vMotion for NVIDIA GRID vGPU Virtual Machines: Case Study of vMotion Using MLaaS

Hari Sivaraman, Dimitrios Skarlatos
Lan Vu, Uday Kurkure

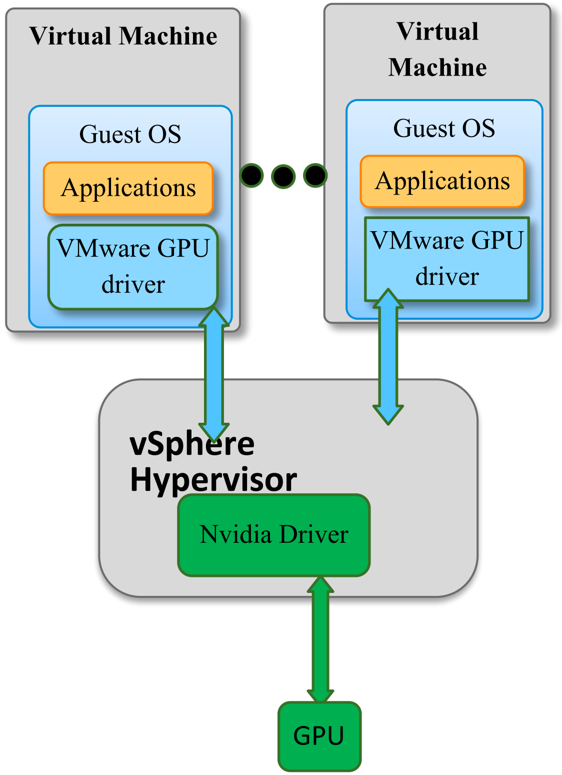
GTC 2019

vMotion for NVIDIA GRID vGPU - Agenda

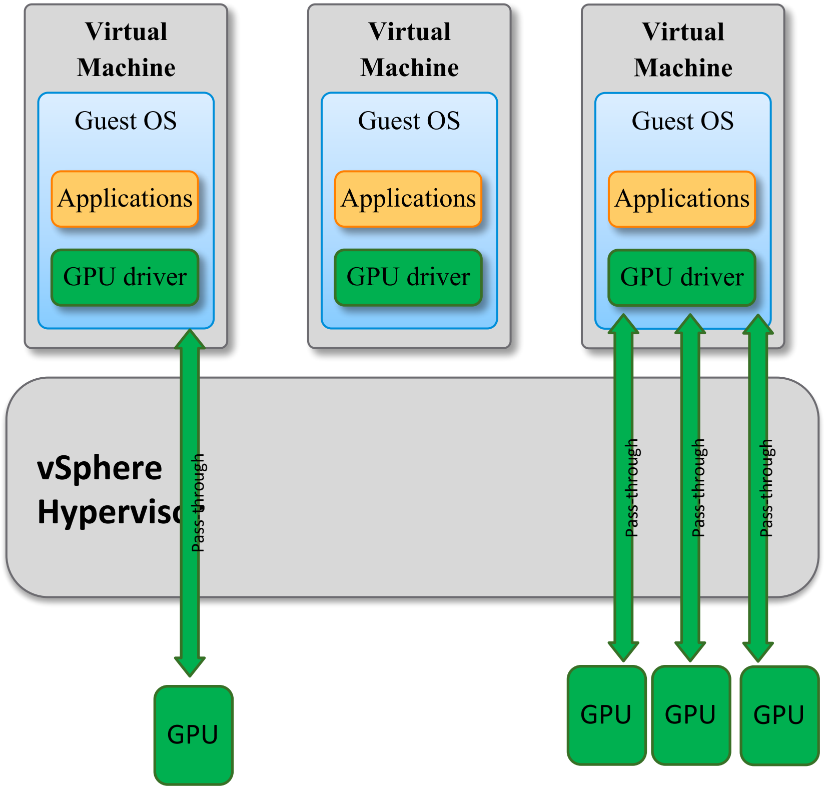
- GPUs in vSphere.
- vMotion for vGPU Architecture.
- Performance of vMotion for vGPU.
- MLaaS – a case study for vMotion performance.
- Conclusions and future work.

vMotion for NVIDIA GRID vGPU – GPUs in vSphere

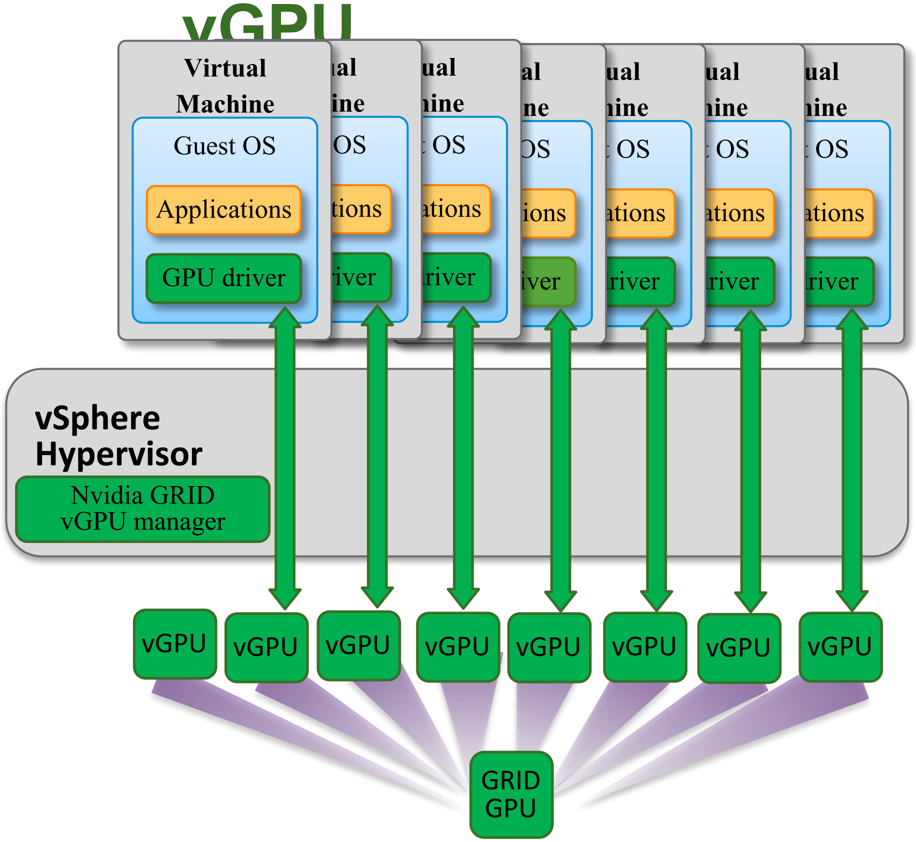
vSGA



VMware DirectPath I/O



Nvidia GRID vGPU



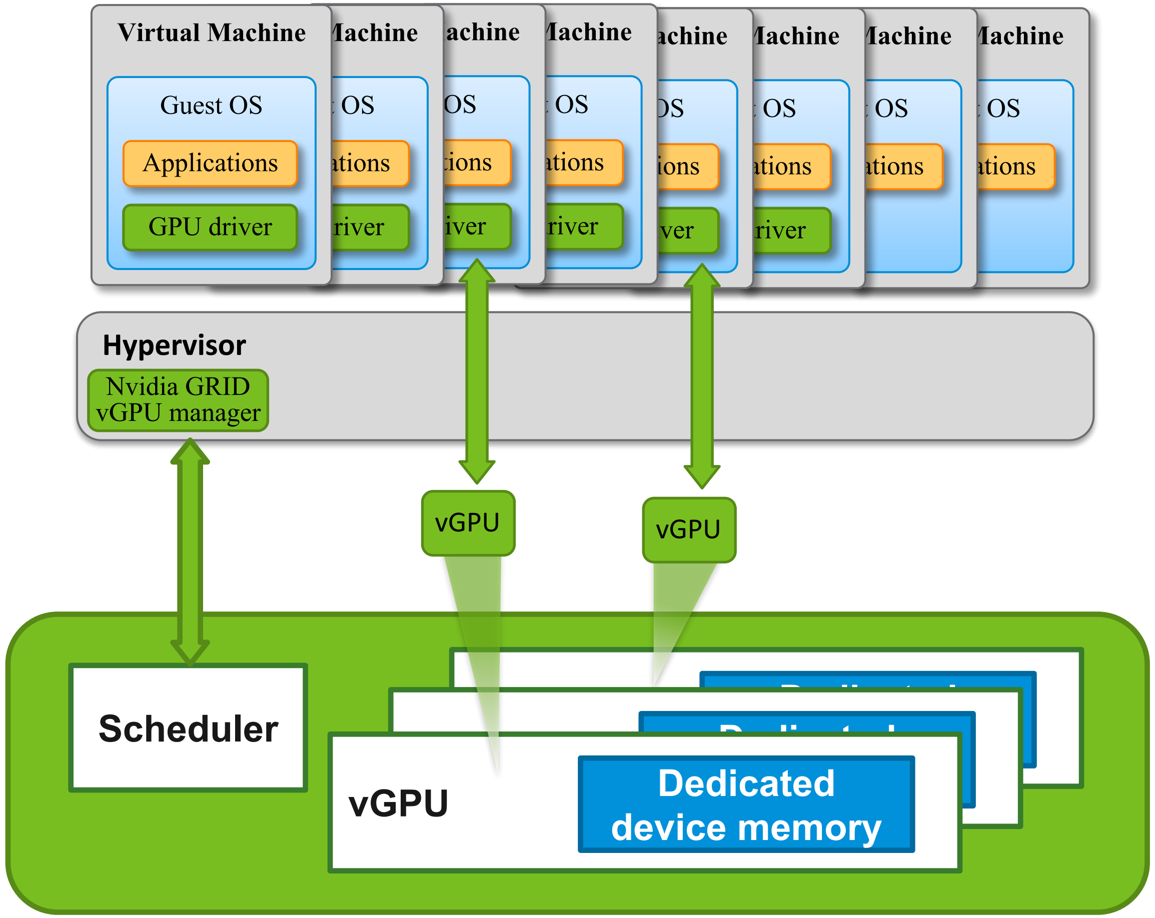
☑
☑ vMotio
n
Sharin

☒
☒ vMotio
n
Sharin

☑
☑ vMotio
n
Sharin

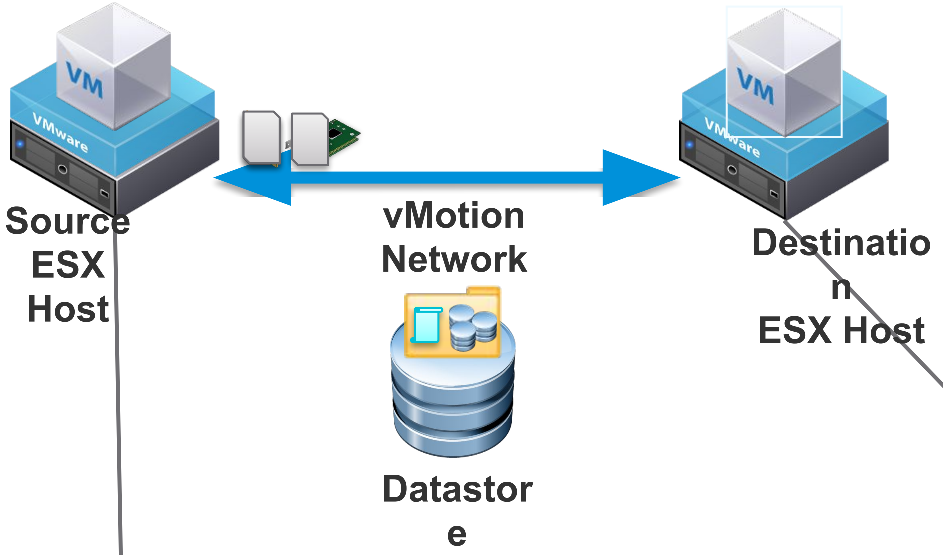
vMotion for NVIDIA GRID vGPU – vGPU

Nvidia GRID vGPU



- GPU Memory is statically shared
- CUDA cores are time-shared
- GPU memory per VM is called vGPU Profile
- **For example**: P40-1q profile for P40 GPU
  - vGPU has 1GB of device memory
  - **24** vGPUs per **1** physical P40

vMotion for NVIDIA GRID vGPU – Types of vMotion

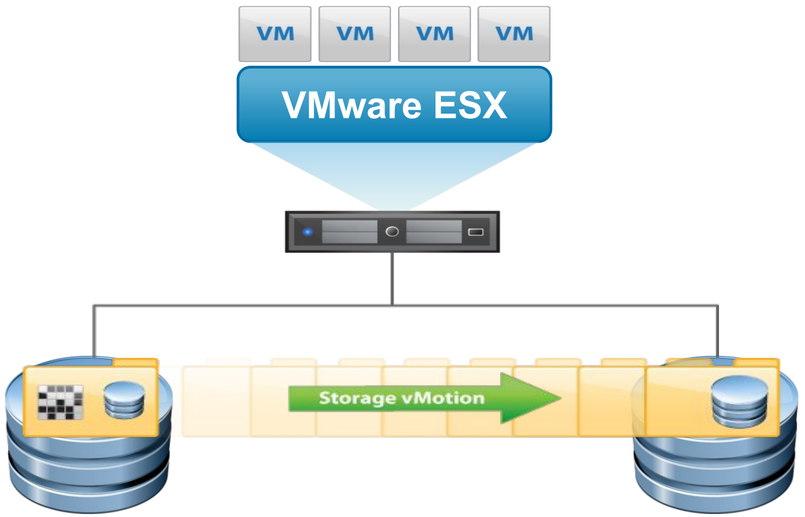


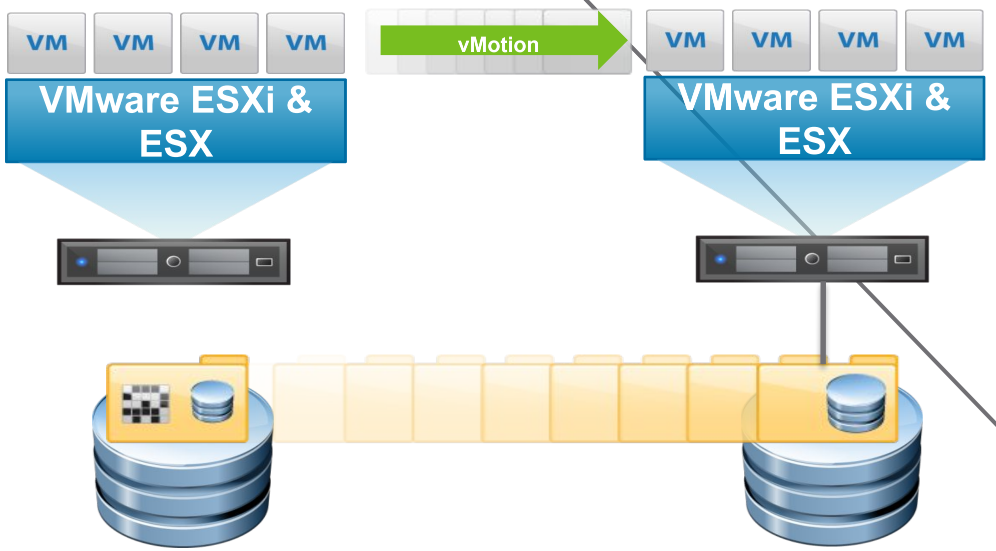
Source
ESX
Host

vMotion
Network

Destinatio
n
ESX Host

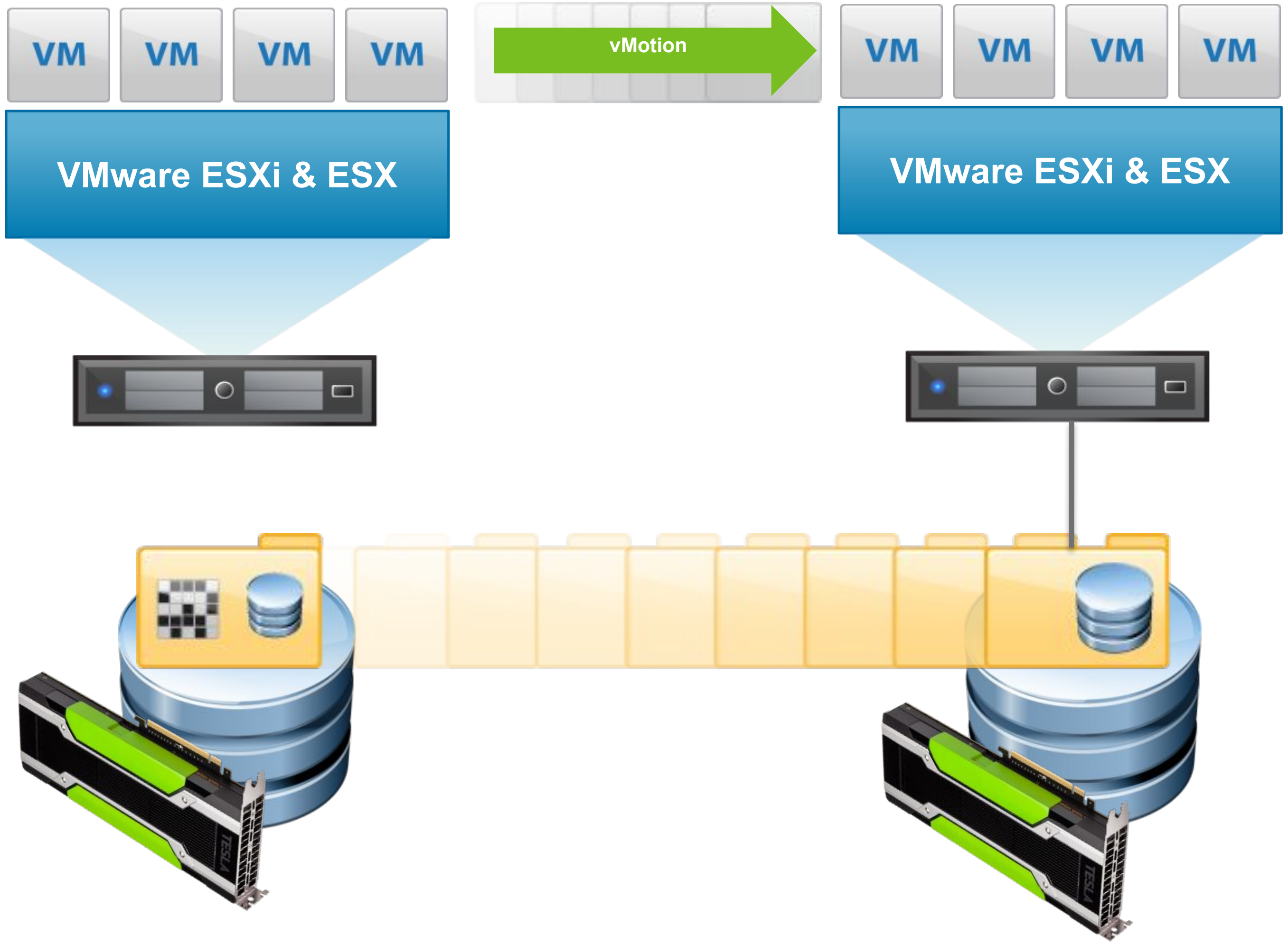
Datastor
e





vMotion for NVIDIA GRID vGPU – vMotion

1. pre-copy memory pages
2. Stun the VM
3. Checkpoint devices
4. Xfer device checkpoint data (includes vGPU memory data)
5. **Power on VM & xfer pages from main memory**



vMotion for NVIDIA GRID vGPU - Agenda

- GPUs in vSphere.

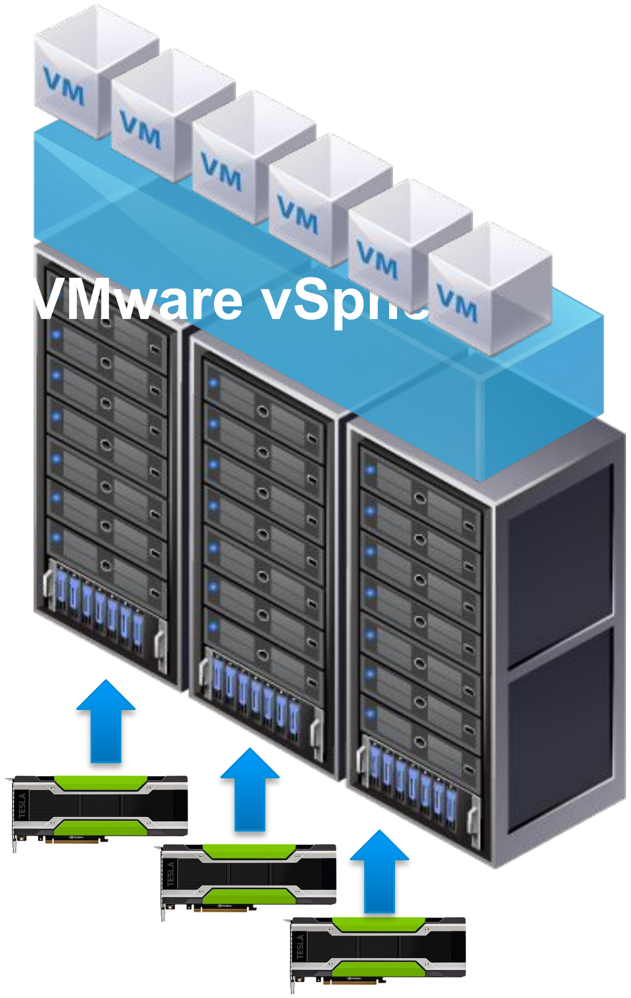
- vMotion for vGPU Architecture.

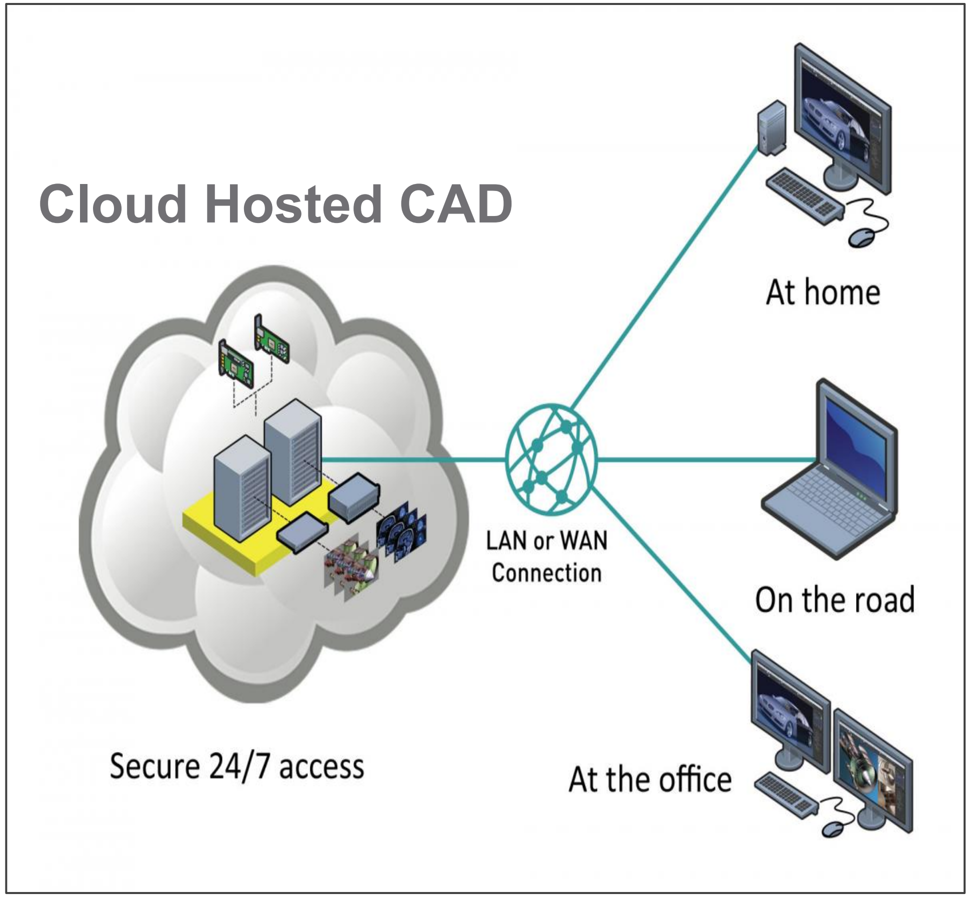
- **Performance of vMotion for vGPU.**

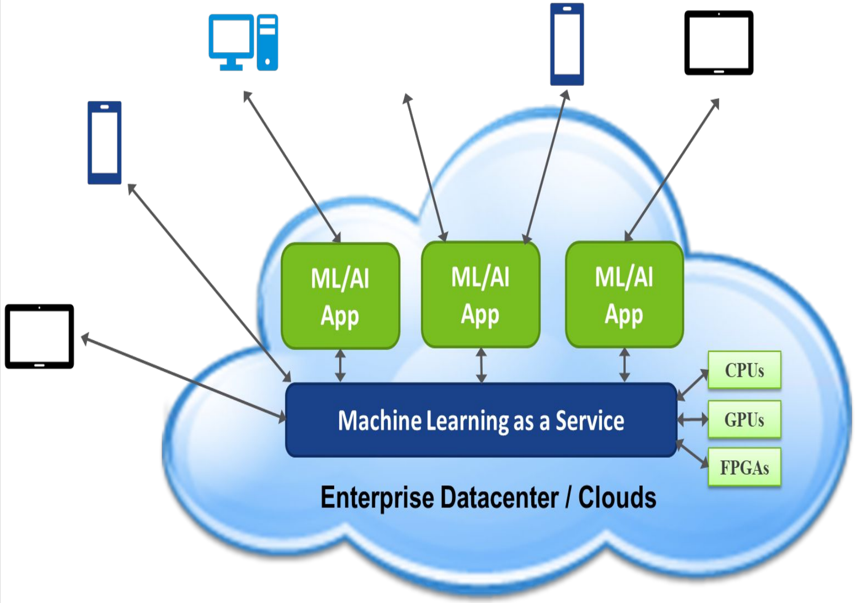
- **MLaaS – a case study for vMotion performance.**

- **Conclusions and future work.**

vMotion for NVIDIA GRID vGPU - Workloads





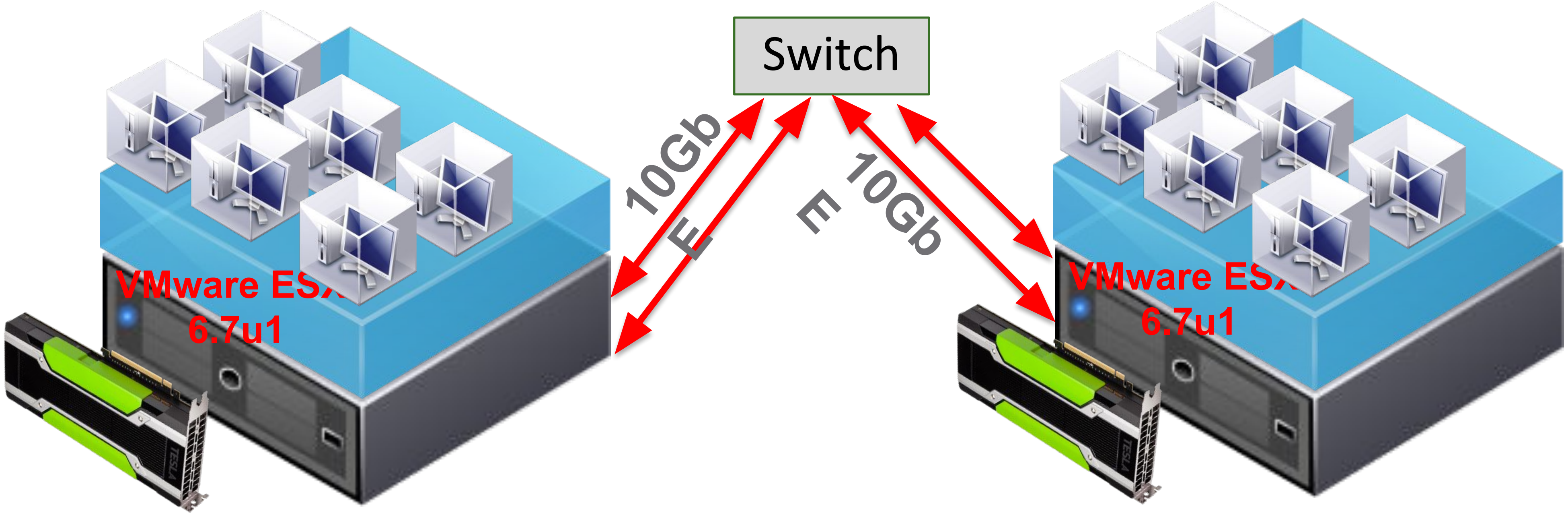


VD
I

Cloud Hosted
CAD

MLaa
S

vMotion for NVIDIA GRID vGPU – Test-bed

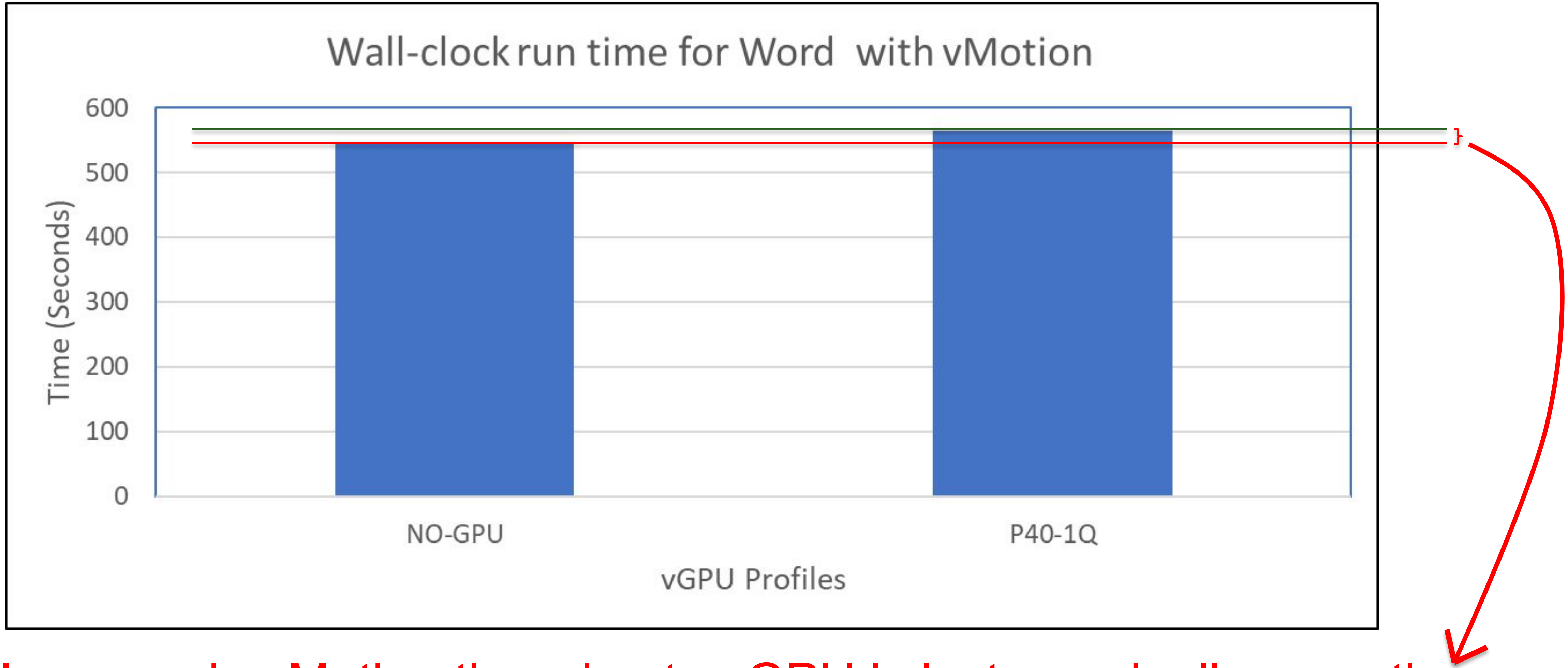


Dell R730 – Intel Broadwell CPUs + 1 x NVidia GRID P40
40 cores (2 x 20-core socket) E5-2698 v4
768 GB RAM

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

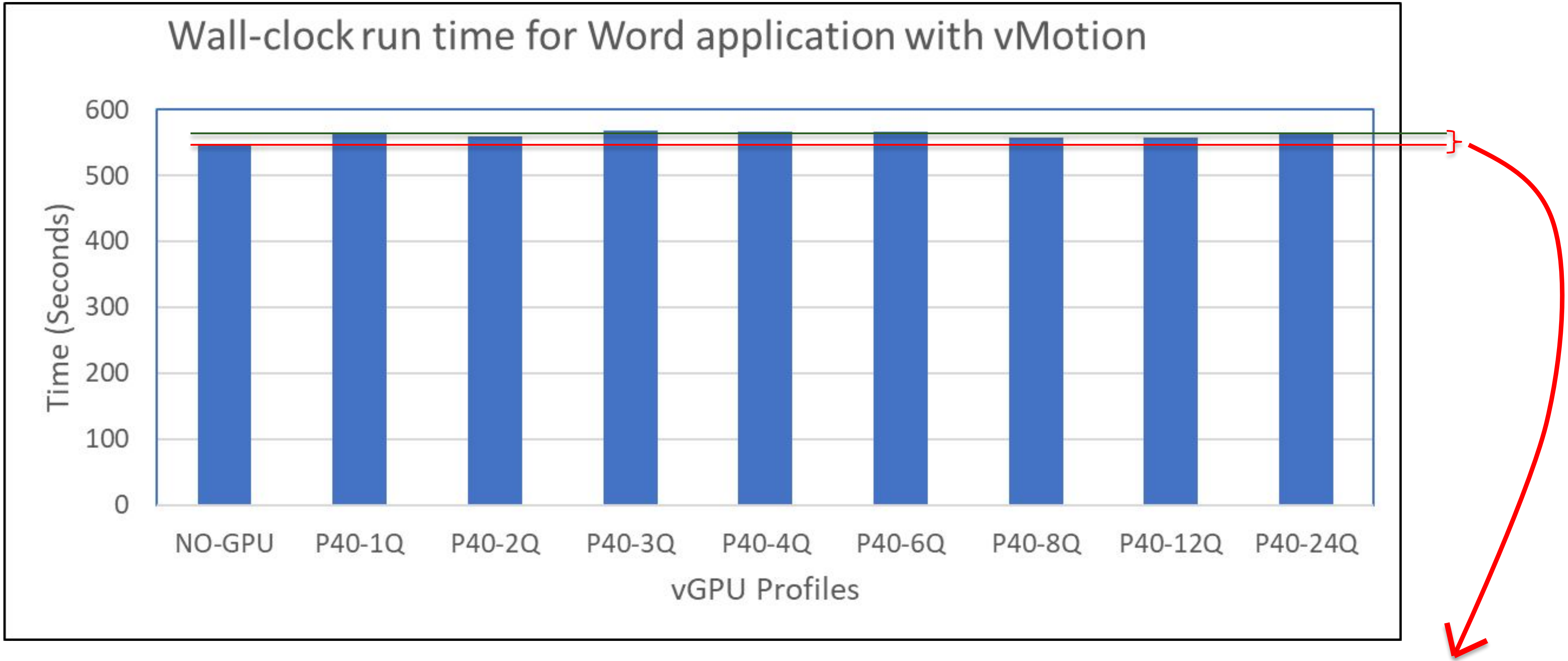
- **ESX**: 6.7u1 **Nvidia Driver**: 410.68

vMotion for NVIDIA GRID vGPU – Performance of Word



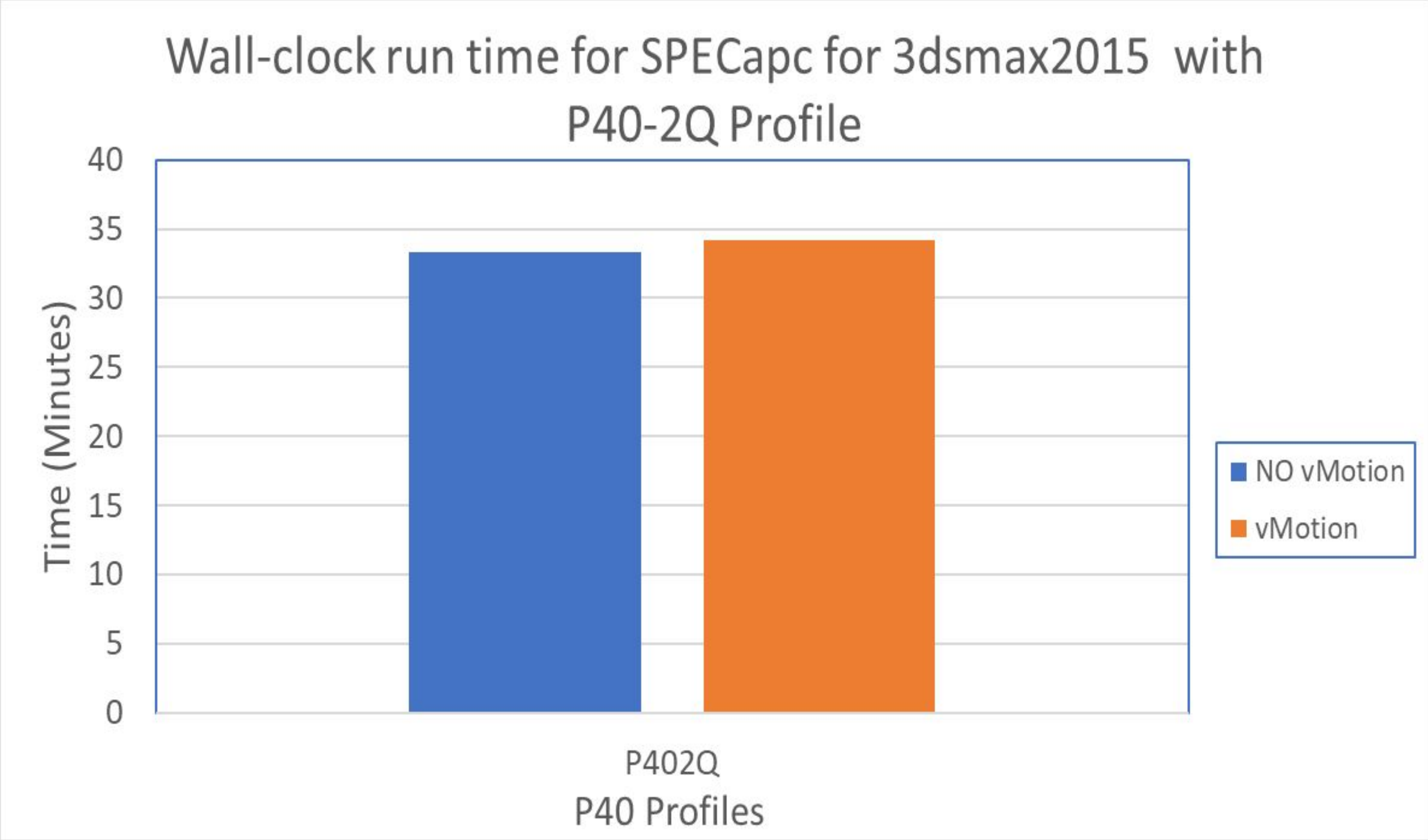
Increase in vMotion time due to vGPU is just marginally more than measurement noise.

vMotion for NVIDIA GRID vGPU – Performance of Word



Increase in vMotion time due to vGPU is just marginally more than measurement noise.

vMotion for NVIDIA GRID vGPU – Performance of SPECapc for 3dsmax 2015

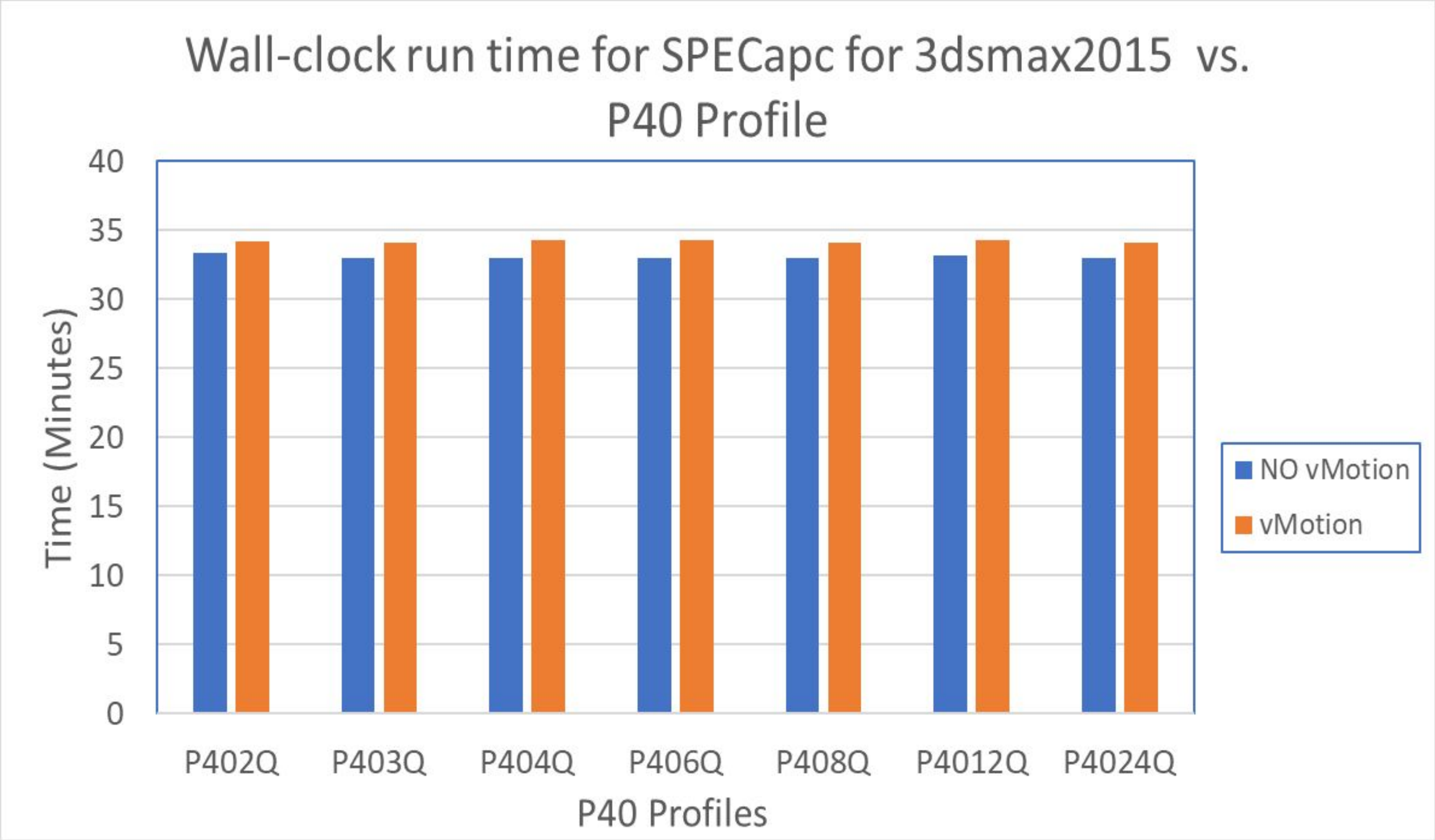


Benchmark: SPEcapc for 3dsmask 2015

Software: Autodesk 3dsmax 2015

Negligible increase in run-time due to vMotion!

vMotion for NVIDIA GRID vGPU – Performance of SPECapc for 3dsmax 2015

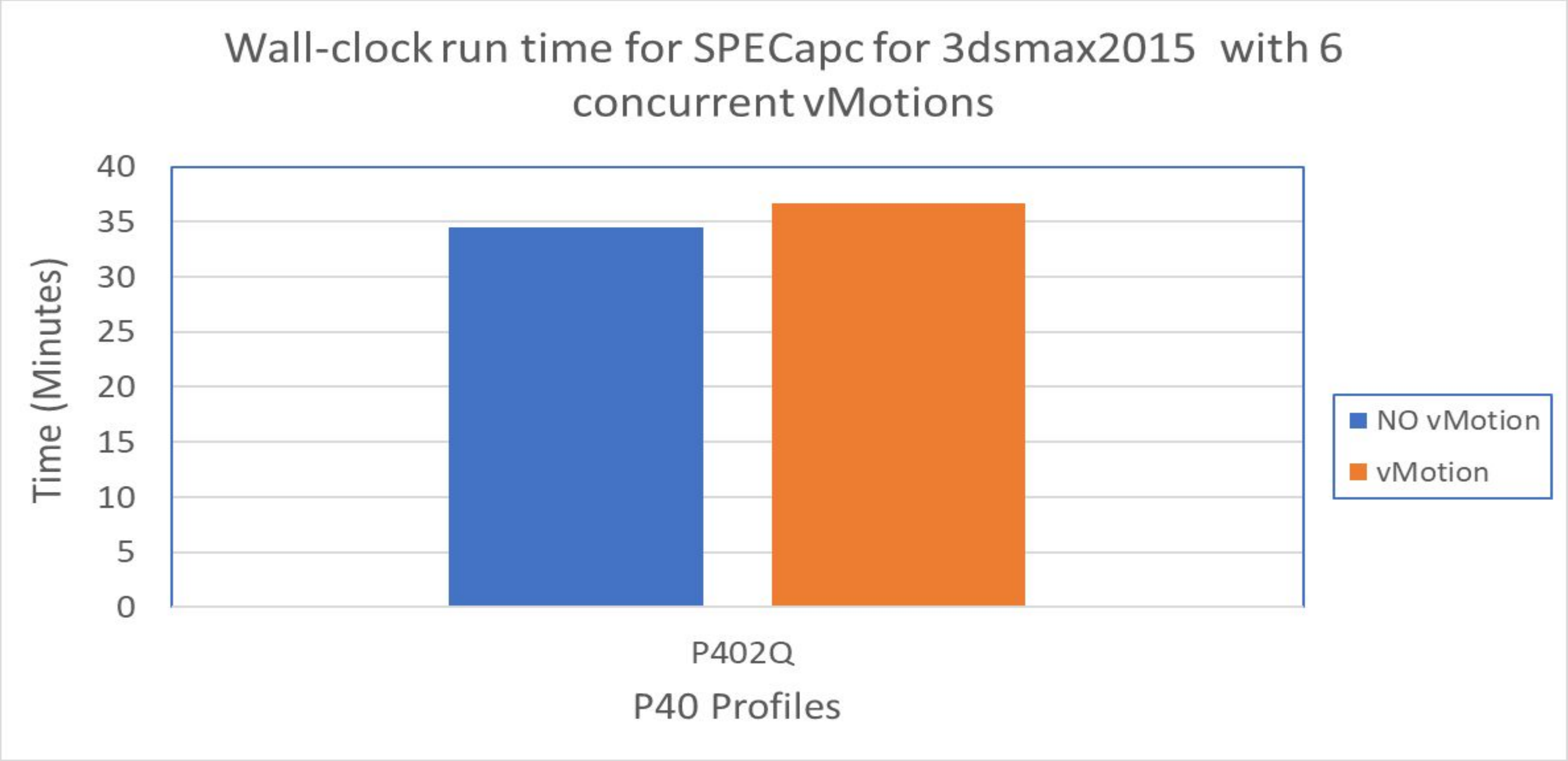


Benchmark: SPEcapc for 3dsmask 2015

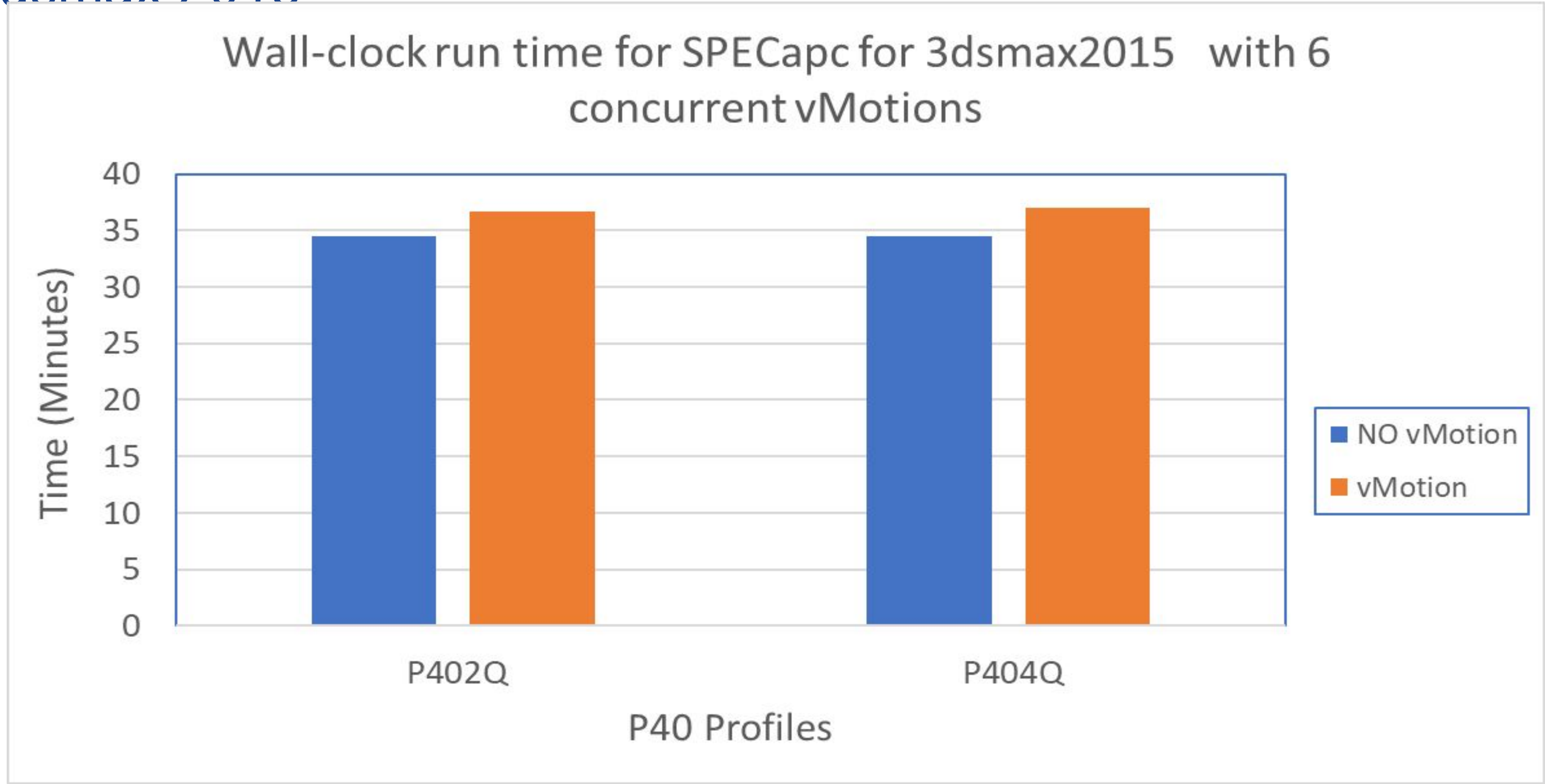
Software: Autodesk 3dsmax 2015

Negligible increase in run-time due to vMotion!

vMotion for NVIDIA GRID vGPU – Performance of SPECapc for 3dsmax 2015



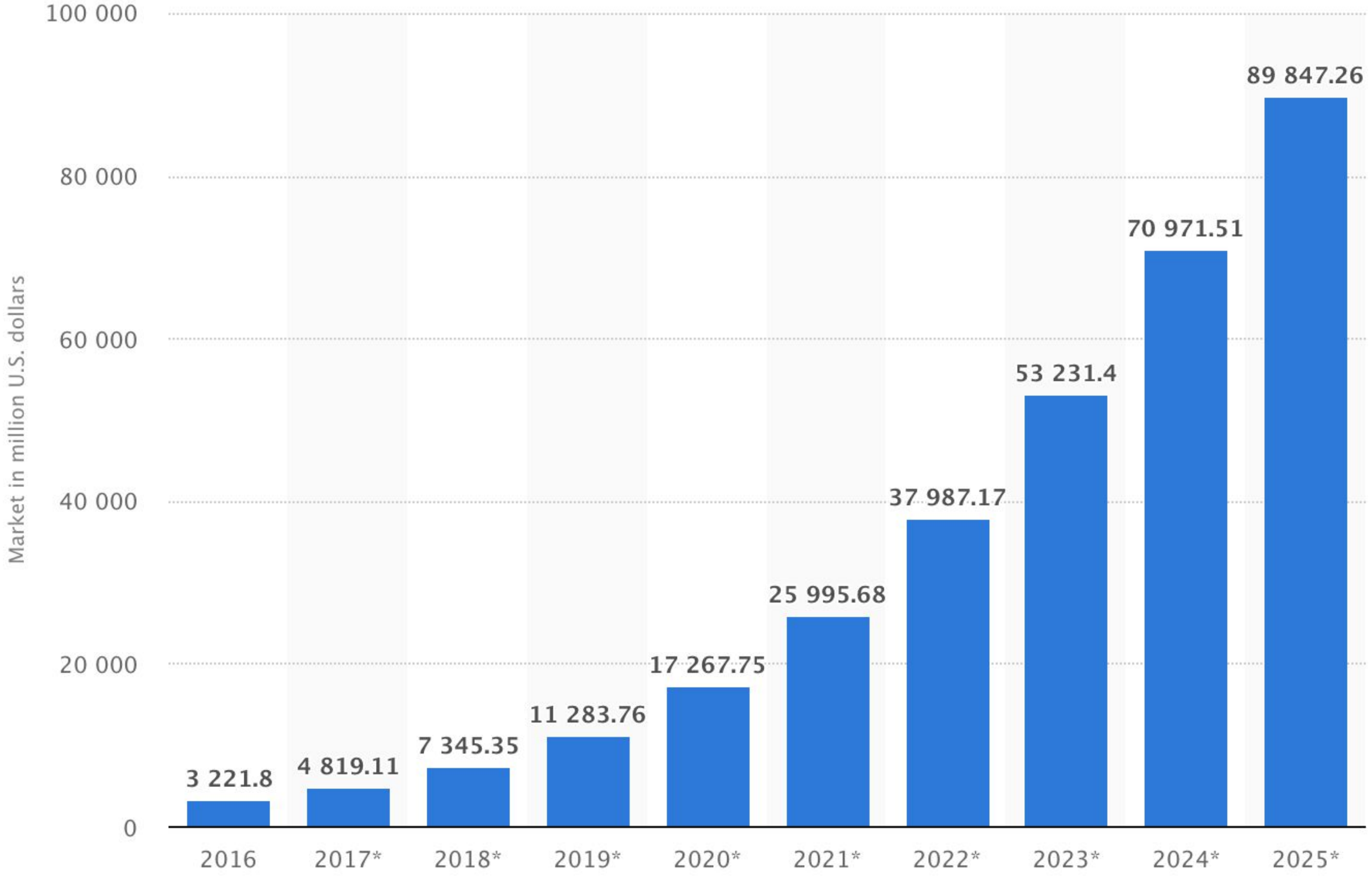
vMotion for NVIDIA GRID vGPU – Performance of SPECapc for 3dsmax 2015



vMotion for NVIDIA GRID vGPU - Agenda

- GPUs in vSphere.
- vMotion for vGPU Architecture.
- Performance of vMotion for vGPU.
- **MLaaS – a case study for vMotion performance.**
- **Conclusions and future work.**

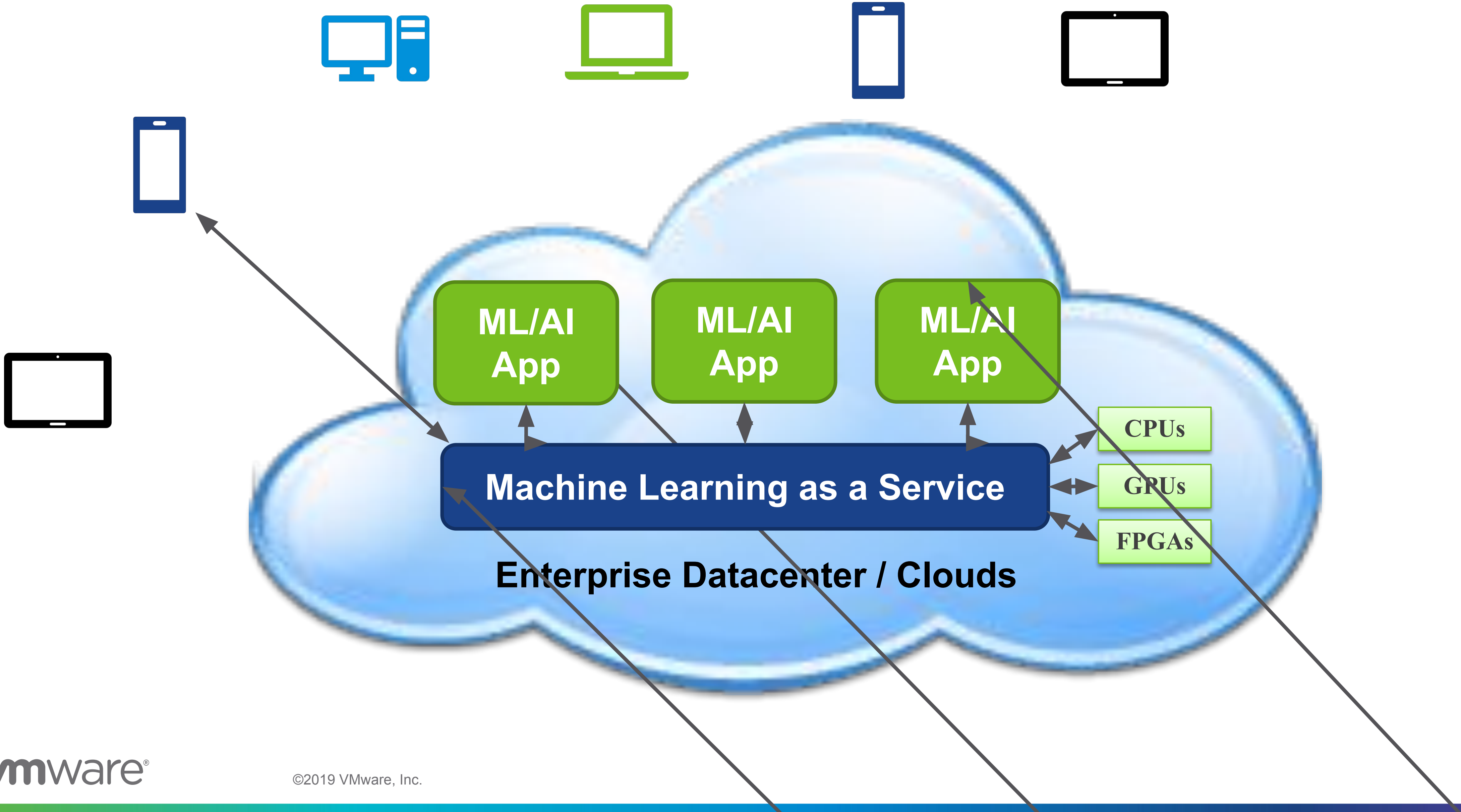
Revenues from the Artificial Intelligence (AI) market worldwide from 2016 to 2025



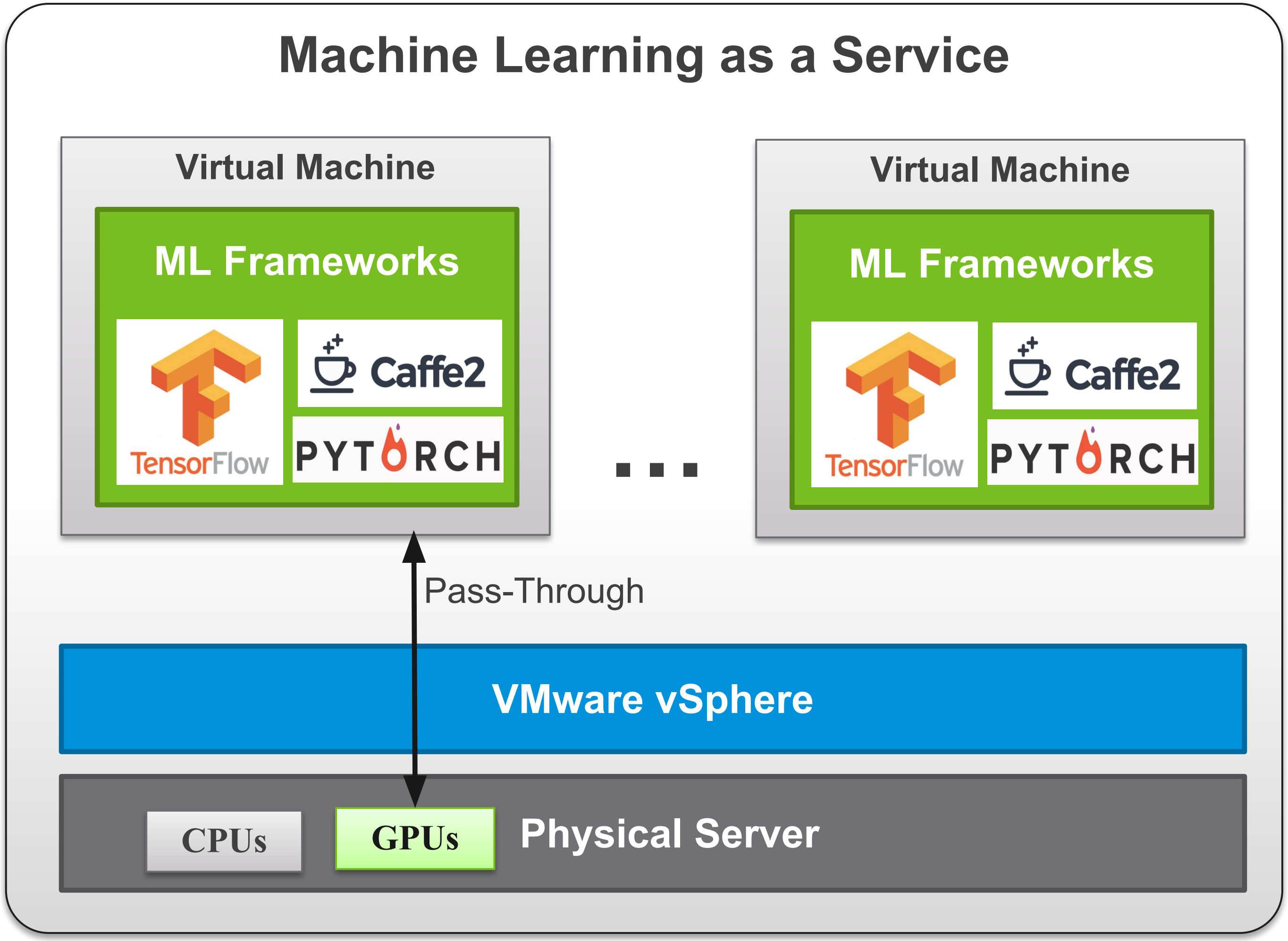
The largest proportion of revenues come from the ML/AI Enterprise Applications

© Statista 2019

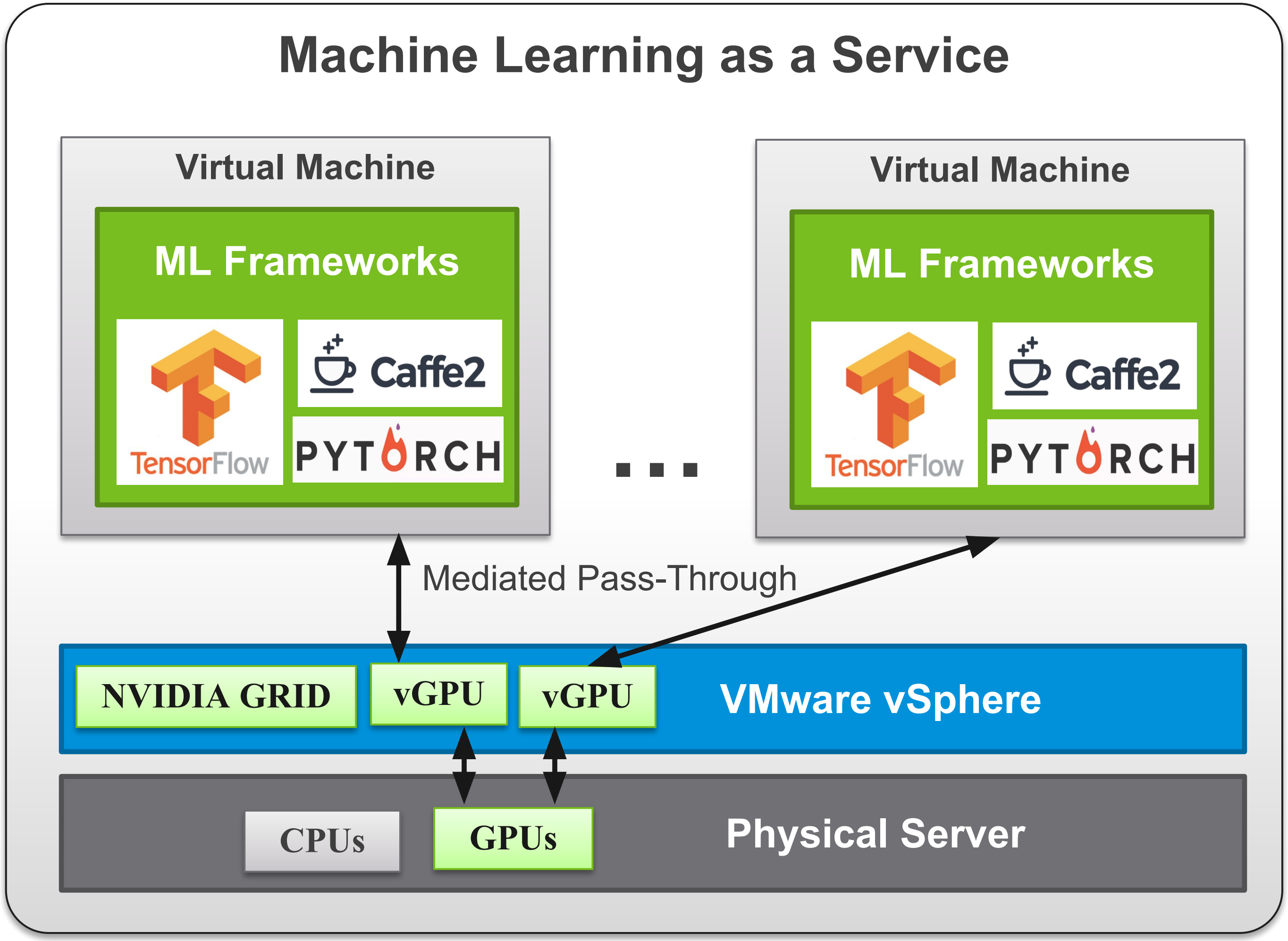
ML/AI Enterprise Application Deployment



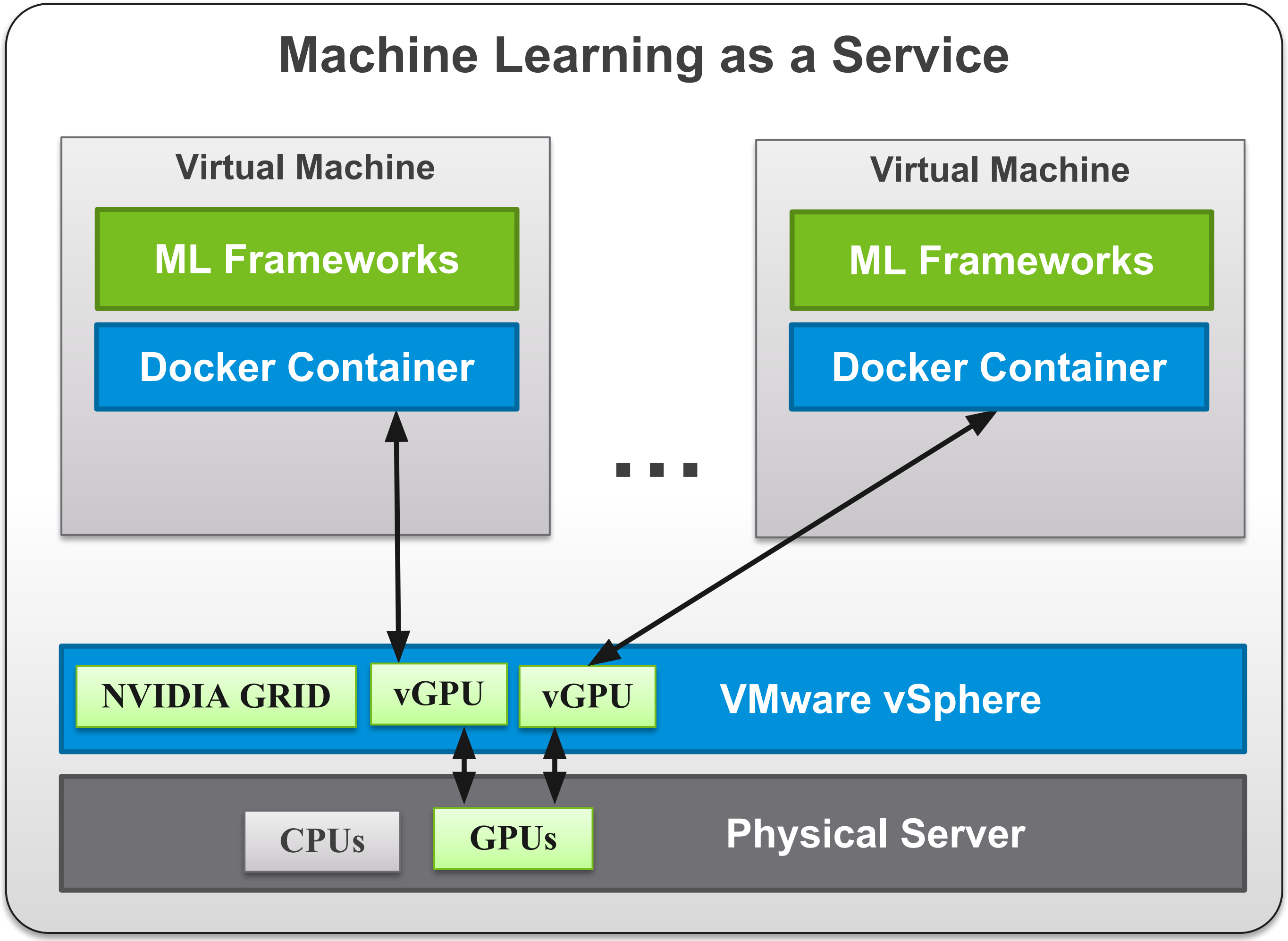
Example #1 of deploying MLaaS on VMware vSphere



Example #2 of deploying MLaaS on VMware vSphere



Example #3 of deploying MLaaS on VMware vSphere with Container



Machine Learning as a Service

Virtual Machine

ML Frameworks

Docker Container

Virtual Machine

ML Frameworks

Docker Container

...

NVIDIA GRID

vGPU

vGPU

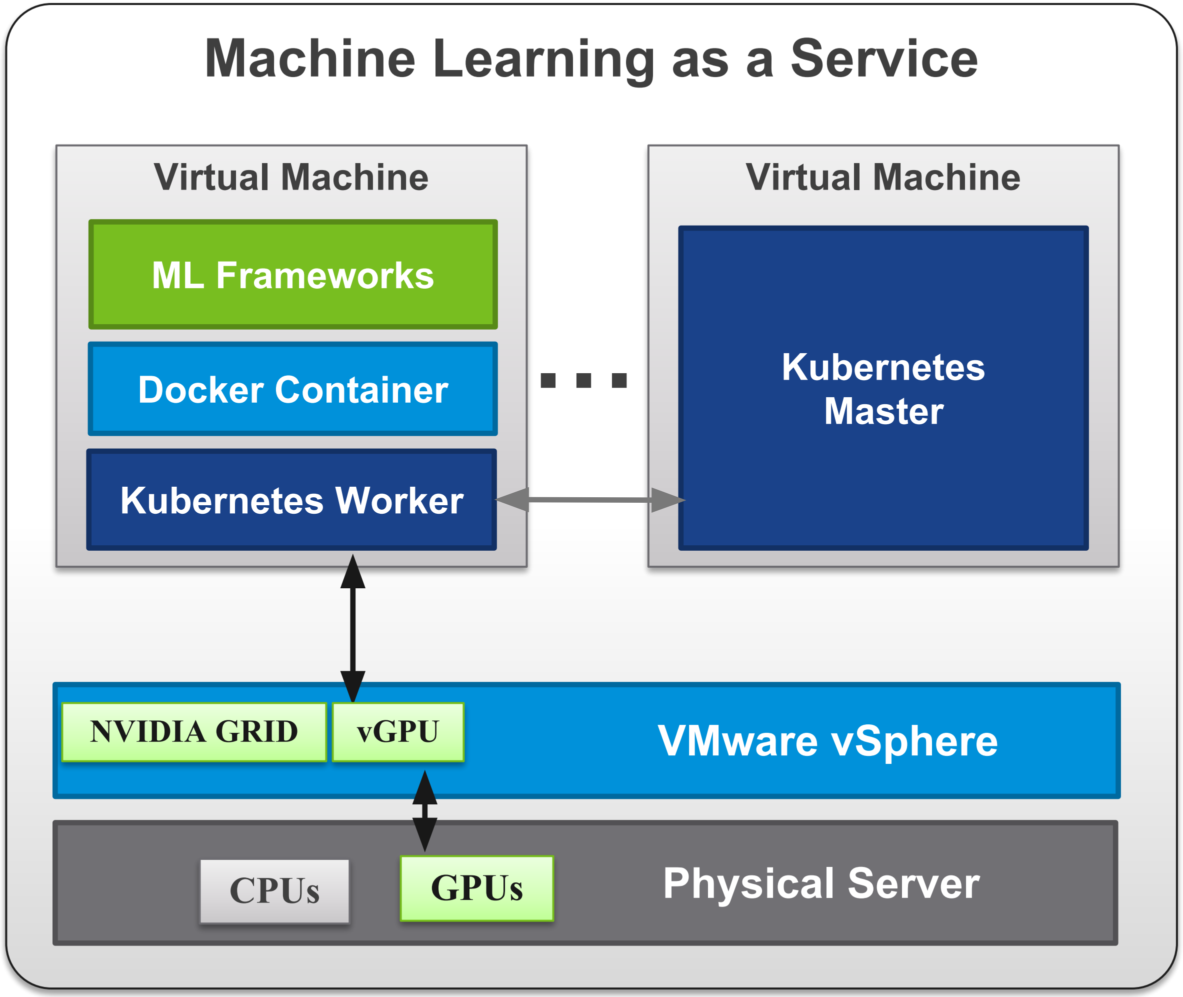
VMware vSphere

CPUs

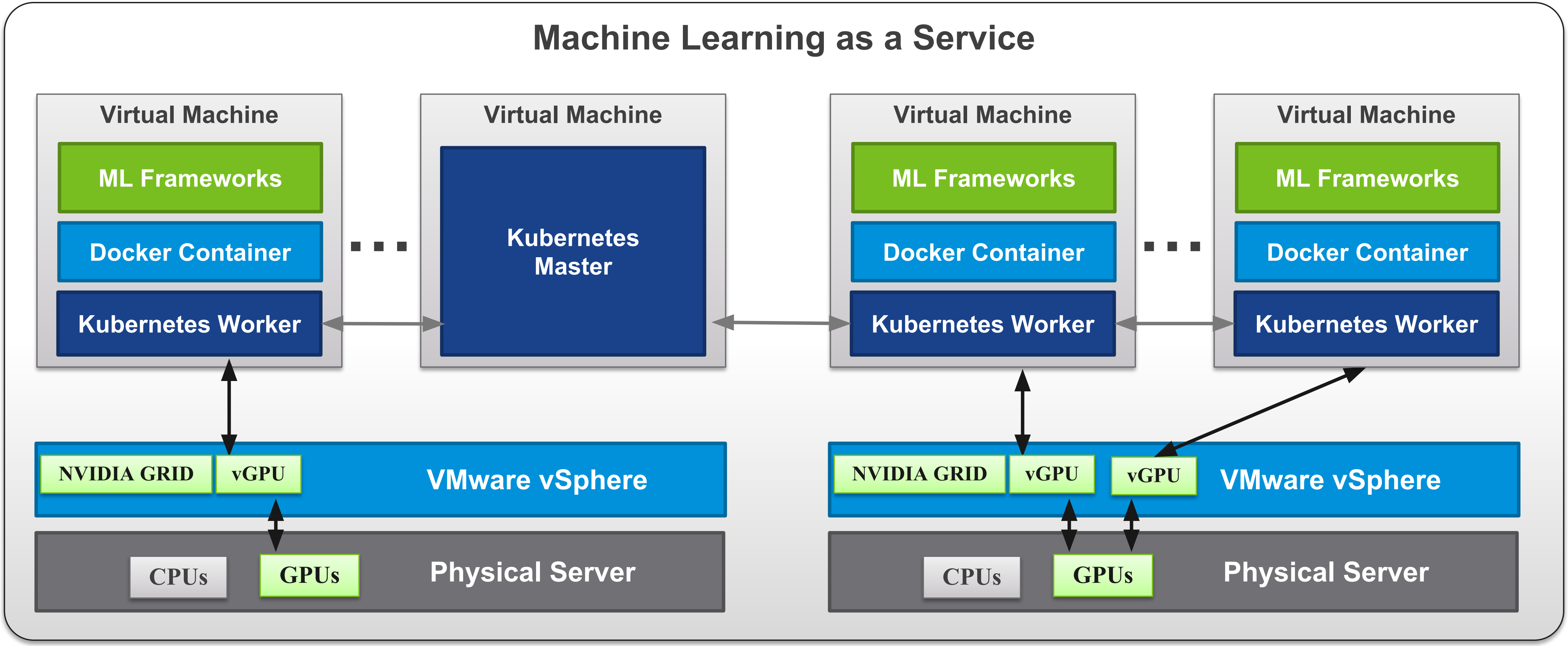
GPUs

Physical Server

Example #4 of deploying MLaaS on VMware vSphere with Container & Kubernetes

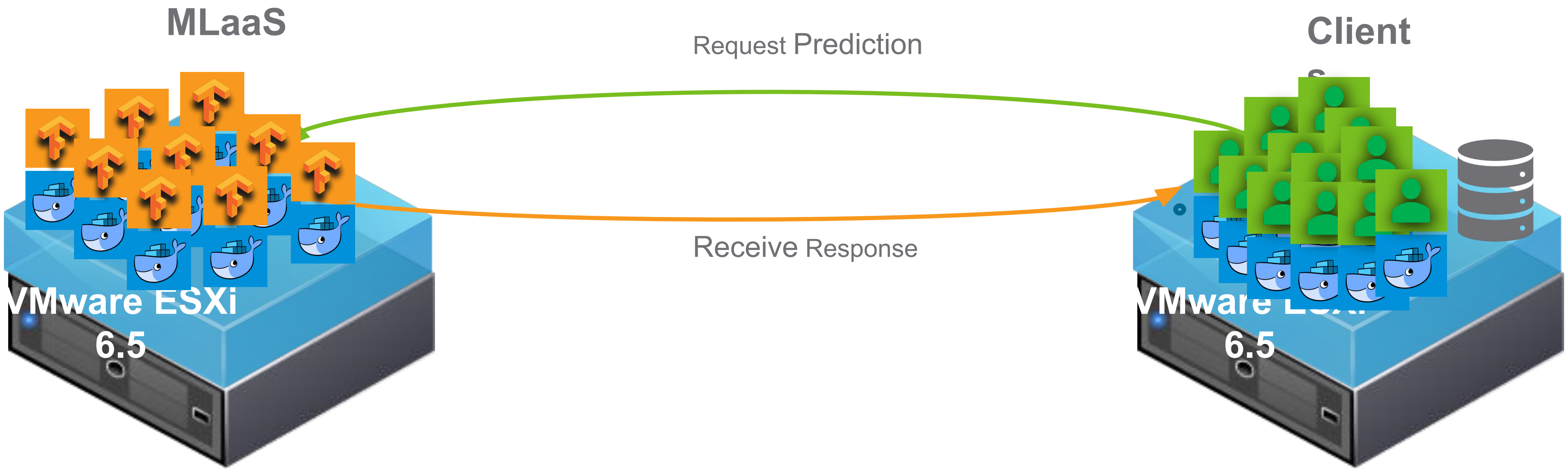


Example #4 of deploying MLaaS on VMware vSphere with Container & Kubernetes



Experiments of MLaaS on VMware vSphere

Hardware and Software



Dell R730 with Intel Haswell CPUs (36 cores)
+ NVIDIA P40 GPU

Intel Haswell CPUs
1VM with 18 vCPU

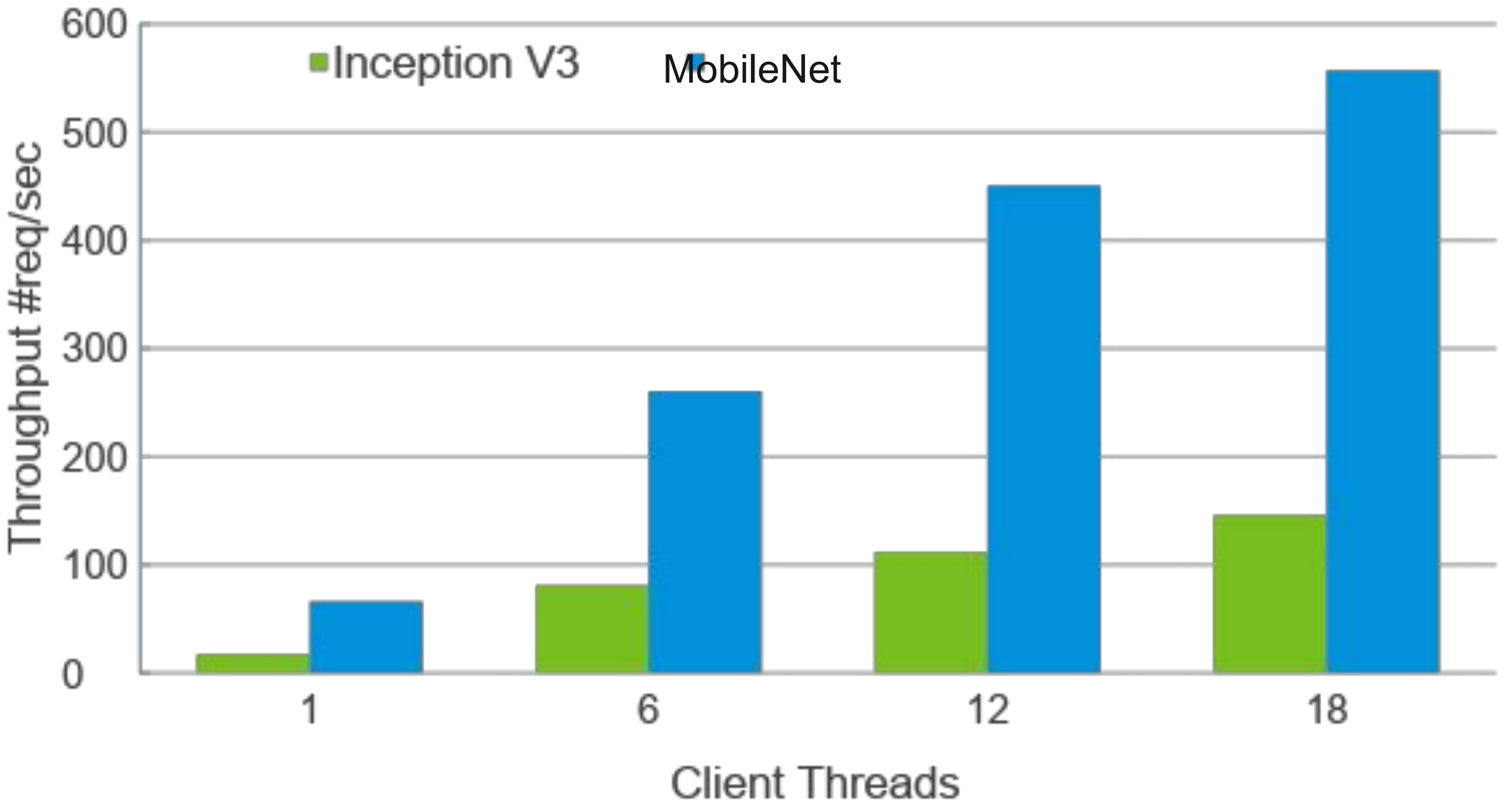
Experiment #1: Inference Throughput

Deep Neural Network: Inception V3 vs. MobileNet – Higher is better

<u>Models:</u>

Inception V3
48 Layers
5000 Million MAC

MobileNet:
28 Layers
569 Million MAC



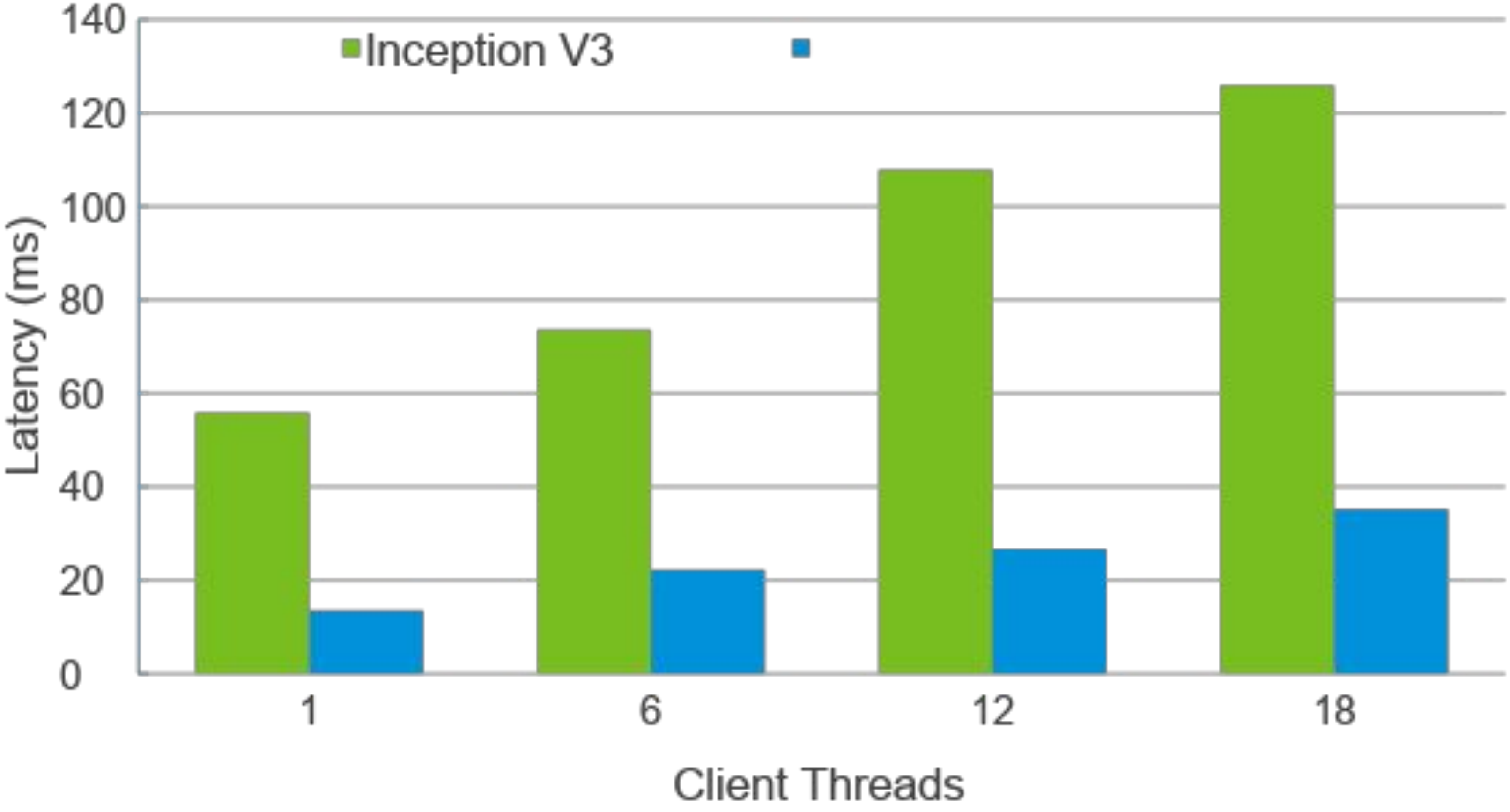
Experiment #1: Inference Mean Latency

Deep Neural Network: Inception V3 vs. MobileNet

Models:

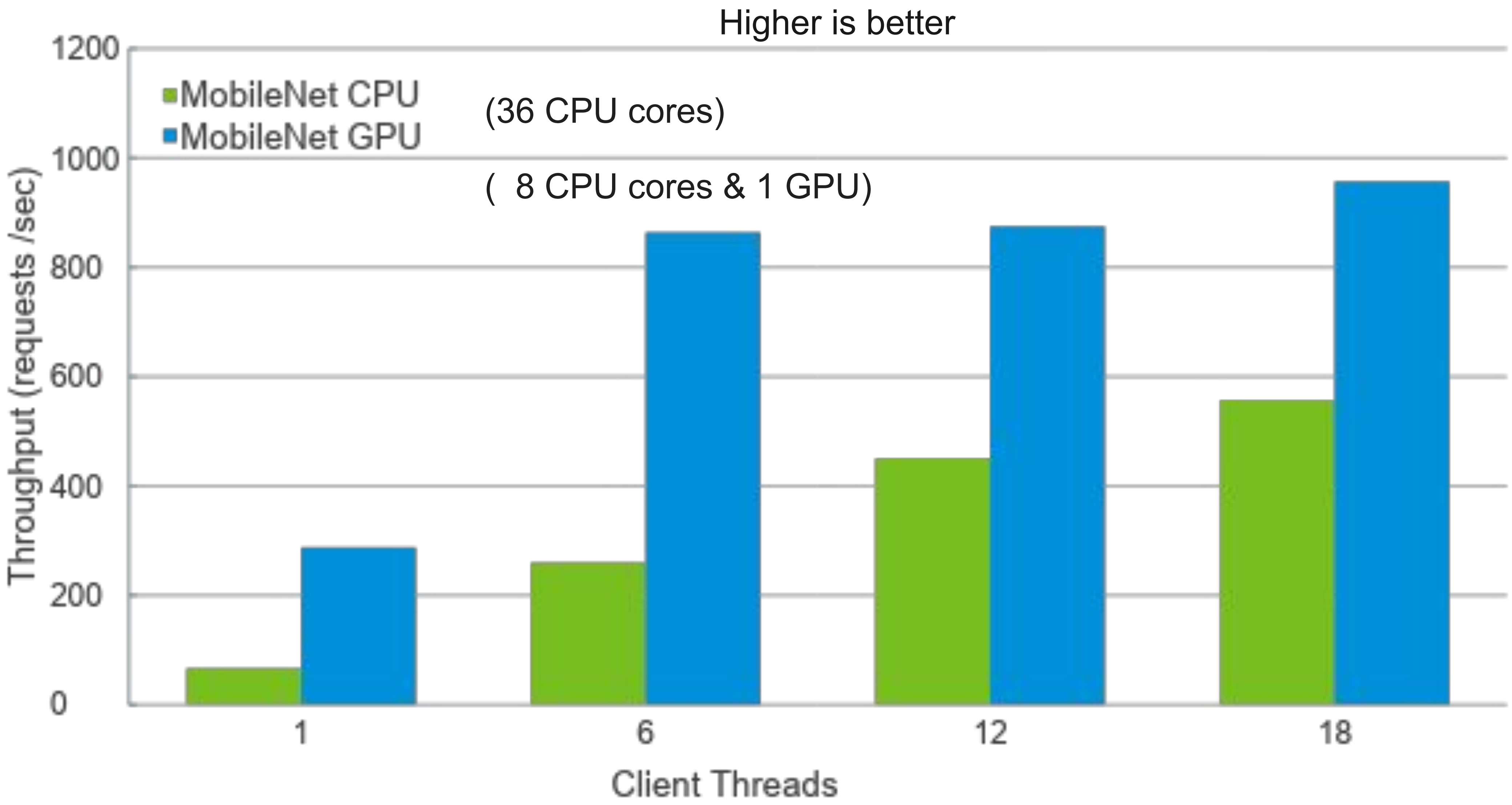
Inception V3
48 Layers
5000M MAC

MobileNet:
28 Layers
569 Million MAC



vmware®

Experiment #2: Inference Throughput



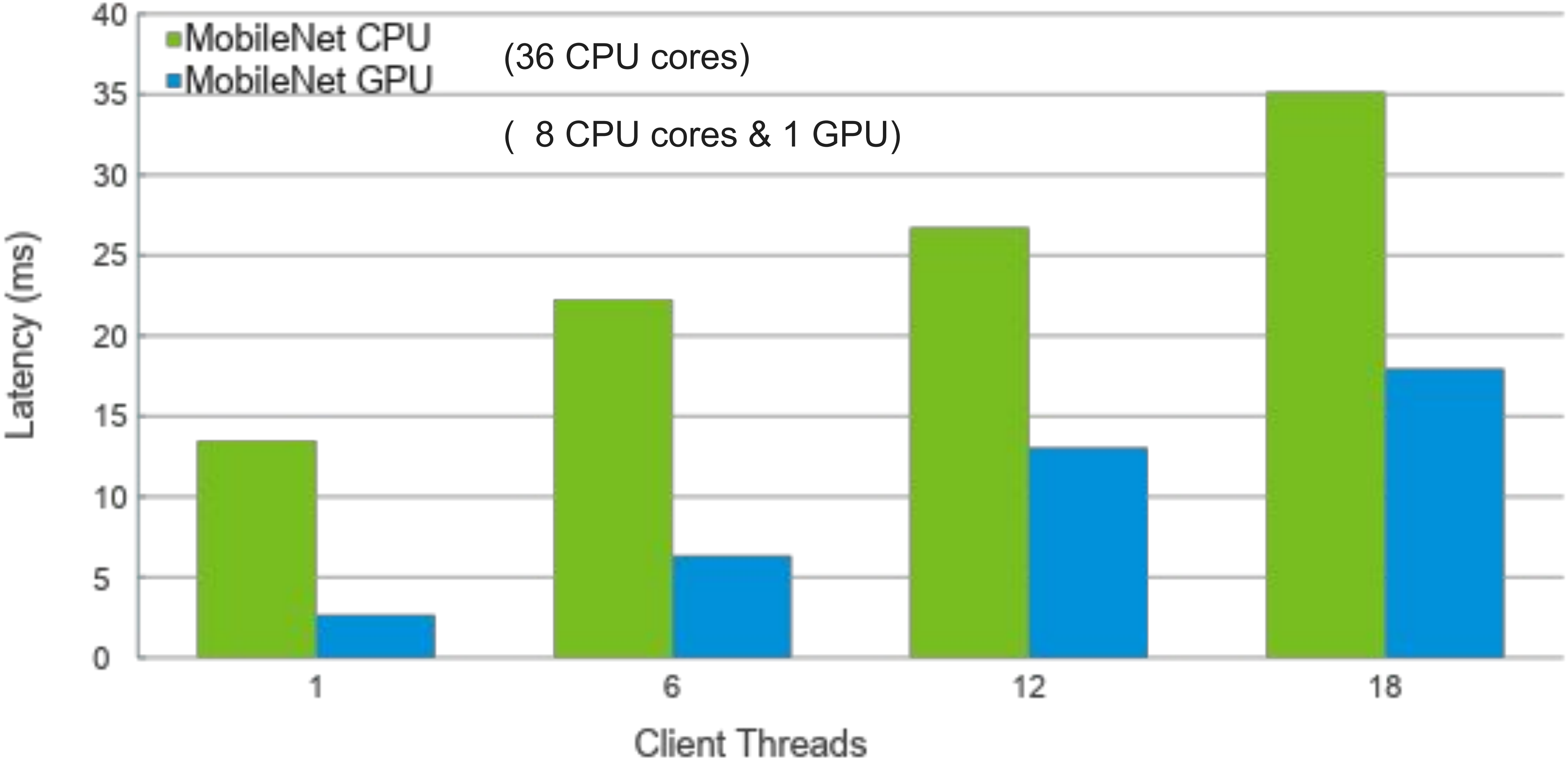
Higher is better

(36 CPU cores)

(8 CPU cores & 1 GPU)

Experiment #2: Mean Inference Latency

Lower is better

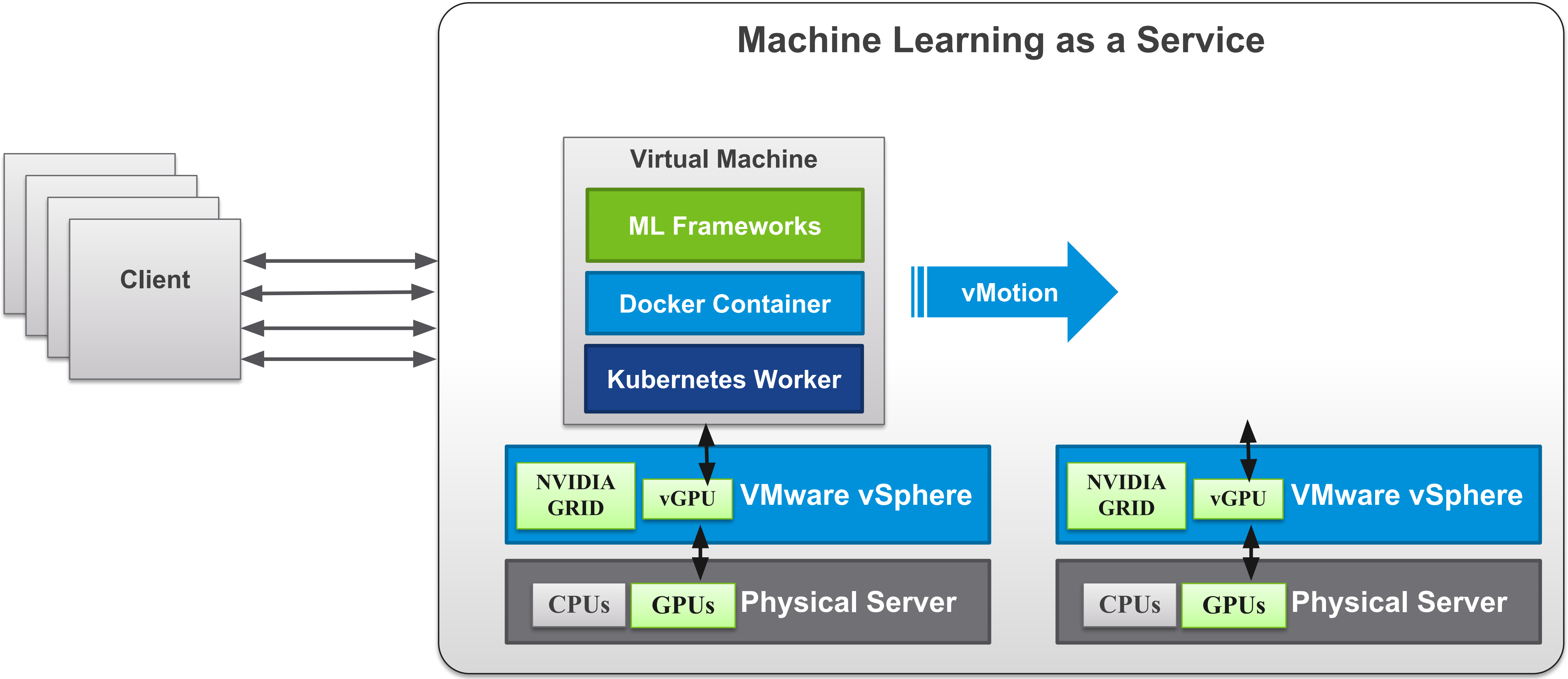


(36 CPU cores)

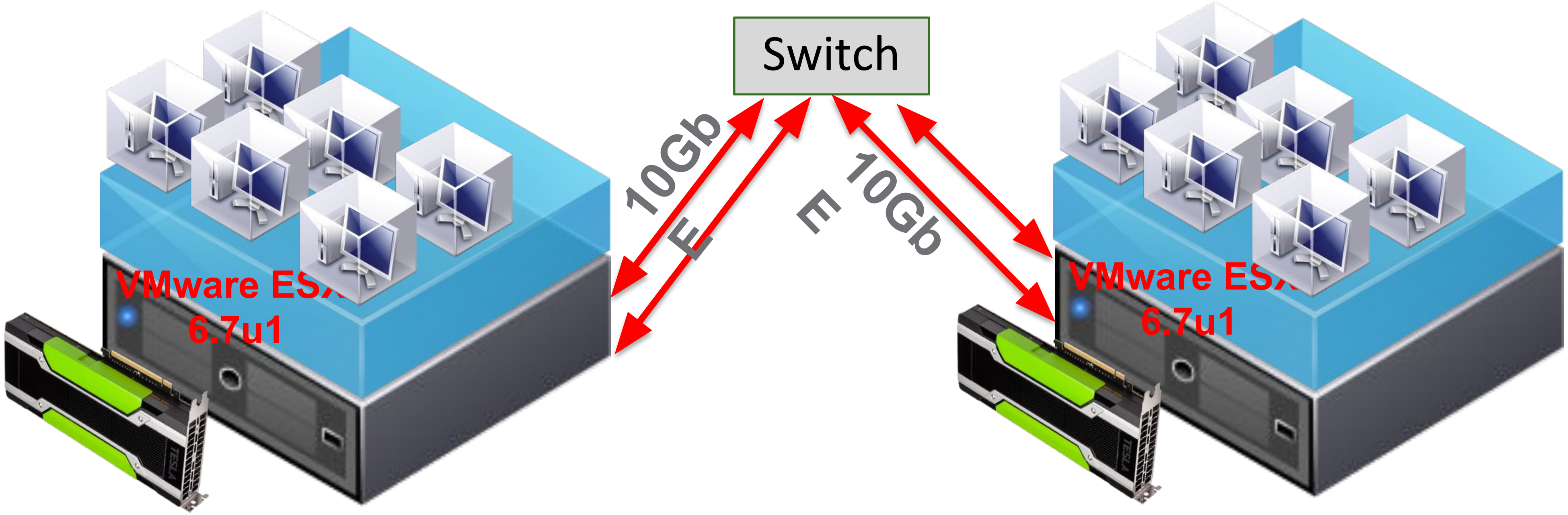
(8 CPU cores & 1 GPU)

©2019 VMware, Inc.

vMotion for NVIDIA GRID vGPU - MLaaS



vMotion for NVIDIA GRID vGPU – Test-bed

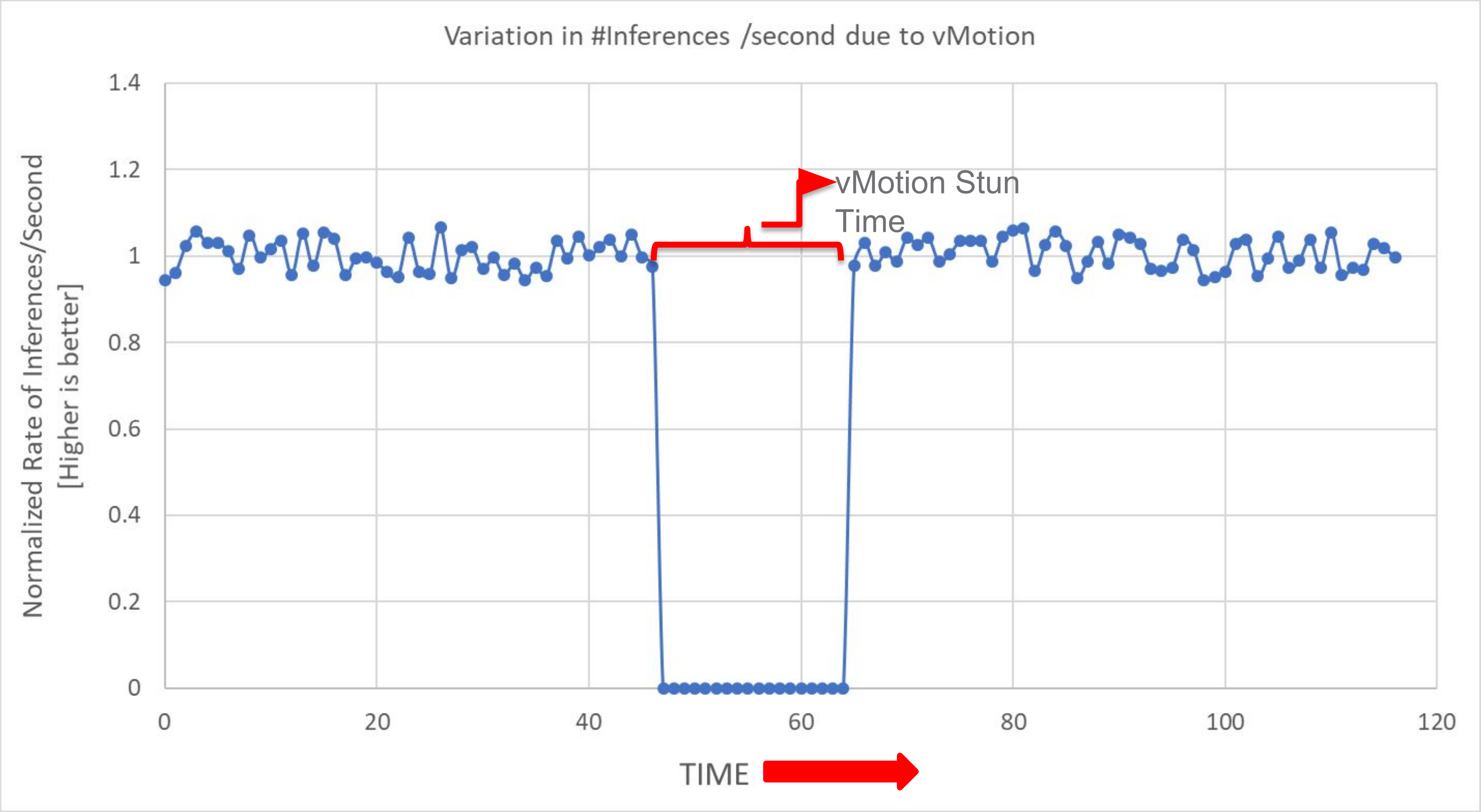


Dell R730 – Intel Broadwell CPUs + 1 x NVidia GRID P40
40 cores (2 x 20-core socket) E5-2698 v4
768 GB RAM

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

- **ESX**: 6.7u1 **Nvidia Driver**: 410.68

vMotion for NVIDIA GRID vGPU - MLaaS



vMotion for Nvidia GRID vGPU: Conclusions and Upcoming Improvements

Conclusions:

- vMotion for Nvidia GRID vGPU is now available
- The performance impact of vMotion on VDI, CAD and ML applications is negligible or small.
- The performance impact of multiple vMotions running concurrently is small.

Upcoming Improvements:

- Speedup xfer rate of device checkpoint and vGPU memory data.
- Pre-copy vGPU memory data to reduce stun time to meet or exceed vMotion’s standard of 1 second.