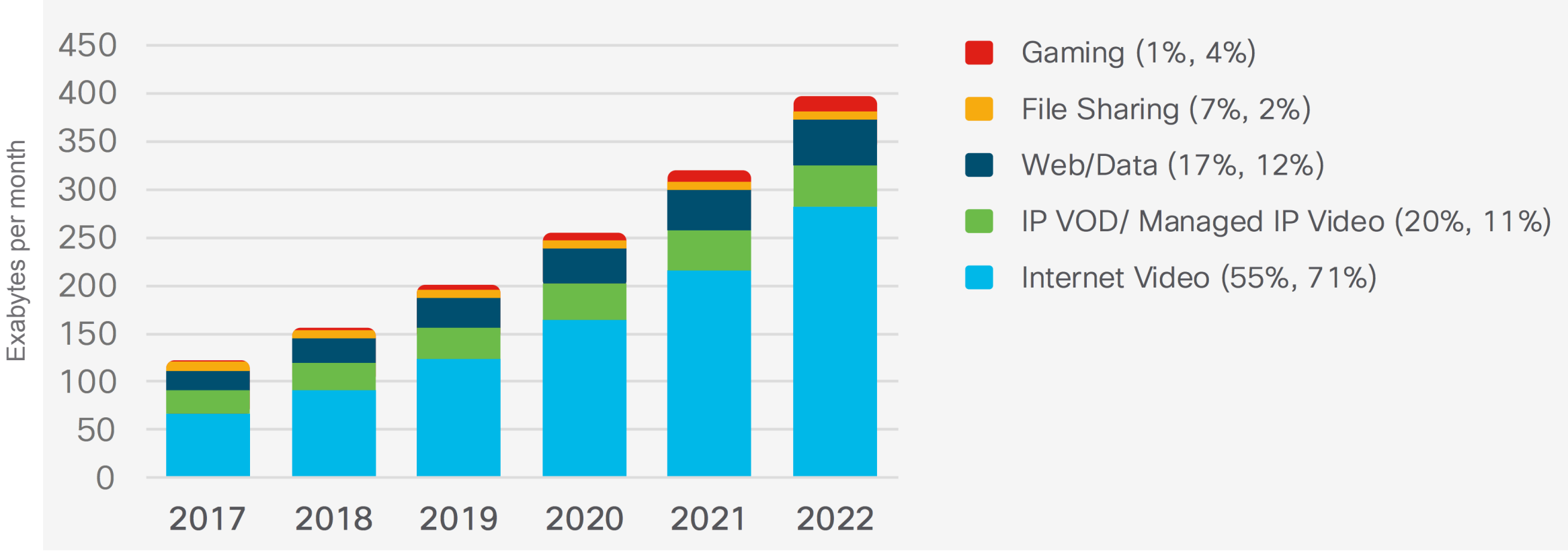
Large-Scale Video Audio Quality Assessment on VMware Platform with NVIDIA GPUs

Lan Vu, Hari Sivaraman

GTC 2019

Global Internet Traffic

Video Audio content accounts for **80 – 90%** of total IP traffic



Source: Cisco Visual Networking Index (VNI) forecast projects global IP traffic (2017 to 2022)

Video Audio Streaming Quality

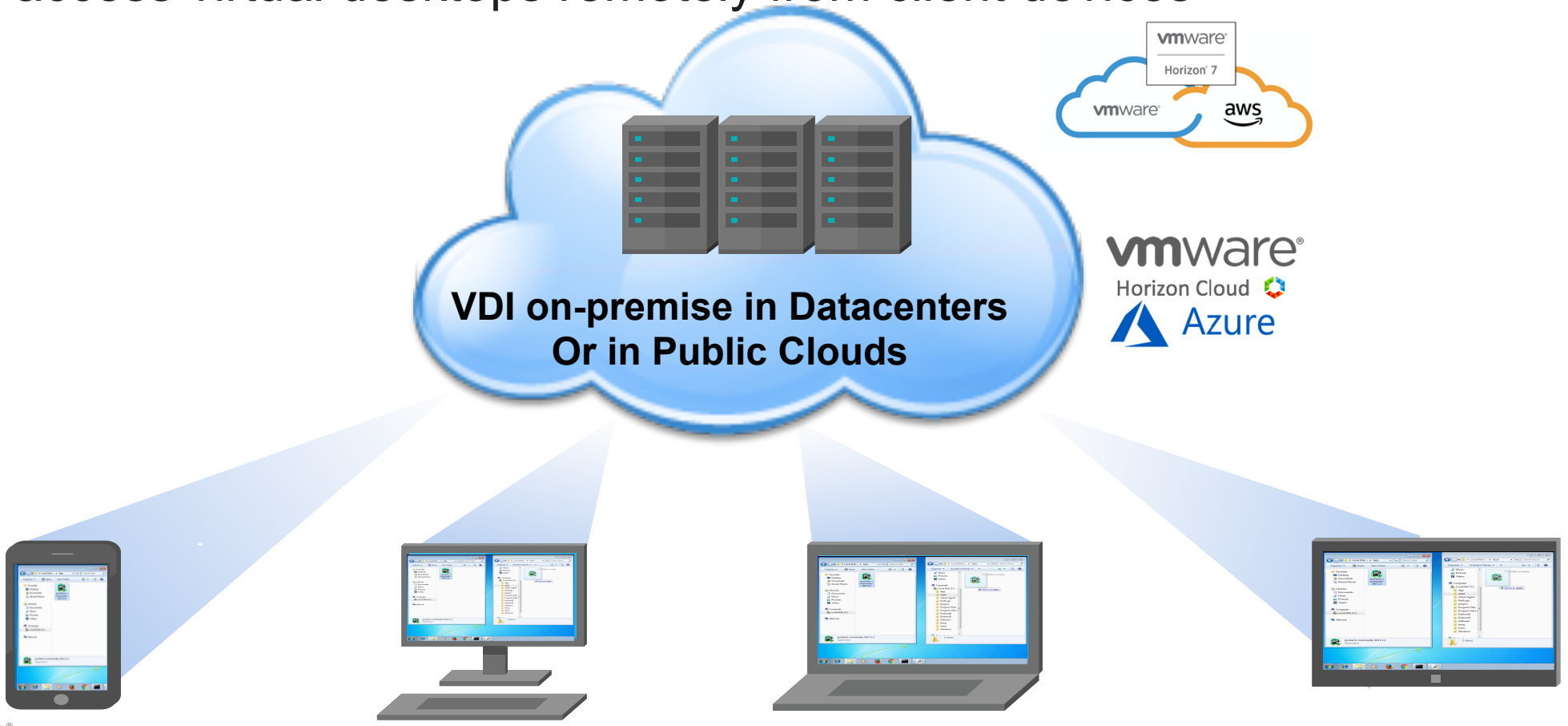
Depend on the network conditions, client load & server load

GOOD network (my office) VS. BAD network (on the bus)

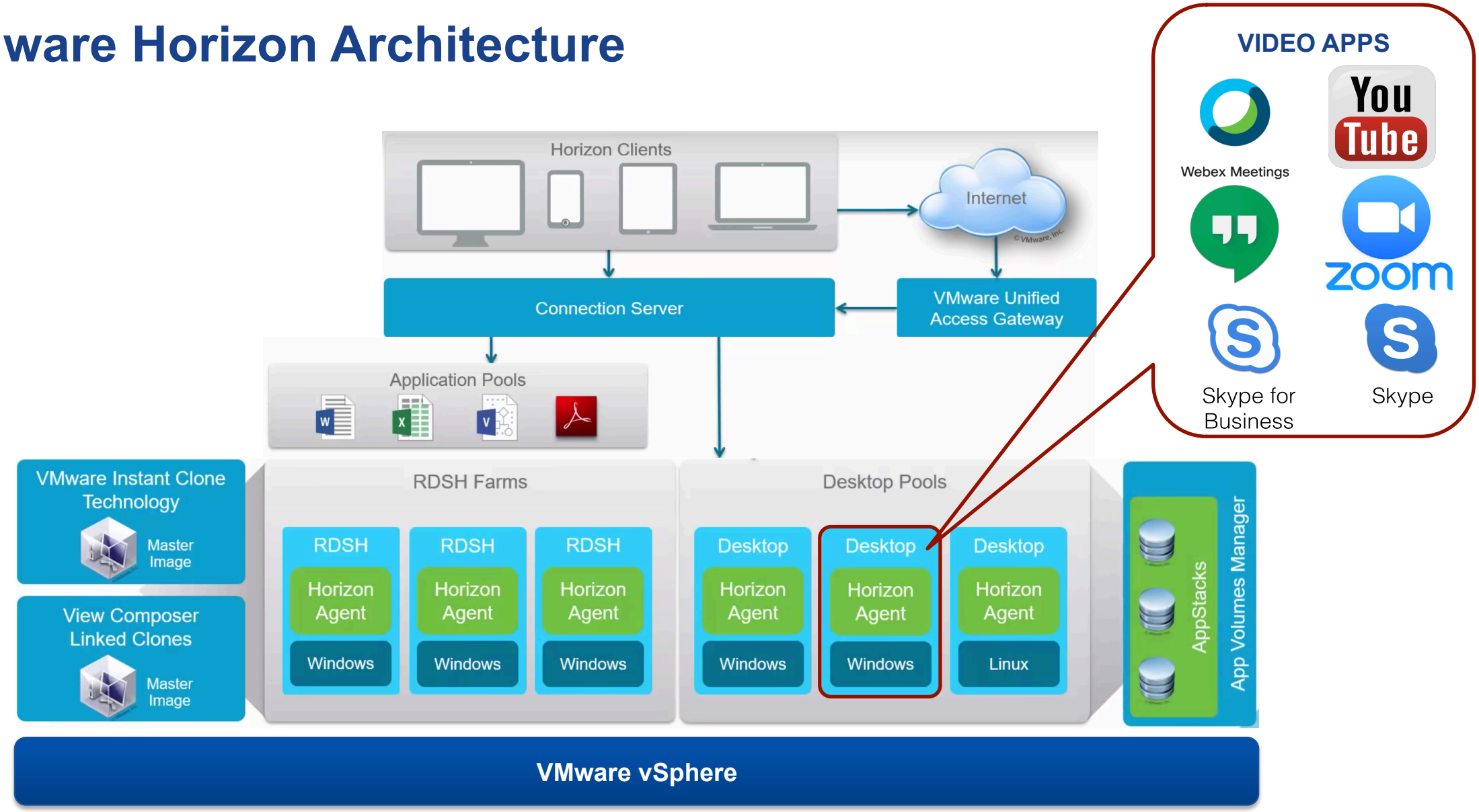
Blurry Images, Low FPS

VMware Horizon

- VDI (Virtual Desktop Infrastructure) software solution
- Desktops are virtualized & hosted in the cloud / datacenter
- Users access virtual desktops remotely from client devices

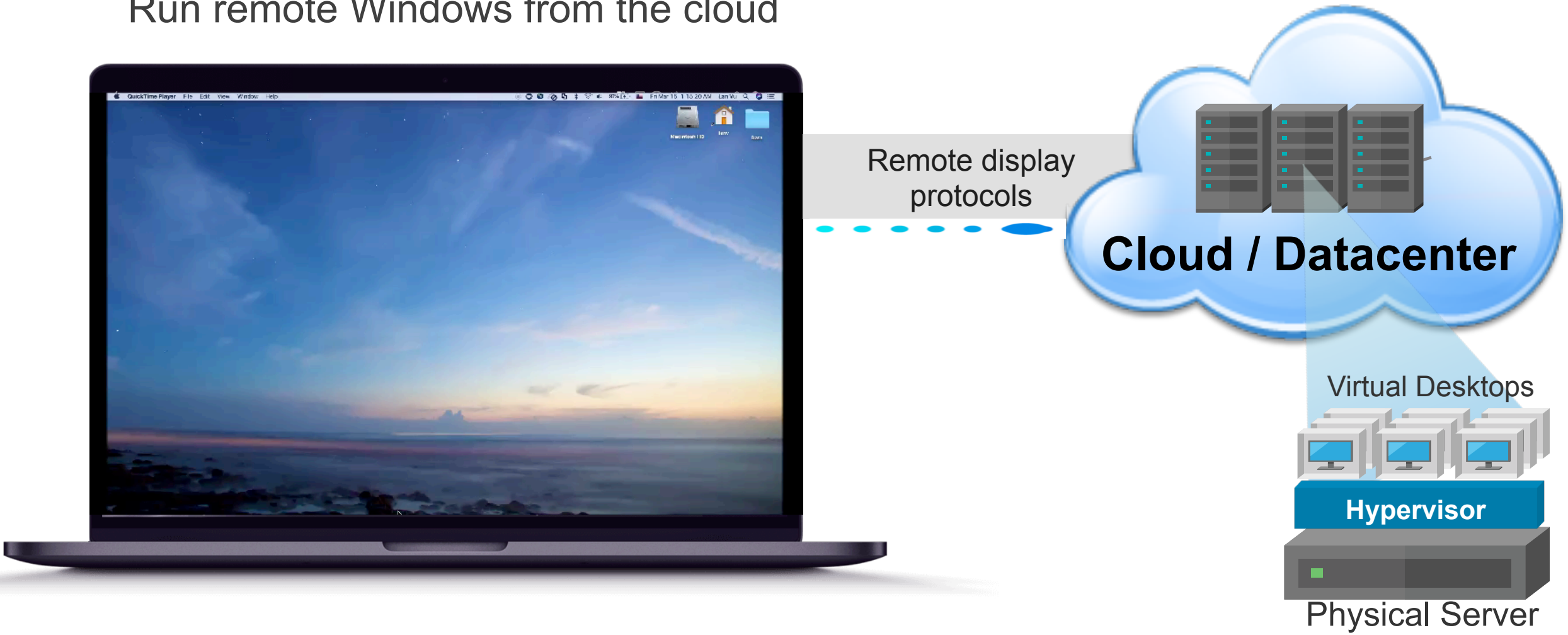


VMware Horizon Architecture



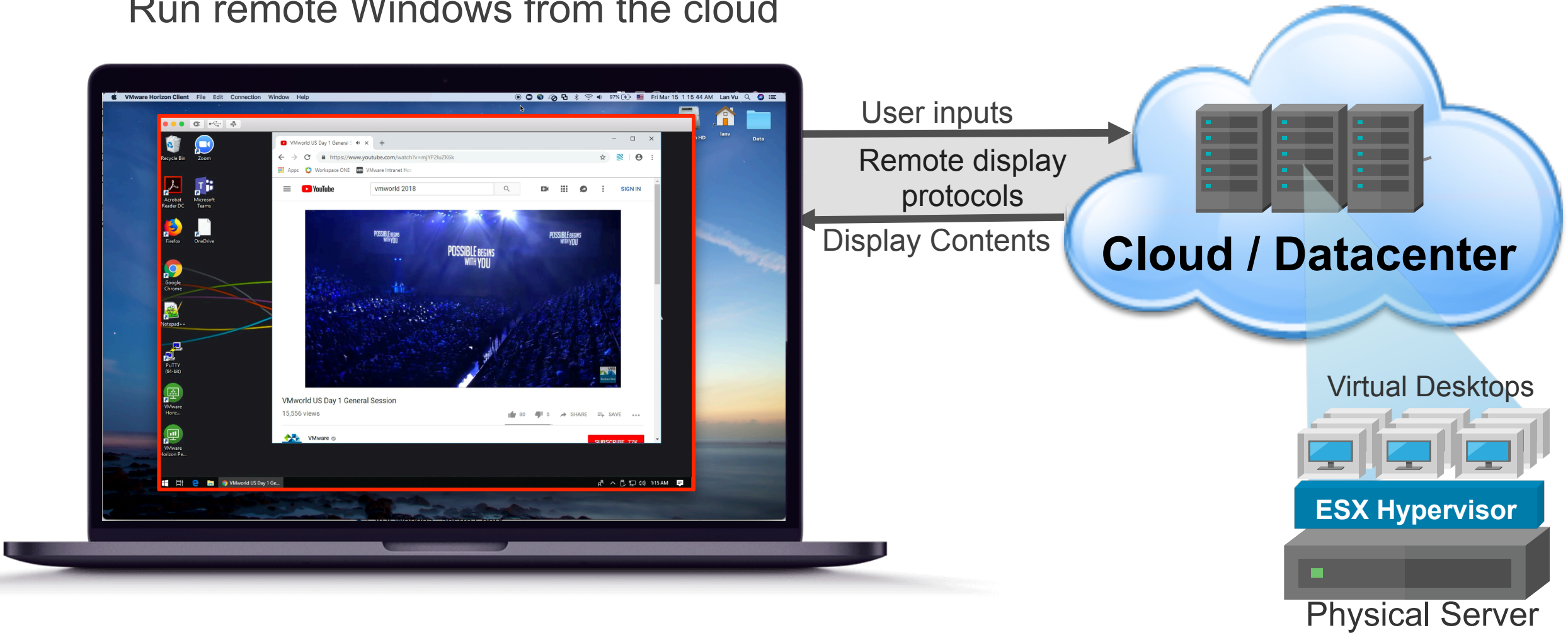
How does VMware Horizon work?

Run remote Windows from the cloud



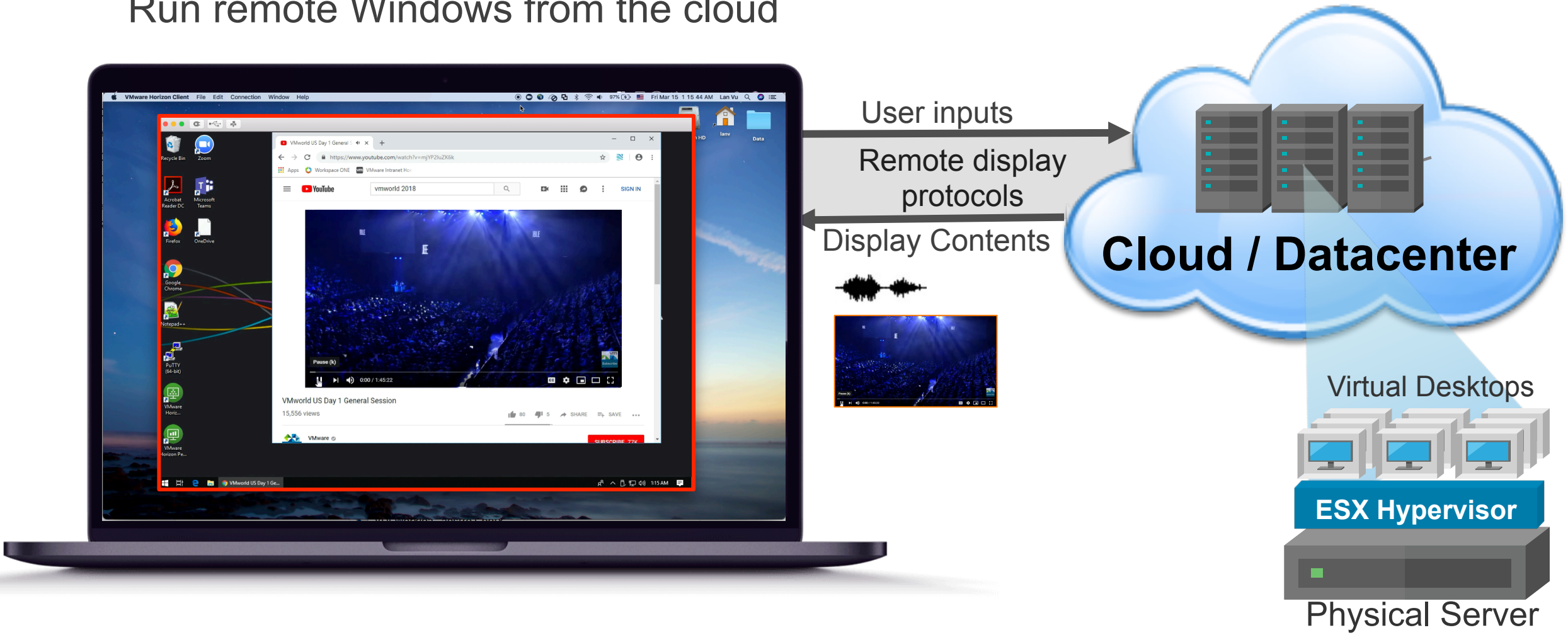
How does VMware Horizon work?

Run remote Windows from the cloud



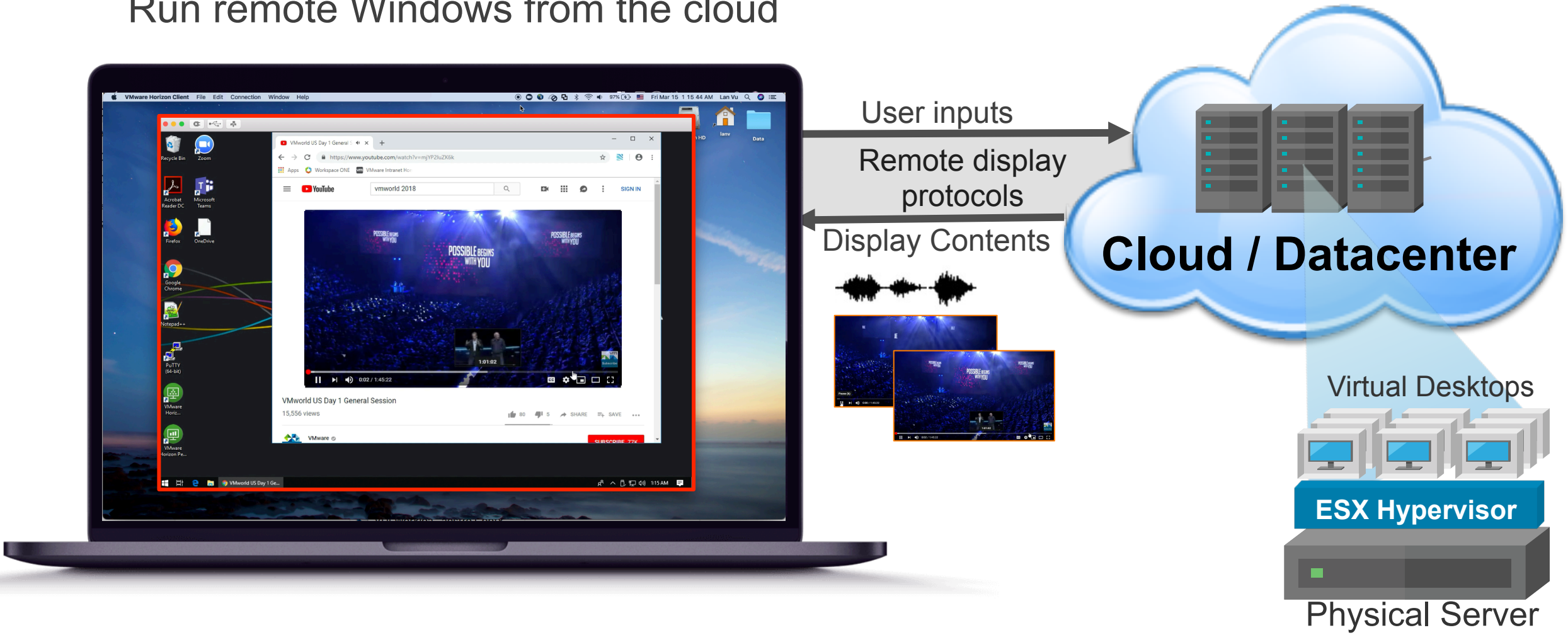
How does VMware Horizon work?

Run remote Windows from the cloud



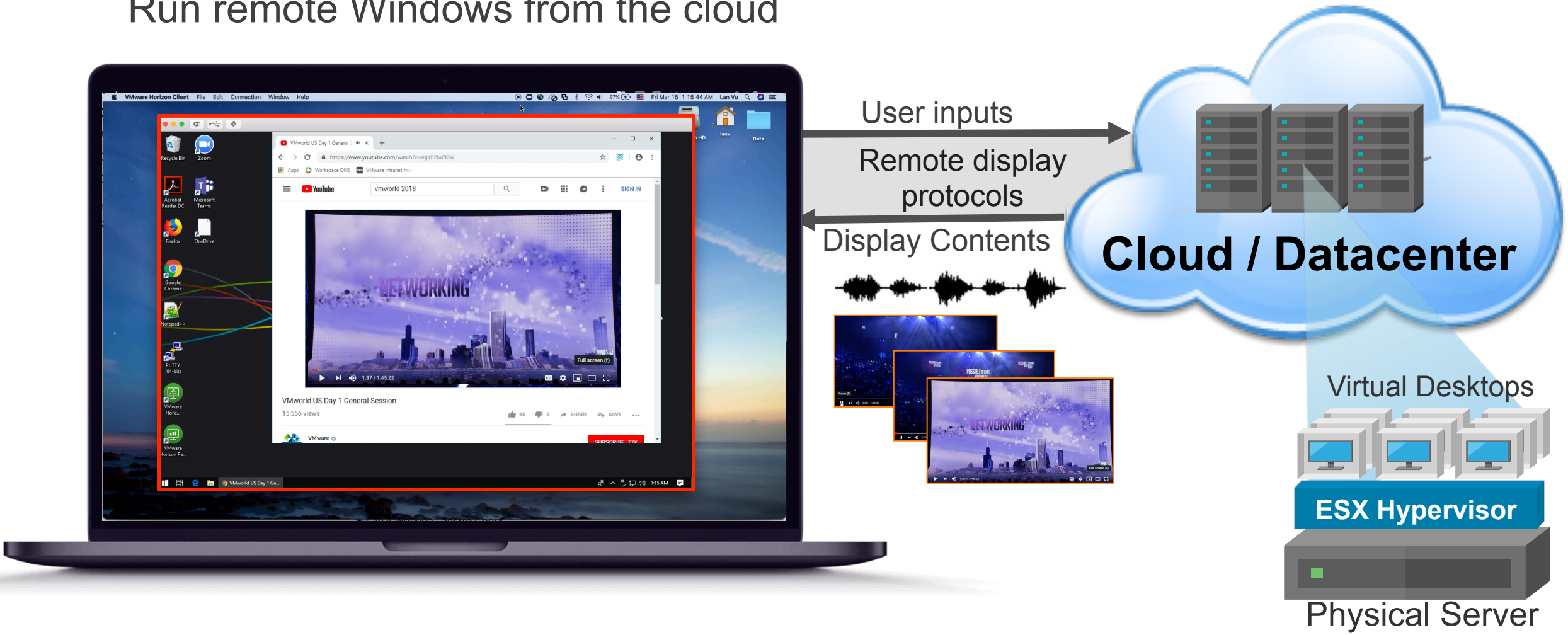
How does VMware Horizon work?

Run remote Windows from the cloud



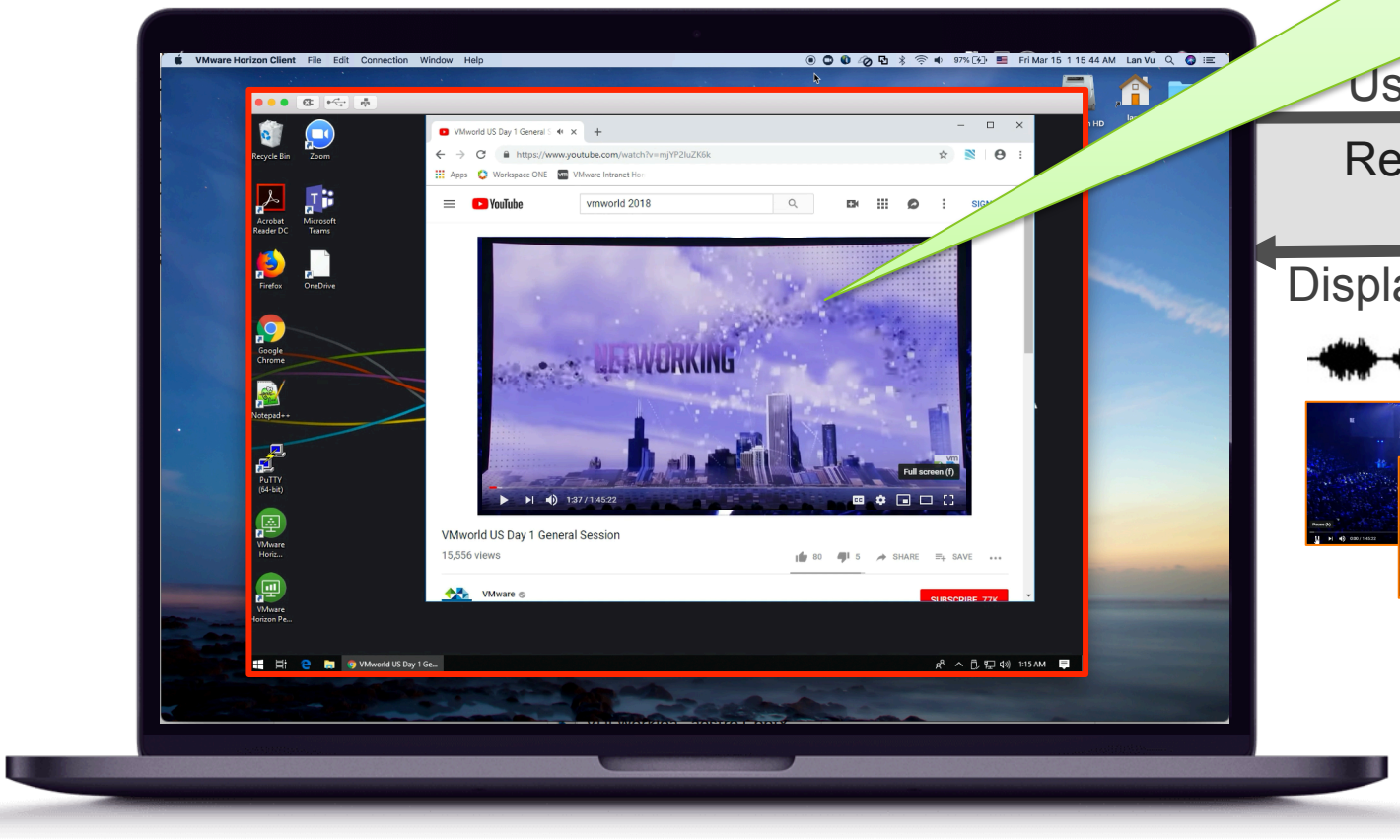
How does VMware Horizon work?

Run remote Windows from the cloud



Video Audio Quality Assessment is important

Run remote Windows from the cloud

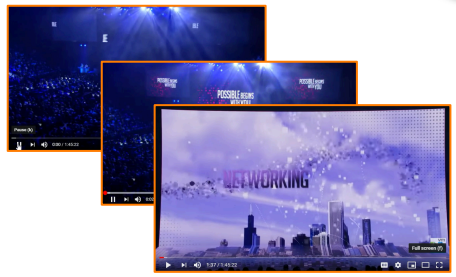


Ensure user satisfaction

User inputs

Remote display protocols

Display Contents



Cloud / Datacenter

Virtual Desktops

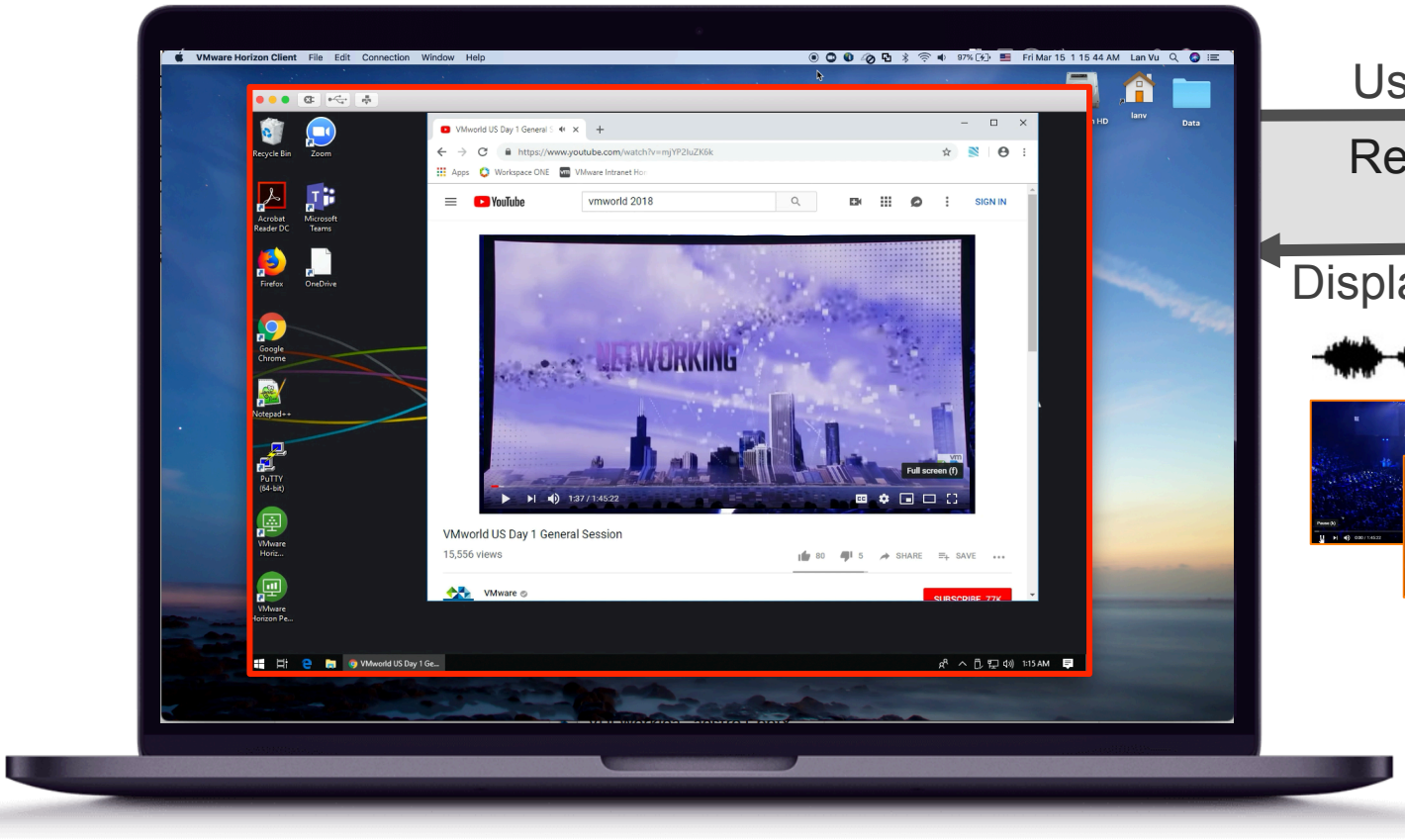
ESX Hypervisor

Physical Server

Video Audio Quality Assessment is important

Help to scale & optimize VDI deployment

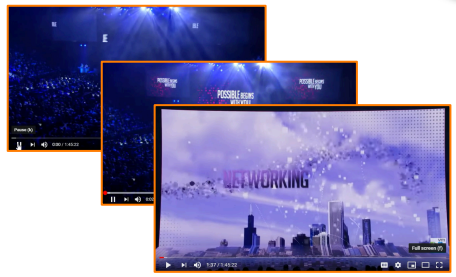
Run remote Windows from the cloud



User inputs

Remote display protocols

Display Contents



Cloud / Datacenter

Virtual Desktops

ESX Hypervisor

Physical Server

Our Approach for Video Quality Assessment

- Combine both subjective & objective assessment
- Use deep neural networks to recognize the quality

GOOD **or** **BAD**

Train Video Quality Recognition Model

Select Model

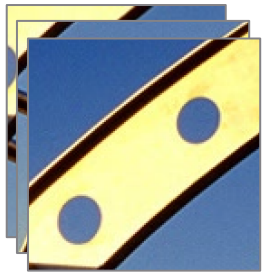
Train DNN models

Data Augmentation

Categorize Data

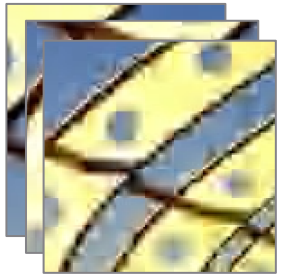
Collect Real & Simulated Data

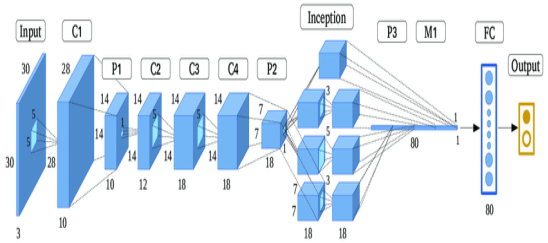
Good



Average

Bad



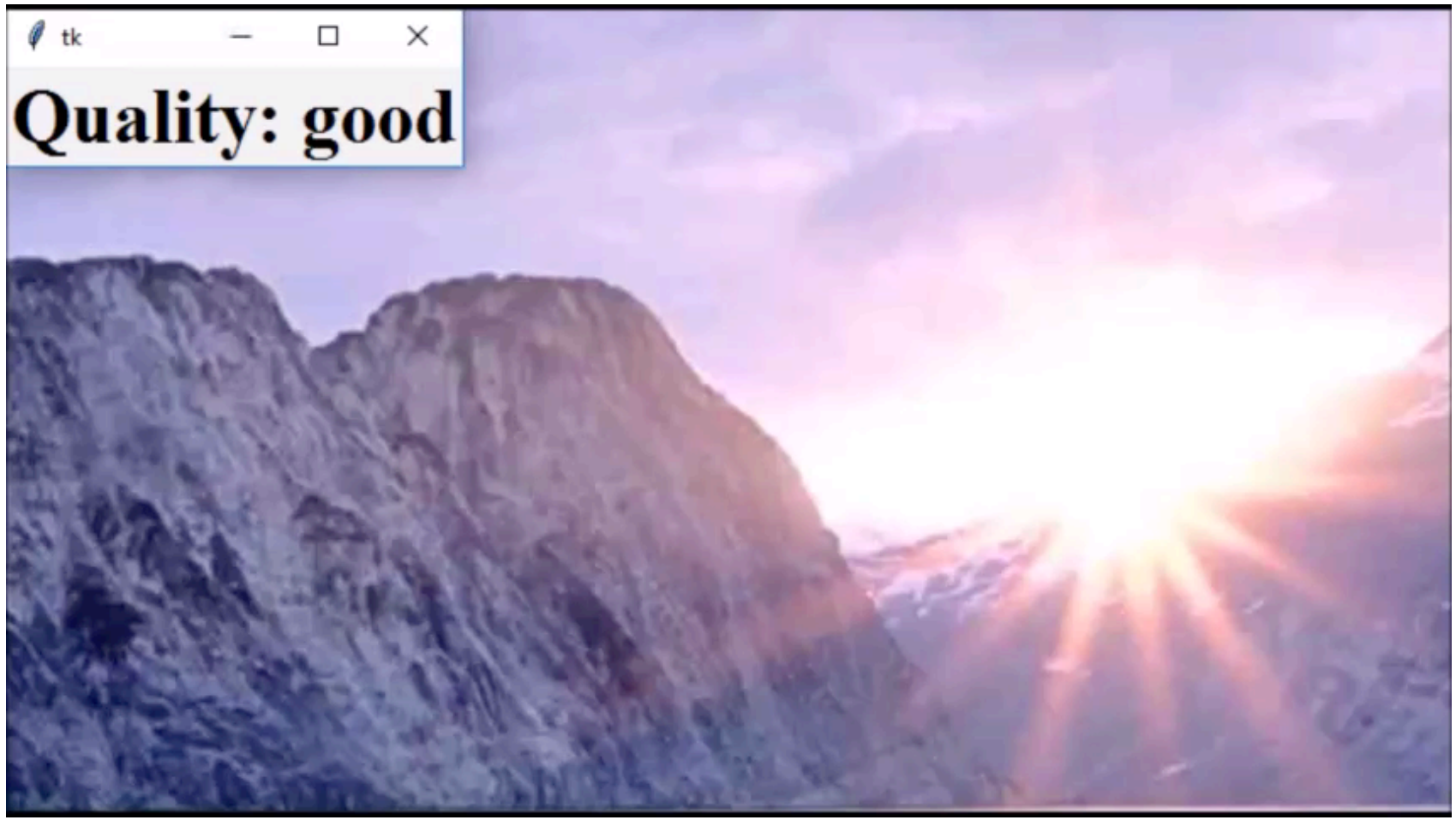


Good

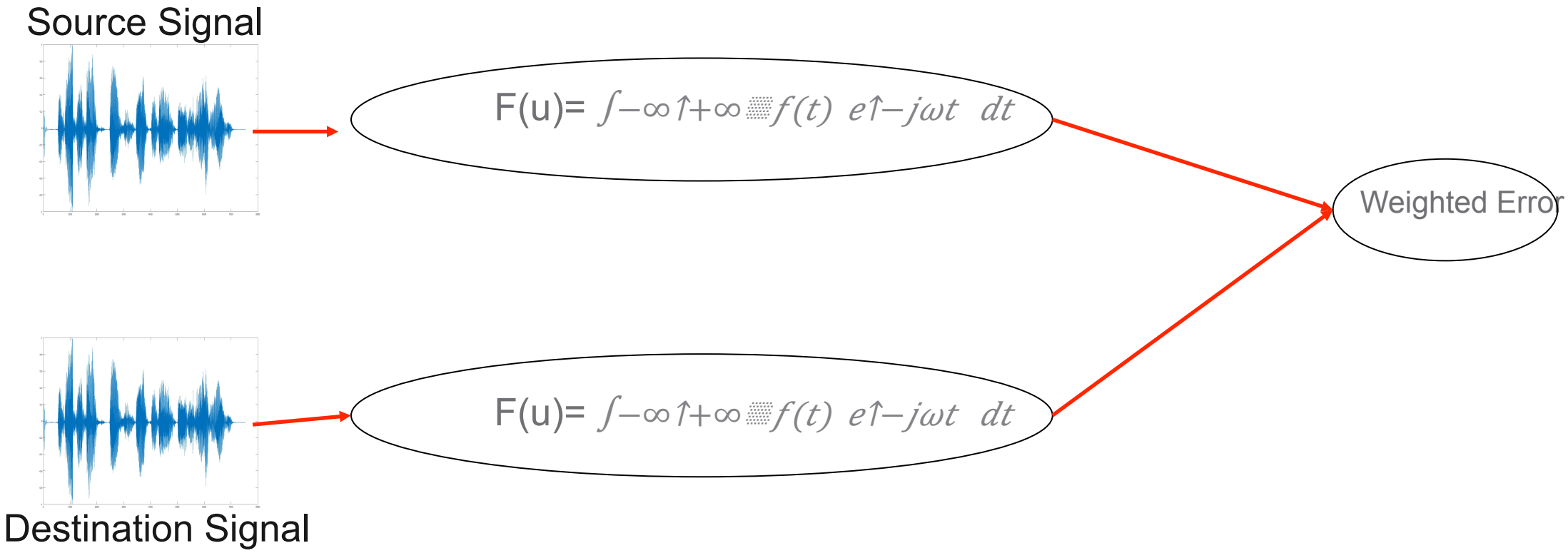
Average

Bad

Demo of Real Time Video Quality Assessment



Audio Quality Assessment



Why not just use this approach?

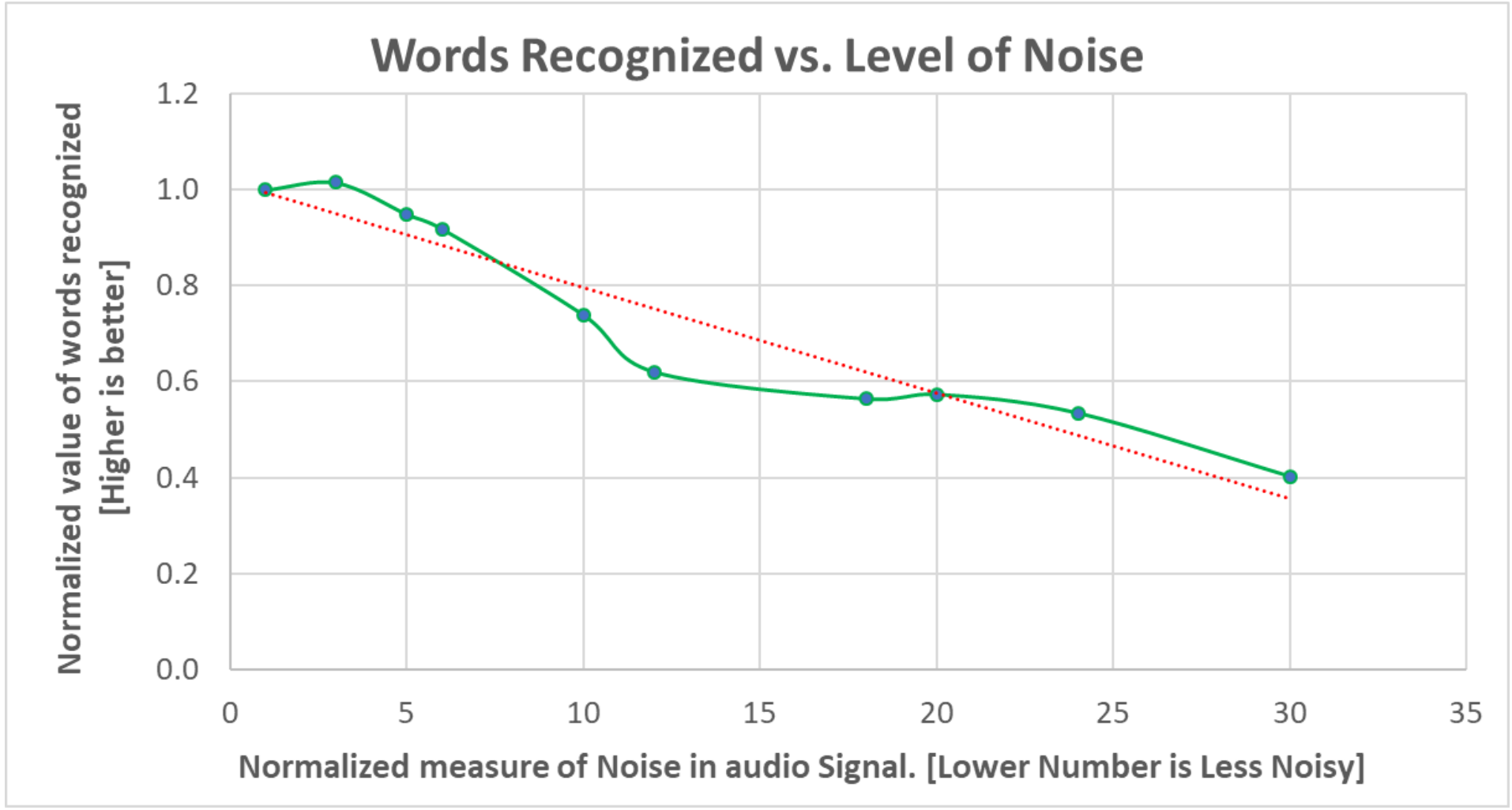
Our Approach for Audio Quality Assessment

- CNNs excel at extracting structural information to identify words in audio stream.

- Train the audio recognition model in TensorFlow to count % of words recognized.

But does % of words recognized vary with noise?

Results of Audio Quality Assessment



Audio Quality Assessment - Future Work

- Need to train the model with a large vocabulary typical of corporate meetings.

- Test with various types of noise

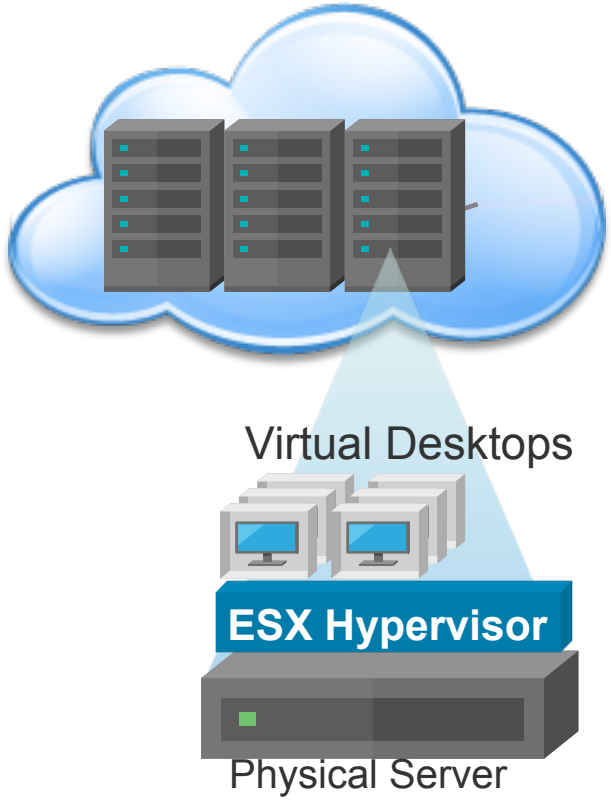
Large Scale Video / Audio Quality Assessment

A Use case to Optimize Horizon View Deployment

A Use Case of Video Audio Quality Assessment

How many desktops / server and how many server for VDI deployment?

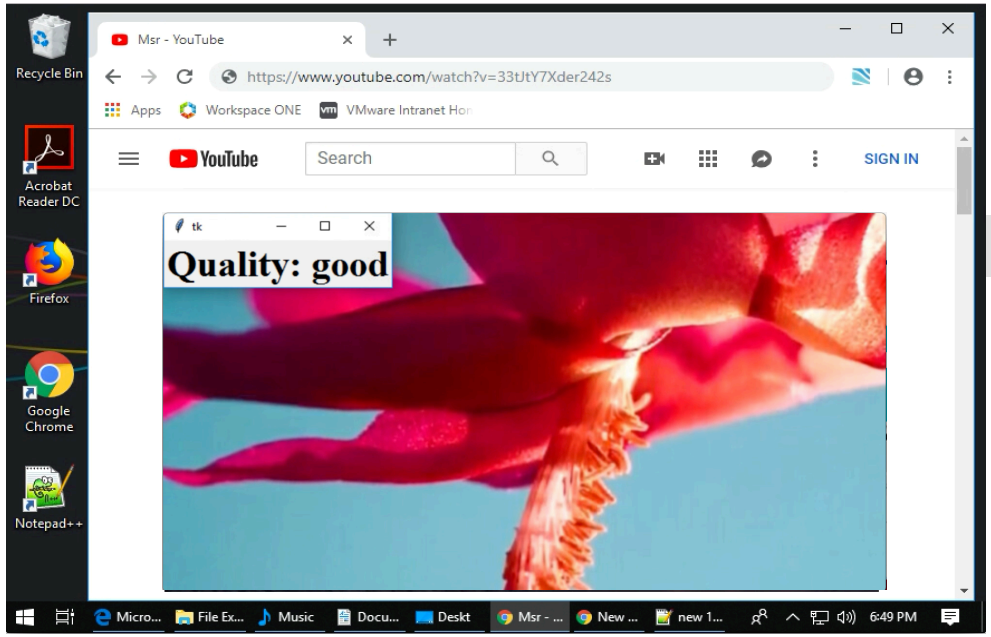
Monitoring the app quality → Sizing VDI deployment → Select the optimal deployment configuration



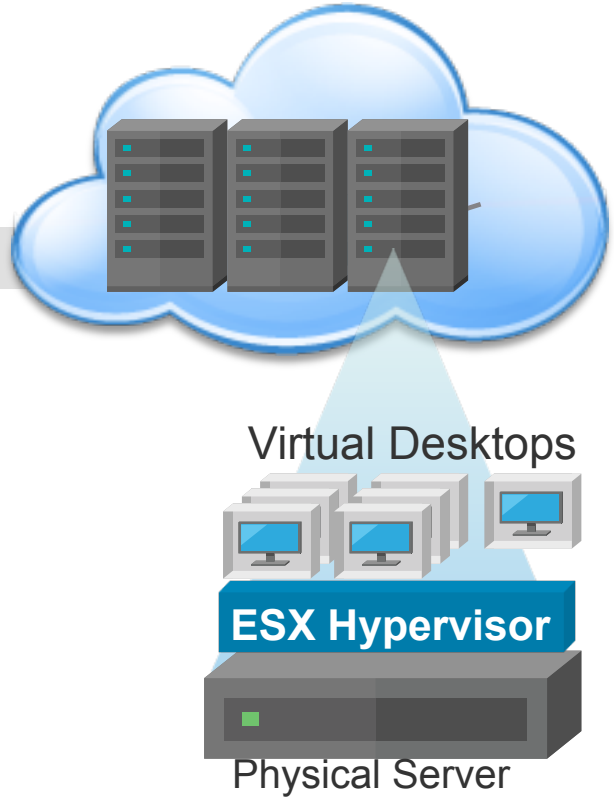
A Use Case of Video Audio Quality Assessment

How many desktops / server and how many server for VDI deployment?

Monitoring the app quality → Sizing VDI deployment → Select the optimal deployment configuration



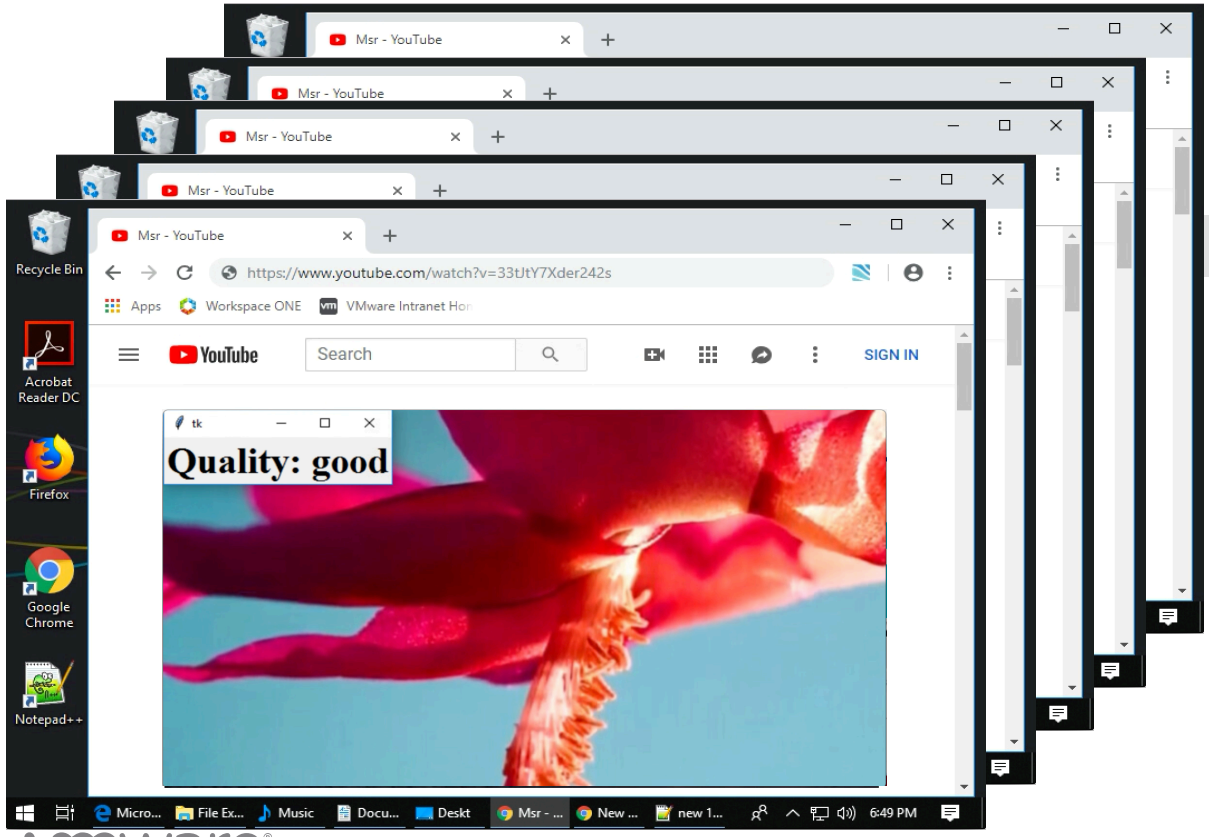
Remote display protocols

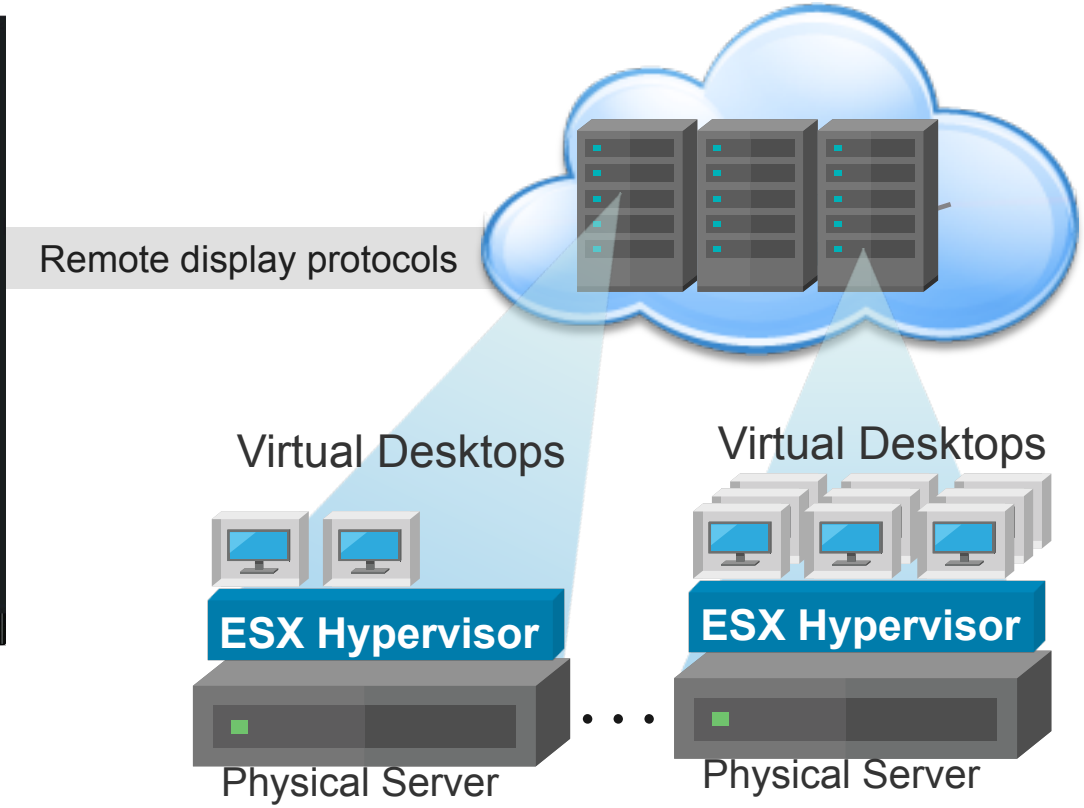


A Use Case of Video Audio Quality Assessment

How many desktops / server and how many server for VDI deployment?

Monitoring the app quality → **Sizing VDI deployment** → Select the optimal deployment configuration

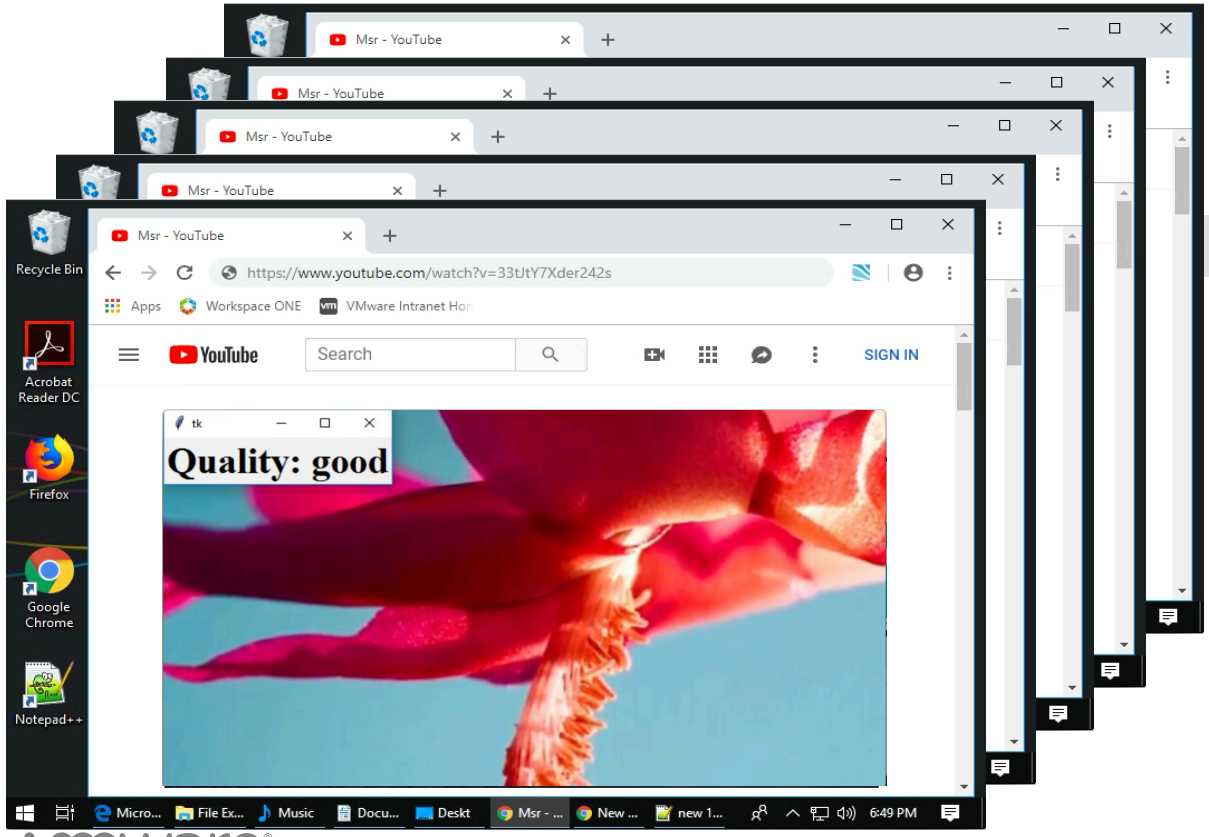


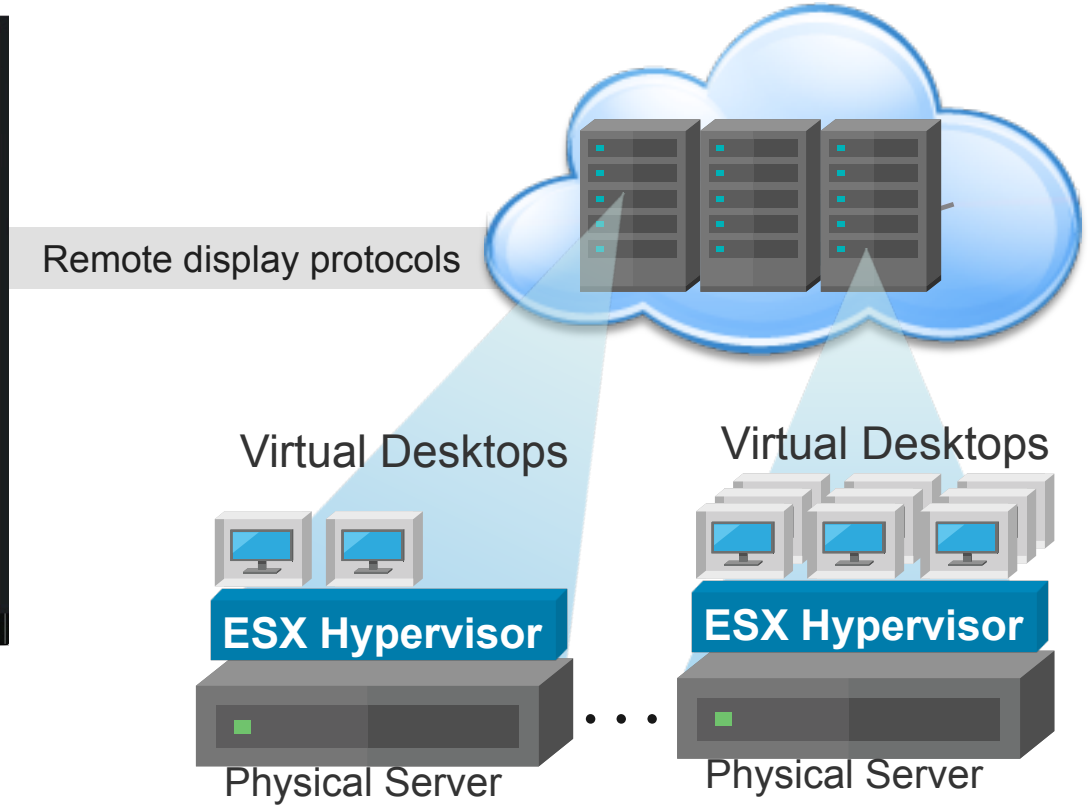


A Use Case of Video Audio Quality Assessment

How many desktops / server and how many server for VDI deployment?

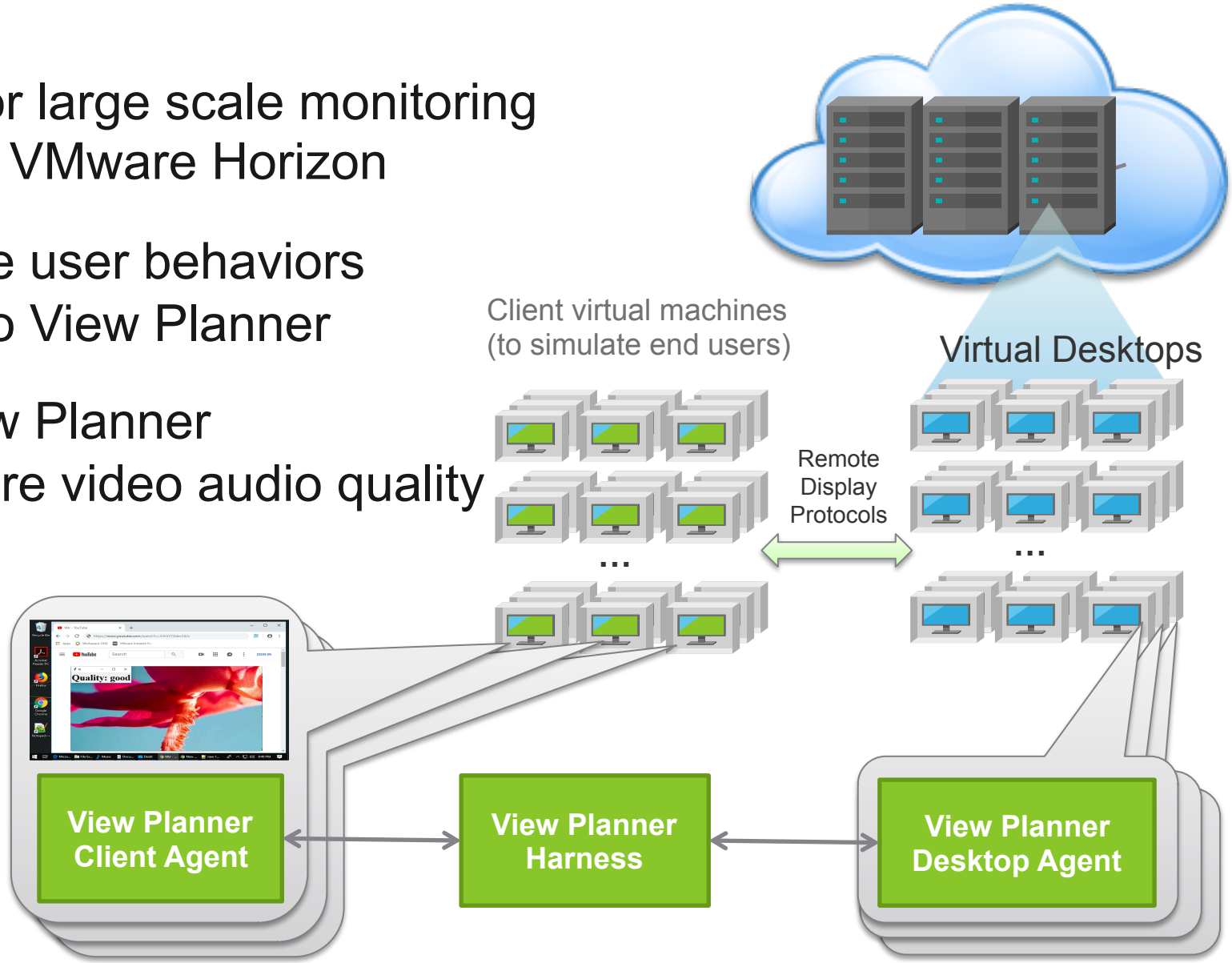
Monitoring the app quality → Sizing VDI deployment → **Select the optimal deployment configuration**





Large Scale Video Audio Quality Assessment with VMware View Planner

- **View Planner**: framework for large scale monitoring app performance in VDI like VMware Horizon
- Create workloads to emulate user behaviors of playing video audio add to View Planner
- Add these workloads to View Planner
  Run at large scale to measure video audio quality



Improving Performance of Inference ML Models

Improving Inference Performance

Enable large scale video audio quality measurement

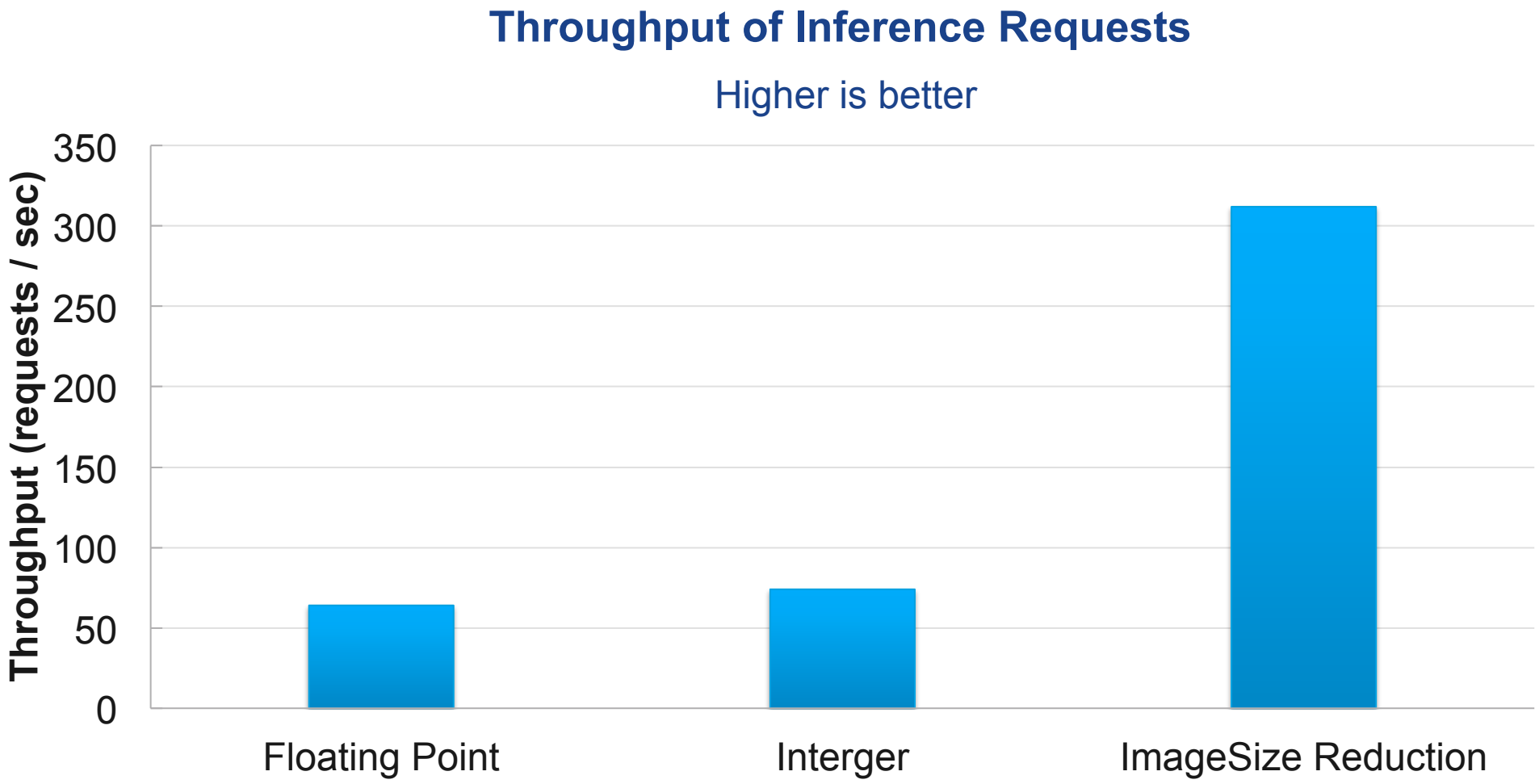
- Increase Throughput
- Reduce Latency

Tuning Neural Networks

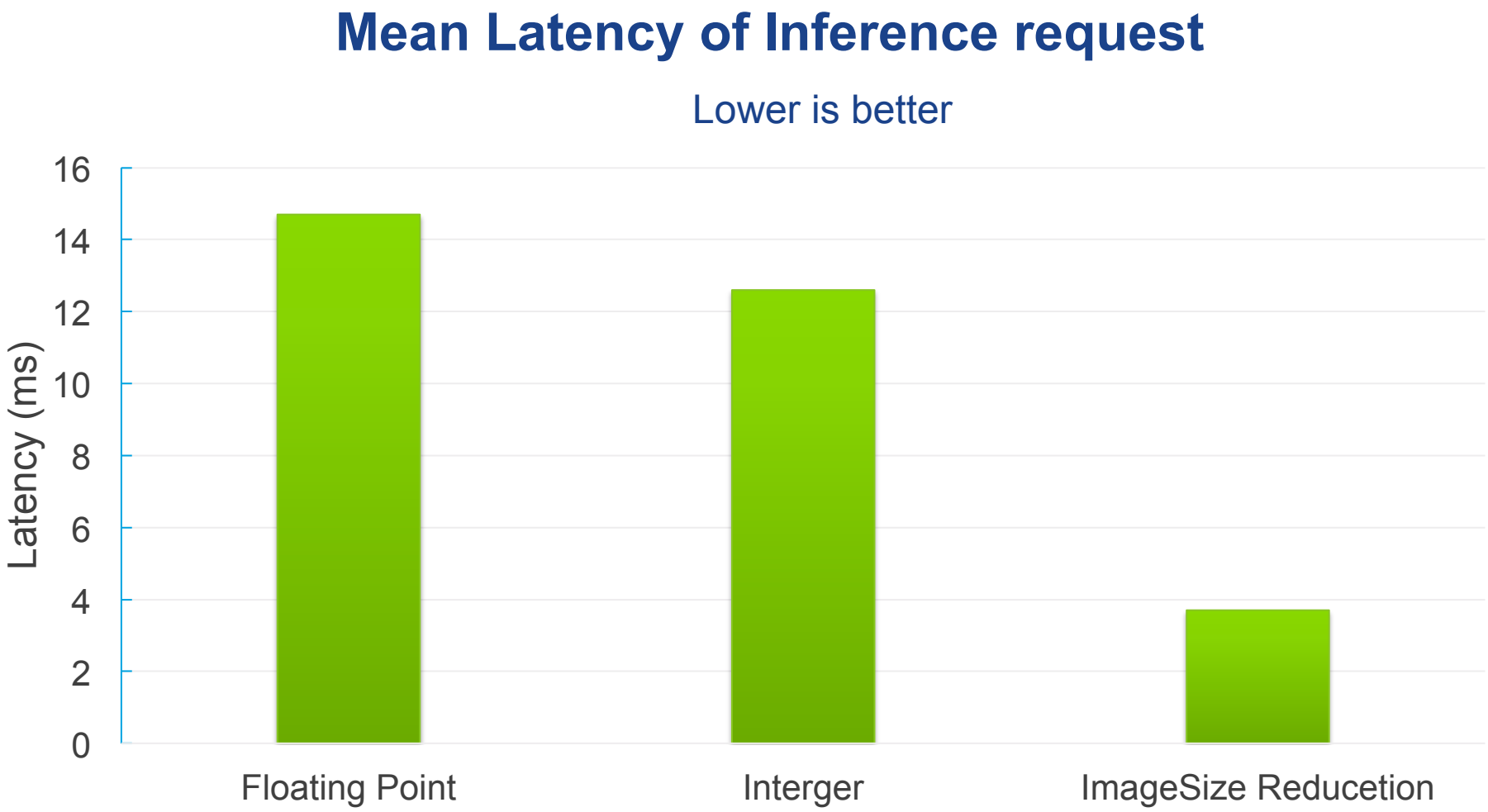
- Quantization: Float → Integer
- Image Size: Large → Small

Using Remote Inference Service

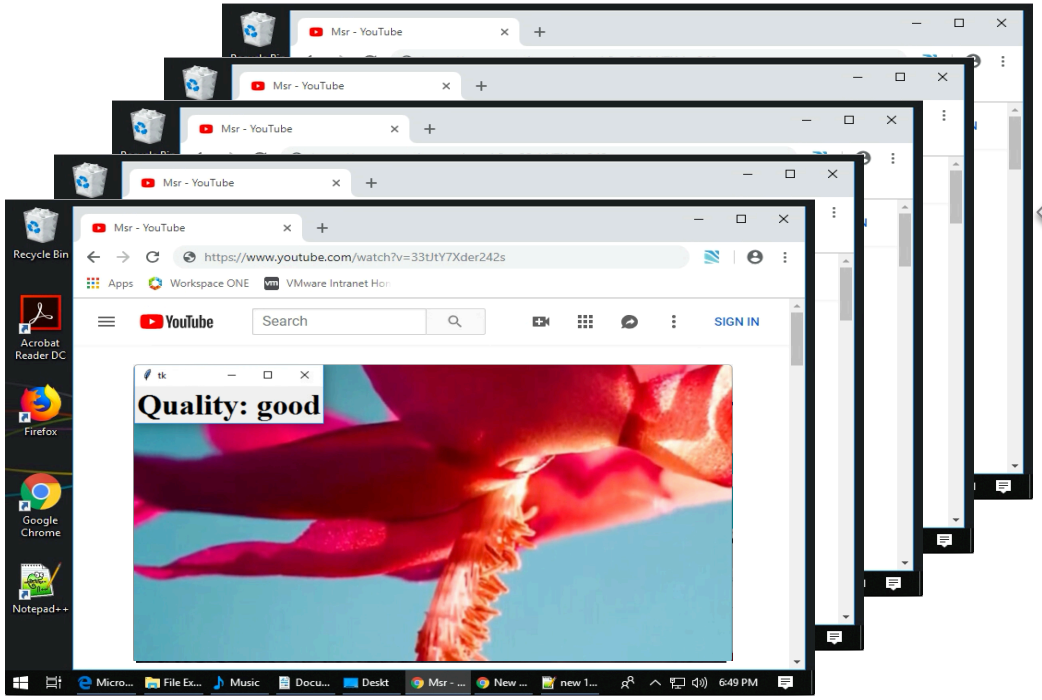
Tuning Neural Networks to Improve Inference Performance



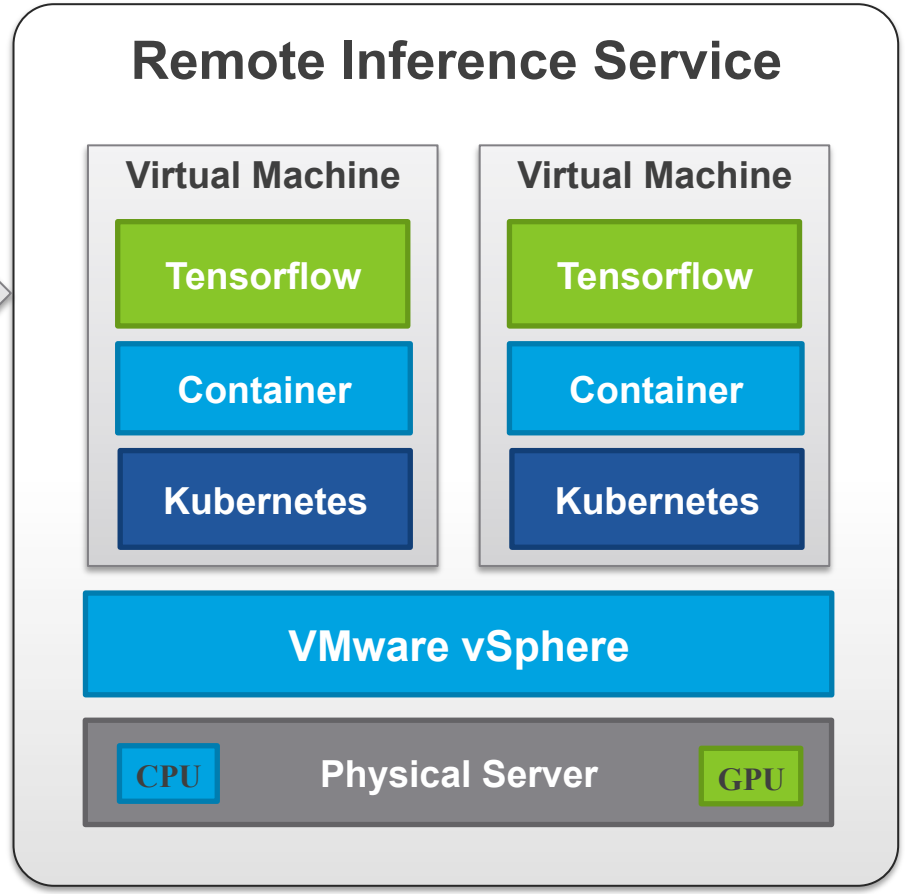
Tuning Neural Networks to Improve Inference Performance



Remote Inference Service to Improve Inference Performance



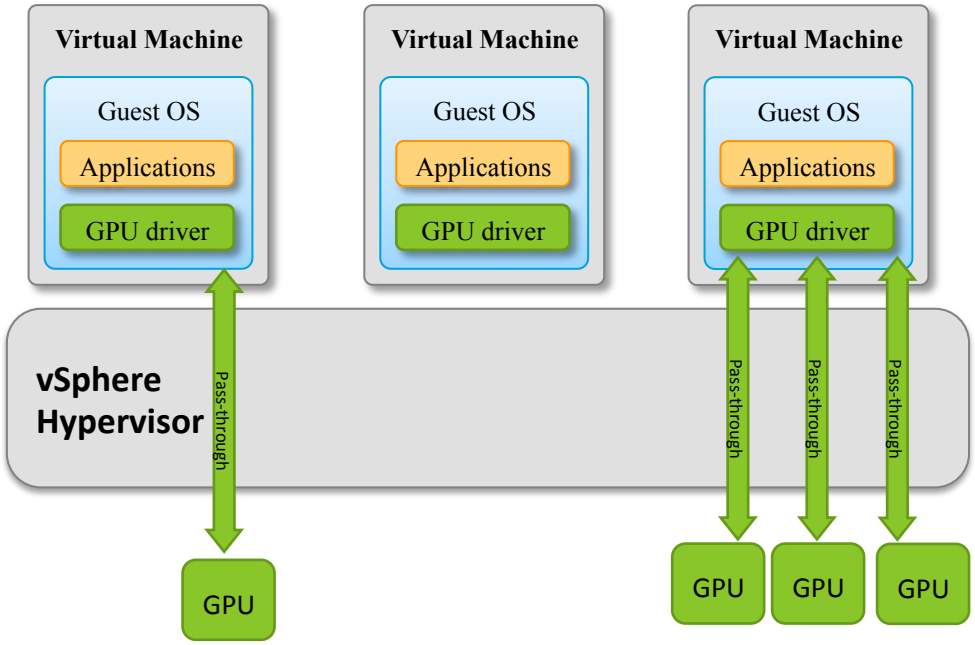
Inference requests



Dell R730 with Intel Haswell CPUs + Nvidia Pascal GPU

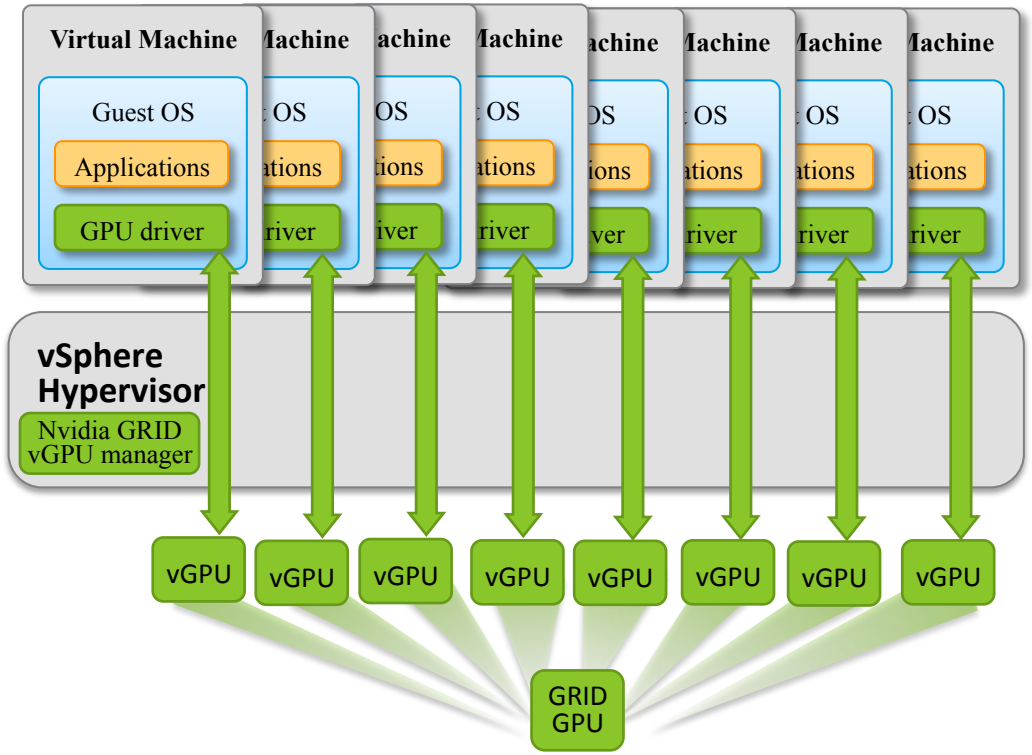
Machine Learning on VMware vSphere using Nvidia GPUs

VMware DirectPath I/O



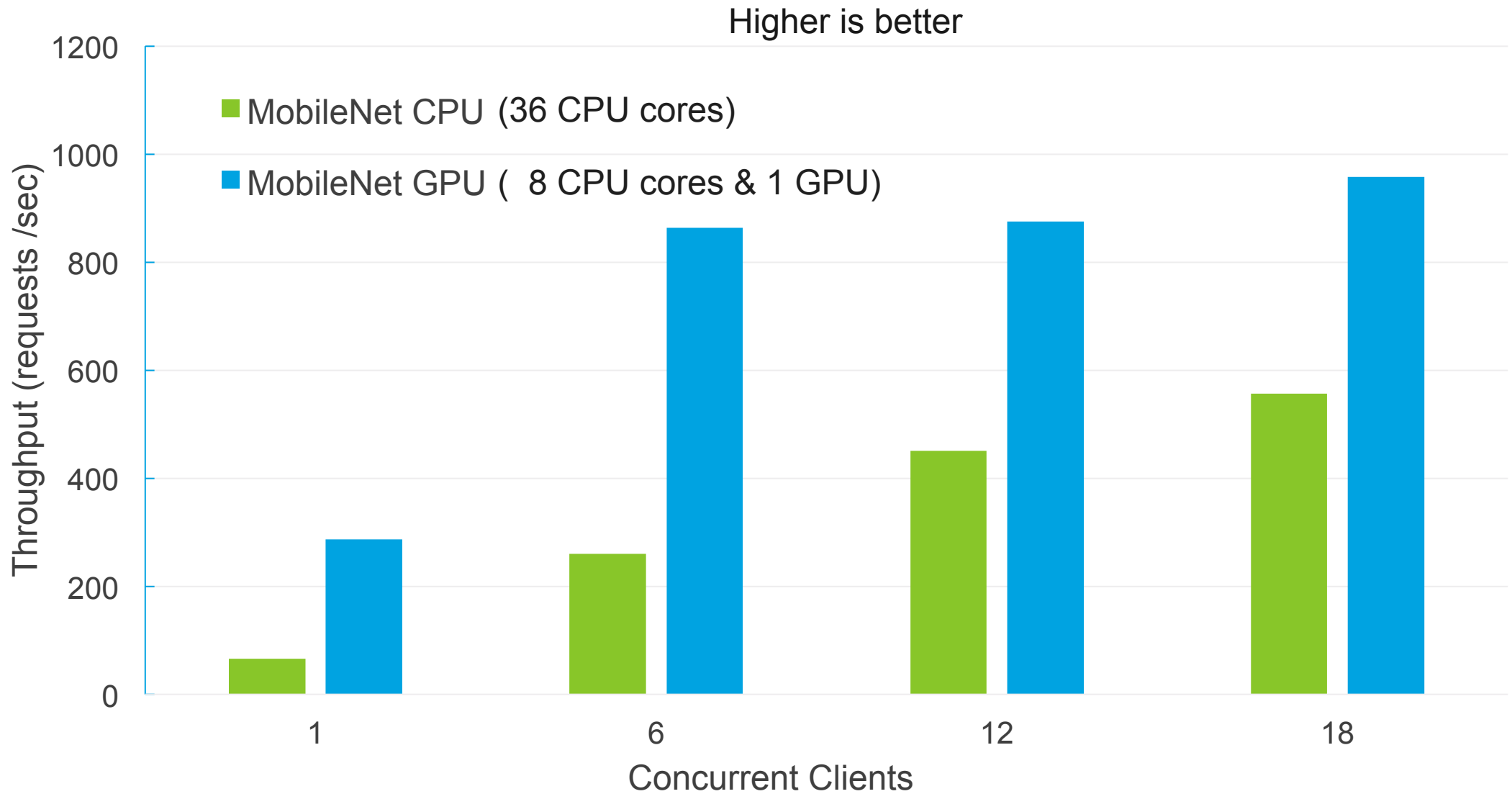
- **One VM** per GPU (no GPU sharing)
- Allow **multiple GPUs** per VM

Nvidia GRID vGPU

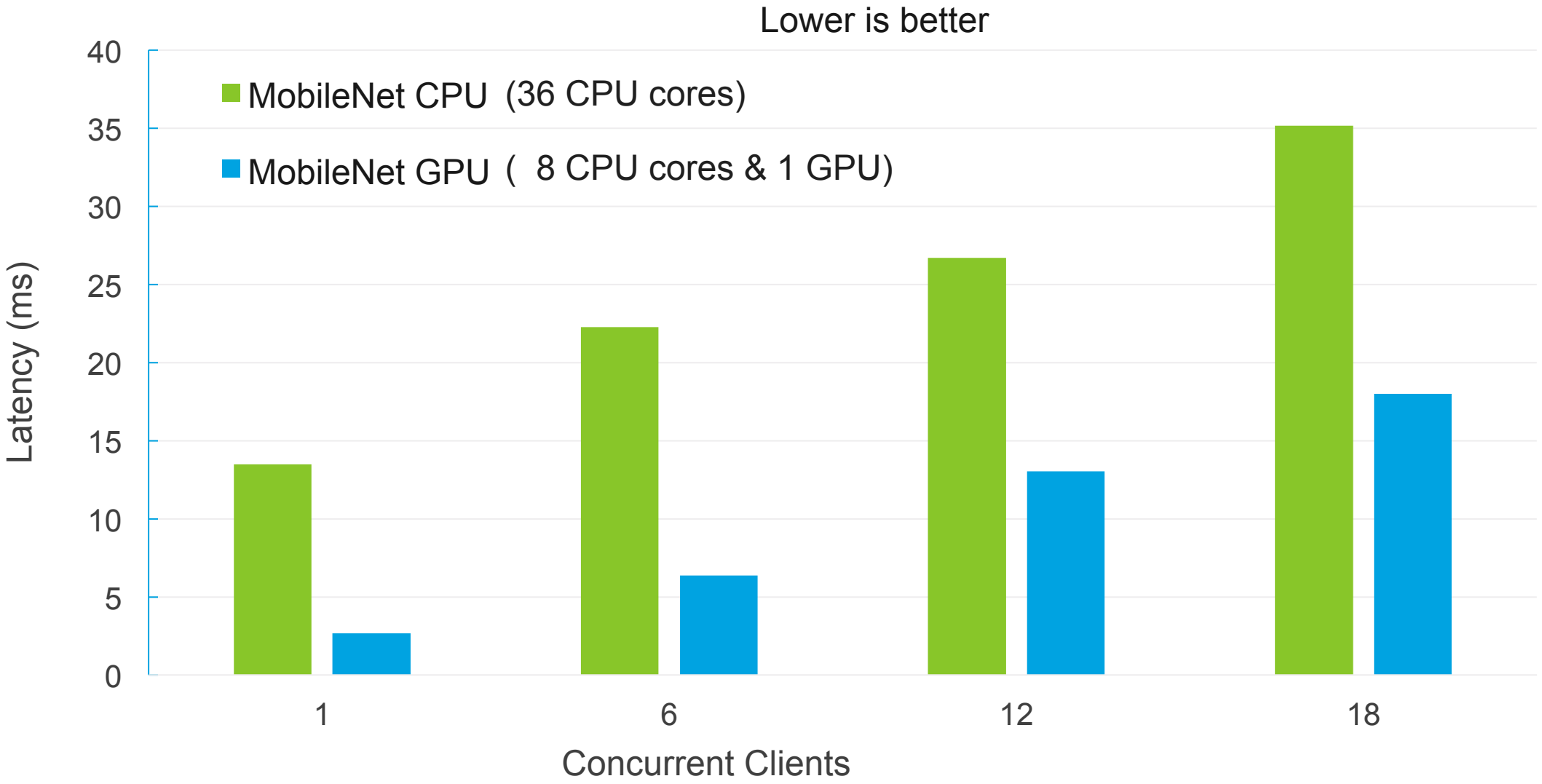


- **Multiple VMs can** share GPU (using GRID vGPU)
- Allow **one vGPU** per VM

Inference Throughputs



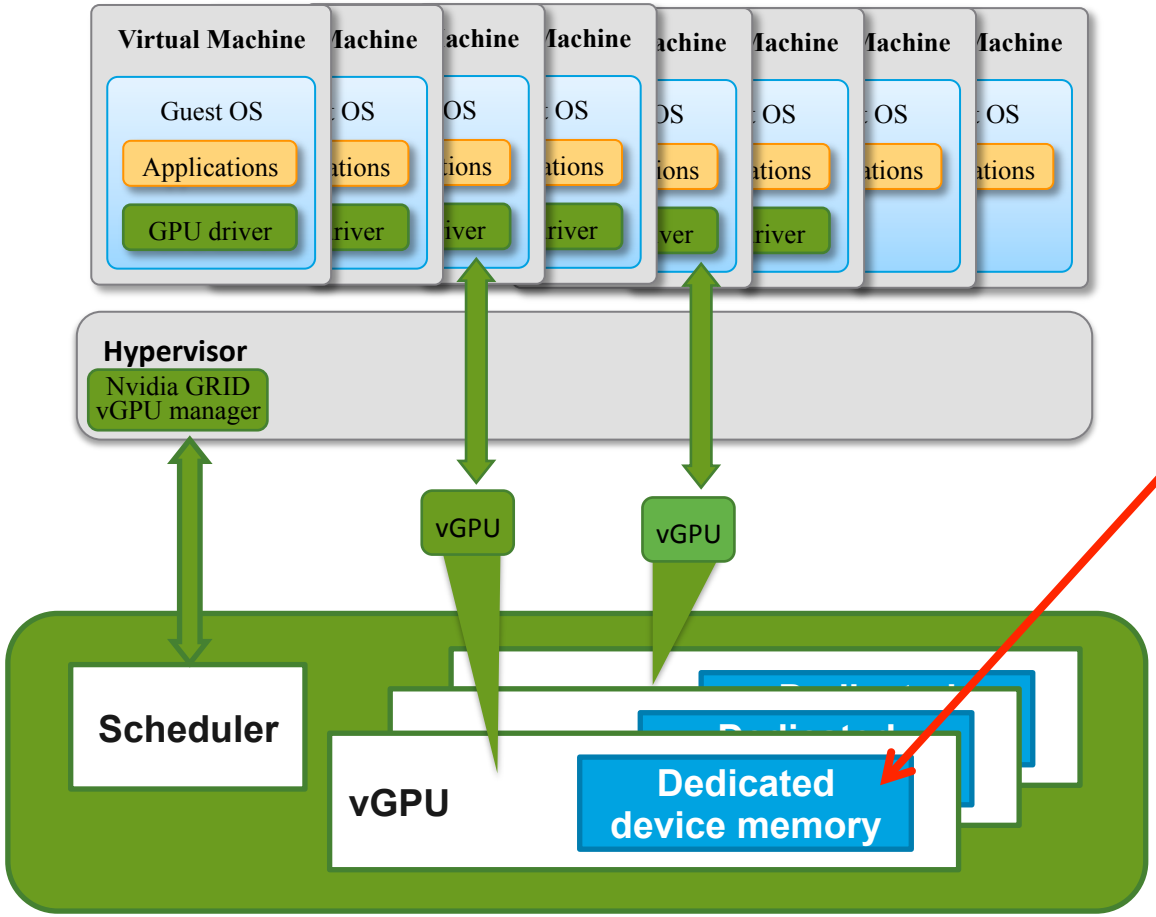
Mean Inference Latency



Sharing GPU for Training with NVIDIA GRID on VMware vSphere

Sharing GPU among VMs with vGPU on VMware vSphere

Nvidia GRID vGPU



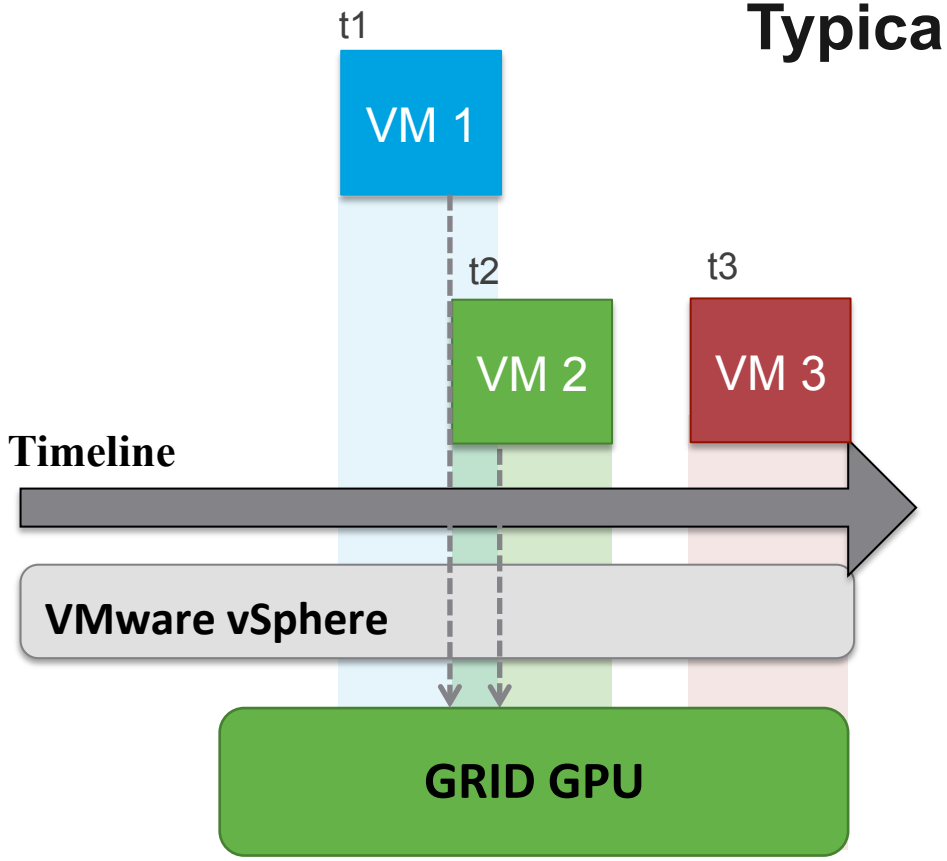
- **Each vGPU is assigned with a profile**
- **Profile defines**:
  - Size of memory each vGPU has
  - How many vGPUs per physical GPU

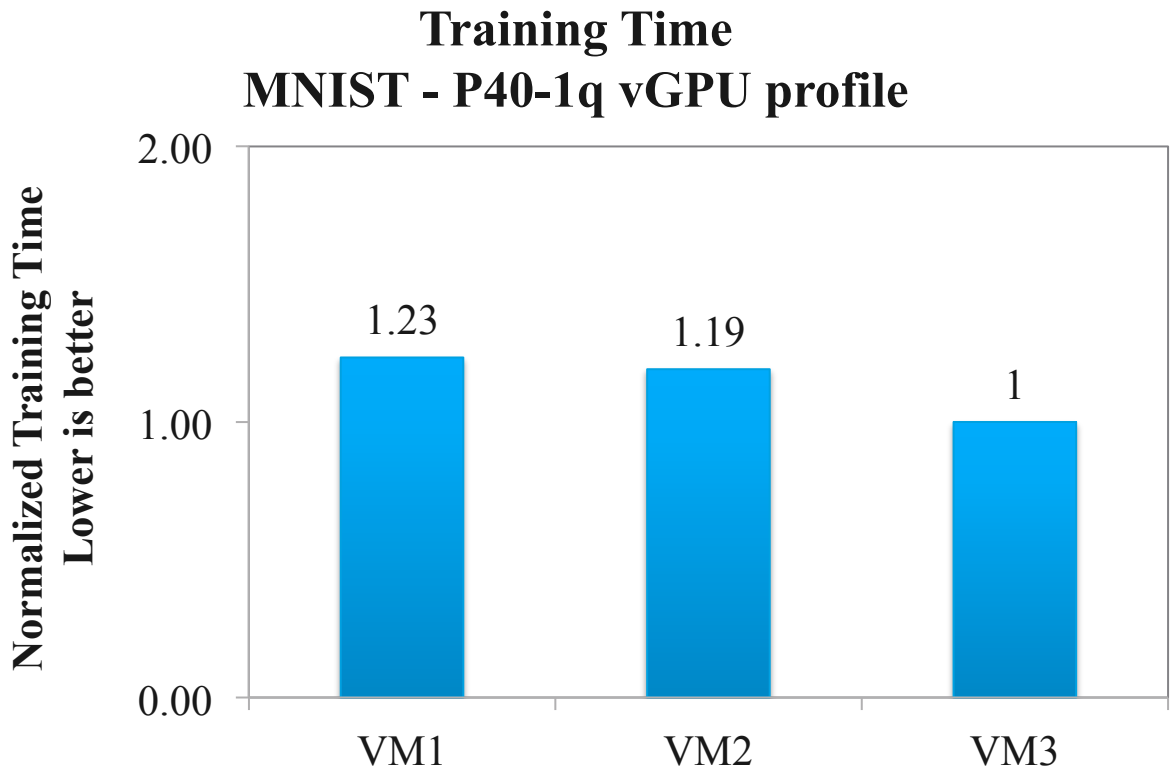
- **For example**: P40-1q profile for P40 GPU
  - vGPU has 1GB of device memory
  - **24** vGPUs per **1** physical P40

Sharing GPU among VMs with vGPU on VMware vSphere

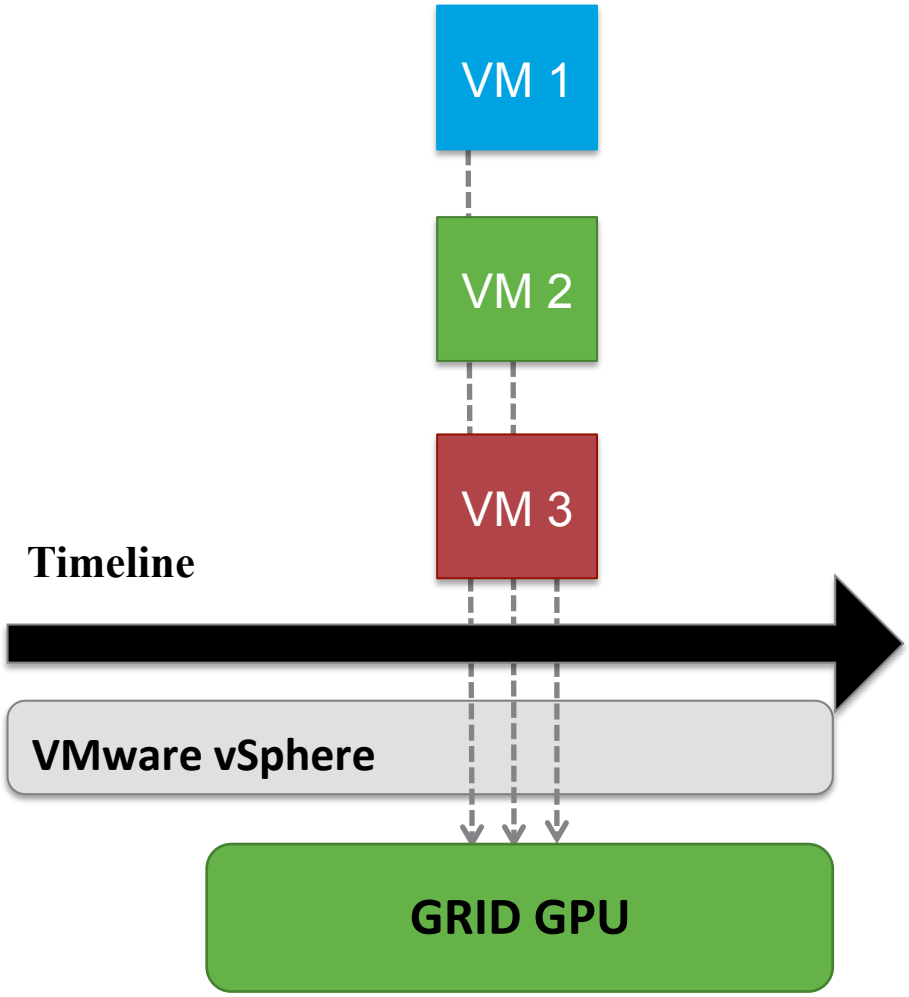
- Many users (e.g. data scientists / engineer) do not use GPUs 24/7
- Sharing GPU reduces hardware cost and increases system utilization

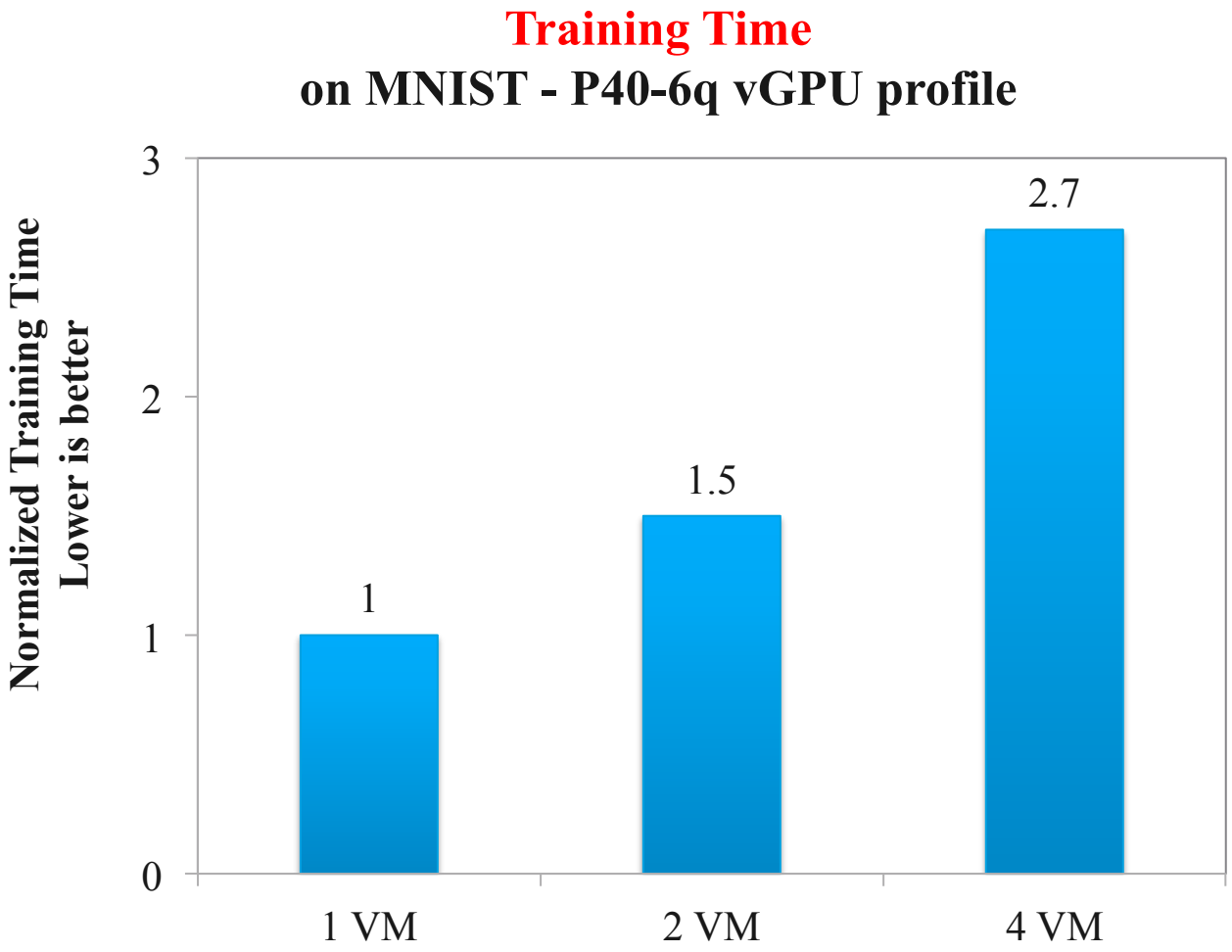
Typical cases of GPU sharing



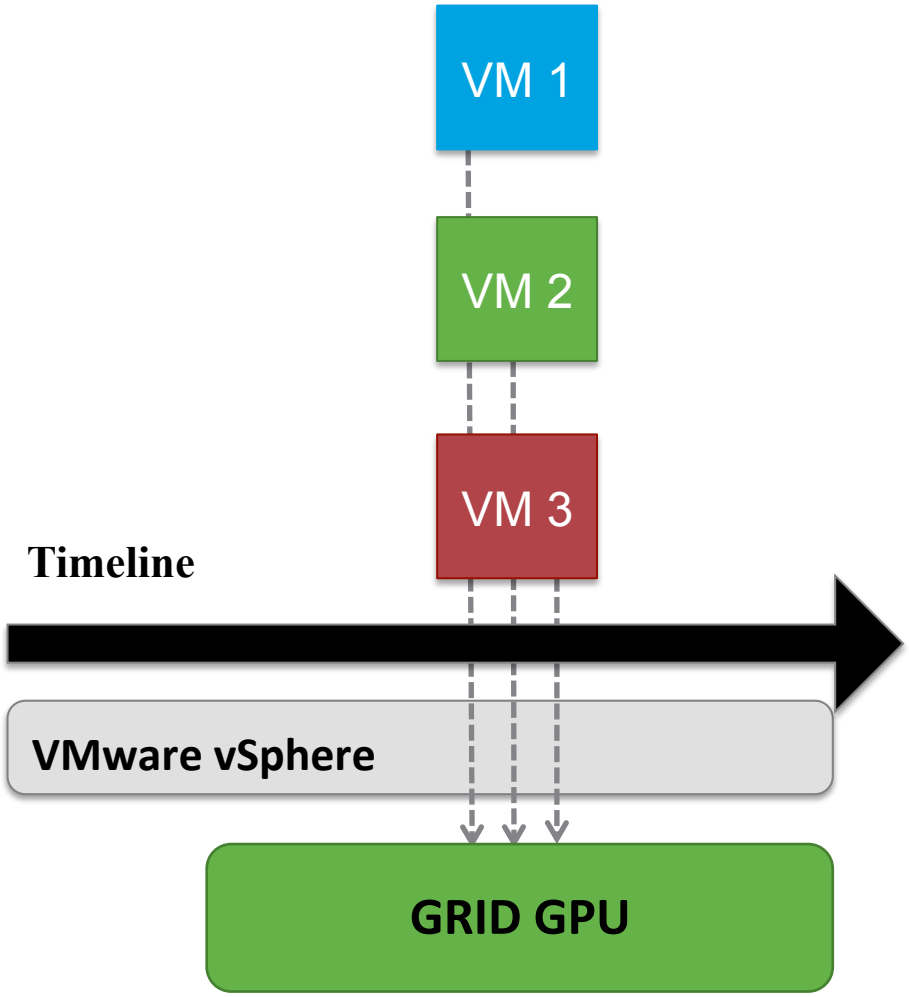


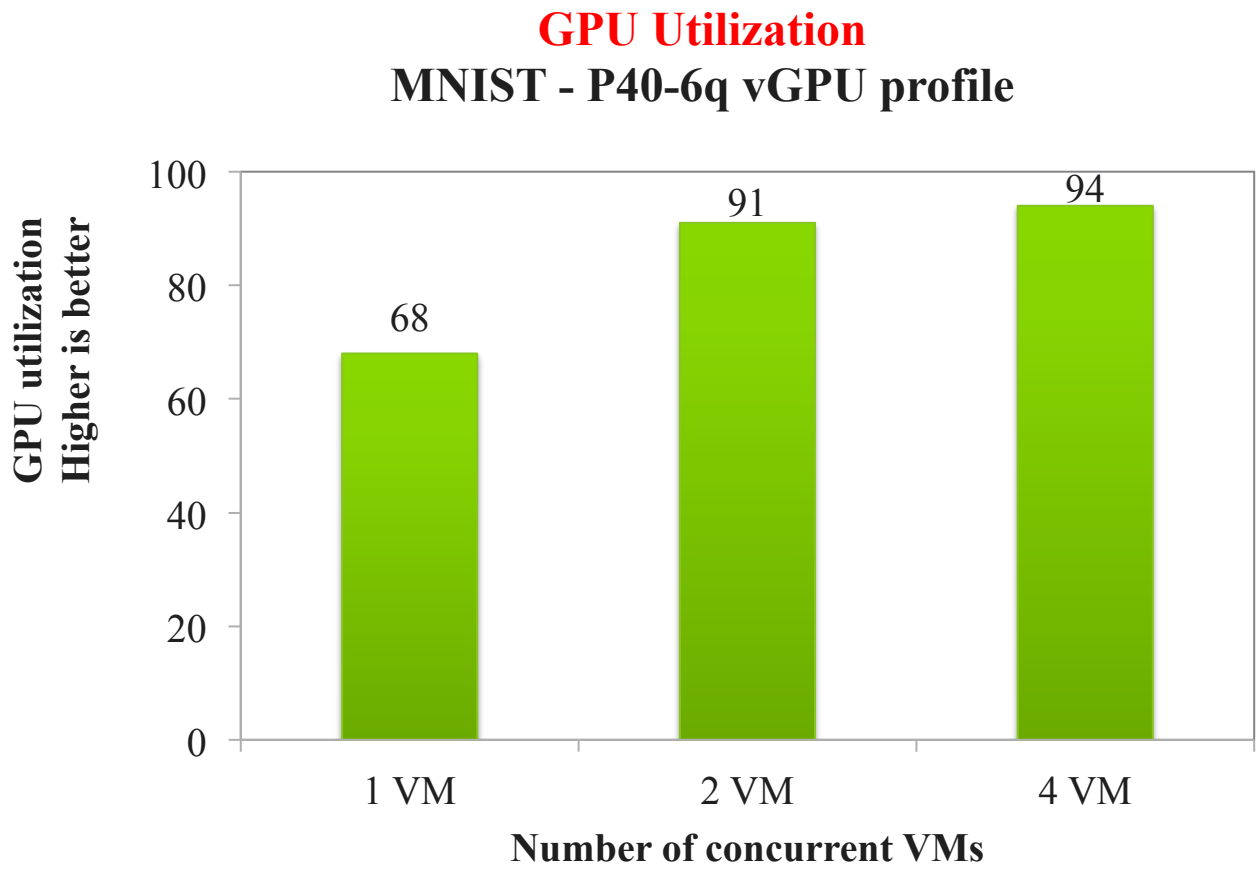
When all VMs use GPU concurrently for TRAINING





When all VMs use GPU concurrently for TRAINING





Conclusion

- Video Audio are important workloads in cloud / datacenter
- Deep Learning is a good solution for Video Audio Quality Assessment
- Large scale Video Audio quality measurement for Horizon using View Planner
- Performance improvement of ML workloads on VMware vSphere with NVIDIA GPUs

Thank you

Contributors

VMware

- Lan Vu (lanv@vmware.com)
- Hari Sivaraman (hsivaraman@vmware.com)
- Uday Kurkure (ukurkure@vmware.com)
- Aravind Bappanadu (abappanadu@vmware.com)

University of Illinois Urbana Champaign

- Dimitrios Skarlatos