Microsoft

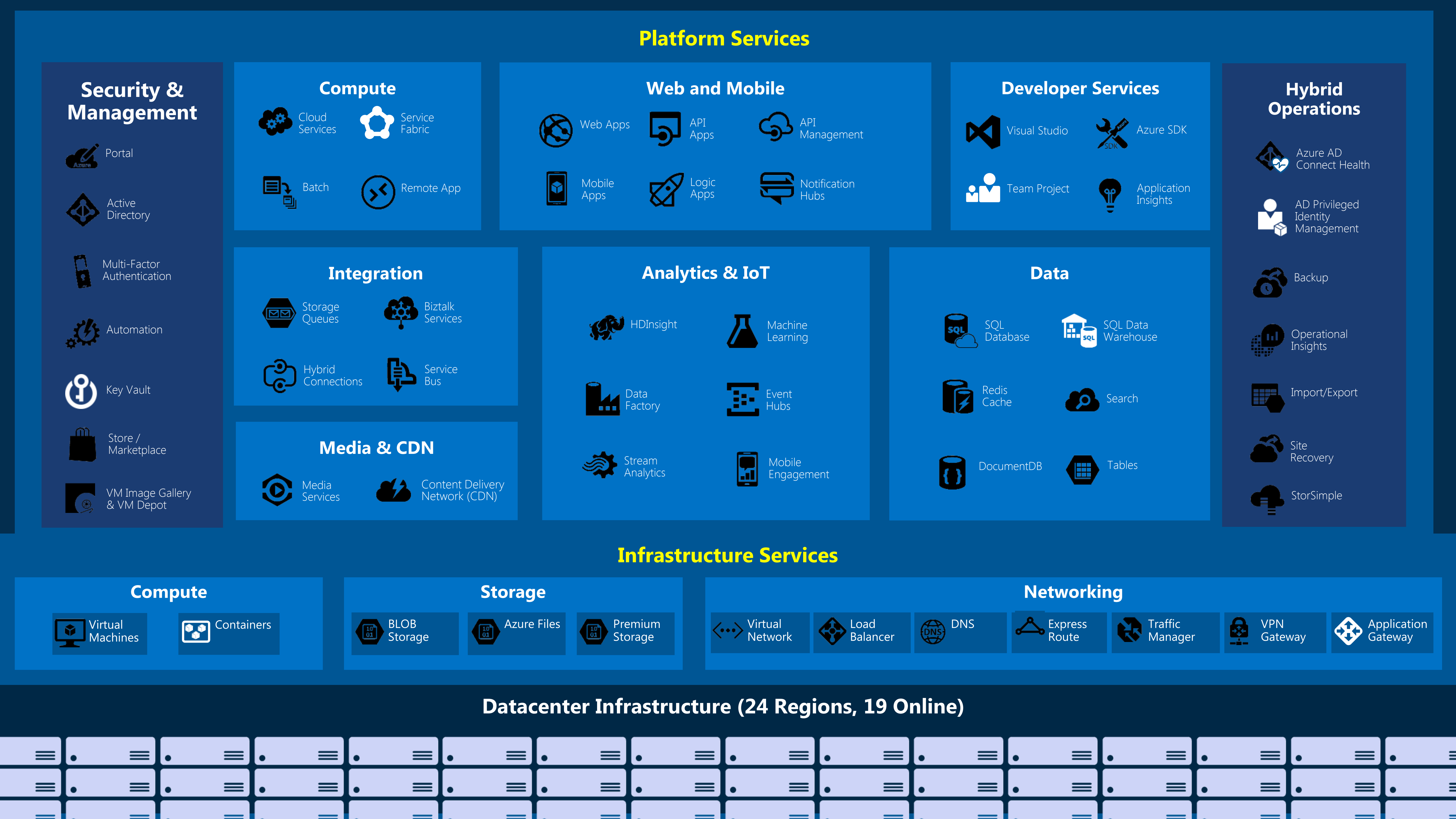
November 17th 2015

Bringing NVIDIA® GPUs to Azure

Mark S. Staveley, PhD

Senior Program Manager

Azure High Performance Computing



Platform Services

Security & Management

- Portal
- Active Directory
- Multi-Factor Authentication
- Automation
- Key Vault
- Store / Marketplace
- VM Image Gallery & VM Depot

Compute

- Cloud Services
- Service Fabric
- Batch
- Remote App

Web and Mobile

- Web Apps
- API Apps
- API Management
- Mobile Apps
- Logic Apps
- Notification Hubs

Developer Services

- Visual Studio
- Azure SDK
- Team Project
- Application Insights

Hybrid Operations

- Azure AD Connect Health
- AD Privileged Identity Management
- Backup
- Operational Insights
- Import/Export
- Site Recovery
- StorSimple

Integration

- Storage Queues
- Biztalk Services
- Hybrid Connections
- Service Bus

Media & CDN

- Media Services
- Content Delivery Network (CDN)

Analytics & IoT

- HDInsight
- Machine Learning
- Data Factory
- Event Hubs
- Stream Analytics
- Mobile Engagement

Data

- SQL Database
- SQL Data Warehouse
- Redis Cache
- Search
- DocumentDB
- Tables

Infrastructure Services

Compute

- Virtual Machines
- Containers

Storage

- BLOB Storage
- Azure Files
- Premium Storage

Networking

- Virtual Network
- Load Balancer
- DNS
- Express Route
- Traffic Manager
- VPN Gateway
- Application Gateway

Datacenter Infrastructure (24 Regions, 19 Online)

Vision and Design

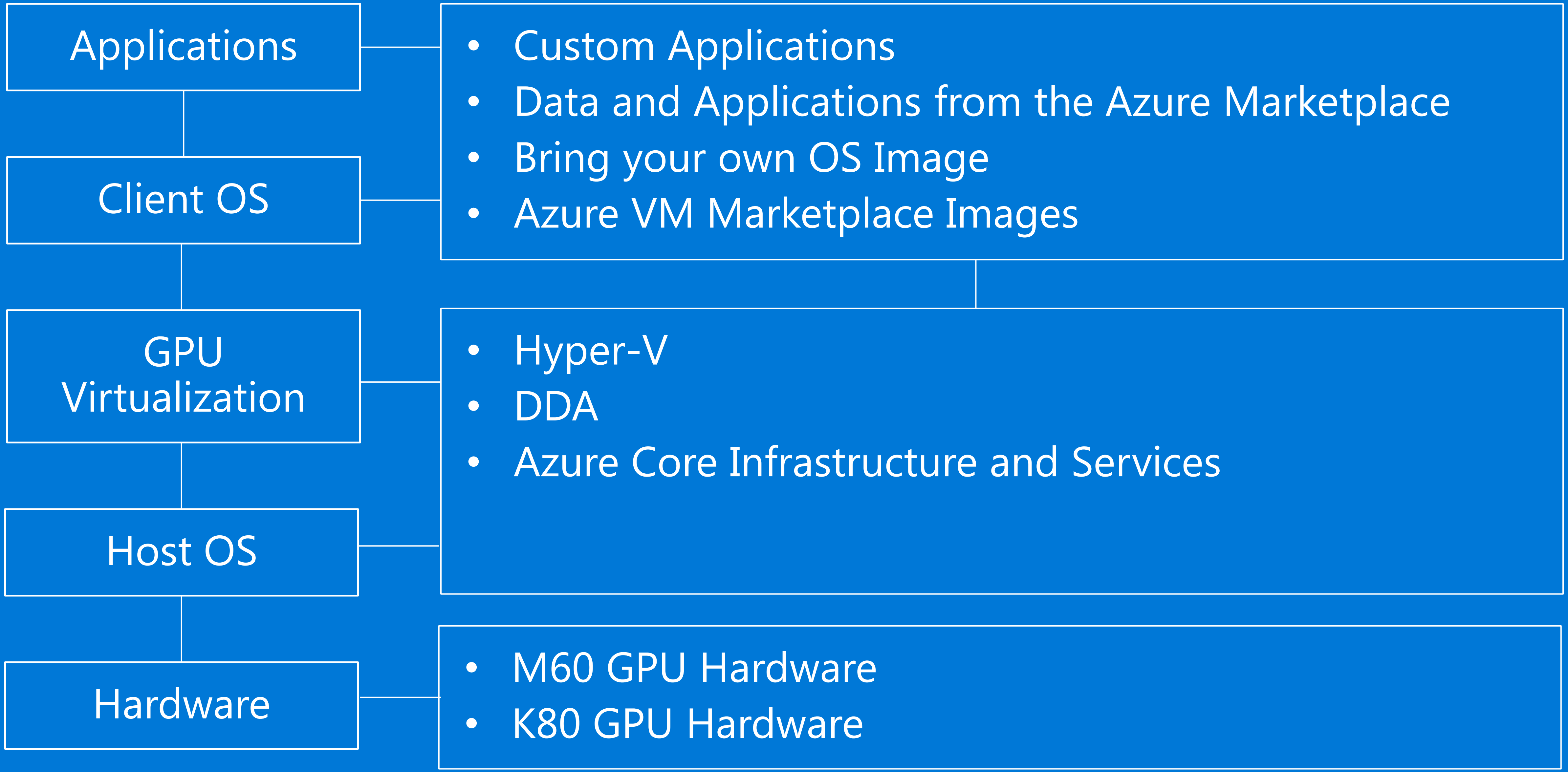
Vision

- ✓ Integrating GPU capabilities into Azure Infrastructure
- ✓ Competitive Price and Performance
- ✓ Supporting both Compute and High-End Visualization
- ✓ Partnership with NVIDIA

Core Scenarios

- ✓ Cloud-based Streaming and Gaming
- ✓ Video Processing / Encoding Workloads
- ✓ Accelerated Desktop Applications (OpenGL and DirectX)
- ✓ GPU Compute (CUDA and OpenCL) - *single and multiple machine workloads*

Cloud Architecture



Applications

Client OS

GPU Virtualization

Host OS

Hardware

- Custom Applications
- Data and Applications from the Azure Marketplace
- Bring your own OS Image
- Azure VM Marketplace Images

- Hyper-V
- DDA
- Azure Core Infrastructure and Services

- M60 GPU Hardware
- K80 GPU Hardware

GPU VM Offerings (N-Series)

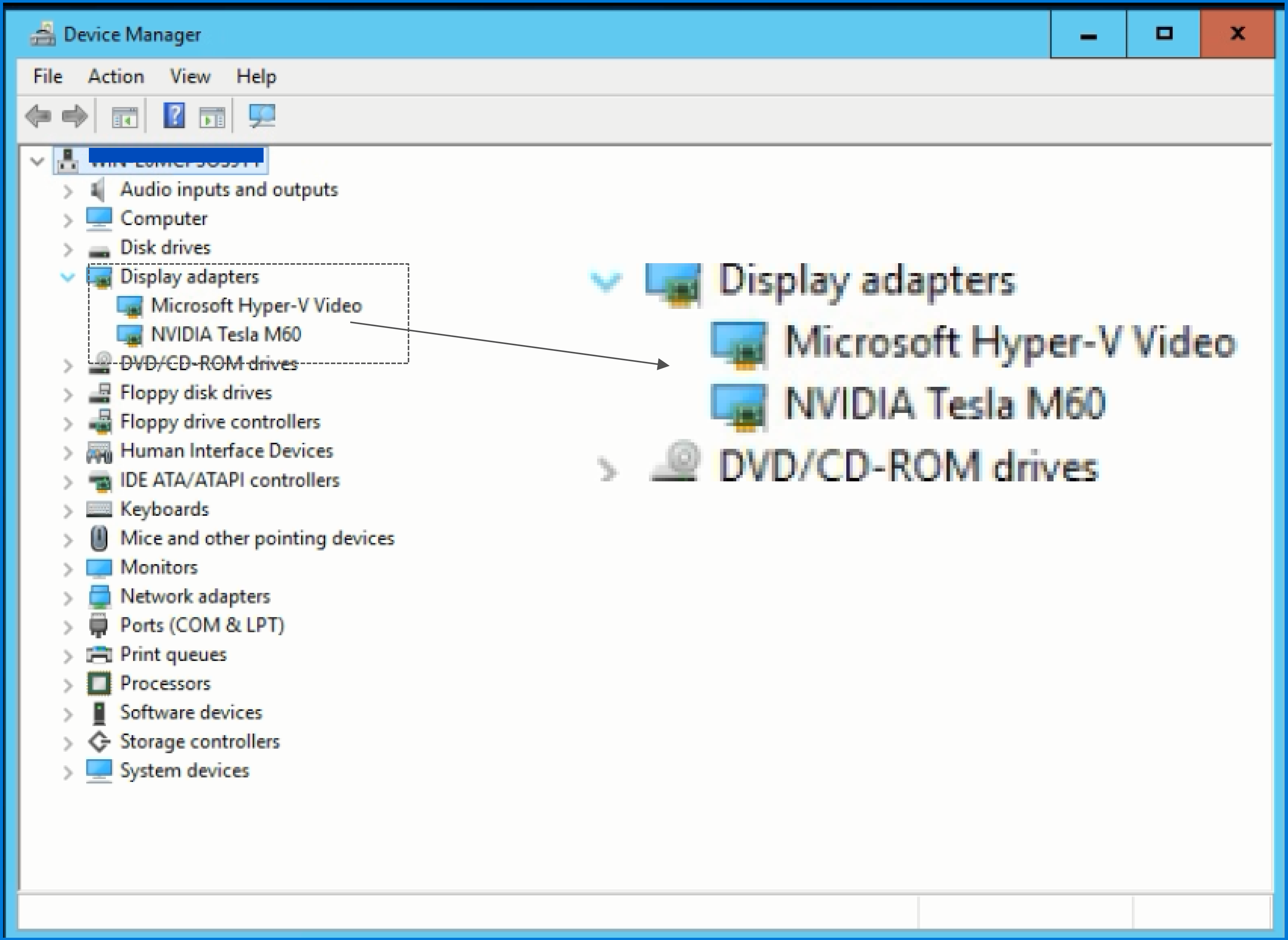
	N1	N2	N10	N11	N12	N21
CPU Cores (E5-2690v3)	6	24	6	12	24	24
RAM (GB)	64	256	64	128	256	256
SSD (TB)	~0.5	~2.0TB	~0.5	~1.0TB	~2.0TB	~2.0TB
Network	Azure Network	Azure Network	Azure Network	Azure Network	Azure Network	Azure Network *RDMA Dedicated Back End*
GPU Resources	1 x M60 GPU (1/2 Physical Card)	4 x M60 GPU (2 Physical Cards)	1 x K80 GPU (1/2 Physical Card)	2 x K80 GPUs (1 Physical Card)	4 x K80 GPUs (2 Physical Cards)	4 x K80 GPUs (2 Physical Cards)

Visualization Capabilities

(N1 & N2)

GPU VM Offerings (N-Series)

	N1	**N2**
CPU Cores (E5-2690v3)	6	24
RAM (GB)	64	256
SSD (TB)	~0.5	~2.0TB
Network	Azure Network	Azure Network
GPU Resources	1 x M60 GPU (1/2 Physical Card)	4 x M60 GPU (2 Physical Cards)

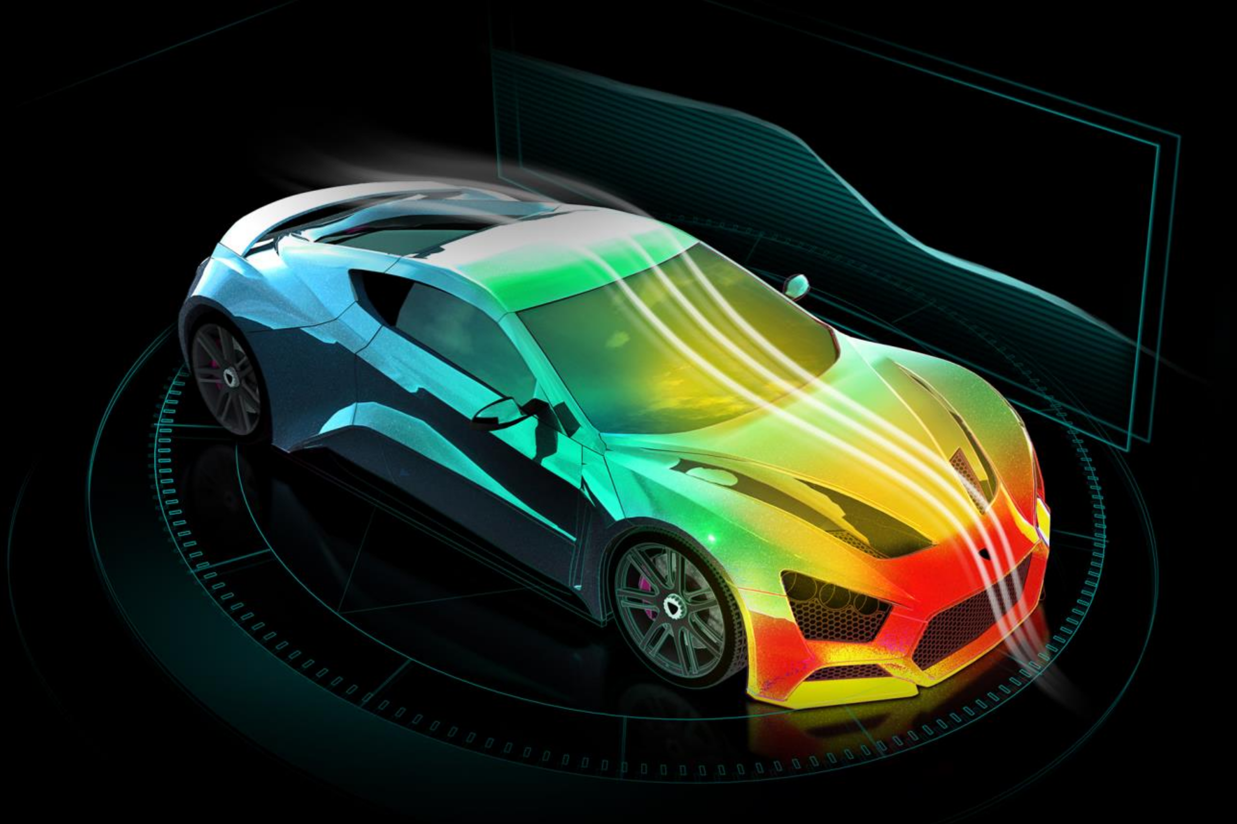


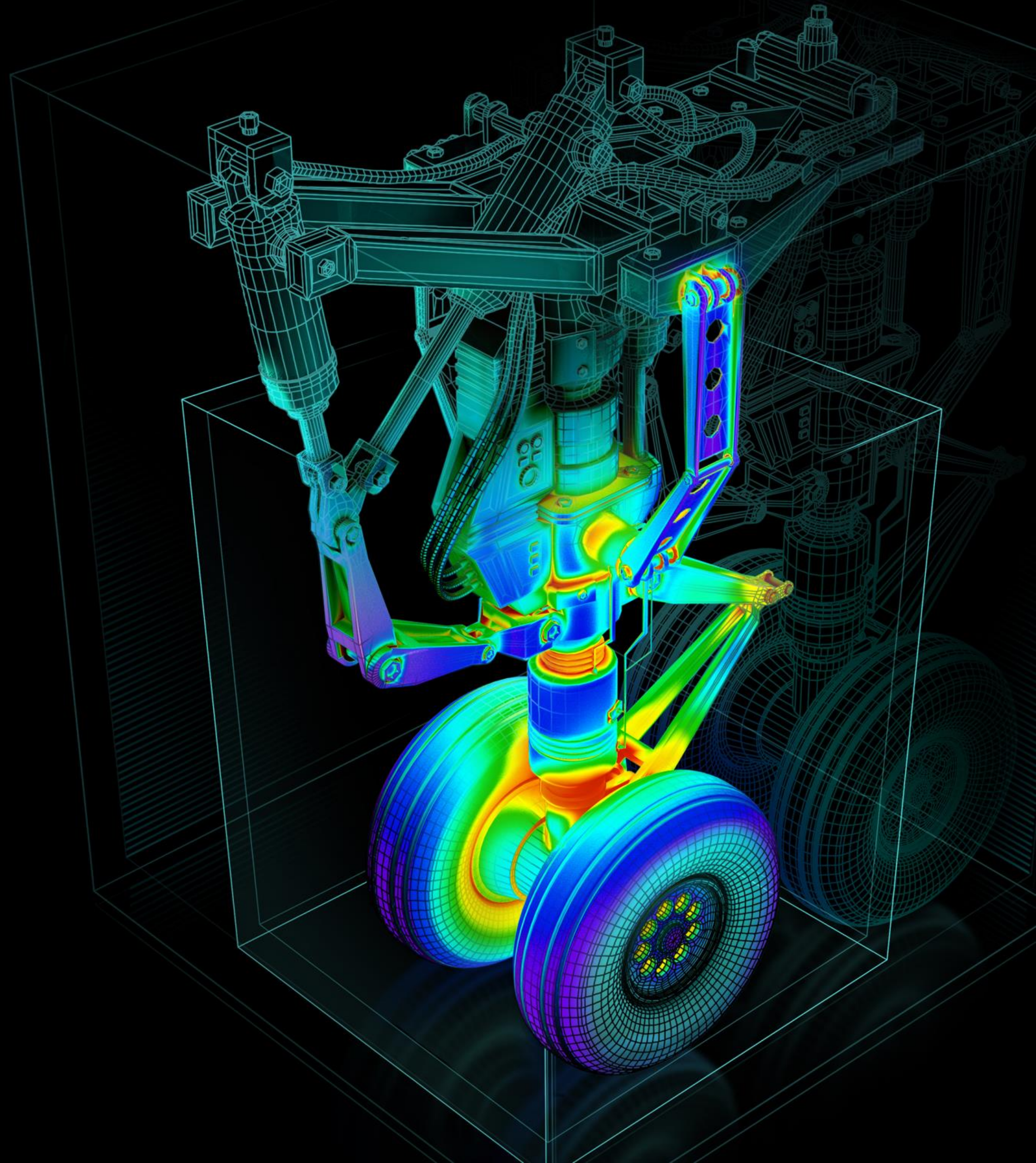
[redacted]
- Audio inputs and outputs
- Computer
- Disk drives
- Display adapters
  - Microsoft Hyper-V Video
  - NVIDIA Tesla M60
- DVD/CD-ROM drives
- Floppy disk drives
- Floppy drive controllers
- Human Interface Devices
- IDE ATA/ATAPI controllers
- Keyboards
- Mice and other pointing devices
- Monitors
- Network adapters
- Ports (COM & LPT)
- Print queues
- Processors
- Software devices
- Storage controllers
- System devices

Display adapters
Microsoft Hyper-V Video
NVIDIA Tesla M60
DVD/CD-ROM drives

Customer and Partner Impact

- ✓ Enterprise Class Visualization + Azure Infrastructure
- ✓ Diverse Application Support
- ✓ Remote Desktop Services on IaaS



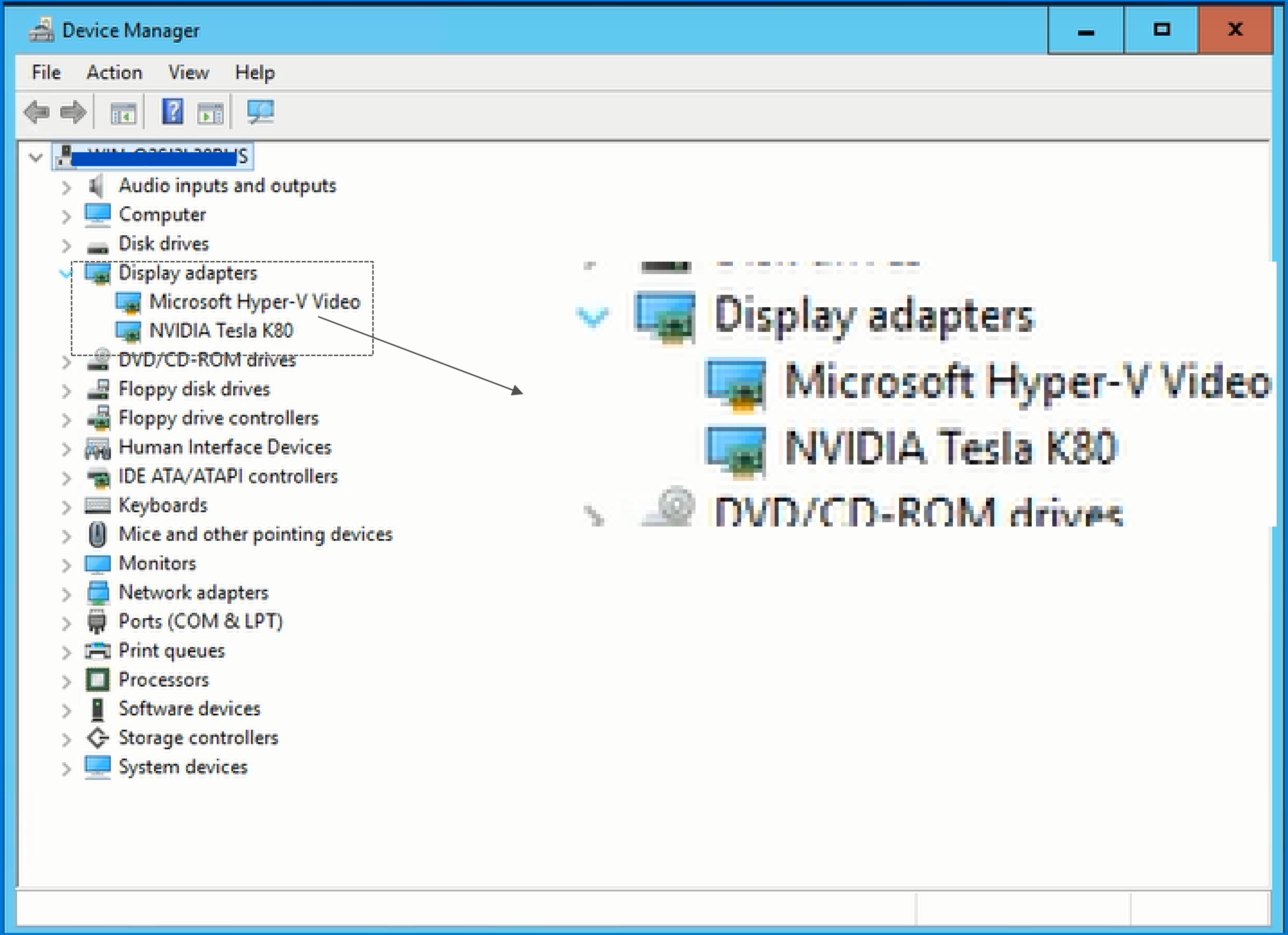


GPU Compute Single Machine

(N10, N11, N12)

GPU VM Offerings (N-Series)

	N10	**N11**	**N12**
CPU Cores (E5-2690v3)	6	12	24
RAM (GB)	64	128	256
SSD (TB)	~0.5	~1.0TB	~2.0TB
Network	Azure Network	Azure Network	Azure Network
GPU Resources	1 x K80 GPU (1/2 Physical Card)	2 x K80 GPUs (1 Physical Card)	4 x K80 GPUs (2 Physical Cards)

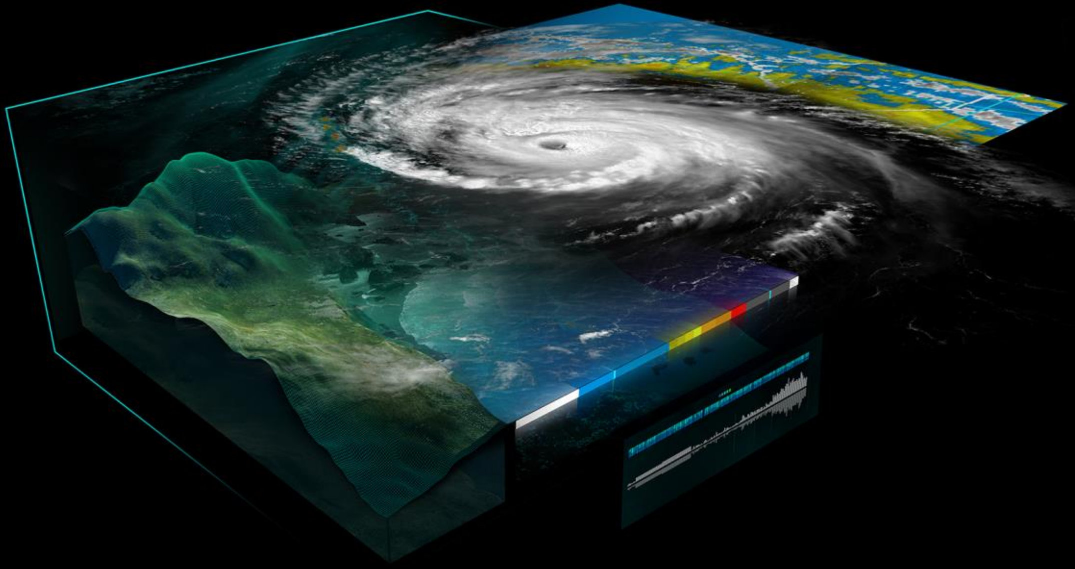


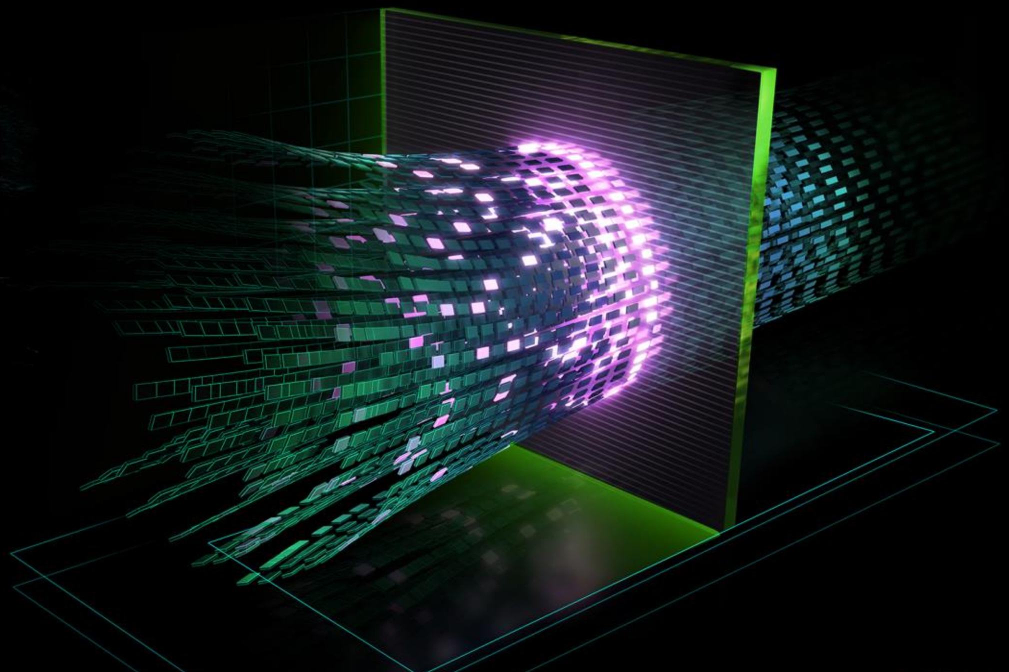
[redacted]
- Audio inputs and outputs
- Computer
- Disk drives
- Display adapters
  - Microsoft Hyper-V Video
  - NVIDIA Tesla K80
- DVD/CD-ROM drives
- Floppy disk drives
- Floppy drive controllers
- Human Interface Devices
- IDE ATA/ATAPI controllers
- Keyboards
- Mice and other pointing devices
- Monitors
- Network adapters
- Ports (COM & LPT)
- Print queues
- Processors
- Software devices
- Storage controllers
- System devices

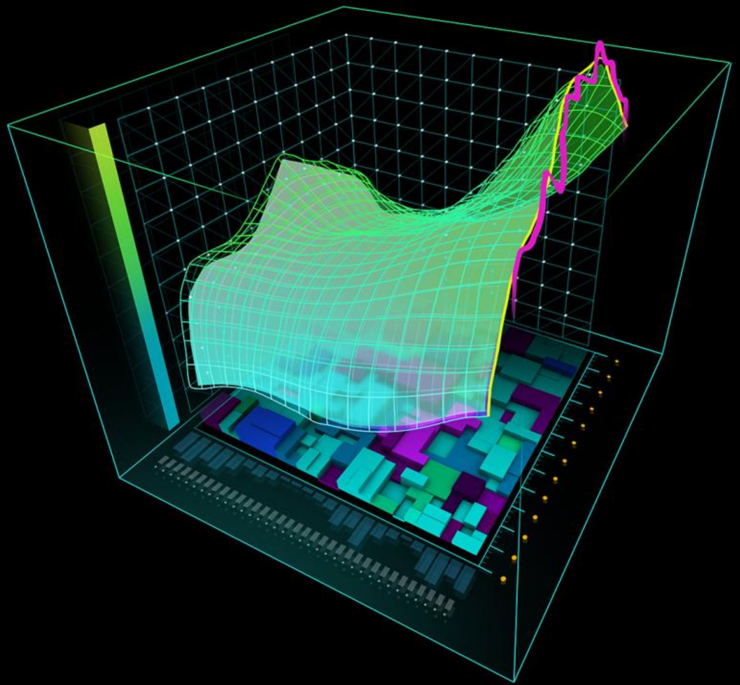
Display adapters
Microsoft Hyper-V Video
NVIDIA Tesla K80
DVD/CD-ROM drives

Customer and Partner Impact

- ✓ Azure ML provides access to state-of-the-art machine learning in the cloud
- ✓ GPUs are the most preferred platform for Deep Neural Network training
- ✓ AzureML allows composing sophisticated experiments with many stages and transforms
- ✓ Integration with existing DB and Hadoop Infrastructure on Azure.







GPU Compute Multi-Machine

(N21)

GPU VM Offerings (N-Series)

	N21
CPU Cores (E5-2690v3)	24
RAM (GB)	256
SSD (TB)	~2.0TB
Network	Azure Network *RDMA Dedicated Back End*
GPU Resources	4 x K80 GPUs (2 Physical Cards)

Customer and Partner Impact

- ✓ Build your own GPU Cluster on Azure
- ✓ Impact on Time to Innovation
- ✓ Why is this special for our customers?

GPUs + Azure + MS Research = *Endless Possibilities*

GPU VM Offerings (N-Series)

	N21	Azure GPU Research Labs
CPU Cores (E5-2690v3)	24	
RAM (GB)	256	
SSD (TB)	~2.0TB	
Network	Azure Network *RDMA Dedicated Back End*	
GPU Resources	4 x K80 GPUs (2 Physical Cards)	

Azure GPU Research Labs

Coming Soon

- ✓ Azure GPU service specialized for distributed DNN training
- ✓ The same services we use internally for large scale training
- ✓ Ability to support single jobs with hundreds of GPUs
- ✓ Big data, intensive algorithms: Speech, Image, Text: LSTM, ASGD

GPU Program Summary

GPU Program Summary

- ✓ Private Preview for N-Series GPUs coming in the next few months.
- ✓ Working closely with partners to support Visualization and Compute Workloads.
- ✓ Plans to support Windows and Linux OS's for N-Series Virtual Machines.
- ✓ Research Partners will also have an opportunity to work with Azure GPU Research Labs

Microsoft