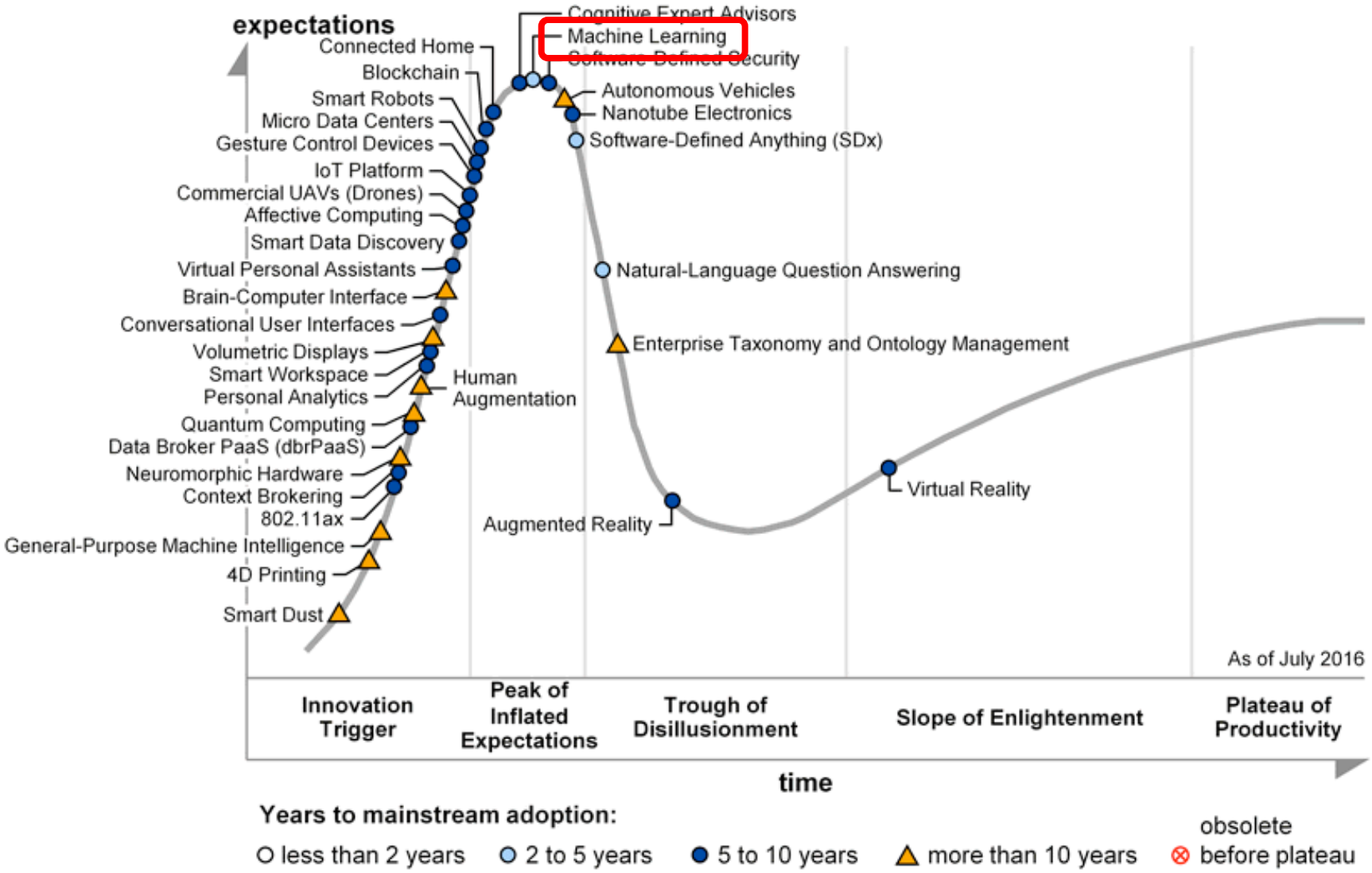
Machine Learning on VMware vSphere with NVIDIA GPUs

Uday Kurkure, Hari Sivaraman, Lan Vu

GPU Technology Conference 2017

vmware

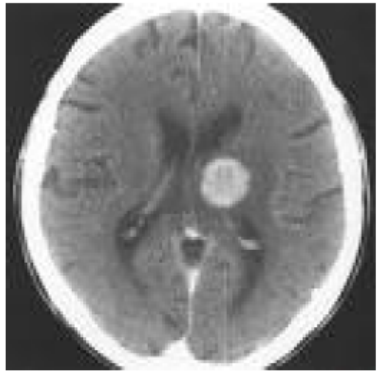
Gartner Hype Cycle for Emerging Technology



Machine Learning Application Areas

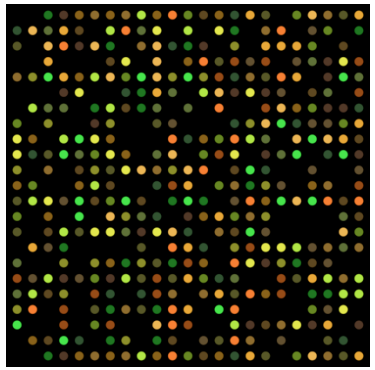
Assets Management

Medical Imaging



Computational Biology

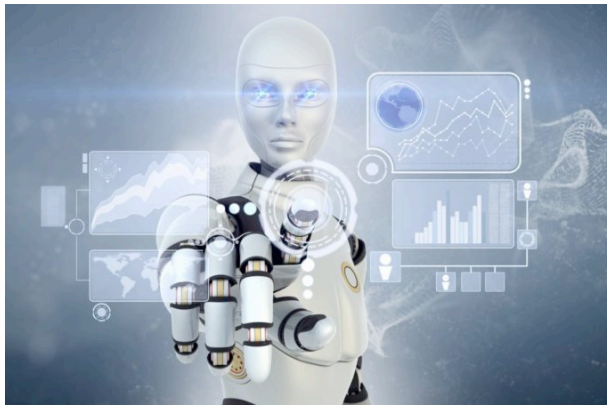
Gene Sequencing



Microarray

Business Analysis

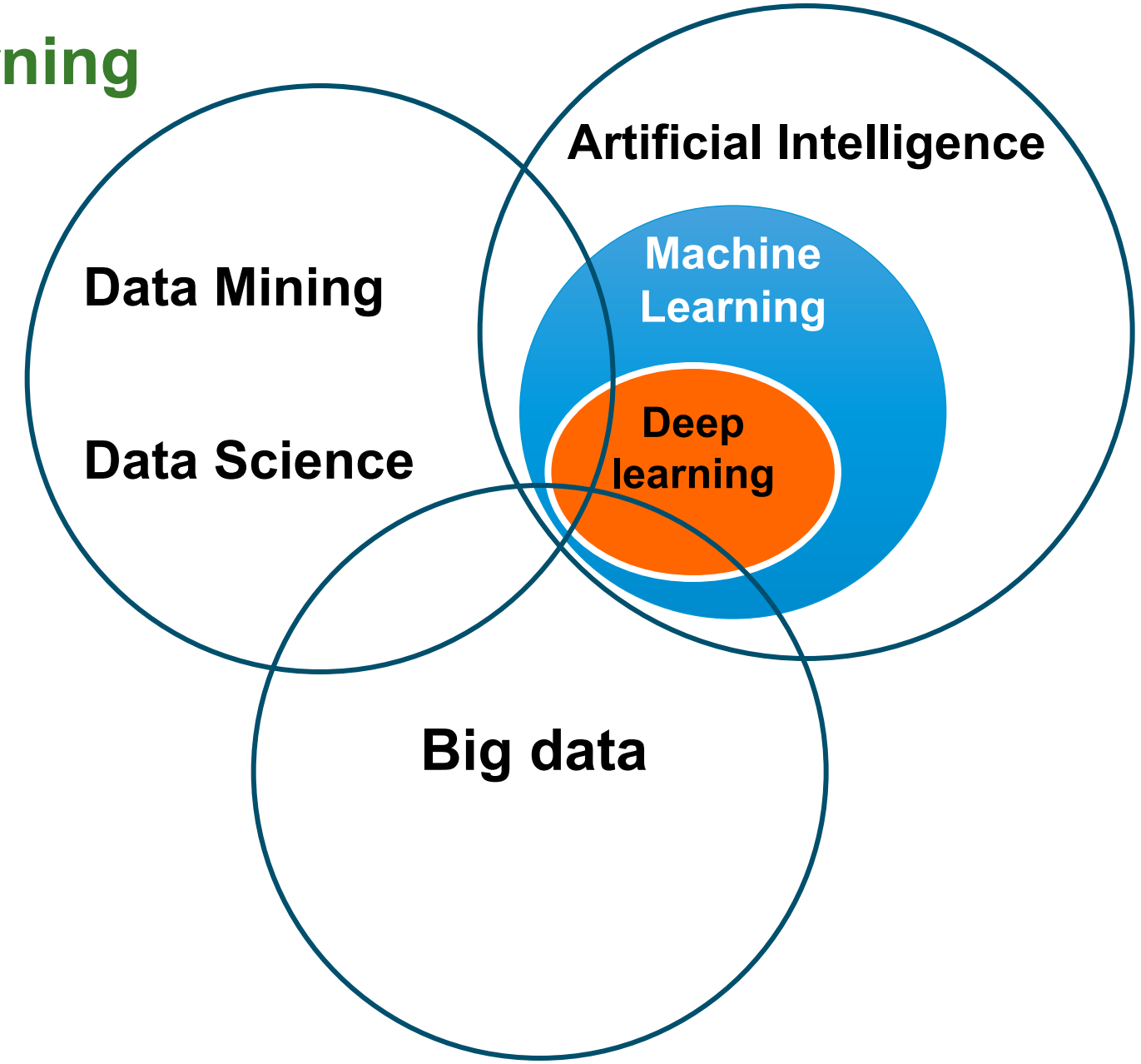
Automation



Social Networks

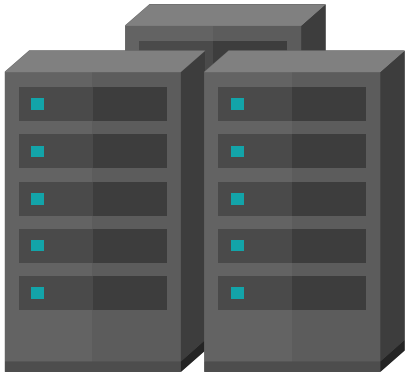
IoT

Machine Learning

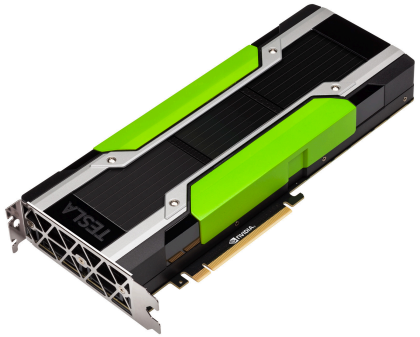


Adapted from http://www.kdnuggets.com/2016/03/data-science-puzzle-explained.html

High Performance Computing for Machine Learning

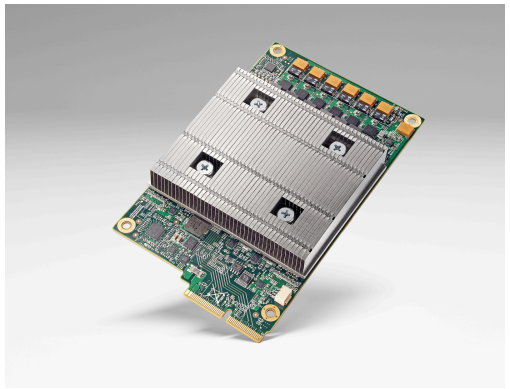


Multi-core CPU Systems



GPU (Graphics Process Unit)

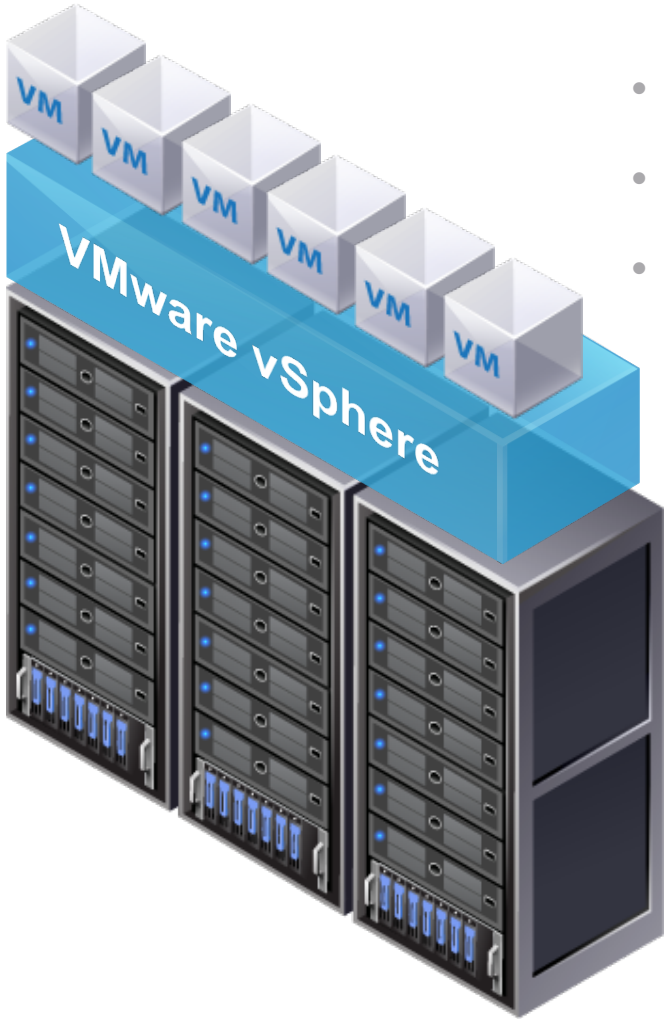
FPGA (A field-programmable Gate Array)



TPU (Tensor Processing Unit)

vmware®

Why Machine Learning on VMware vSphere with Nvidia GPUs?



- VMware vSphere hypervisor efficiently manages servers of data center
- Machine learning workloads are becoming important in cloud environment
- VMware vSphere supports multiple GPU virtualization solutions

Capabilities of GPUs supported by VMware vSphere

- Accelerating 2D/3D Graphics workloads for VMware VDI
- VMware Blast Extreme protocol encoding / decoding for VDI
  - H.264 Based (MPEG-4)
- General Purpose GPU (GPGPU)
  - **Machine learning / Deep Learning**
  - Other high performance computing workloads

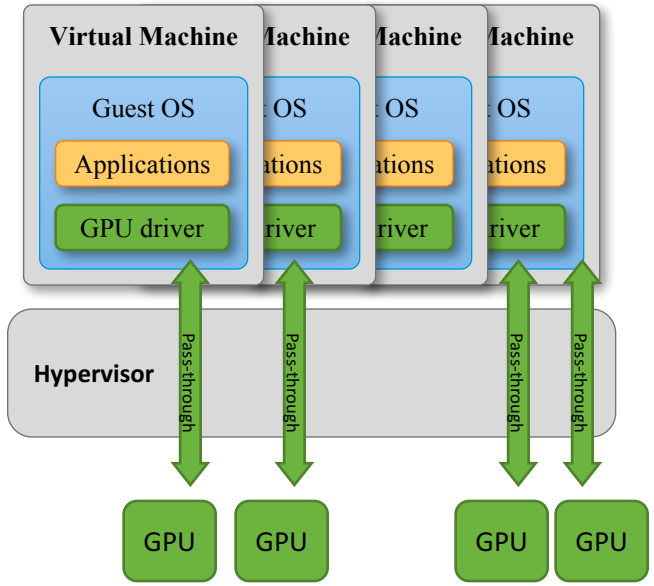
Developing Machine Learning Applications with GPUs

- For Nvidia GPUs, we can use CUDA with cuDNN and other libraries

Machine Learning on VMware vSphere using Nvidia GPUs

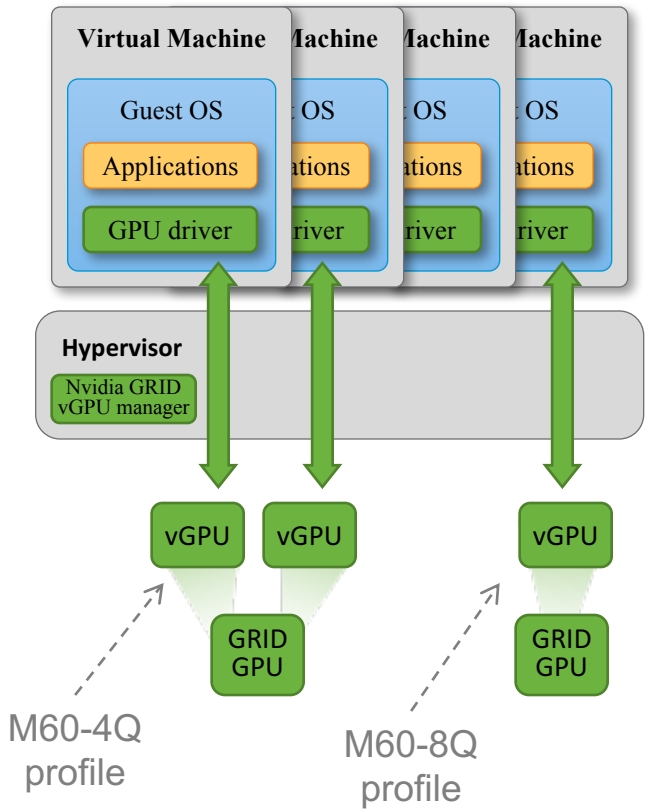
- VMware DirectPath IO
- NVIDIA GRID vGPU

VMware DirectPath IO (Pass-through)



- Support a large set of GPU cards including GRID GPU
- Allow direct access to the physical GPUs
- A virtual machine (VM) allows one or multiple GPUs

NVIDIA GRID vGPU



- Support NVIDIA GRID GPUs
- Each virtual machine owns a vGPU
- vGPU profile specifies the frame buffer size
- **Graphics mode**: One or multiple vGPU per physical GPU
- **Compute mode** (for GPGPU applications like CUDA)
  - Require the highest vGPU profile
  - One vGPU per physical GPU

- Ex: M60 GPU Card has 2 Maxwell GM 204 GPUs
  - Up to 2 CUDA VMs
  - Up to 32 Graphics VMs

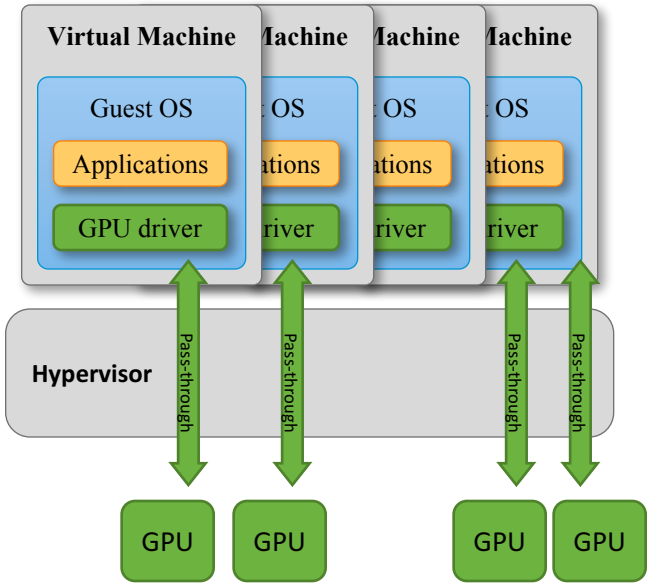
Nvidia GRID vGPU configurations

Card	Physical GPUs	GRID Virtual GPU	Intended Use Case	Frame Buffer (Mbytes)	Virtual Display Heads	Max Resolution per Display Head	Maximum vGPUs	
							Per GPU	Per Board
Tesla M60	2	M60-8Q	Designer	8192	4	3840x2160	1	2
		M60-4Q	Designer	4096	4	3840x2160	2	4
		M60-2Q	Designer	2048	4	2560x1600	4	8
		M60-1Q	Power User, Designer	1024	2	2560x1600	8	16
		M60-0Q	Power User, Designer	512	2	2560x1600	16	32
		M60-2B	Power User	2048	2	2560x1600	4	8
		M60-1B	Power User	1024	2	2560x1600	8	16
		M60-0B	Power User	512	2	2560x1600	16	32

Benefits of each solution

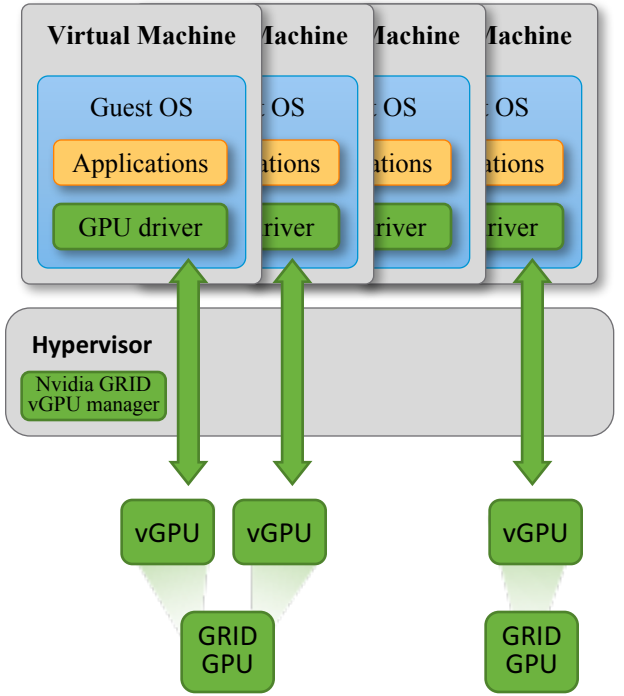
VMware DirectPath I/O

- Support more GPU cards
- Allow multiple GPUs per VM
- Low overhead



Nvidia GRID vGPU

- Allow multiple VMs per physical GPU
- More users / VMs with vGPU
- Flexibility of management



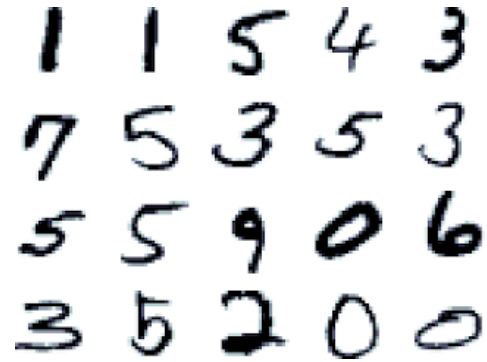
Exploring Machine Learning Workload Performance

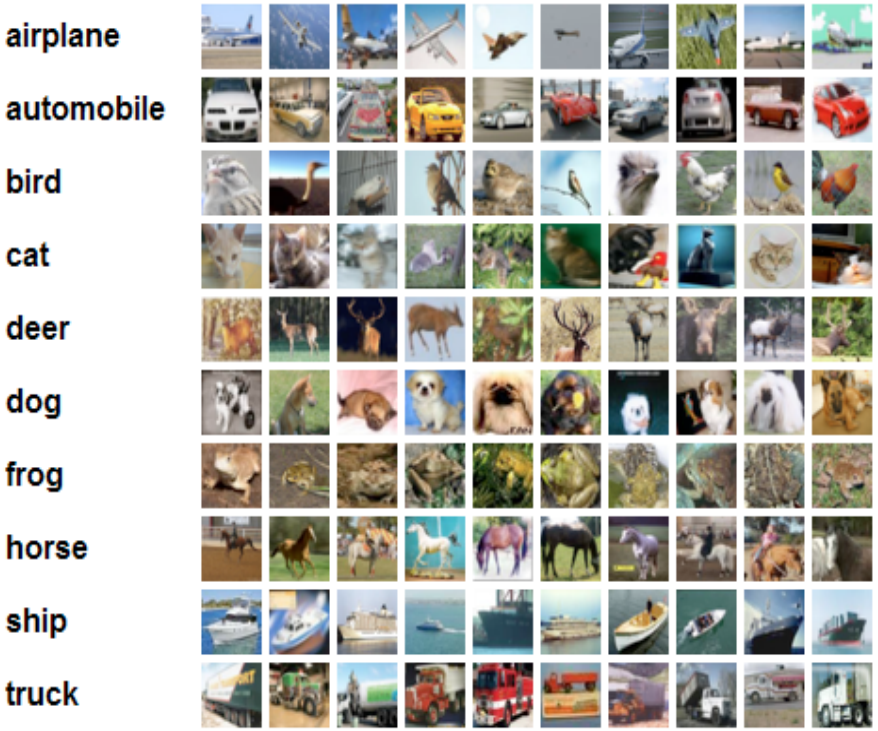
- **Performance: Native vs. Virtual, GPU vs. CPU, etc.**
- **Scalability: with # of user and # of GPUs**
- **Mixed Workloads: ML, 3D workloads**

PERFORMANCE

Machine Learning Framework and Neural Networks

- Machine Learning Framework: TensorFlow
- Nvidia CUDA 7.5 , Nvidia cuDNN 5.1
- Neural Network Architectures and Workloads:
  - Recurrent Neural Network (RNN):
    - Language Modeling on Penn Tree Bank
  - Convolutional Neural Network (CNN)
    - Handwriting Recognition with MNIST
    - Image Classifier with CIFAR-10





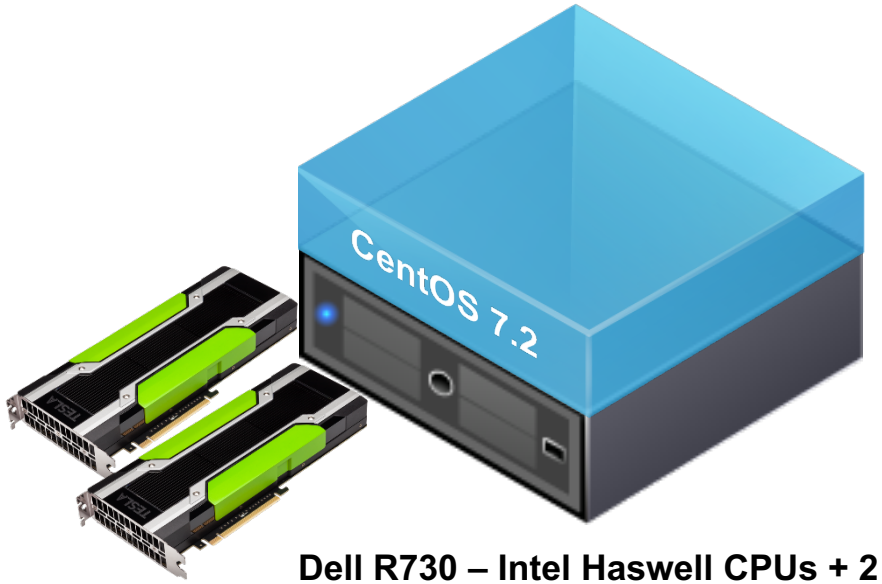
Performance:

Native vs Virtual

Testbed Configurations for Native vs Virtual Comparison

Native Configuration

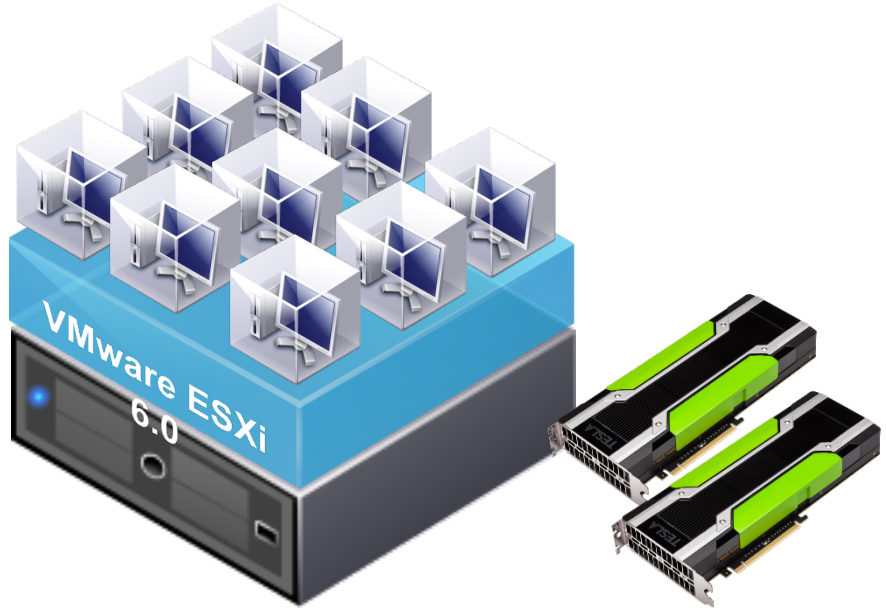
- **TensorFlow 0.10**
- CentOS 7.2
- Hyperthreading ON
- SSD Hard Disk



Dell R730 – Intel Haswell CPUs + 2 x NVidia GRID M60
24 cores (2 x 12-core socket) E5-2680 V3
768 GB RAM

Virtual Linux VMs

- **TensorFlow 0.10**
- CentOS 7.2
- **ESX 6.X**
- 12vCPU, 16 GB RAM, 96GB HD
- SSD hard disk



Dell R730 – Intel Haswell CPUs + 2 x NVidia GRID M60
24 cores (2 x 12-core socket) E5-2680 V3
768 GB GB RAM

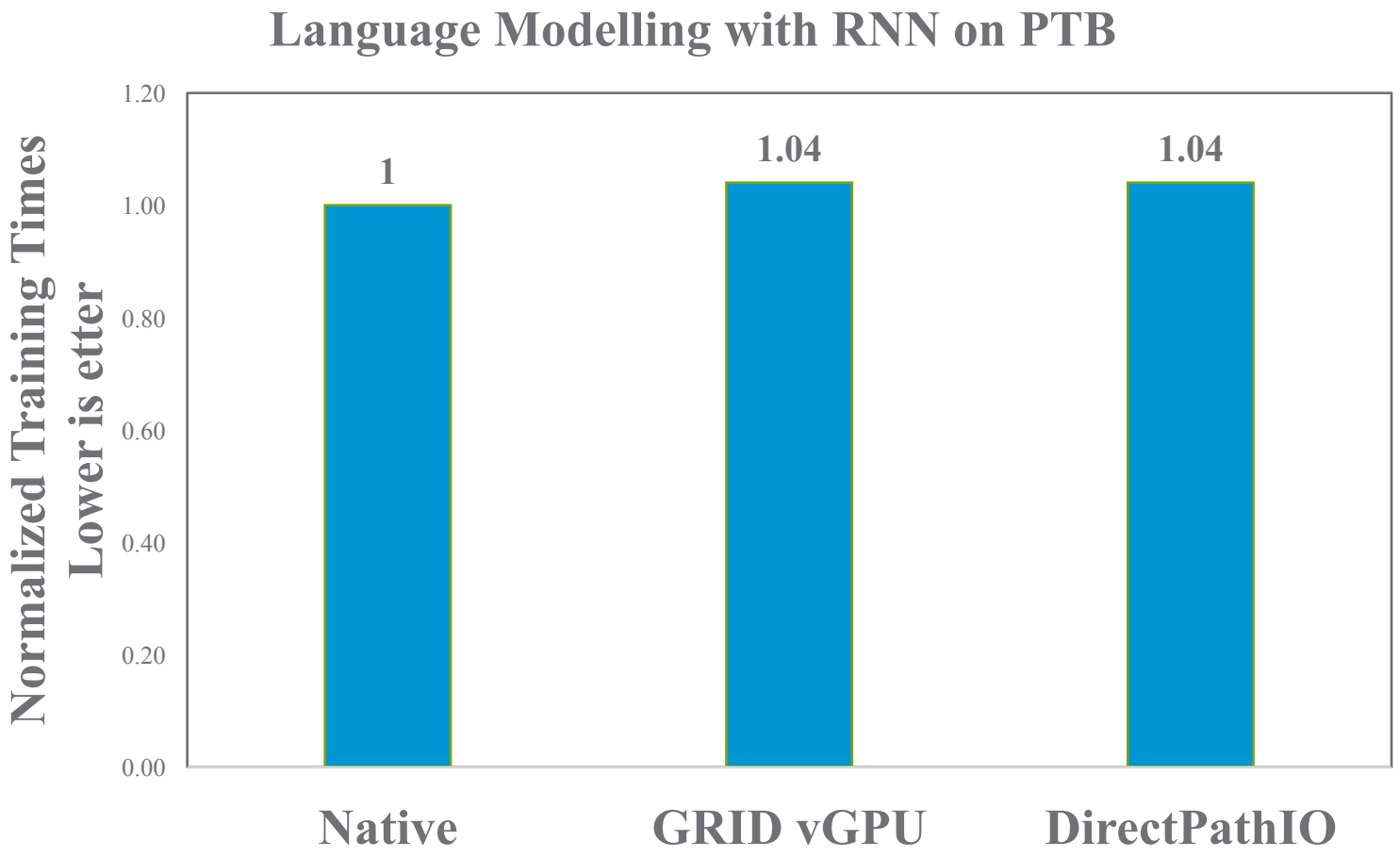
vmware

Performance: Native vs Virtualized GPU

- **Neural Network Type: Recurrent Neural Network**
  - Large Model
    - 1500 LSTM units /layer
- **Workload**
  - Complex Language Modelling Using Recurrent Neural Network
    - Word Level Prediction
  - Penn Tree Bank (PTB) Database:
    - 929K training words
    - 73K validation words
    - 82K test words
    - 10K vocabulary

Performance: Training Times on native GPU vs virtualized GPU

4% of overhead for both DirectPath I/O vs. GRID vGPU compared to native GPU



Performance:

GPUs vs CPUs

Performance: CPU vs GPU

	Intel Haswell	Intel Broadwell	GM204 (M60)	Implications
Cores	12 (24 with hyperthreading)	22 (44 with hyperthreading)	2048	Task Parallelism vs Data Parallelism
TeraFlop	~0.26	~0.5	~4.X	8x to 15x the number of Flops
Clock Rate GHz	2.5 and 3.30 (Turbo Boost)	2.2 and 3.6(Turbo Boost)	1.12 and 1.2(Boost)	0.5 the speed of Broadwell
TDP	120 Watts	125 Watts	165 Watts	
Memory BW			224 GB/s	
Technology	22nm	14nm	28nm	

- **In spite of CPU's process technology & clock rate, GPUs outperform CPUs on ML workloads.**
- GPUs: **More Silicon is devoted to increase the number of ALUs**

vmware

Performance: GPU vs. CPU on virtualized server

Two VM Configurations:

- 1 VM with 1 vGPU (M60-8q) vs 1 VM without a GPU
- Each VM has 12 vCPUs, 60GB memory, 96GB of SSD storage, CentOS 7.2

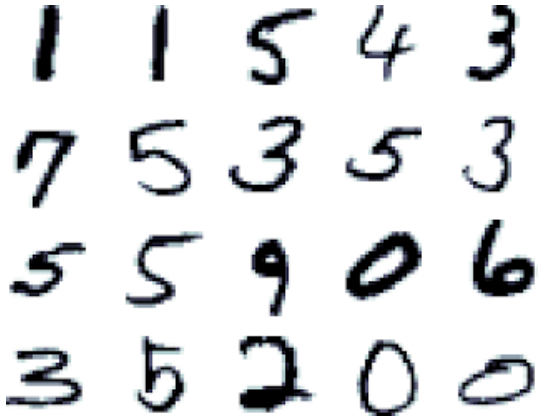
- **Neural Networks Types**: CNN and RNN

Dataset for CNN: MNIST database of handwritten digits

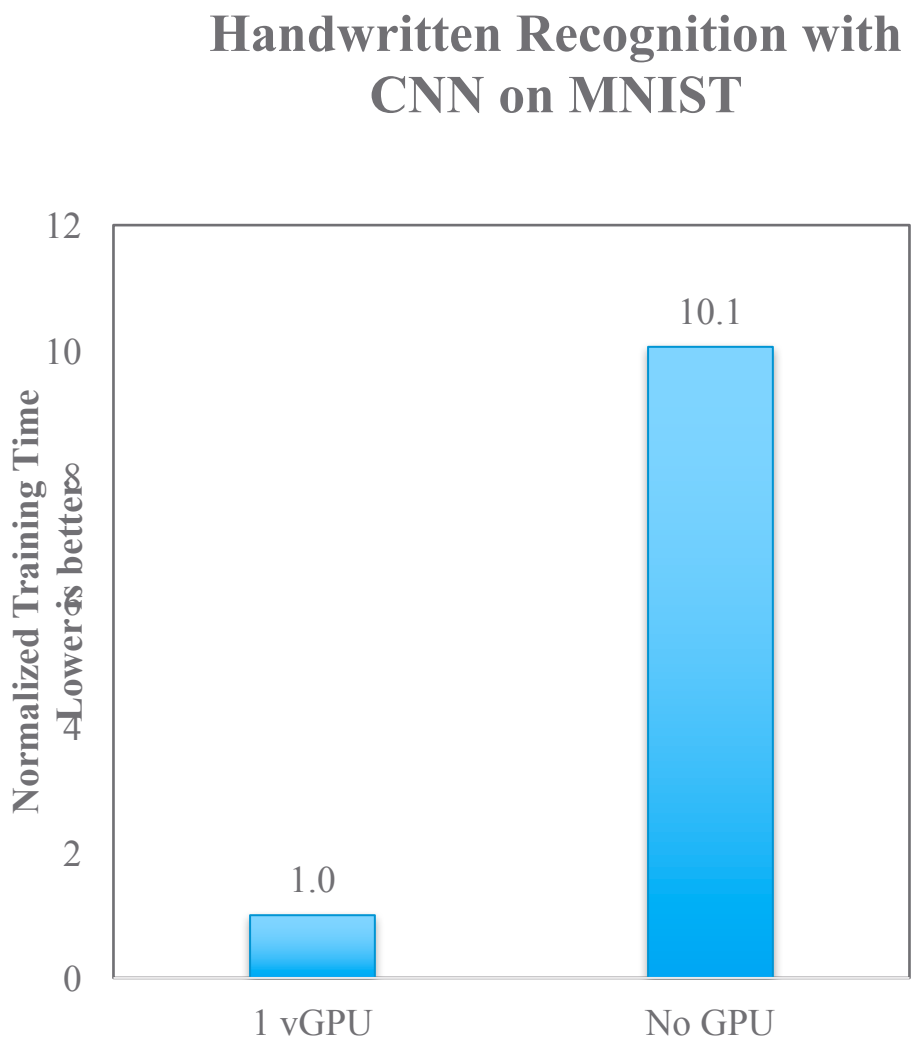
Training set: 60,000 examples

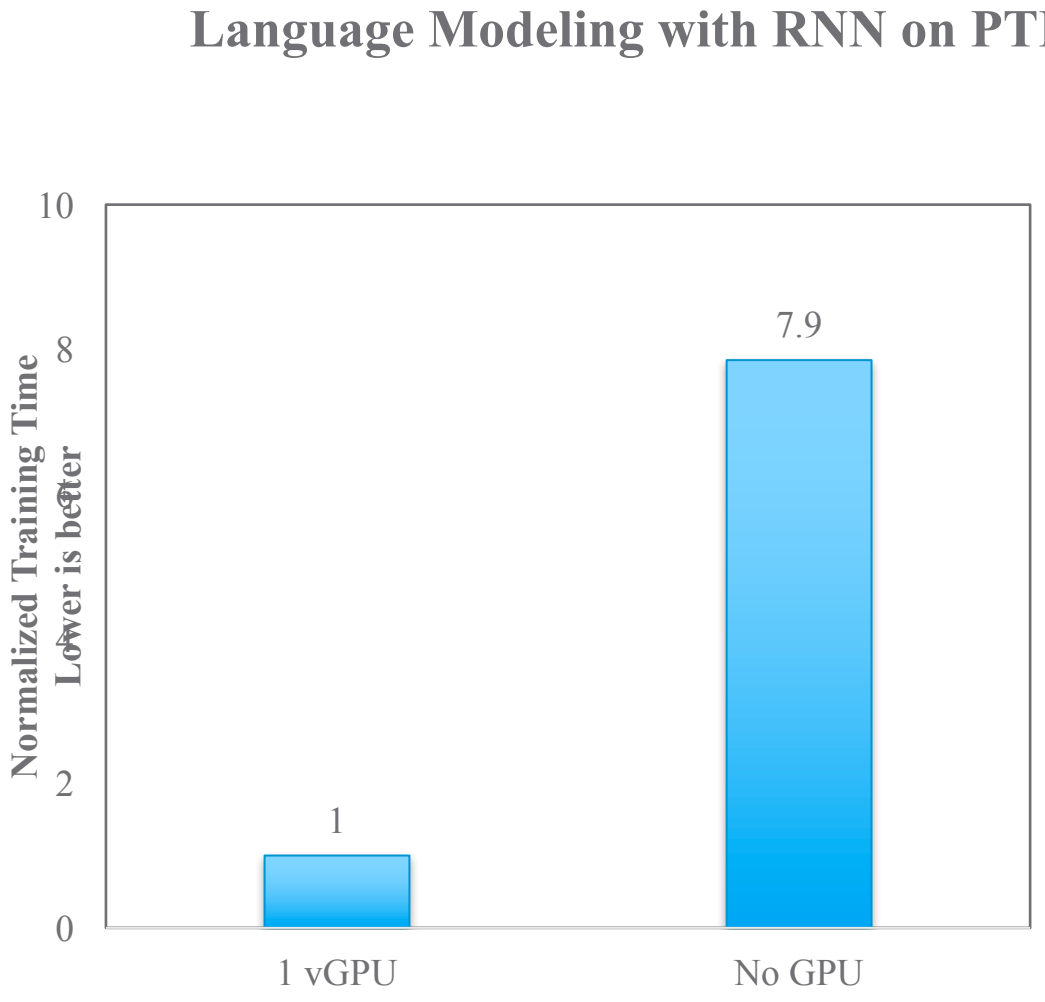
Test set: 10,000 examples.

Dataset for RNN: PTB



Training Times with GPU vs. without GPU on virtualized server

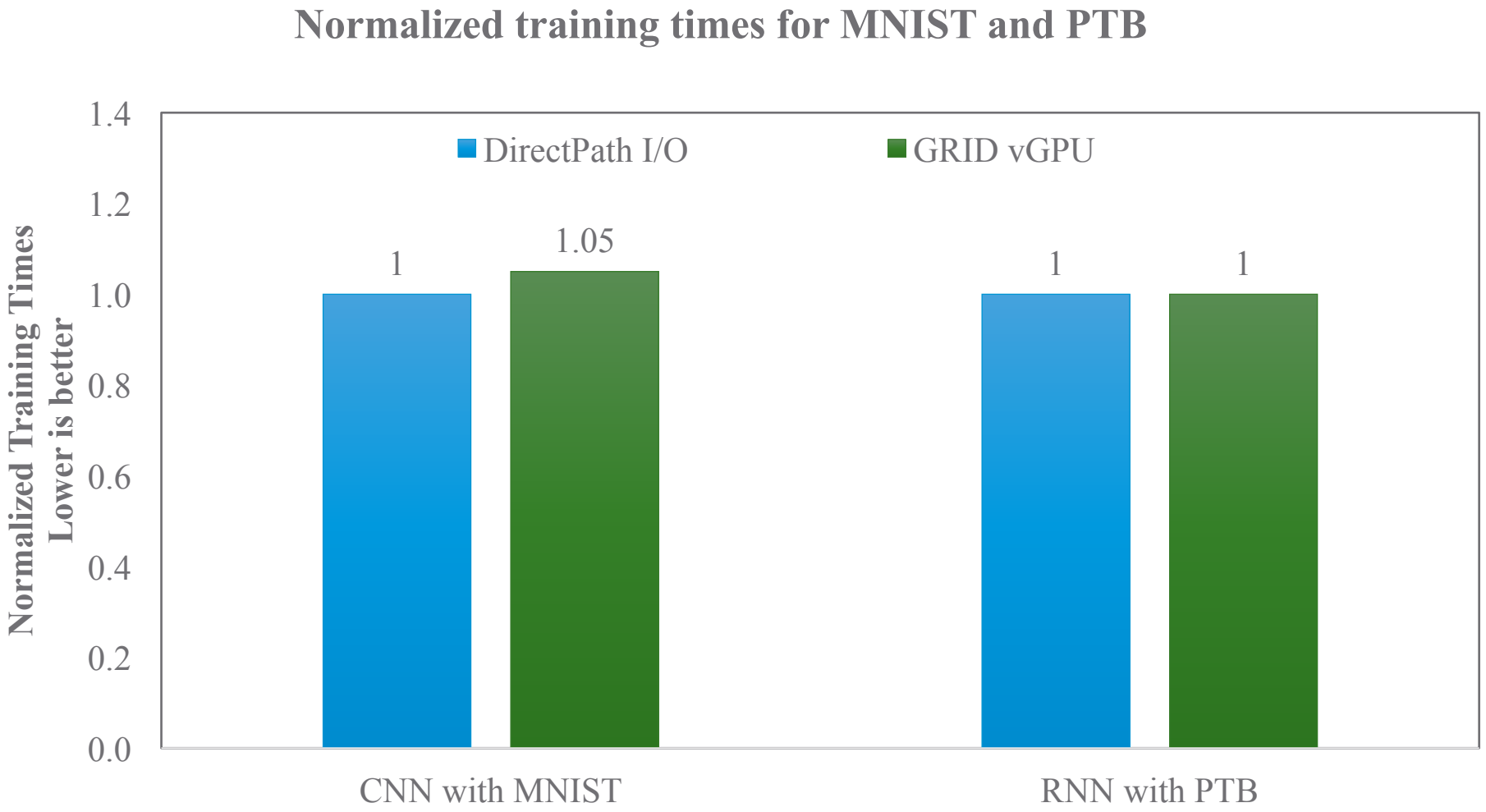




Performance:

DirectPath IO vs GRID vGPU

Performance: DirectPath I/O and GRID vGPU



Scalability- VMs or Users

Scalability: Scaling number of VMs/server

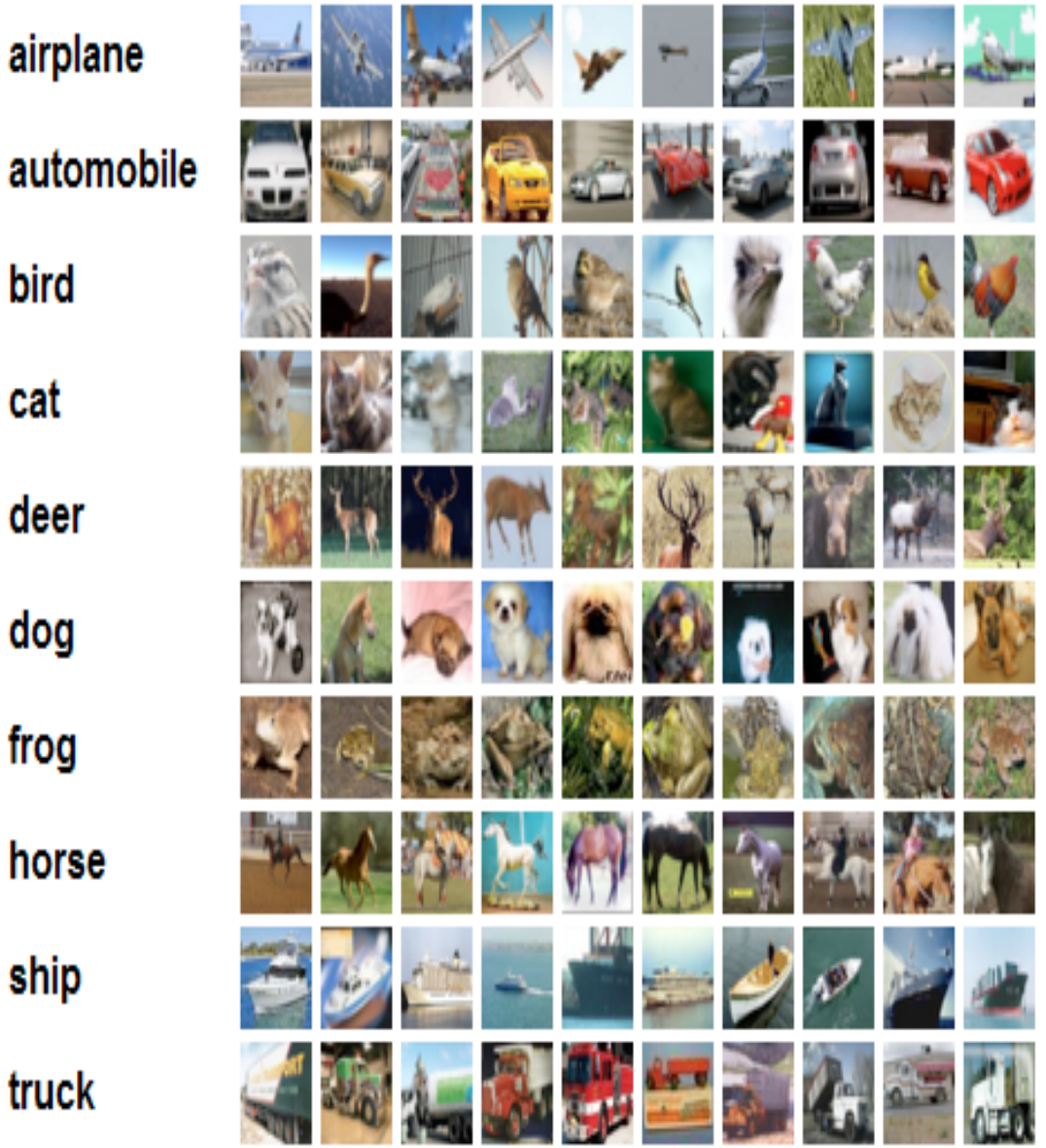
The number of users or VMs with ML workload per server

One VM with one GPU

Four VMs with one gpu each

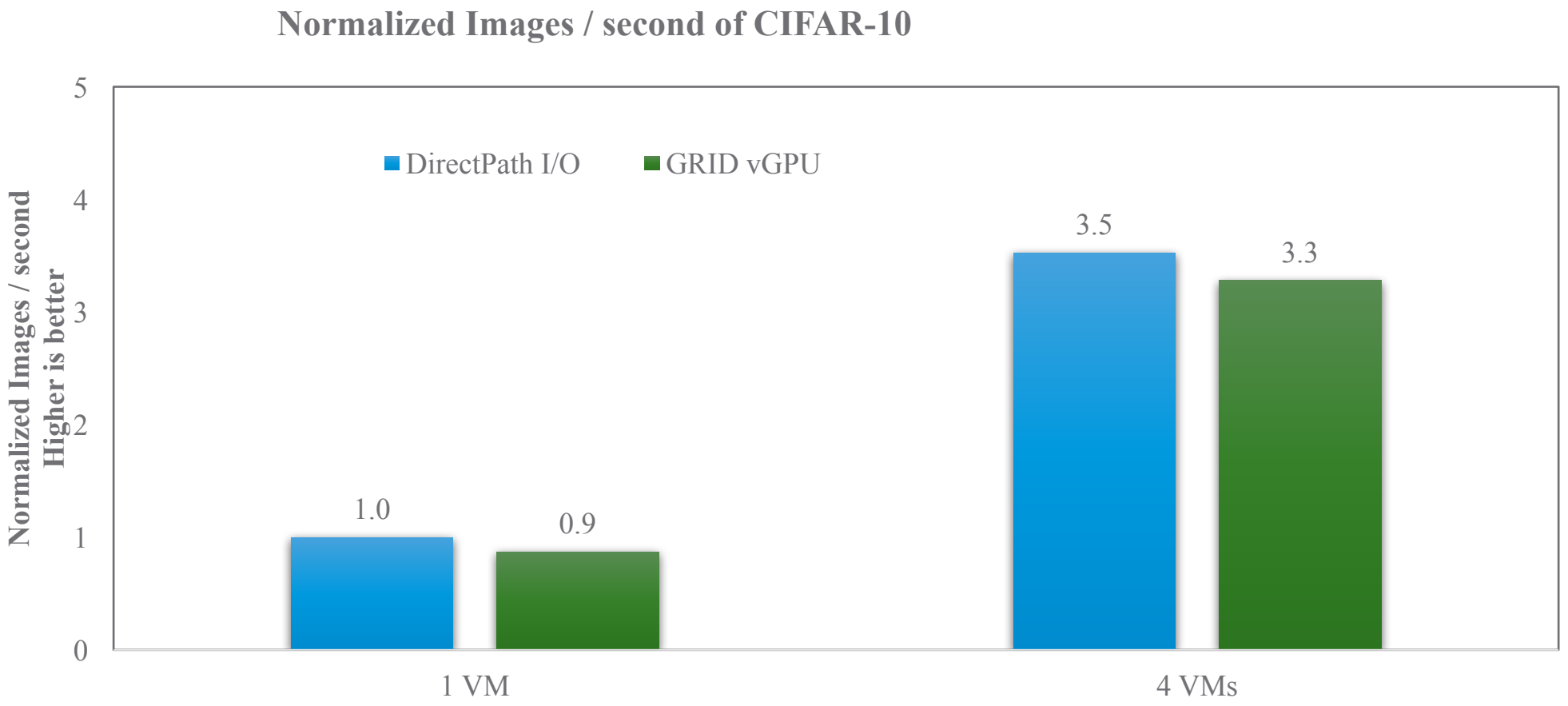
Scalability: CIFAR-10 Workload

- Workload for Convolutional Neural Network
  - Image Classifier
  - 60K images
    - 50K training and 10K image
- Convolutional Neural Network
  - ~ One Million learning parameters
  - ~19 million multiply-add to compute inference on a single image



Scalability – VMs or Users

- Performance Scales almost linearly with number of VMs / users
- Performance of DirectPath I/O and GRID vGPU are comparable

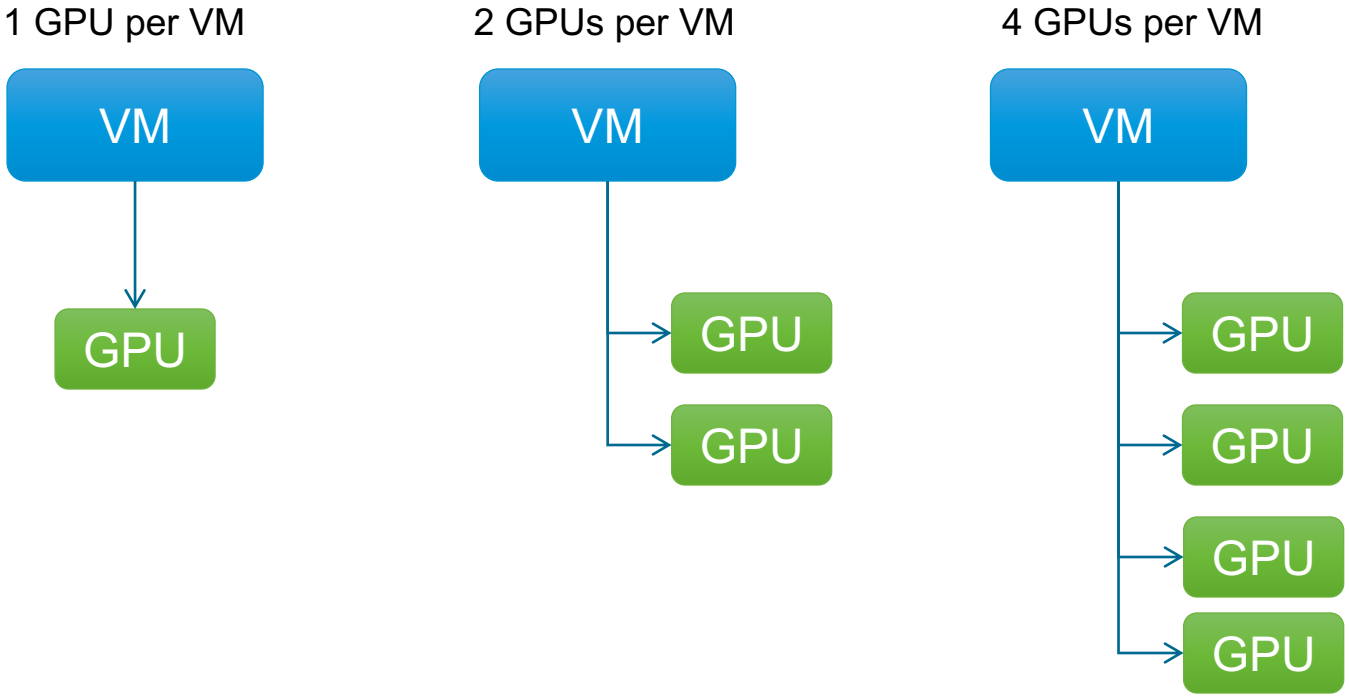


vmware

Scalability - GPUs

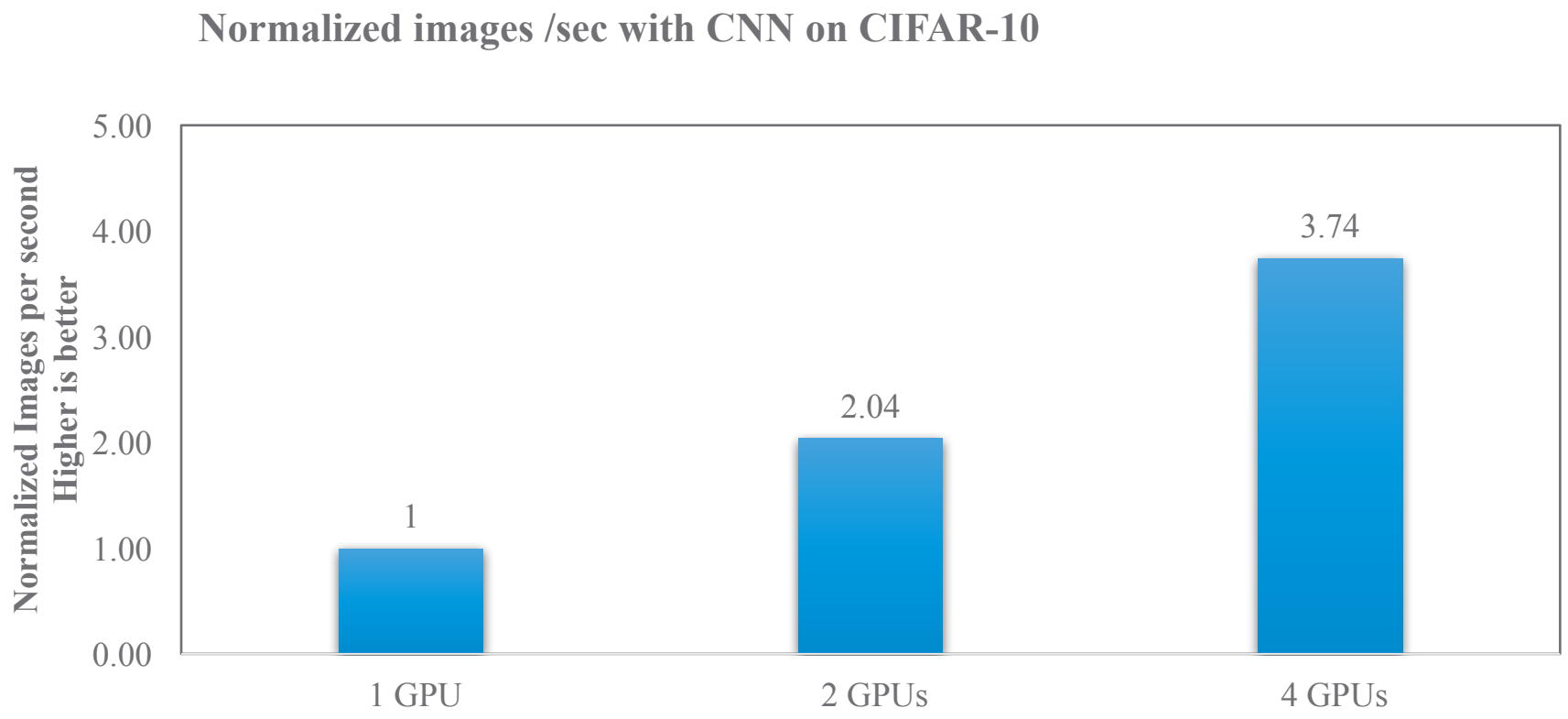
GPU Scalability

- How a ML application performs as increasing number of GPUs up to maximum available on the server



GPU Scalability

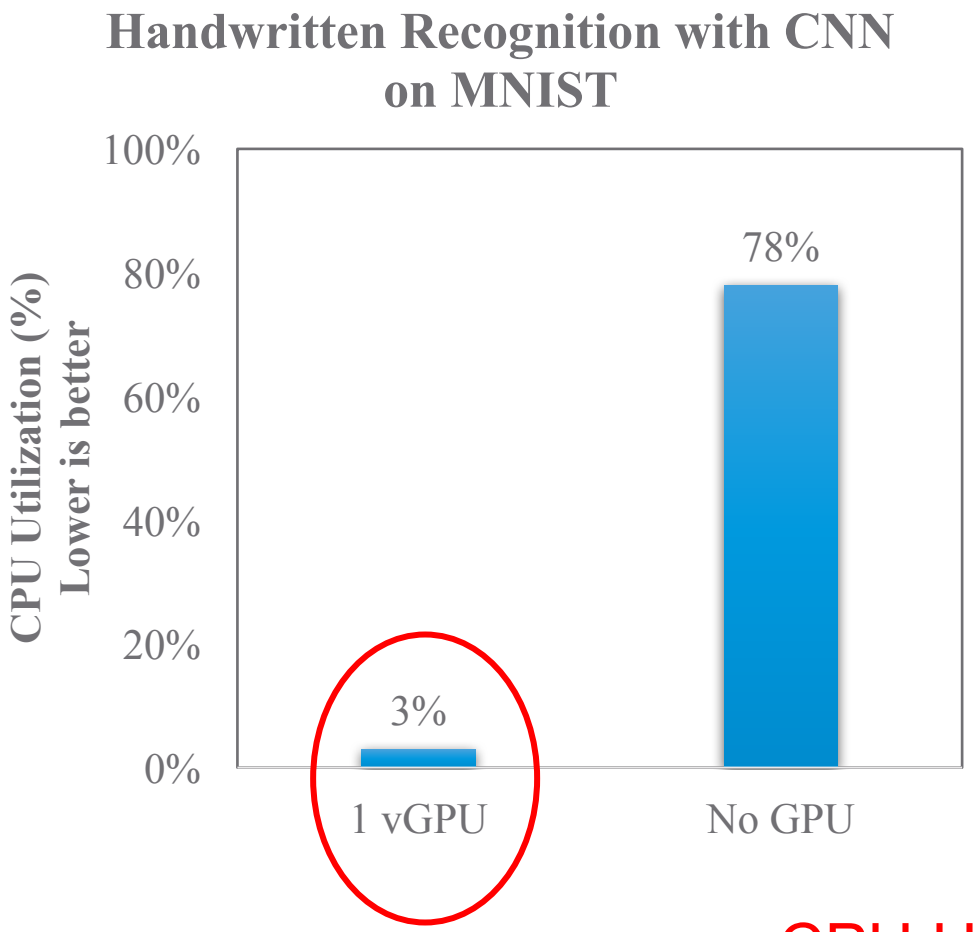
- Scale almost linearly with number of GPUs
- Only DirectPath I/O supports multiple GPUs per VM

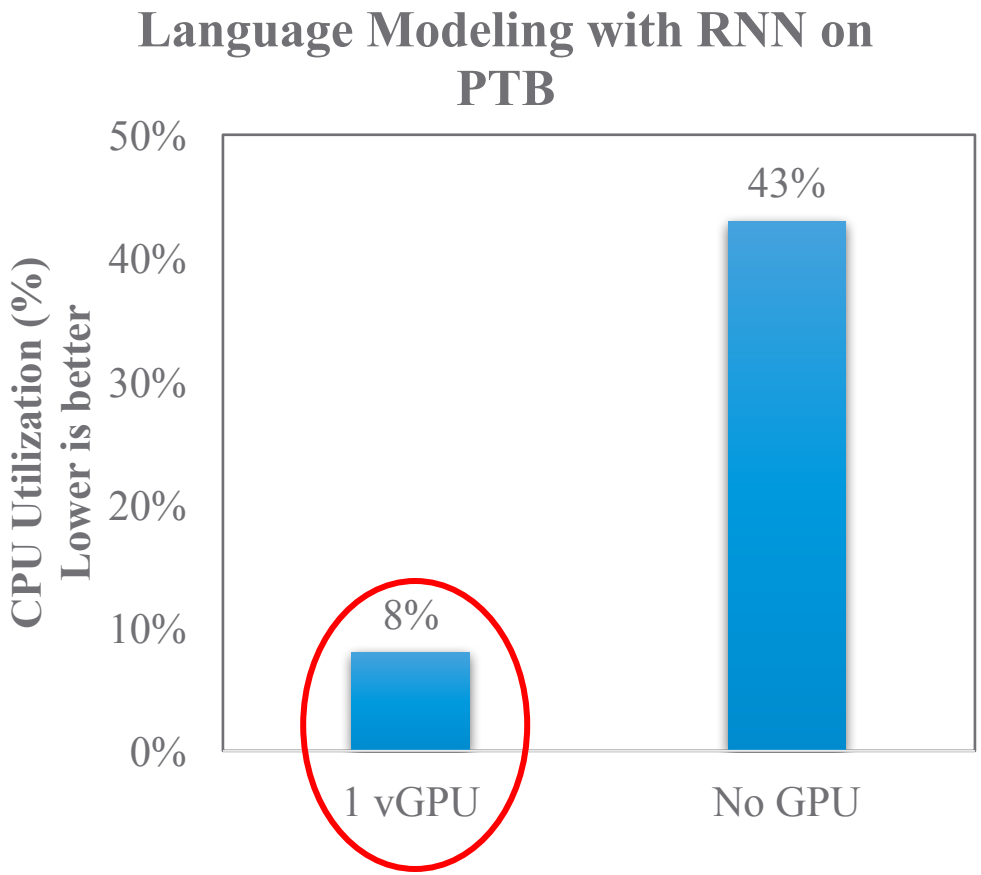


vmware

Resource Utilization

CPU Underutilization with Machine Learning with GPUs





CPU Underutilization

Implications of CPU Under-utilization

- Have more users / VMs per server
  - Mix GPU and non-gpu workloads on the same server
  - Mix Workloads
    - Mix CUDA and Graphics Workloads

Mixed Workloads

Mixed Workloads - Motivation

If you

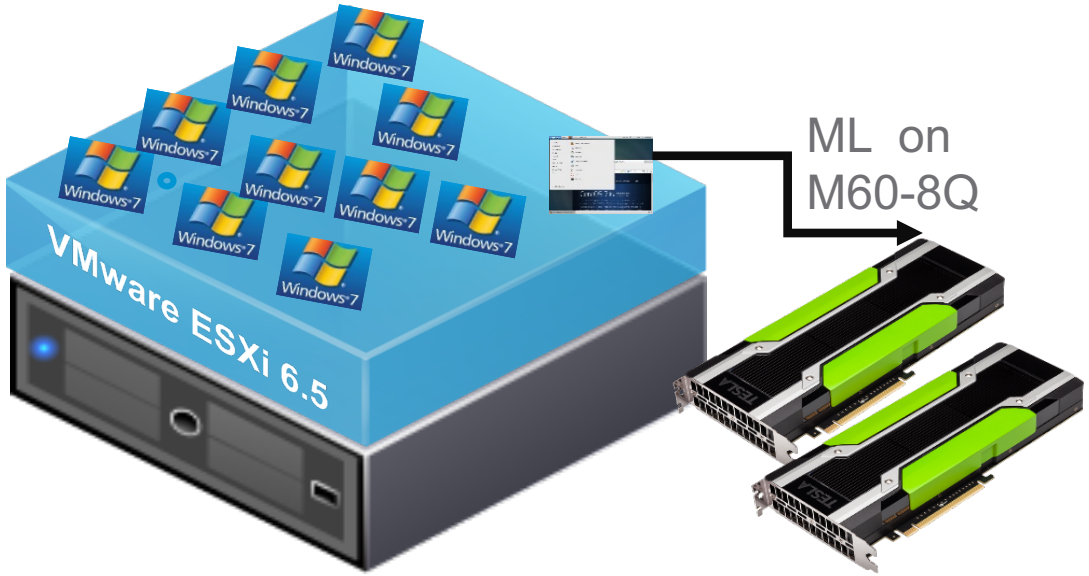
- are setting up an environment with GPUs, consider sharing resources
- have a VDI environment with spare GPU capacity consider sharing resources
- have a setup to run ML, consider sharing GPU resources

WHY?

Better server utilization with minimal performance impact.

Mixed Workloads

Goal: Quantify impact of running 3D CAD & ML on the same server concurrently.



Dell R730 – Intel Haswell CPUs + 2 x NVidia GRID M60
24 cores (2 x 12-core socket) E5-2680 V3
768 GB GB RAM

Benchmarks:

CAD: SPECapc for 3ds Max™ 2015
ML: mnist

CentOS 7.2, **TensorFlow 0.10**
12vCPU, 16 GB RAM, 96GB HD

Win 7, x64, 4 vCPU, 16 GB RAM, 120 GB HD
Autodesk 3ds Max 2015

Mixed Workloads – Experiment Design

1.) First set of runs : Run SPECapc + MNIST concurrently

2.) Second set of runs: Run SPECapc ONLY

3.) Third set of runs: Run MNIST ONLY

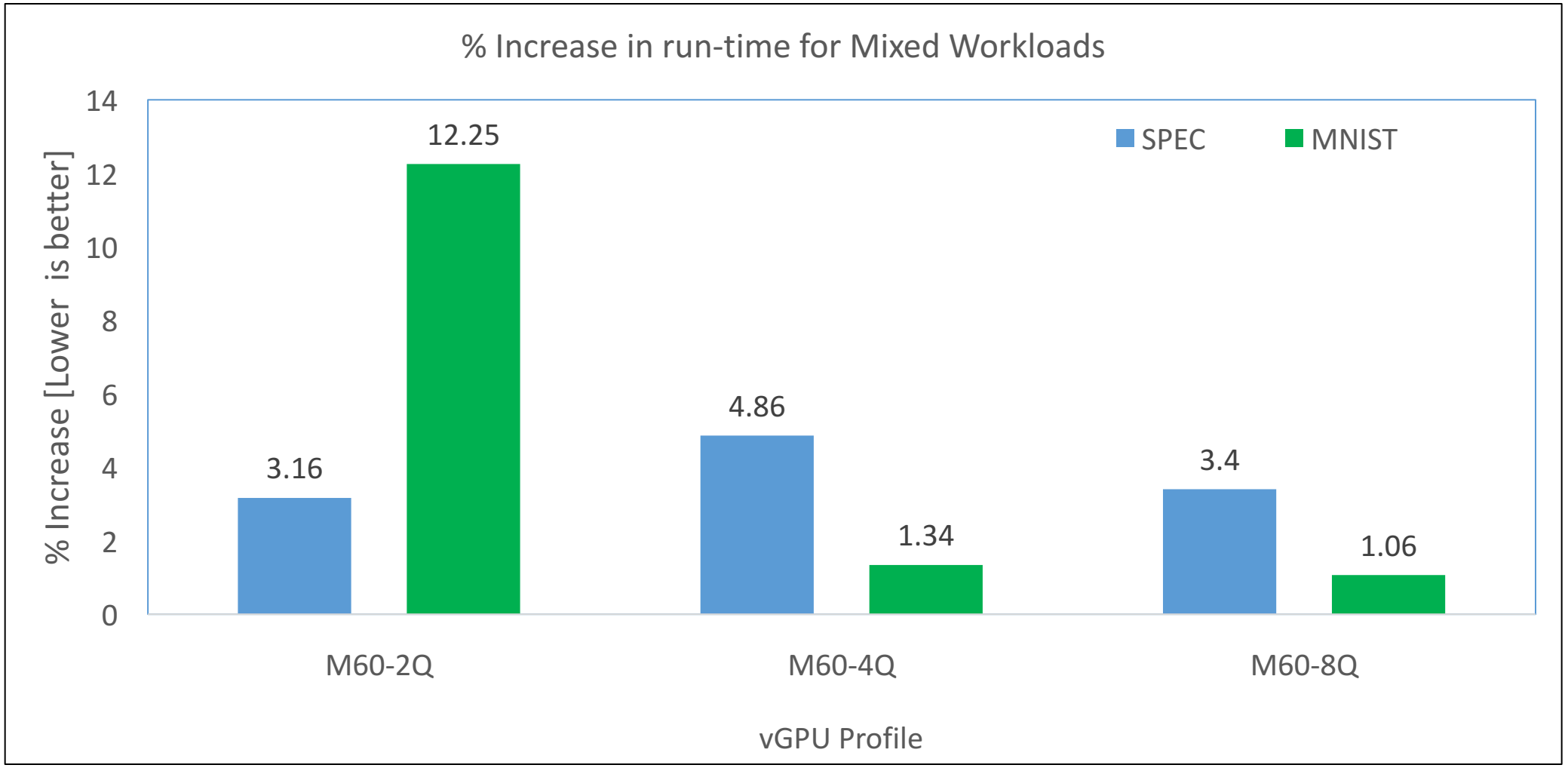
Perf. Metrics:

SPECapc: Geo. Mean of run-times
MNIST: Run time as measured by wall-clock

Please Note:
SPECapc runs on all vGPU profiles
MNIST runs only on M60-8Q profile

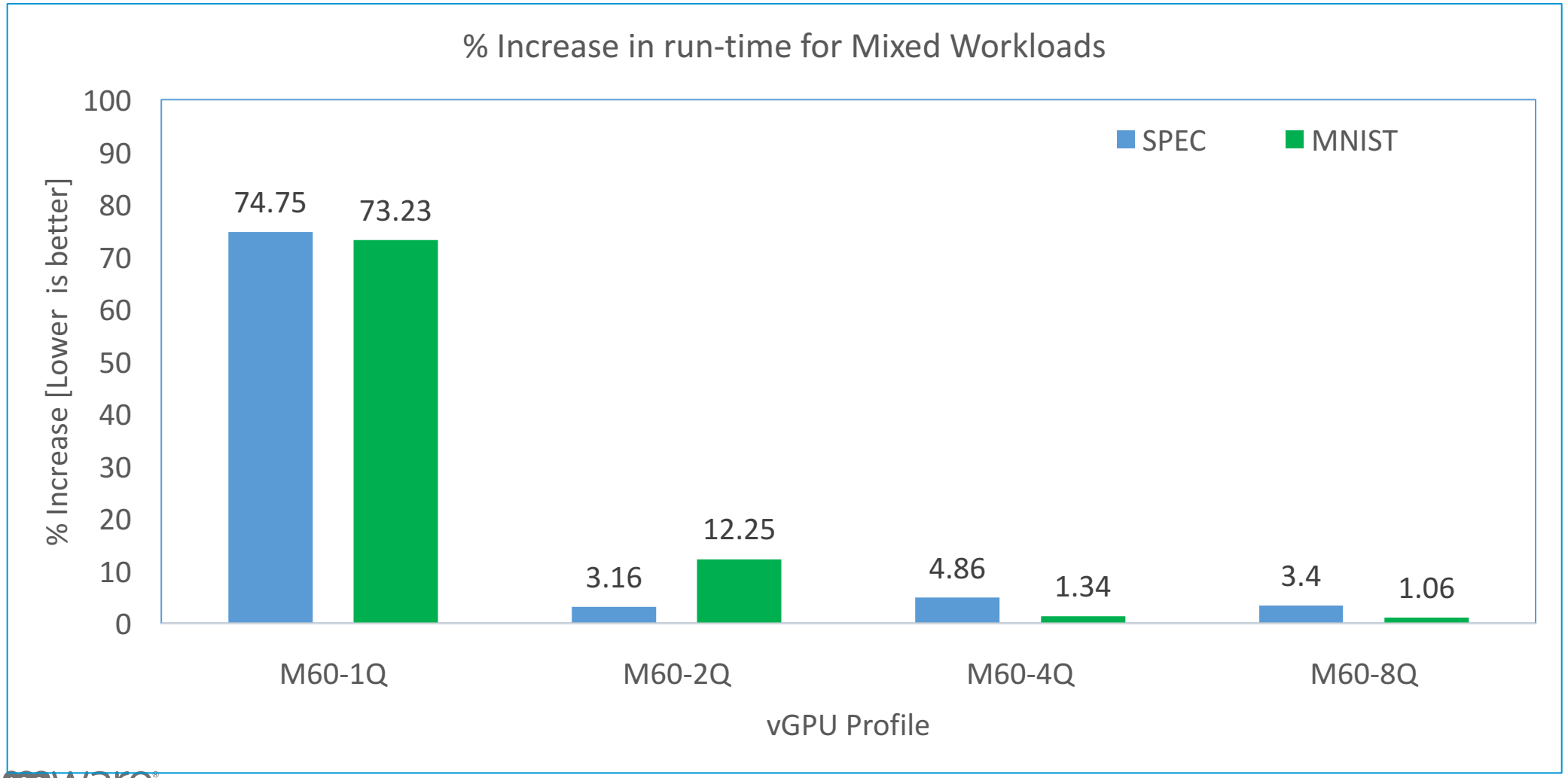
Mixed Workloads - Results

- SPEC < 5% penalty in M60-2Q, M60-4Q, M60-8Q
- MNIST < 15% in M60-2Q, M60-4Q, M60-8Q

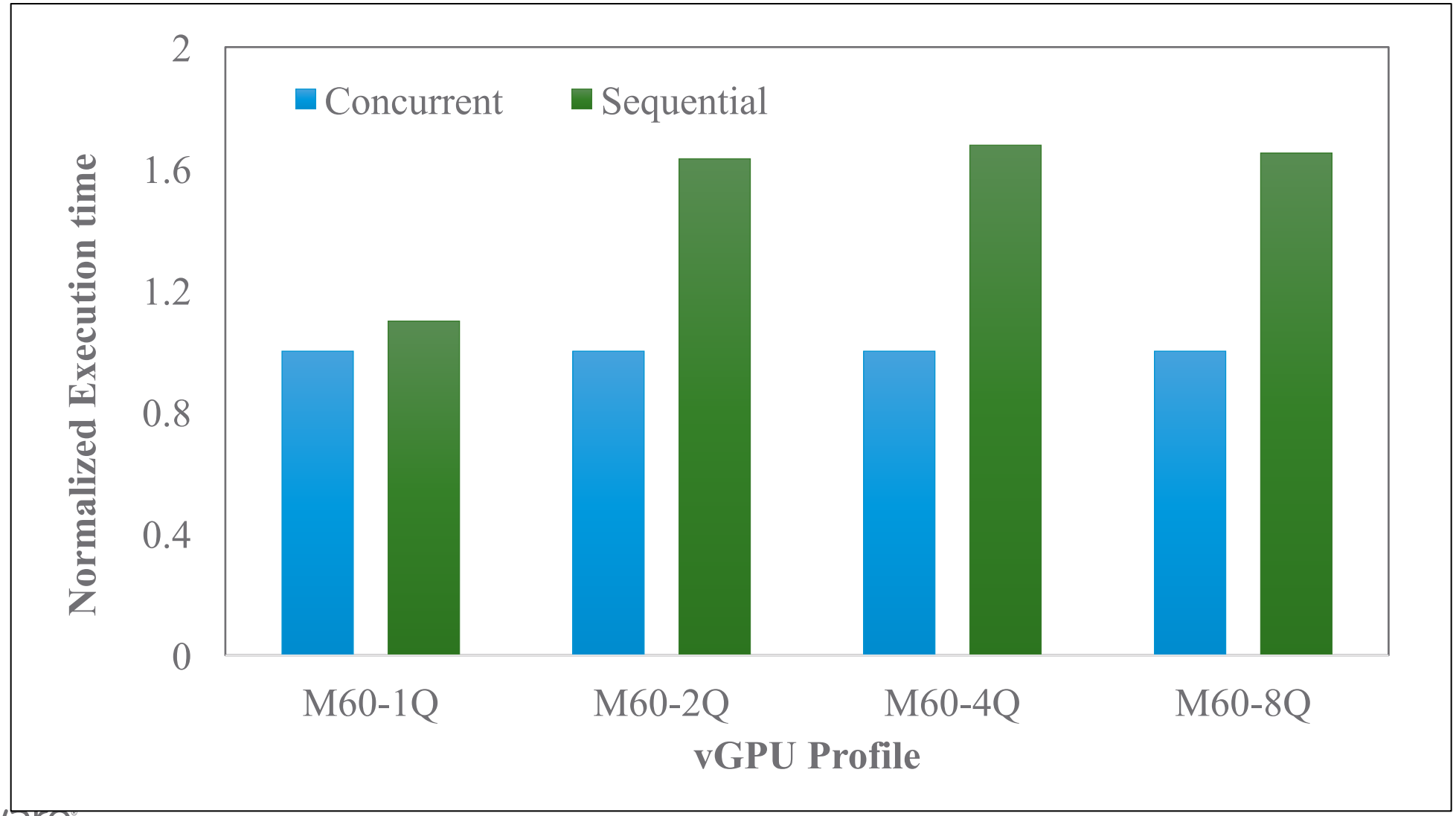


Mixed Workloads - Results

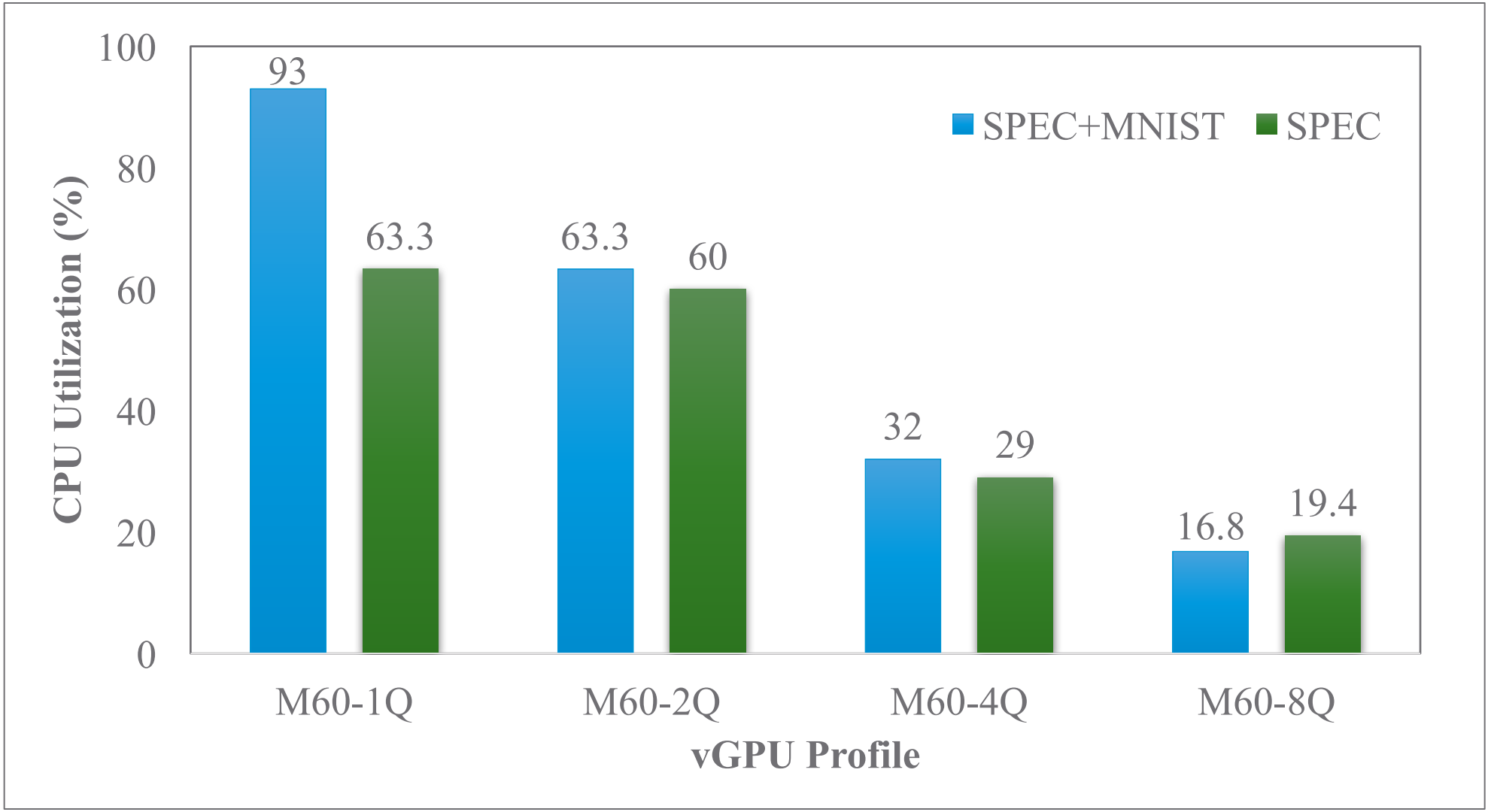
- SPEC < 5% penalty in M60-2Q, M60-4Q, M60-8Q
- MNIST < 15% in M60-2Q, M60-4Q, M60-8Q



Mixed Workloads - Results



Mixed Workloads - Results



Takeaways

- Virtualized GPUs deliver near bare metal Performance for ML workloads
- GPUs can be used in two modes on vSphere: Direct Path IO and NVidia GRID vGPU
- For CUDA/ML workloads, GRID VGPU requires 1 GPU per VM i.e. the highest vGPU profile
- For Multi-GPU ML workloads, use Direct Path IO mode
- For more consolidation of GPU-based workloads, use GRID vGPU
- GRID vGPU combines performance of GPUs and datacenter management benefits of VMware vSphere

Future Work

- Distributed Machine Learning
  - Cluster with Multi-Hosts & Multi-GPUs
- Different Machine Learning Frameworks
  - Bidmach
  - Caffe
  - Torch
  - Theano
  - TensorFlow
- Performance Studies of Containers with GPUs

Q&A

- Thank you !
- Contact

Uday Kurkure

Lan Vu

Hari Sivaraman

{ukurkure,lanv,hsivaraman}@vmware.com

- Thanks to our colleagues
  - Aravind Bappanadu, Ravi Soundararajan, Ziv Kalmanovich, Reza Taheri, Vincent Lin