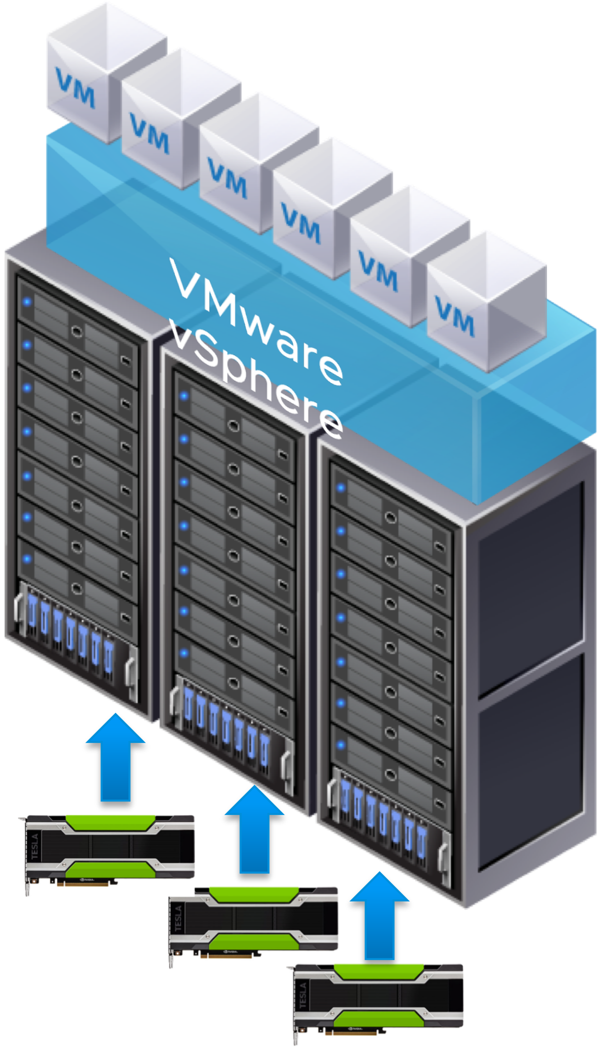
VMware vSphere with Nvidia GPUs



Our customers are using GPUs on VMware vSphere

1. **Accelerating 2D/3D Graphics** workloads for VMware Horizon
2. **Enabling VMware Blast Extreme** protocol
   - Encoding / Decoding H.264 and H.265 Based
3. **General Purpose GPU (GPGPU)**
   - Machine learning / Deep Learning
   - High performance computing workloads

Benefits of vGPUs in VMware vSphere



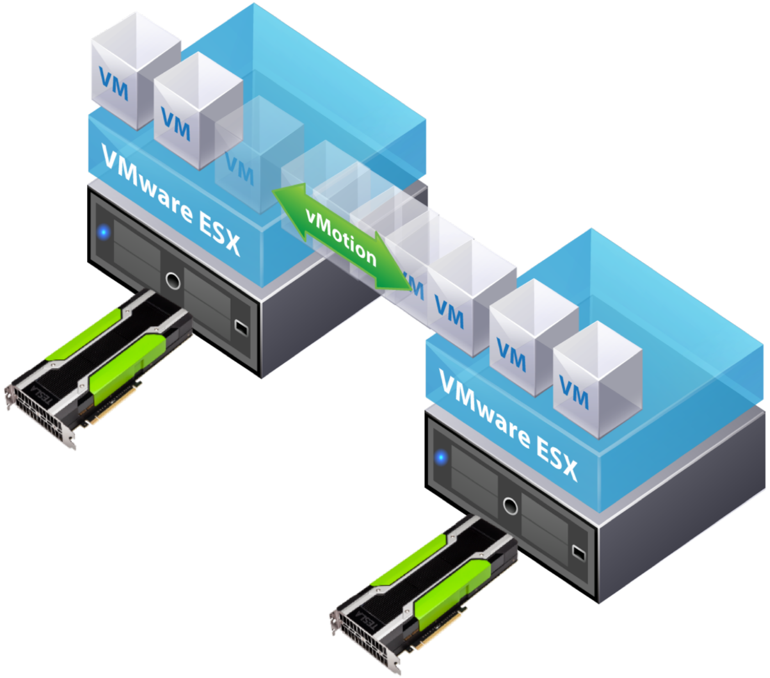
Virtualization Technology efficiently manages servers in the data centers

- Enables Diverse Workloads
  - Windows and Linux VMs running on the same host
- Higher Consolidation Ratios
- Suspend/Resume of Virtualized GPU enabled VMs
  - ML Training at night
  - Interactive CAD jobs during the day
- vMotion of vGPU VMs
  - ML Training or HPC jobs can take days
  - Before the server maintenance, vMotion the VMs to another host and then move them back after the maintenance. Thus, saving days of work
- Combine the Power of GPUs with Management Benefits of Virtualization

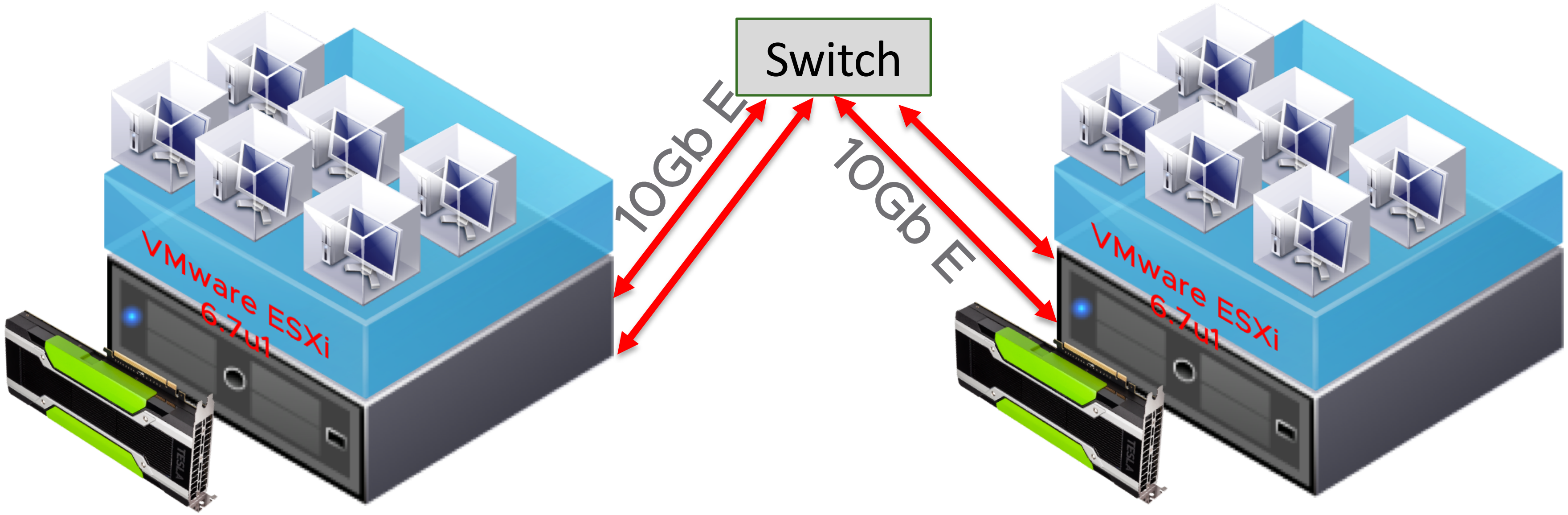
A Typical Customer Scenario

- Leverage GPU investment across different use cases
  - ML Workloads on Linux for Data Scientist/ML researchers
  - Virtual Desktop Infrastructure (VDI) for Office Workers on Windows
  - 3-D CAD Workloads on Windows and Linux for Scientists
  - Simulations on Linux
  - End Users in Different Time Zones using GPUs at different times
  - Improve Data Center Resource Utilization Using vGPUs in Data Centers

Virtualization Performance: vMotion for vGPUs enabled VMs



vMotion for NVIDIA vGPU – Test-bed

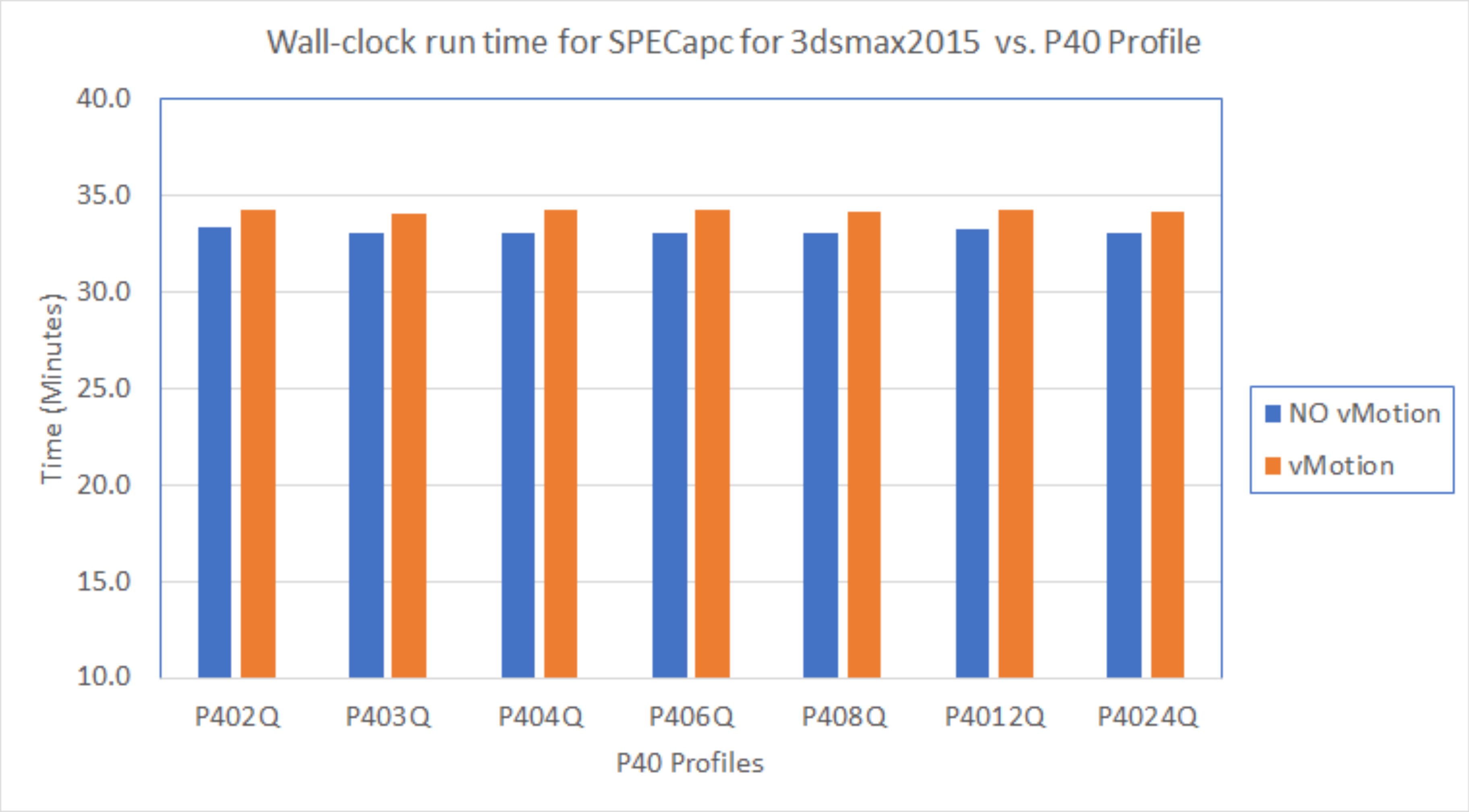


Dell R730 – Intel Broadwell CPUs + 1 x NVidia P40
40 cores (2 x 20-core socket) E5-2698 v4
768 GB RAM

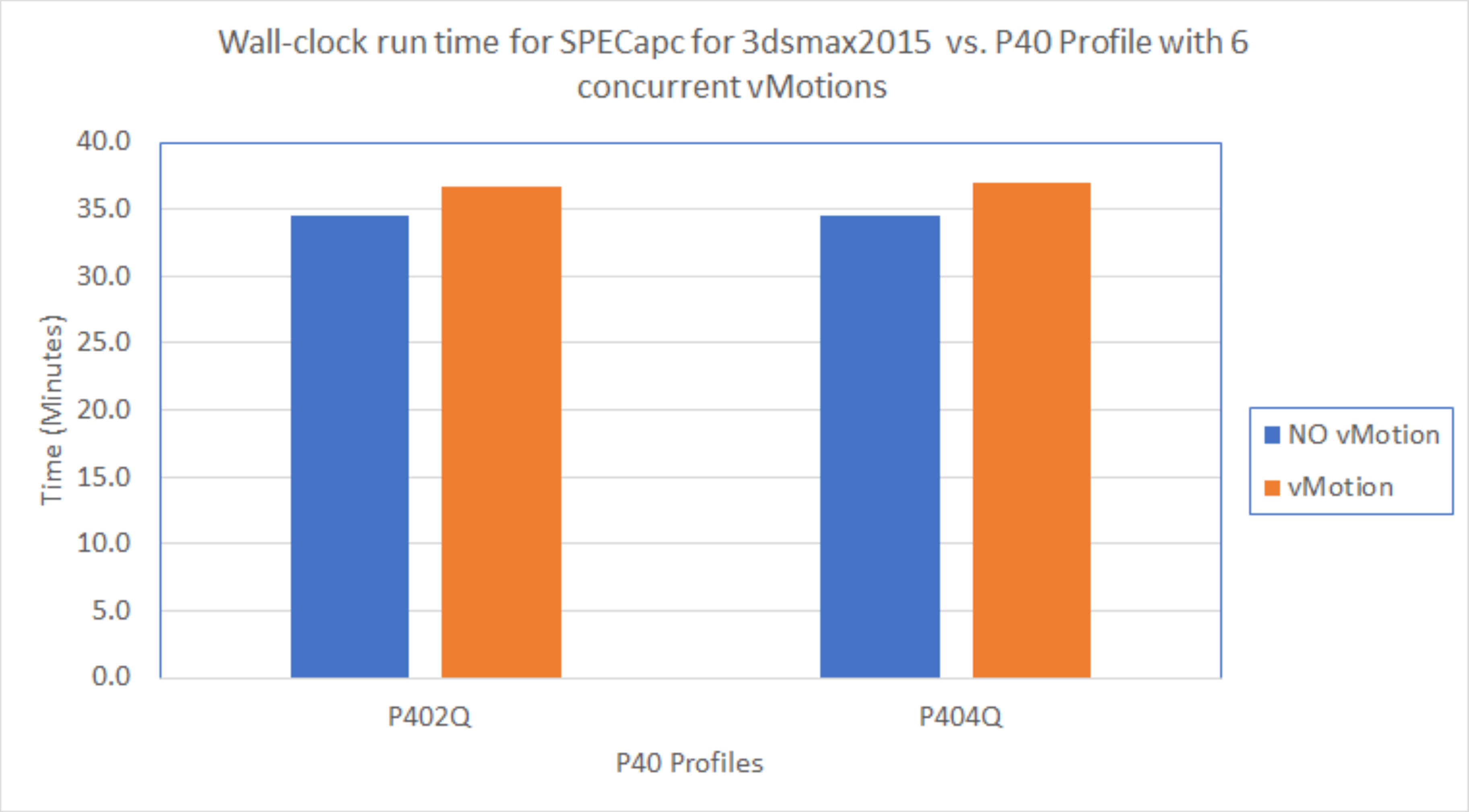
Dell R730 – Intel Broadwell CPUs + 1 x NVidia P40
40 cores (2 x 20-core socket) E5-2698 v4
768 GB RAM

- ESX: 6.7u1   Nvidia Driver: 410.68

vMotioning of different vGPUs running SPECapc

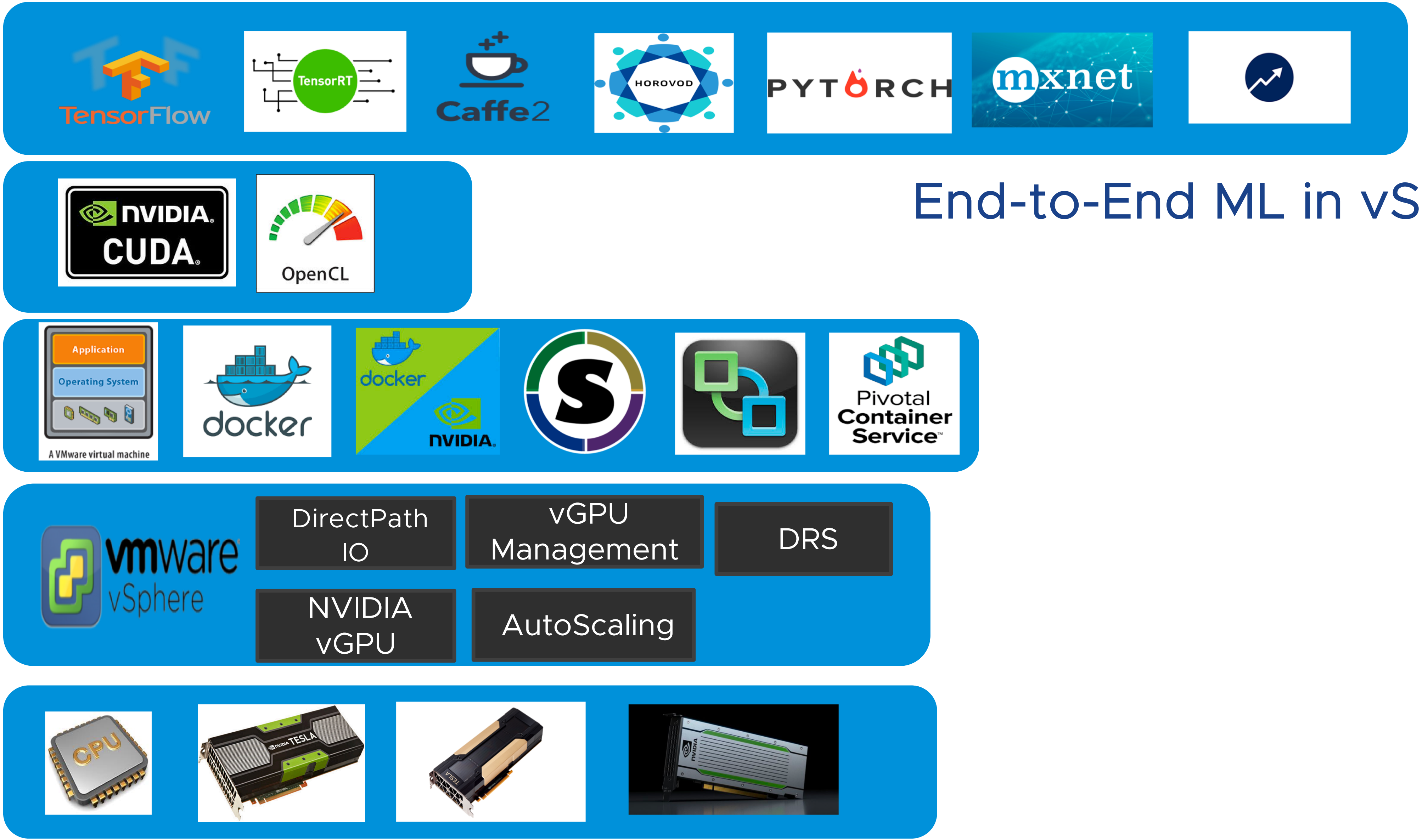


Concurrent vMotions of VMs running SPECapc



End-To-End Machine Learning Using NVIDIA vGPUs in VMware vSphere

End-to-End ML in vSphere



DirectPath IO

vGPU Management

DRS

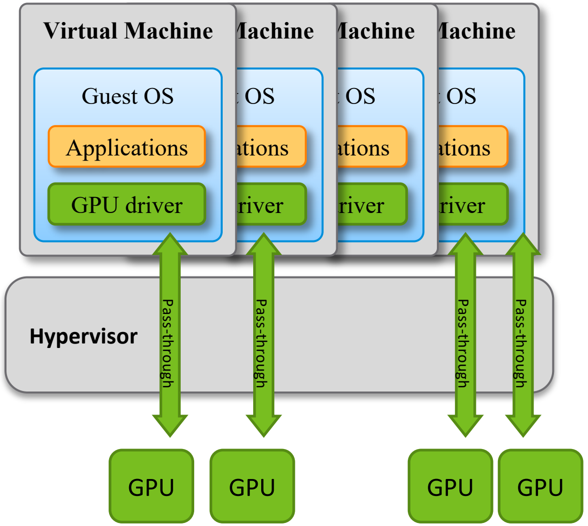
NVIDIA vGPU

AutoScaling

Accessing GPUs from a VM: 2 Solutions

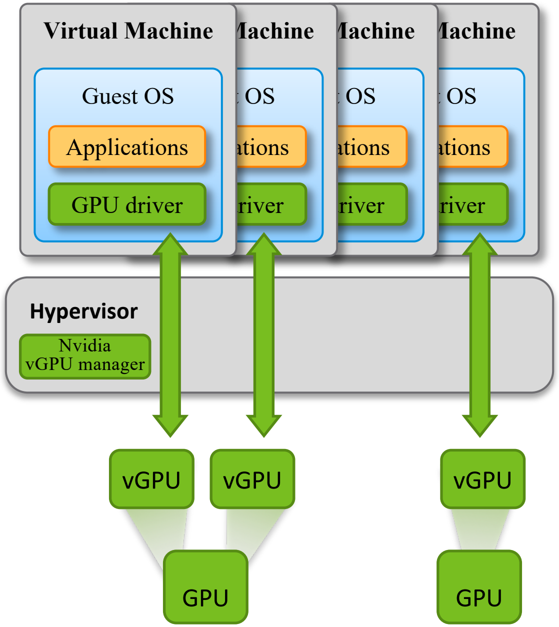
VMware DirectPath I/O

- Allows multiple GPUs per VM
- VMs cannot share GPU



Nvidia vGPU

- Allows multiple VMs per physical GPU
- Allows multiple vGPU profiles
- Management benefits of Virtualization



Performance:
Native GPU vs Virtual GPU

Training Workload: Language Modelling Using RNN

Complex Language Modeling

- Given history of words, predicts next word

- **Neural Network Type: Recurrent Neural Network**
  - Large Model
    - 1500 LSTM units /layer
  - Medium
    - 650 LSTM units /layer
  - Small
    - 200 LSTM units /layer
  - Penn Tree Bank (PTB) Database:
    - 929K training words
    - 73K validation words
    - 82K test words
    - 10K vocabulary

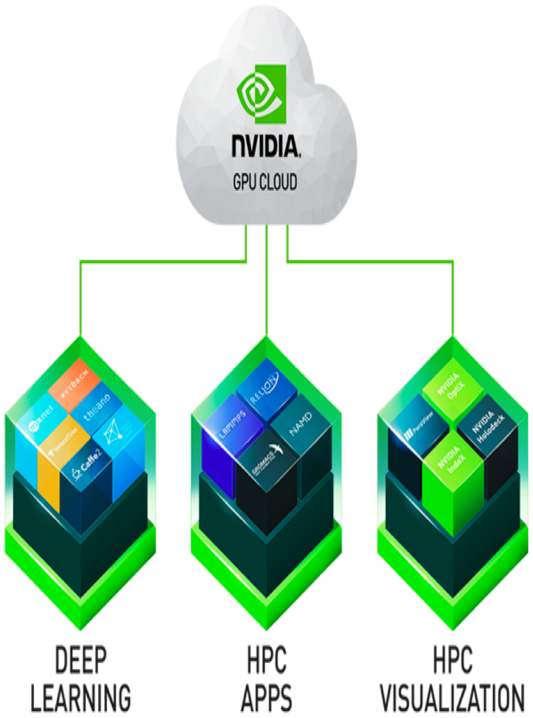
Testbed Configuration

Container in a VM Configuration

- Nvidia Docker: 18.09.1
- vGPU T4-16Q
- CentOS 7.4
- ESX 6.X

Dell R730 – Intel Broadwell CPUs + Turing T4 GPU
40 cores (2 x 12-core socket) E5-2698 V5
768 GB GB RAM

NVIDIA GPU CLOUD

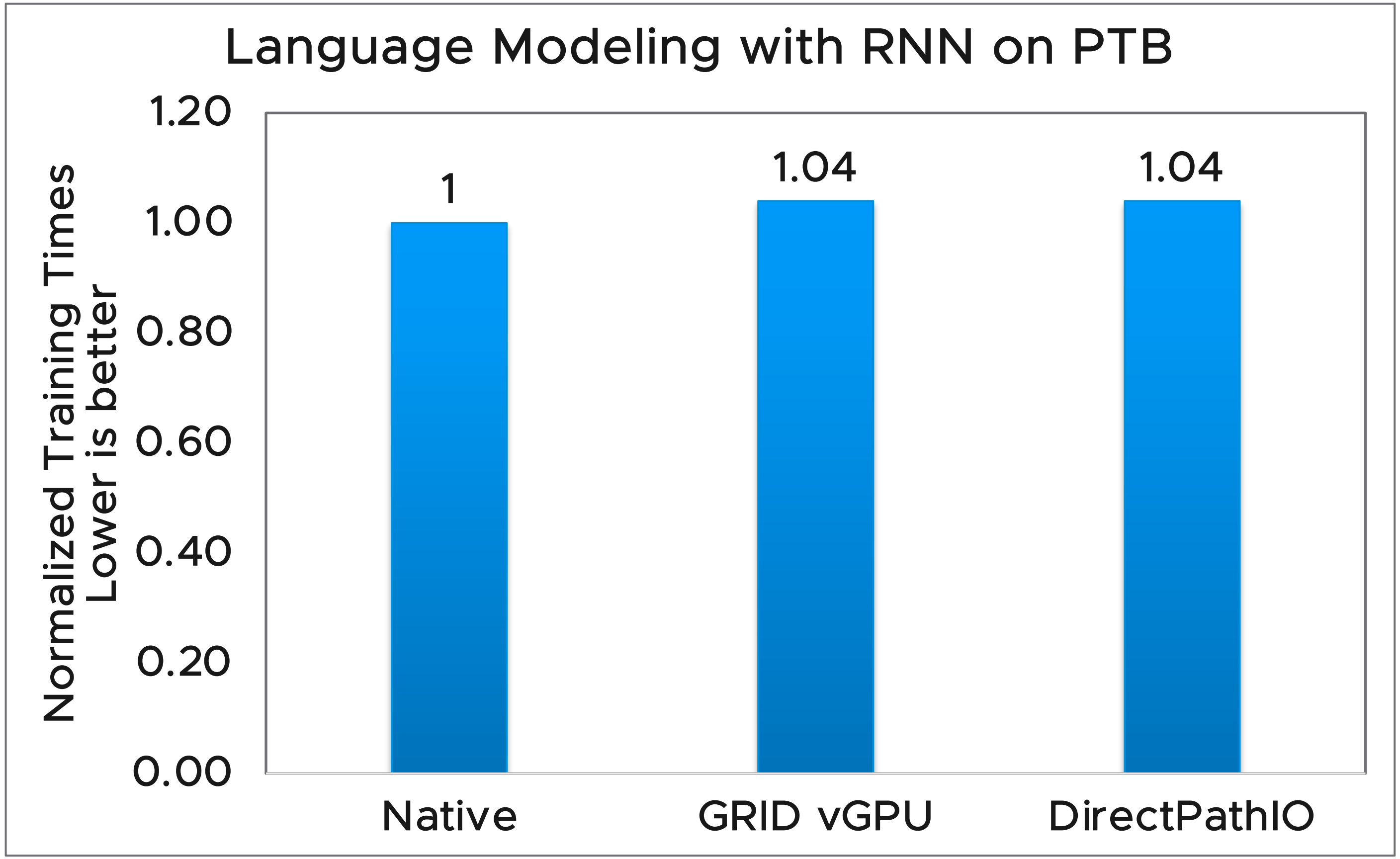


Deep Learning Components

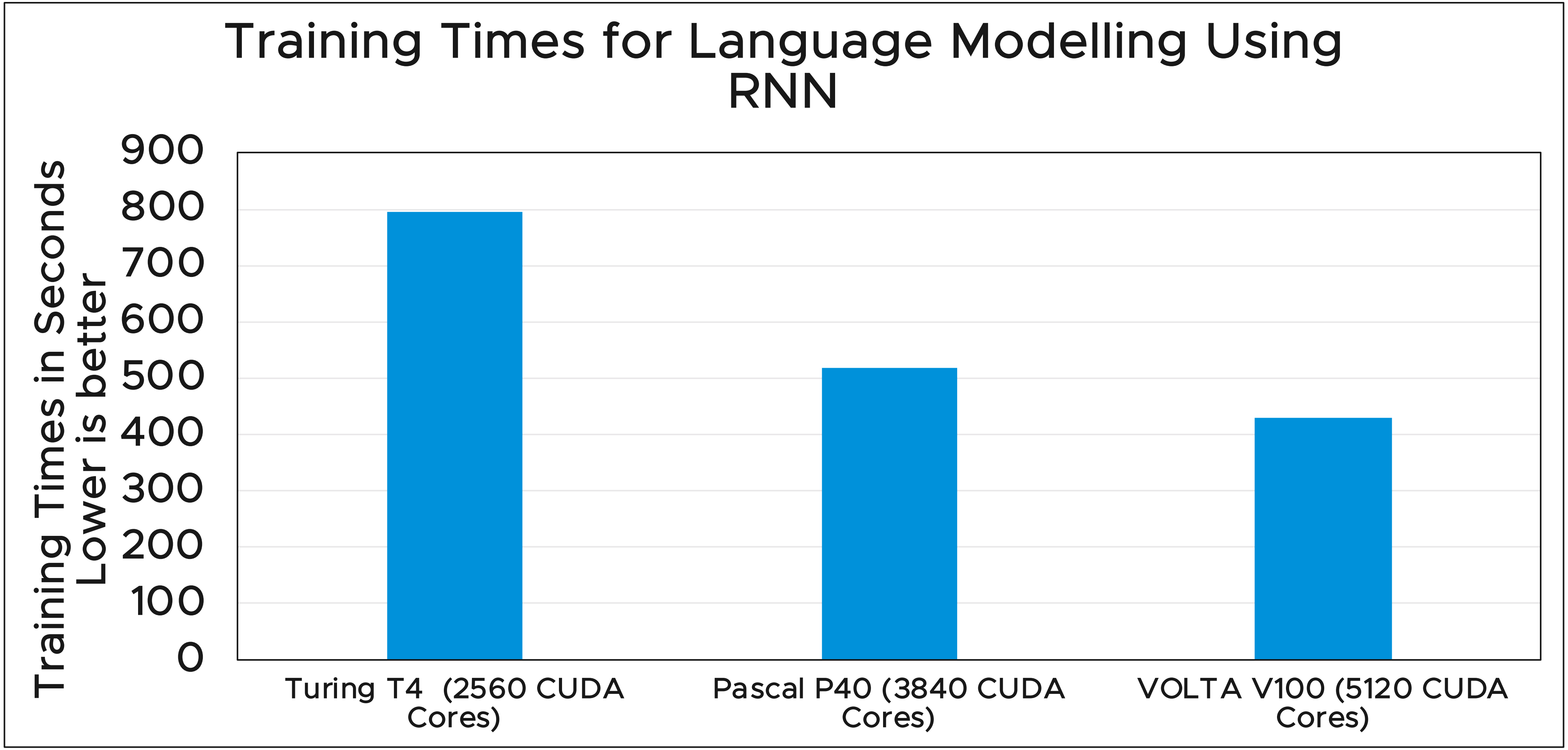
- Machine Learning Workloads
- TensorRT:19.02-py3
- TensorRT-Server: 19.02-py3
- TensorFlow: 1.10

Performance: Training Times on native GPU vs virtualized GPU

4% of overhead for both vGPU & DirectPath I/O compared to native GPU

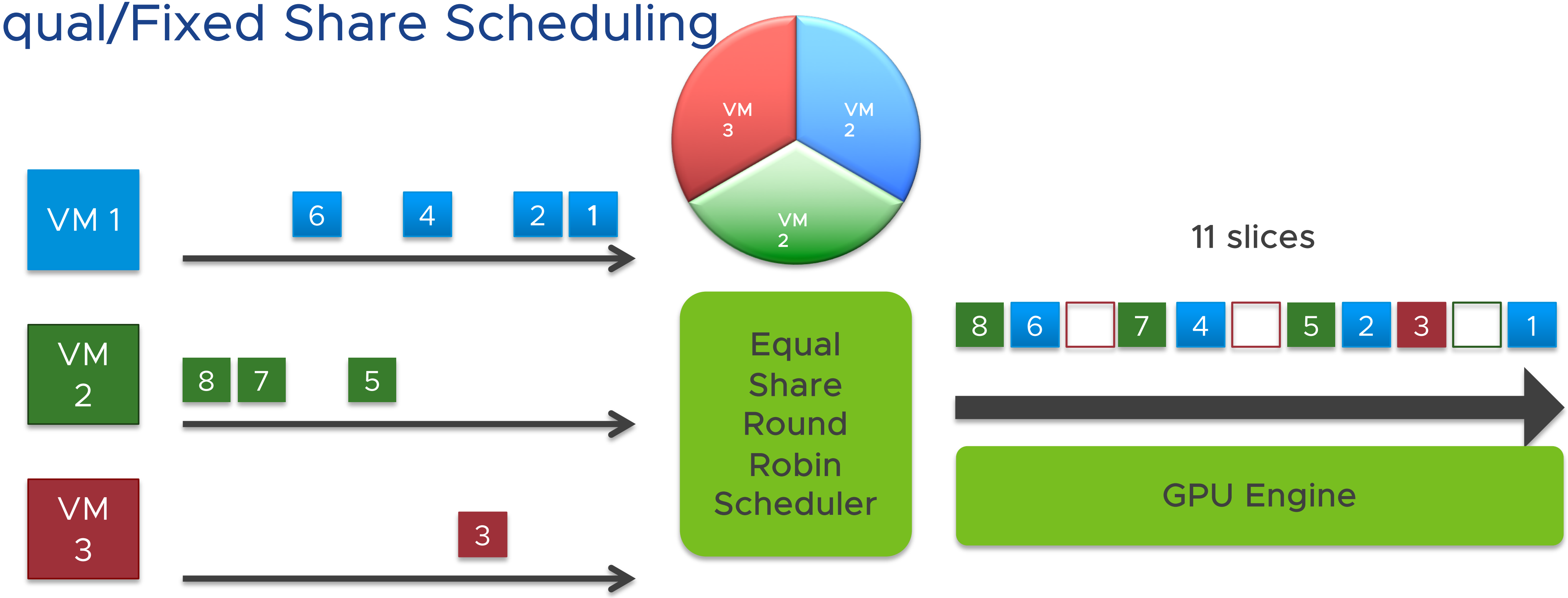


Turing T4 vs Pascal P40 vs Volta V100 Using Highest vGPU Profile



Performance: vGPU Scheduling Policies & VM Scaling for Inferencing

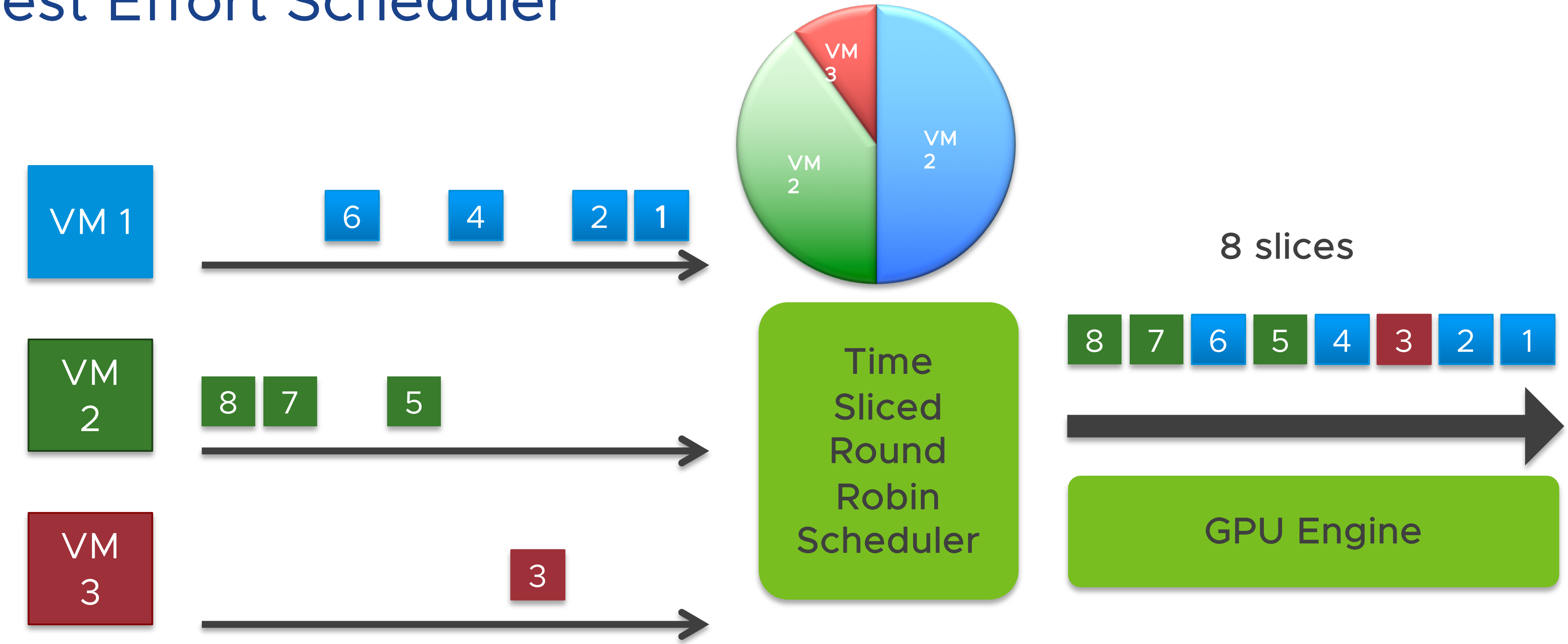
Equal/Fixed Share Scheduling



- Equal Share: Time slice is reserved for every powered ON VM
- Fixed Share: Time slice is determined by the vGPU Profile associated with the GPU.

Content is adapted from NVIDIA materials

Best Effort Scheduler



- If VM has no task or has used its time slice, the scheduler will move to the next VM

Content is adapted from NVIDIA materials

Inferencing Workload: Image Classification Using ResNet50

Workload:

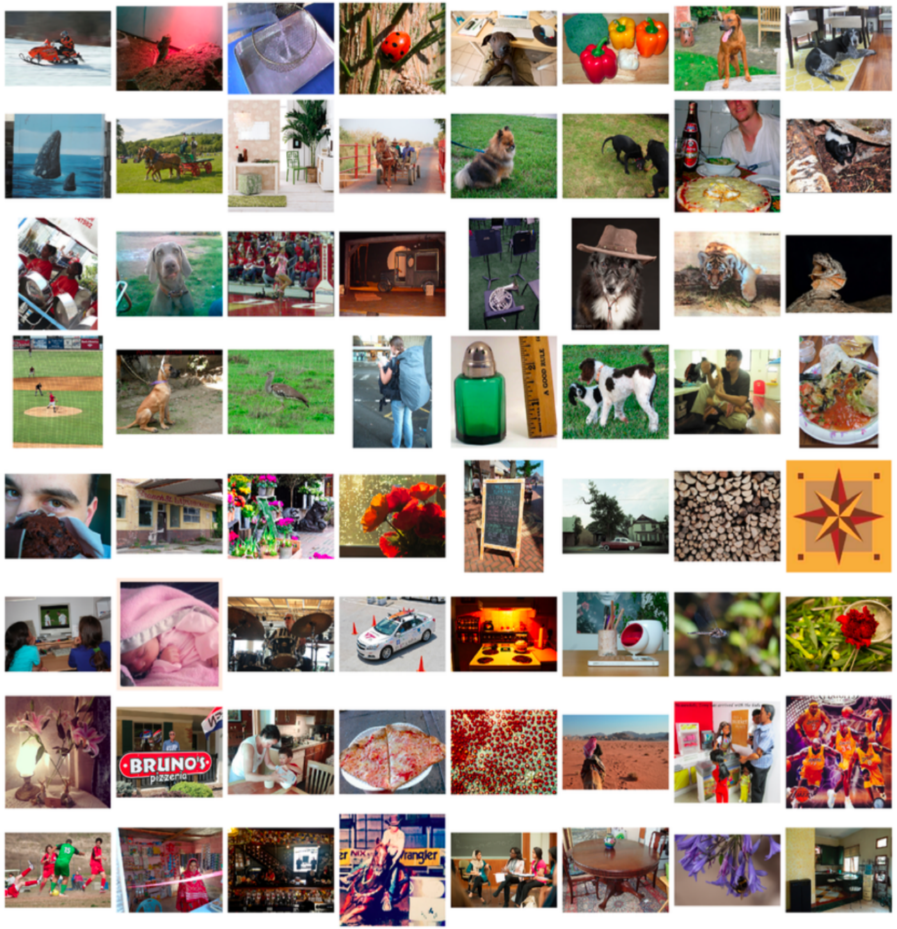
- Image Classification
- 1000 classes/labels

Convolutional Neural Network

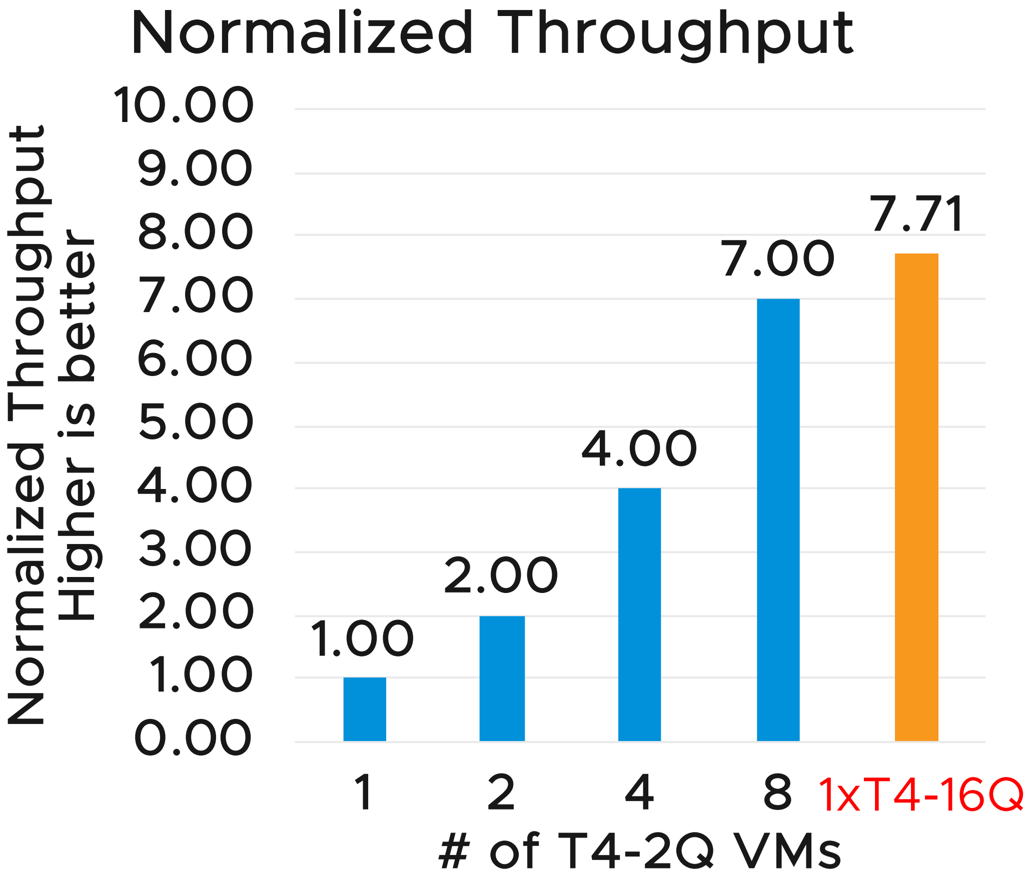
- ResNet: Residual Network
- Precision: FP 32
- 50 Layers
- Human Brain has similar structure

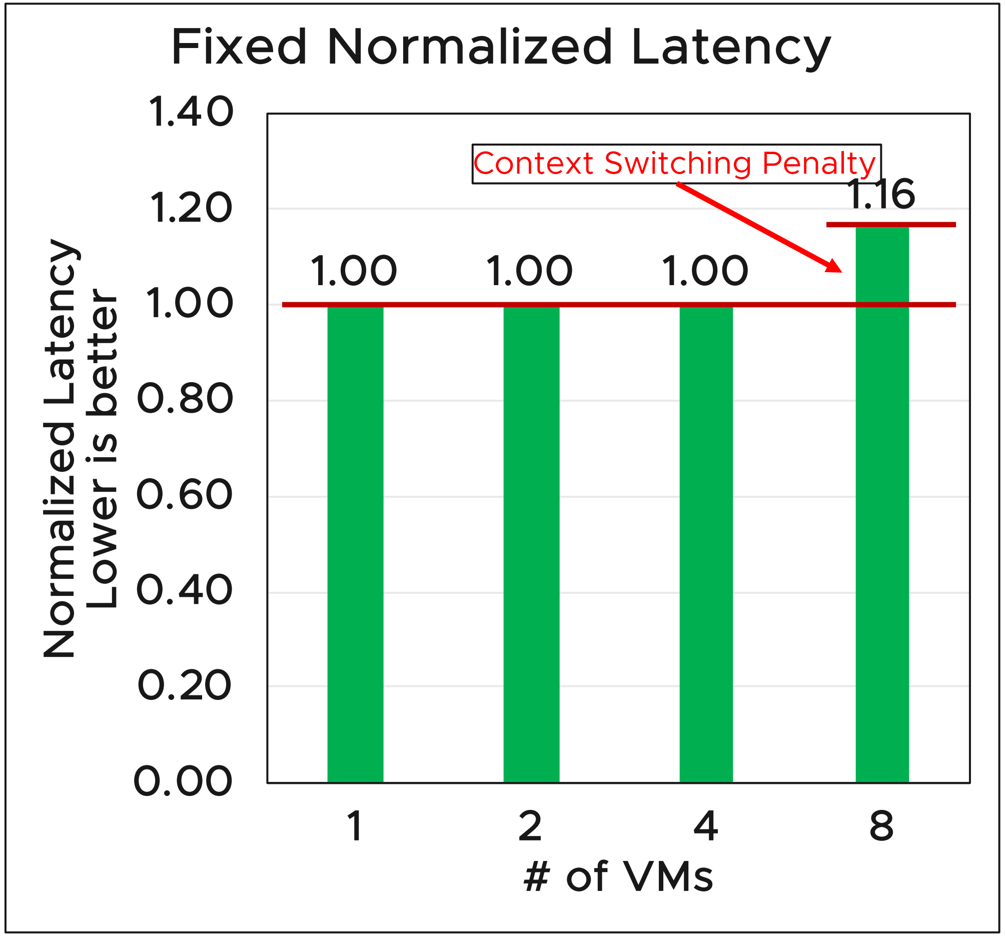
GPU: Turing T4

- ResNet50 FP32 needs at least 2GB
- T4-2Q profile => Max 8 Users Per T4 GPU

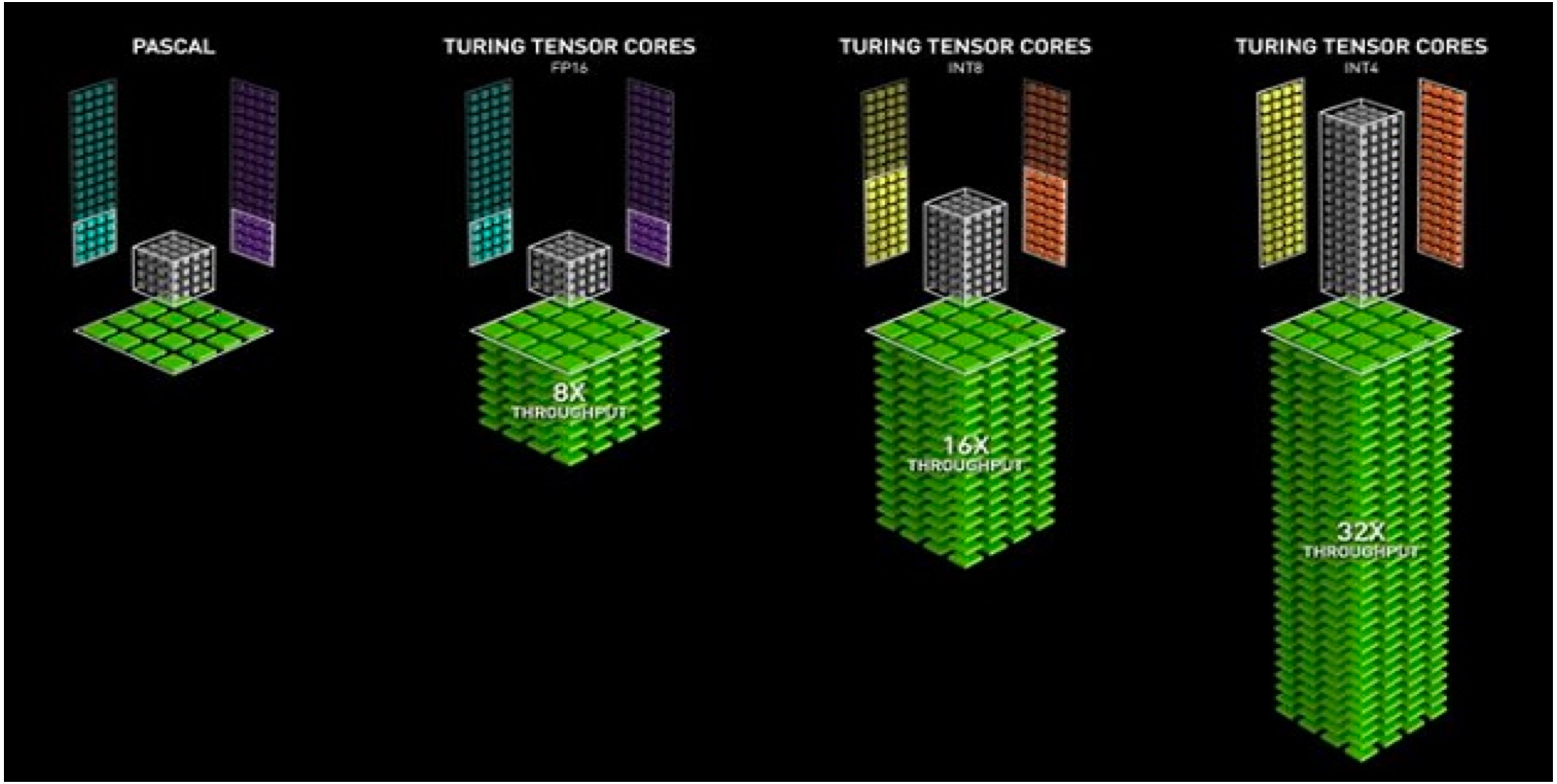


Fixed Share Scheduling: Image Classification Using NVIDIA TensorRT Server

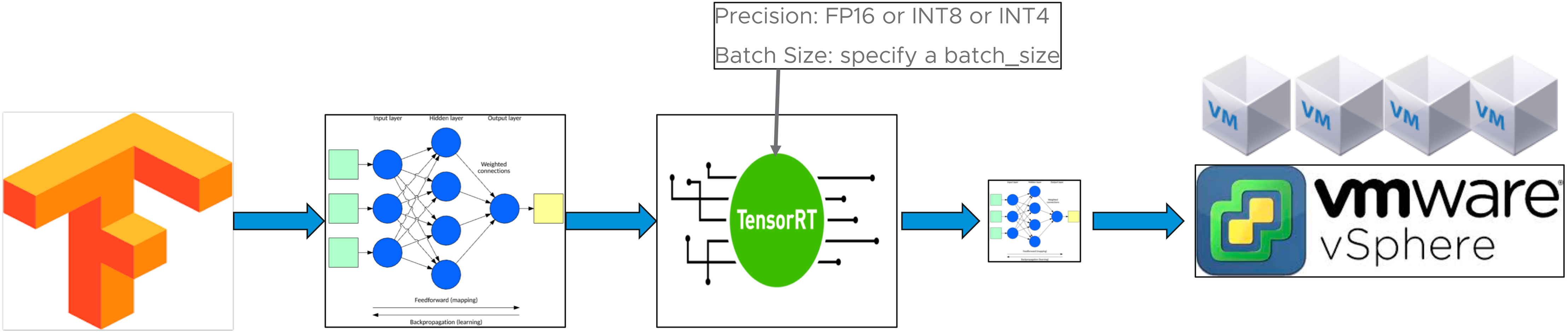




Improving Inferencing Performance With Turing



How to Improve Inference Latency, Throughput and Multi-tenancy using TensorRTand vGPUs?



Precision: FP16 or INT8 or INT4

Batch Size: specify a batch_size

Uses FP32 and Needs 2 GB

Uses T4-2Q

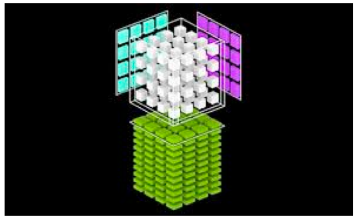
Supports up to 8 Users on T4

Uses FP16 or INT8 or INT4

Needs 1 GB

Supports up to 16 Users on T4

Latency Improvements



Now we can support up to 16 Users on T4 with Major Latency Improvements!

Key Takeways

- Turing T4: A Universal GPU for Virtual Workstations, Knowledge Workers, Rendering, Inferencing and Training.
  - T4 is energy efficient. Takes only 70 Watts of Power!
- Turing, Pascal and Volta support a full spectrum of workloads and users.
- Virtualization and TensorRT magnifies the benefits of lower and mixed precision features of Turing and Volta by improving latency, throughput and multitenancy.
- For Multi-GPU workloads, use Direct PathIO mode
- For more consolidation and multitenancy, use vGPU solution.
- Take Advantage of vMotion and Suspend/Resume feature of vGPU enabled VMs.
- vGPU combines performance of GPUs and data center management features of VMware vSphere!

Presentations by VMware at Nvidia GTC 2019

S9435 - Large Scale Video Audio Quality Assessment on VMware VDI Platform with NVIDIA GPUS
Talk by Hari Sivaraman and Lan Vu
- Wednesday, 3/20/19 | 14:00 - 14:50

S9411 – vMotion for NVIDIA vGPU Virtual Machines: Case Study of vMotion Using MLaaS
Talk by Lan Vu, Hari Sivaraman and Dimitrios Skarlatos
- Wednesday, 3/20/19 | 16:00 – 16:50

S9815 – Maximizing Utilization of NVIDIA Virtual GPUs in VMware vSphere for End-to-End Machine Learning
Talk by Manvendar Rawat and Uday Kurkure
- Tuesday, 3/19/19 | 11:00 – 11:50

Q&A

Thank you NVIDIA GTC for the opportunity

Contact

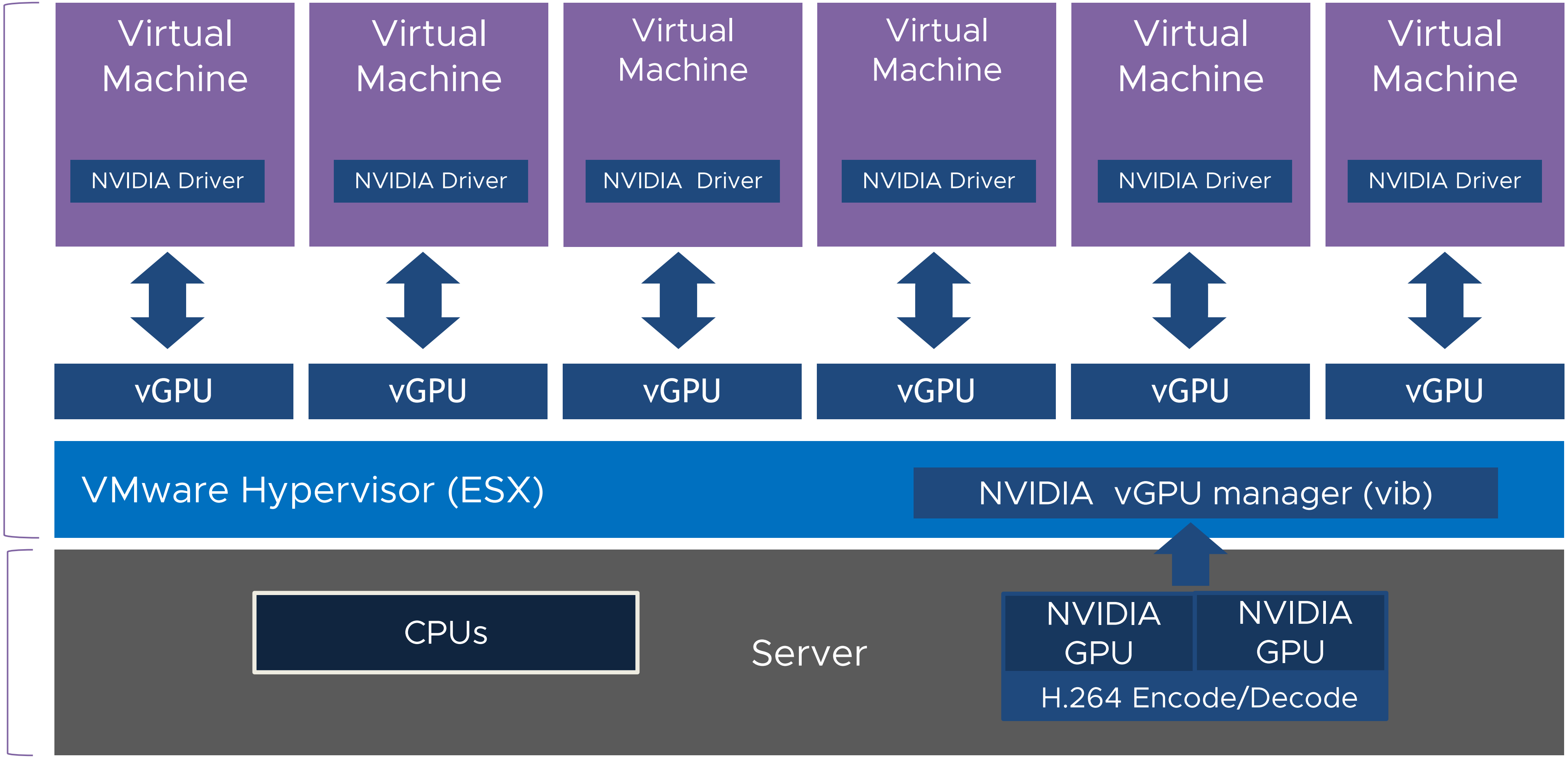
Uday Kurkure ukurkure@vmware.com

Manvendar Rawat mrawat@nvidia.com

Thanks to our colleagues

- Lan Vu, Hari Sivaraman, Juan Garcia-Rovetta, Ravi Soundararjan

What is a virtualized GPU (vGPU) in NVIDIA ?


Virtual Machine
NVIDIA Driver
Virtual Machine
NVIDIA Driver
Virtual Machine
NVIDIA Driver
Virtual Machine
NVIDIA Driver
Virtual Machine
NVIDIA Driver
Virtual Machine
NVIDIA Driver
vGPU
vGPU
vGPU
vGPU
vGPU
vGPU
VMware Hypervisor (ESX)
NVIDIA vGPU manager (vib)
CPUs
Server
NVIDIA GPU
NVIDIA GPU
H.264 Encode/Decode