Microsoft

HPC and more……in Azure

Debora Schampers
Principal Solution Specialist - Microsoft Azure - Cloud Computing & HPC Computing | Microsoft Italy

Microsoft Azure Big Compute

Our mission is to enable researchers, engineers, quants, designers, developers, and data scientists to achieve radically better results by making it easy to do simulation and parallel computing in the cloud at hyper-scale with a variety of VM's

Enterprise-ready

Flexibility & choice

Cloud Solutions

Portfolio of choices

Entry Level VMs

Dev/Test
Workolads

General Purpose VMs

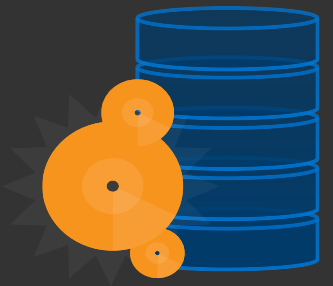
Common Applications,
Web servers etc

Compute Optimized VMs

Gaming, Analytics

Large Memory VMs

Large Databases



>80,000 IOPs Premium Storage

Low latency, high throughput apps

High Performance VMs

HPC simulations with high-bandwidth & low-latency interconnect

Storage optimized VMs

No SQL Databases (Cassandra, MongoDB), Data warehousing

GPU-enabled VMs

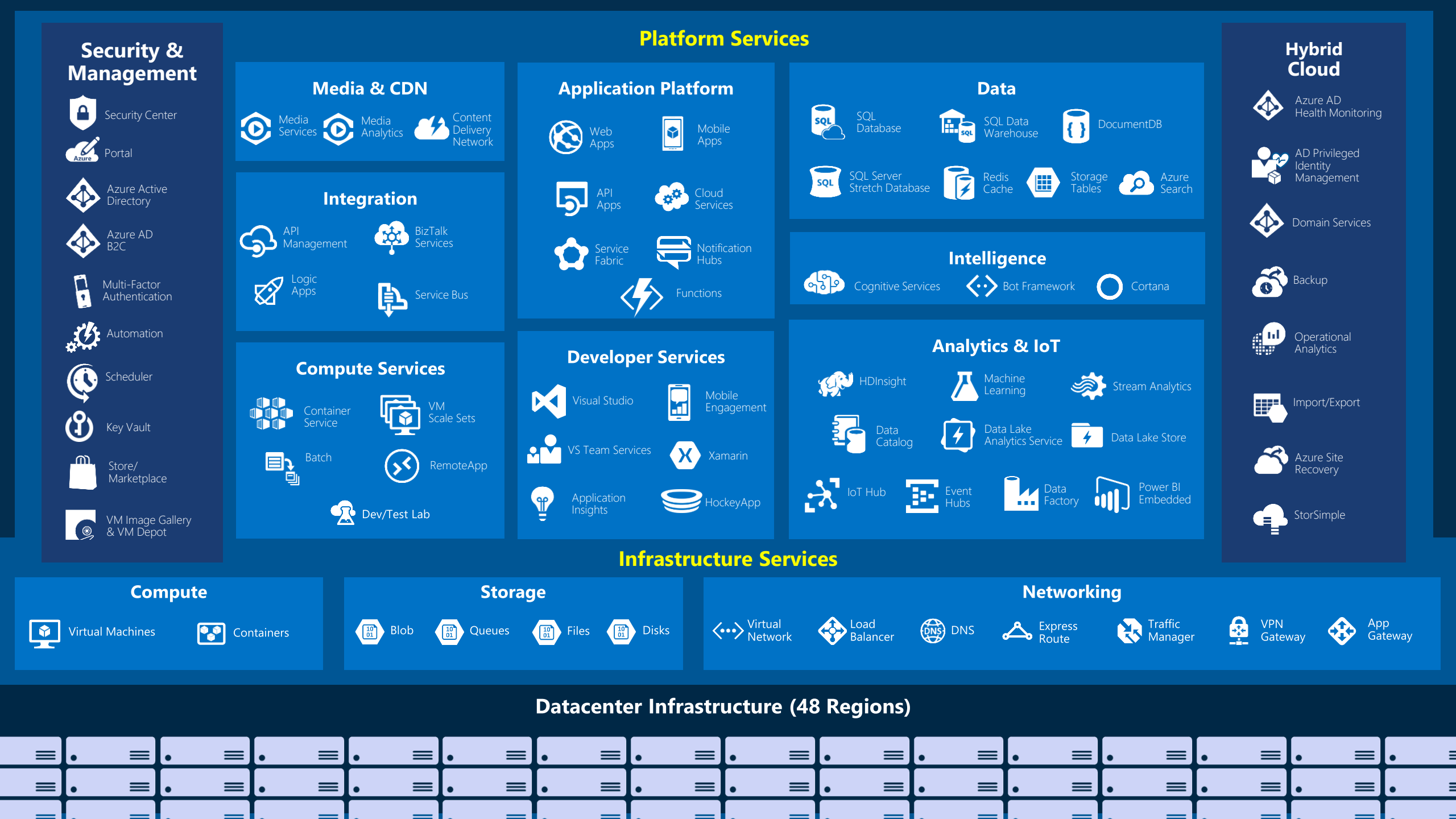
NV: Graphic based applications
NC1: Accelerated Compute (K80s)
NC2: Accelerated Compute + training (P100)
ND1: AI Inferencing + training (P40)
ND2*: AI Training (V100 SXM)

FPGA*

Virtual Machines – HPC
FPGA Microservices* – AI/Edge

Cray Services in Azure*

IB Connected CPU/GPU/Storage available in cloud

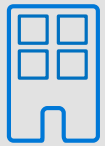

Platform Services
Security & Management
Security Center
Portal
Azure Active Directory
Azure AD B2C
Multi-Factor Authentication
Automation
Scheduler
Key Vault
Store/ Marketplace
VM Image Gallery & VM Depot
Media & CDN
Media Services
Media Analytics
Content Delivery Network
Integration
API Management
BizTalk Services
Logic Apps
Service Bus
Compute Services
Container Service
VM Scale Sets
Batch
RemoteApp
Dev/Test Lab
Application Platform
Web Apps
Mobile Apps
API Apps
Cloud Services
Service Fabric
Notification Hubs
Functions
Developer Services
Visual Studio
Mobile Engagement
VS Team Services
Xamarin
Application Insights
HockeyApp
Data
SQL Database
SQL Data Warehouse
DocumentDB
SQL Server Stretch Database
Redis Cache
Storage Tables
Azure Search
Intelligence
Cognitive Services
Bot Framework
Cortana
Analytics & IoT
HDInsight
Machine Learning
Stream Analytics
Data Catalog
Data Lake Analytics Service
Data Lake Store
IoT Hub
Event Hubs
Data Factory
Power BI Embedded
Hybrid Cloud
Azure AD Health Monitoring
AD Privileged Identity Management
Domain Services
Backup
Operational Analytics
Import/Export
Azure Site Recovery
StorSimple
Infrastructure Services
Compute
Virtual Machines
Containers
Storage
Blob
Queues
Files
Disks
Networking
Virtual Network
Load Balancer
DNS
Express Route
Traffic Manager
VPN Gateway
App Gateway
Datacenter Infrastructure (48 Regions)

Challenges with on-premises

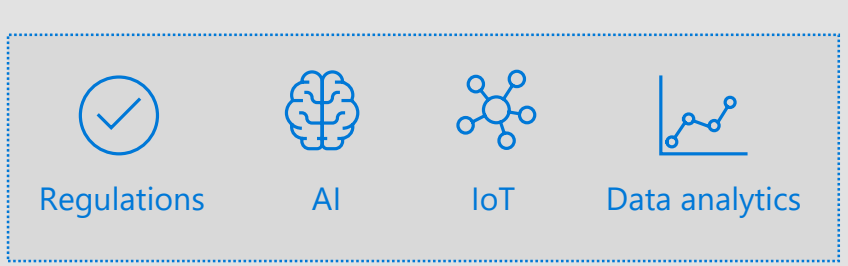
Fixed capacity

Fixed capability

Siloed environments



On-premises



New business demands

Azure for every Big Compute workload

Existing apps

Clone to cloud

Start using cloud without rewriting applications

Hybrid workflows

Simple to optimize infrastructure

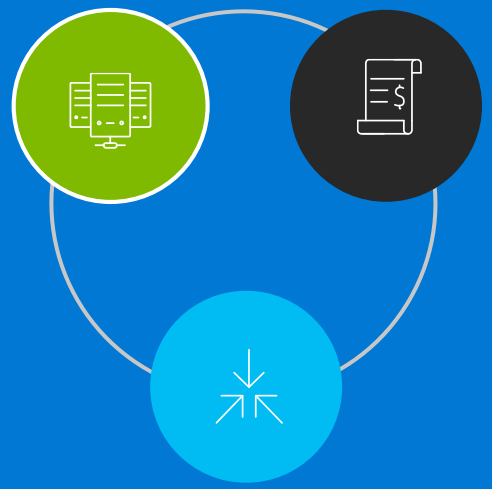
Cloud workflows

End-to-end workflows in the cloud

Cloud-native apps

Create new services and modernize apps that matter

Cray in Azure



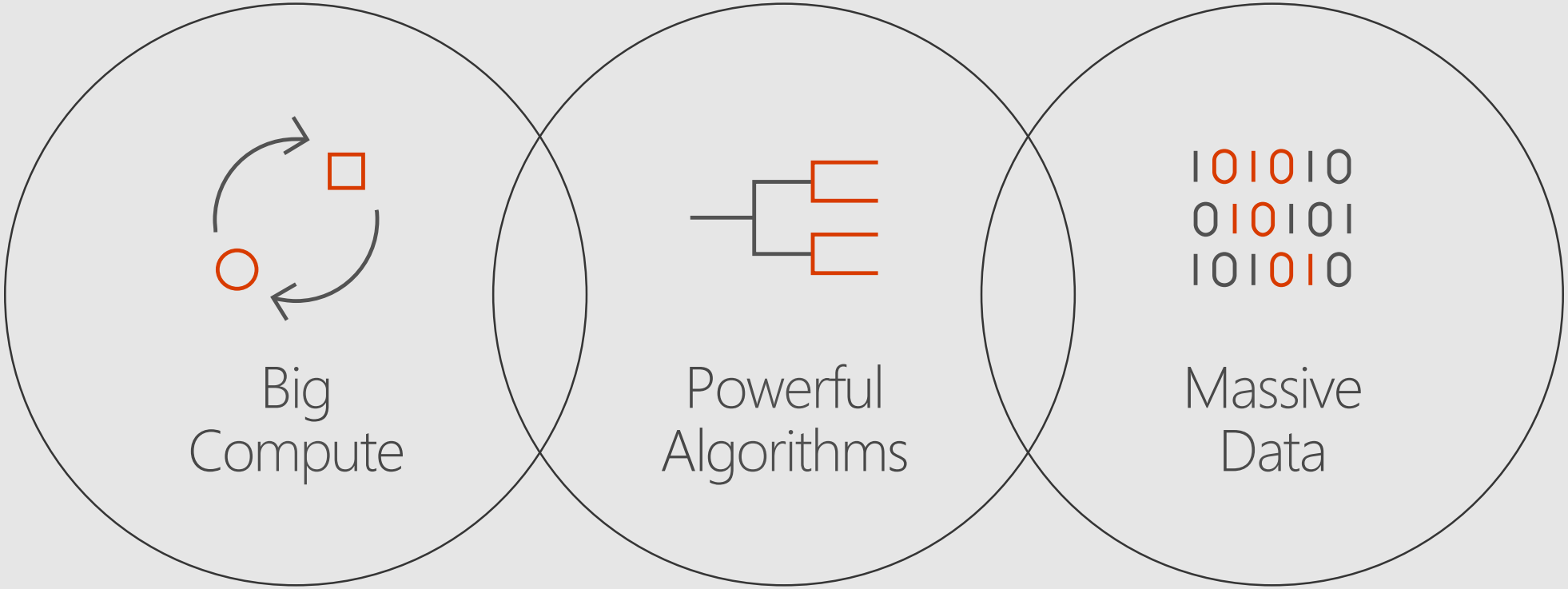
Your Cray supercomputer running in Azure, close to your Azure services.

Rely on a dedicated, built-to-spec Cray XC or SC supercomputer for your most demanding workloads.

Connect to the broad range of Azure services on your Azure Virtual Network.

Access the Cray as a managed service in the cloud as OpEx, instead of maintaining specialized infrastructure with high up-front costs.

Recipe for AI Innovation



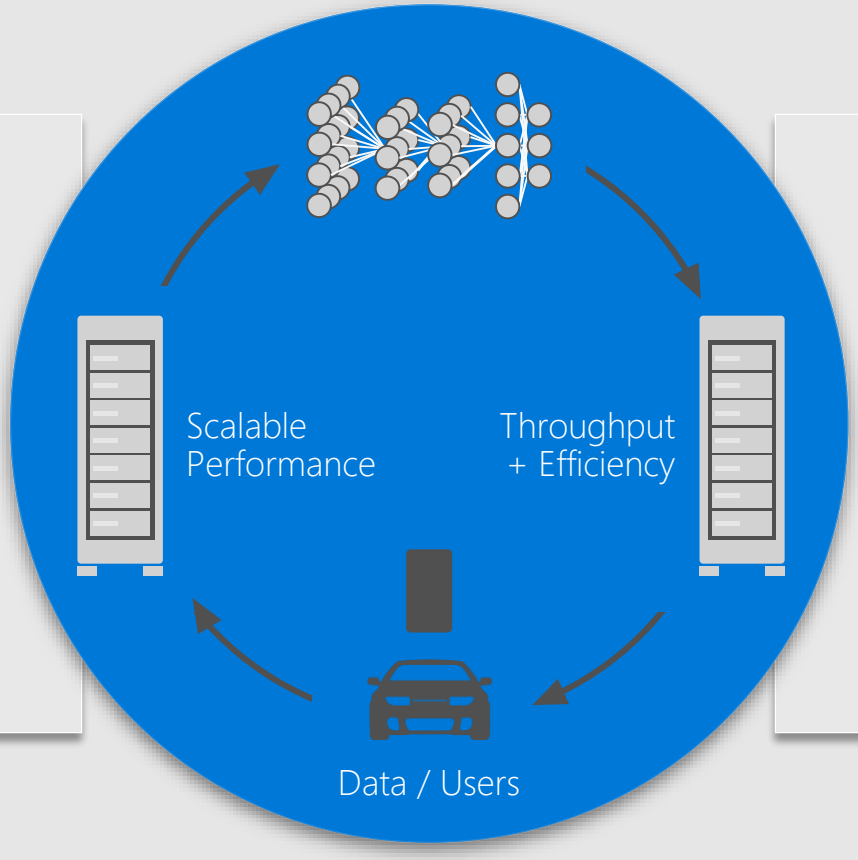
Deep Learning Demands New Class of HPC

TRAINING & TESTING

Billions of TFLOPS per training run

Years of compute-days on Xeon CPU

GPU turns years to days



INFERNECE & SCORING

Billions of FLOPS per inference

Seconds for response on Xeon CPU

GPU for instant response

Job Scheduler Agnostic

- Commercial: Altair PBS, Univa Grid Engine, IBM LSF, Adaptive Computing Moab
- Open Source: SLURM, SGE, Torque
- Financial Services: Tibco Data Synapse, IBM Platform Symphony
- Microsoft Hybrid: Microsoft HPC Pack (Windows and also now Linux)

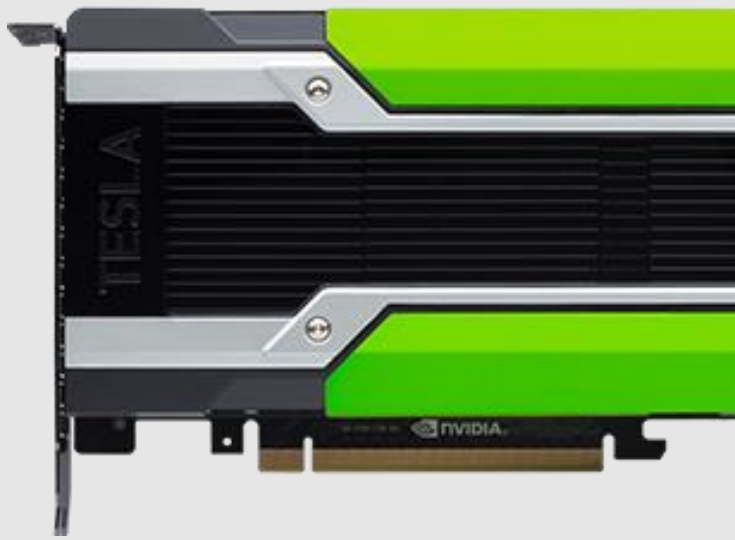
- Deploy in cloud or hybrid via template (see our samples)
- Connect network with VPN or Express Route
- Scripts to scale up and down
- Still need to stage data (think about Azure Data Factory)

- Partner tools: Cycle Computing, Bright Computing, UberCloud, Elastacluster, Altair, Rescale

Visualization Virtual Machines (NV)

Powered by NVIDIA Quadro

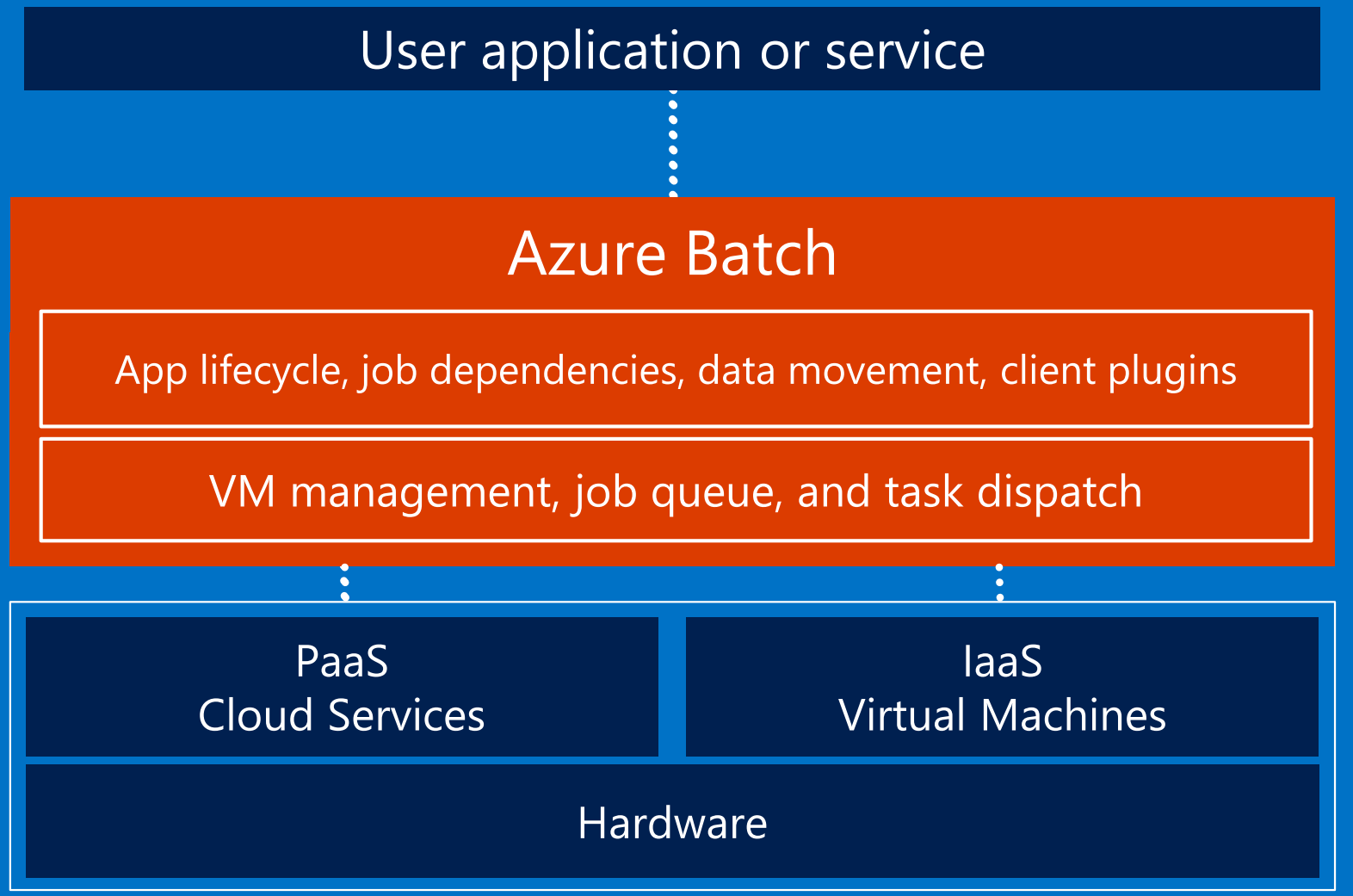
	NV6	NV12	NV24
Cores	6	12	24
GPU	1 M60 GPU (1/2 Physical Card)	2 M60 GPUs (1 Physical Card)	4 M60 GPUs (2 Physical Cards)
Memory	56 GB	112 GB	224 GB
Disk	~380 GB SSD	~680 GB SSD	~1.5 TB SSD
Network	Azure Network	Azure Network	Azure Network
GRID/Quadro Licenses	1	2	4



Compute Virtual Machines (NC)

	NC6	NC12	NC24	NC24r
Cores	6	12	24	24
GPU	1 K80 GPU (1/2 Physical Card)	2 K80 GPUs (1 Physical Card)	4 K80 GPUs (2 Physical Cards)	4 K80 GPUs (2 Physical Cards)
Memory	56 GB	112 GB	224 GB	224 GB
Disk	~380 GB SSD	~680 GB SSD	~1.5 TB SSD	~1.5 TB SSD
Network	Azure Network	Azure Network	Azure Network	InfiniBand

Azure Batch: Job Scheduling as a Service



Cloud-enable applications, made available as SaaS:

- Higher-level set of capabilities
- Minimizes required Azure knowledge
- Foundational batch processing service
- Don't worry about the "plumbing"

New: Low-Priority VM's

Significantly lower priced compute

- Up to 80% discount compared to on-demand price - fixed price
- All Batch VM sizes and regions
- Uses surplus capacity; availability could vary; VMs could be preempted

Suitable workloads

- Distributed parallel jobs - many discrete tasks, interrupt tolerant, shorter task execution times, flexible job completion time
- e.g. Dev, test – regression, scale, load

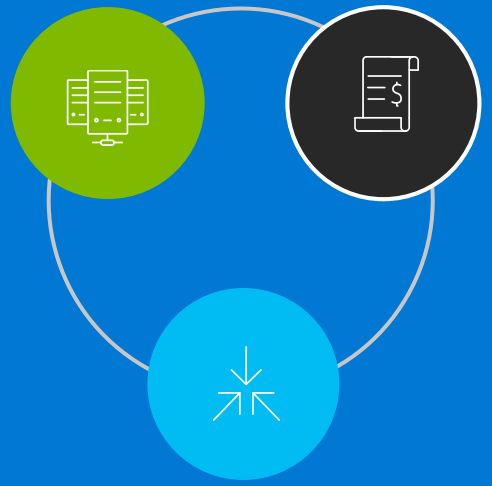
Value

- Get work done for lower cost, faster, or do more for same price

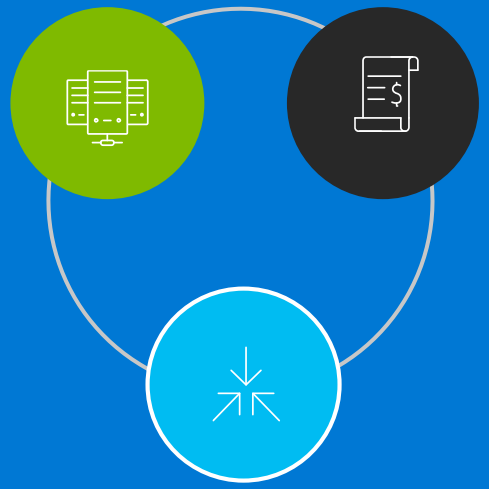
Batch Opportunities in Addition to "HPC"

- Media transcoding
  - Media Services - audio and video transcoding, etc.
  - ISV – video transcoding using ffmpeg
  - XBox video – video pre-processing
- Rendering
  - HoloLens team renders test scenes with Blender
  - ISV rendering plug-in for 3DSMax and V-Ray
- Test execution
  - Azure Engineering CloudValidate service
  - Intune runs 20K tests in 20 mins following check-in's
- Monte Carlo simulations
  - Python, C#, C++, etc.
  - Insurance risk analysis
  - Energy pricing, hotel room pricing
  - Employee pension plan savings analysis
- Deep Learning
  - Training and evaluation
- Genomics:
  - Microsoft Genomics service
- OCR:
  - Use Tesseract to process document images
- Data ingestion and processing / ETL
  - StorSimple ingest and pre-process for HDInsight jobs
  - Nightly data ingestion and process automotive data
  - ADF data copy activities
- R
  - Baseball team - player stats
  - Financial services
- MATLAB:
  - MATLAB clusters
  - Compiled MATLAB scaleout
- Simulations
  - Engineering ISV customers can run apps at scale
  - Aircraft route optimization
- Image processing
  - Analyze MRI scans for signs of disease
  - Analyze blood samples

The largest compliance portfolio in the industry



Support for Microsoft and open source software



redhat

CentOS

Windows Server

FreeBSD

ubuntu

SUSE

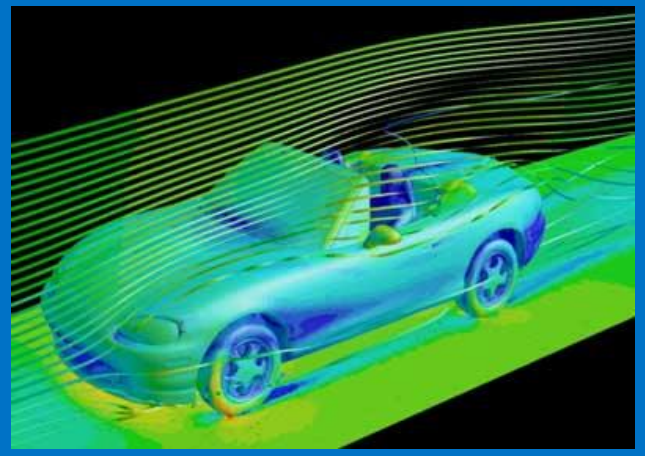
R

python

Java

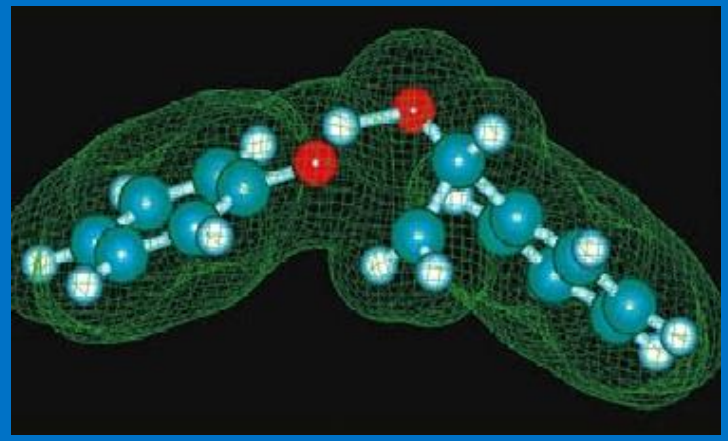
.NET

Example Big Compute Workloads

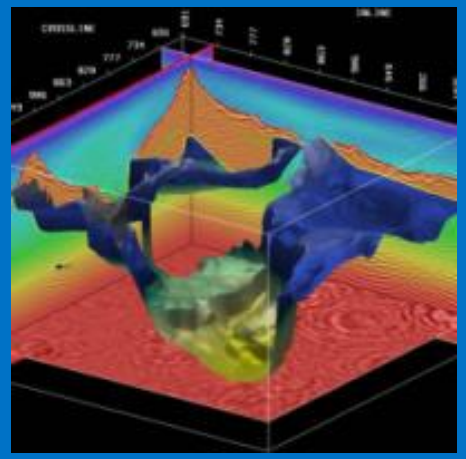


Engineering model and simulation
Car crash simulation

Video rendering



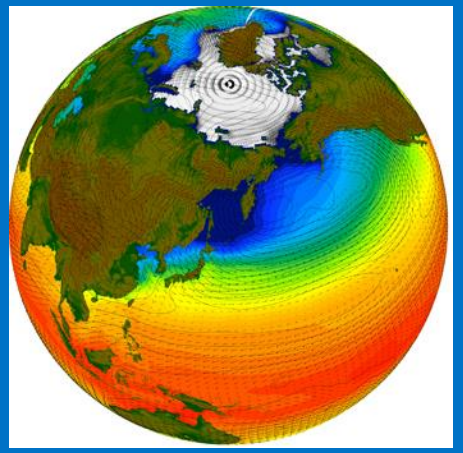
Genomics / Pharma



Oil & gas, seismic, reservoir simulation

Probability		Impact: Very Low 1	Low 2	Medium 4	High 8	Very High 16
Very High	5	5	10	20	40	80
High	4	4	8	16	32	64
Medium	3	3	6	12	24	48
Low	2	2	4	8	16	32
Very Low	1	1	2	4	8	16

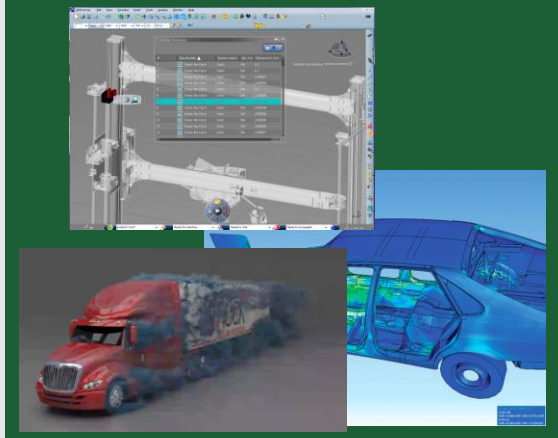
Financial risk analysis



Earth science, climate hydrology

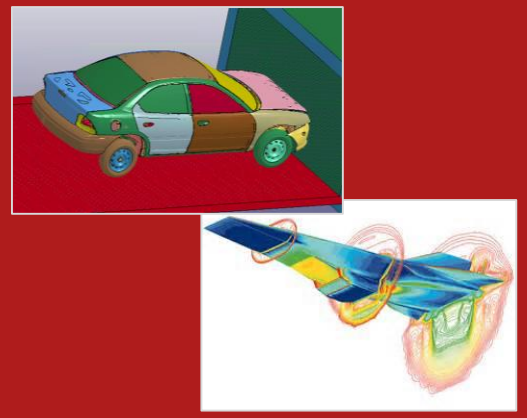
Broad Range of GPU Scenarios

Visualization

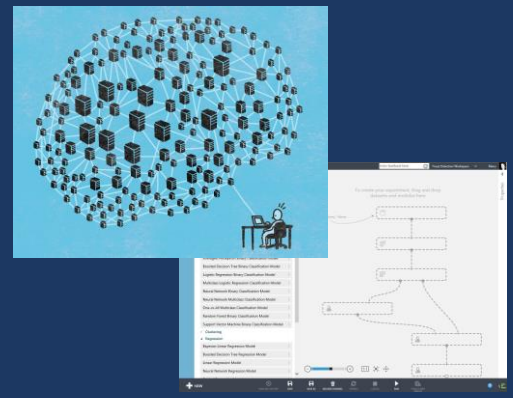


Rendering

HPC/Simulation



Deep-Learning/AI



Media, Entertainment & Gaming

Healthcare & Research

Financial Services

Oil & Gas and Sciences

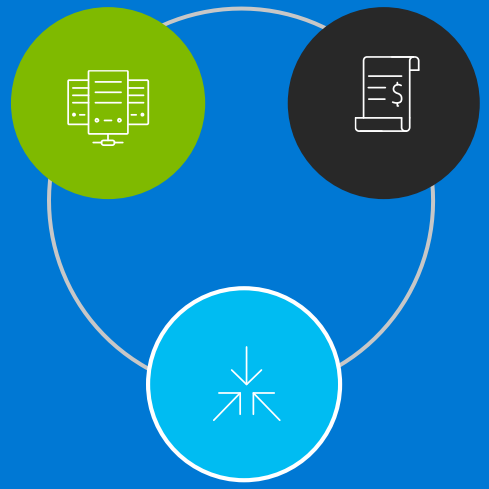
Manufacturing

Automotive

Aerospace

Retail

Robust partner ecosystem



Bright Computing
BizData
teradici
Workspot
rescale
Willis Towers Watson
kinetica SOLUTIONS
Milliman
excelian LUXOFT FINANCIAL SERVICES
Uber Cloud
IHS Markit
ACTIVEVIAM Analysing data. Empowering the future.
Altair
d3View
CD-adapco www.cd-adapco.com
ANSYS
PBS Works
MSC Software
DASSAULT SYSTEMES
LSTC Livermore Software Technology Corp.
HALLIBURTON Landmark
Schlumberger
SIMULIA
JSOL Corporation
esi
SIEMENS
CRAY
intel
Mellanox TECHNOLOGIES
NVIDIA

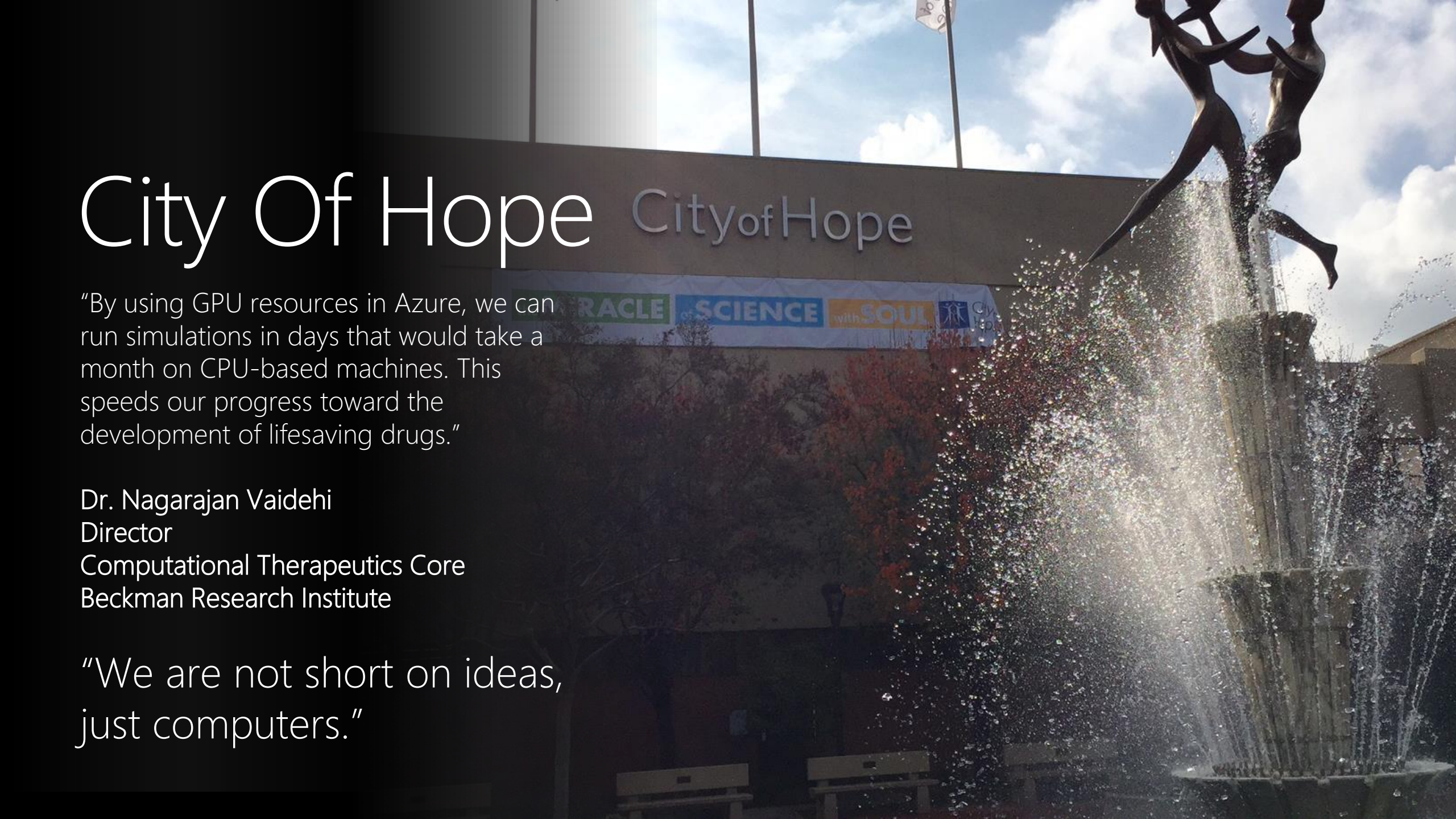
NVIDIA Graphics: Quadro and Grid

NVIDIA Quadro Virtual Workstation Driver

Azure NV/NVIDIA Tesla M60 GPUs

Azure Virtual Machines

City Of Hope

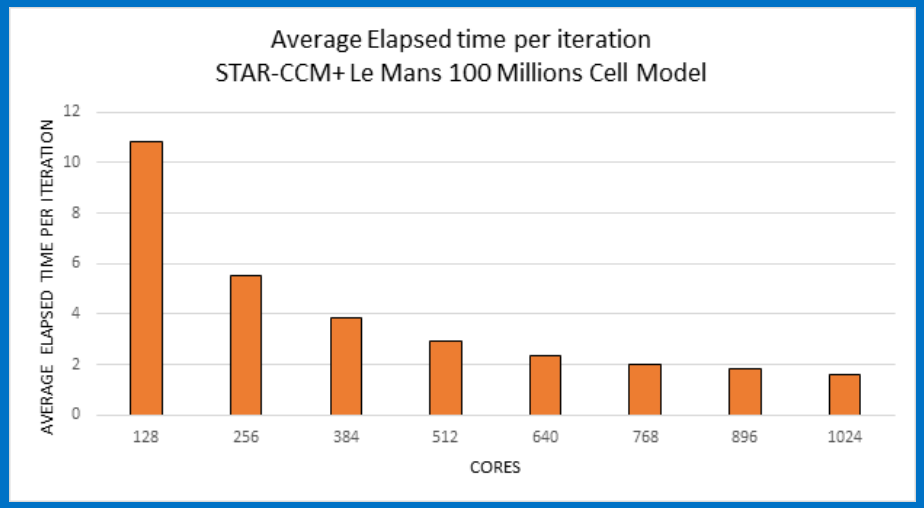


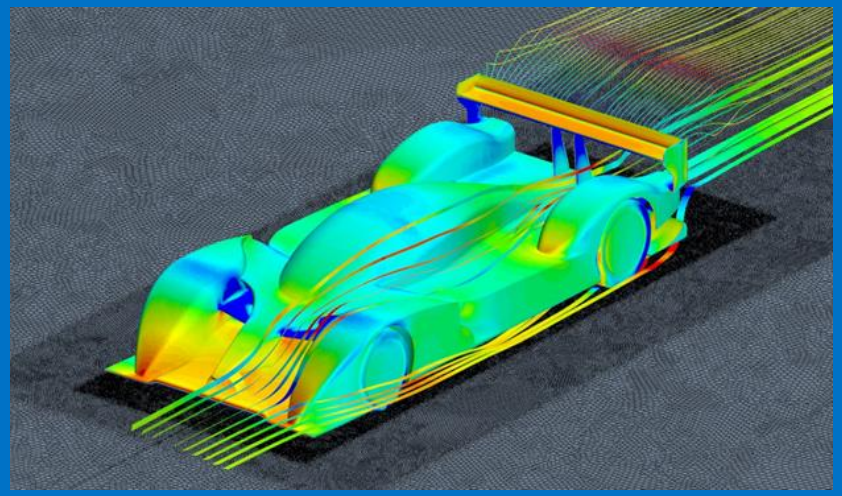
"By using GPU resources in Azure, we can run simulations in days that would take a month on CPU-based machines. This speeds our progress toward the development of lifesaving drugs."

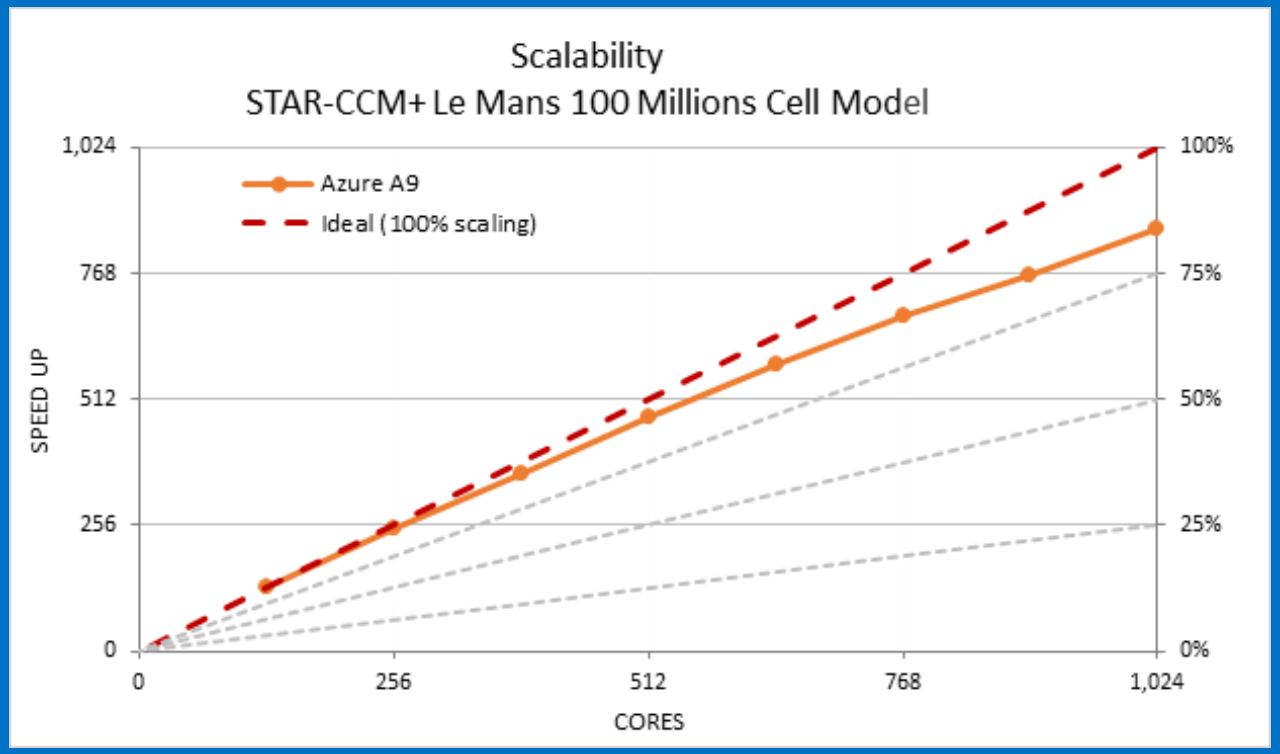
Dr. Nagarajan Vaidehi
Director
Computational Therapeutics Core
Beckman Research Institute

"We are not short on ideas, just computers."

STAR-CCM+ Scalability to 1024 cores



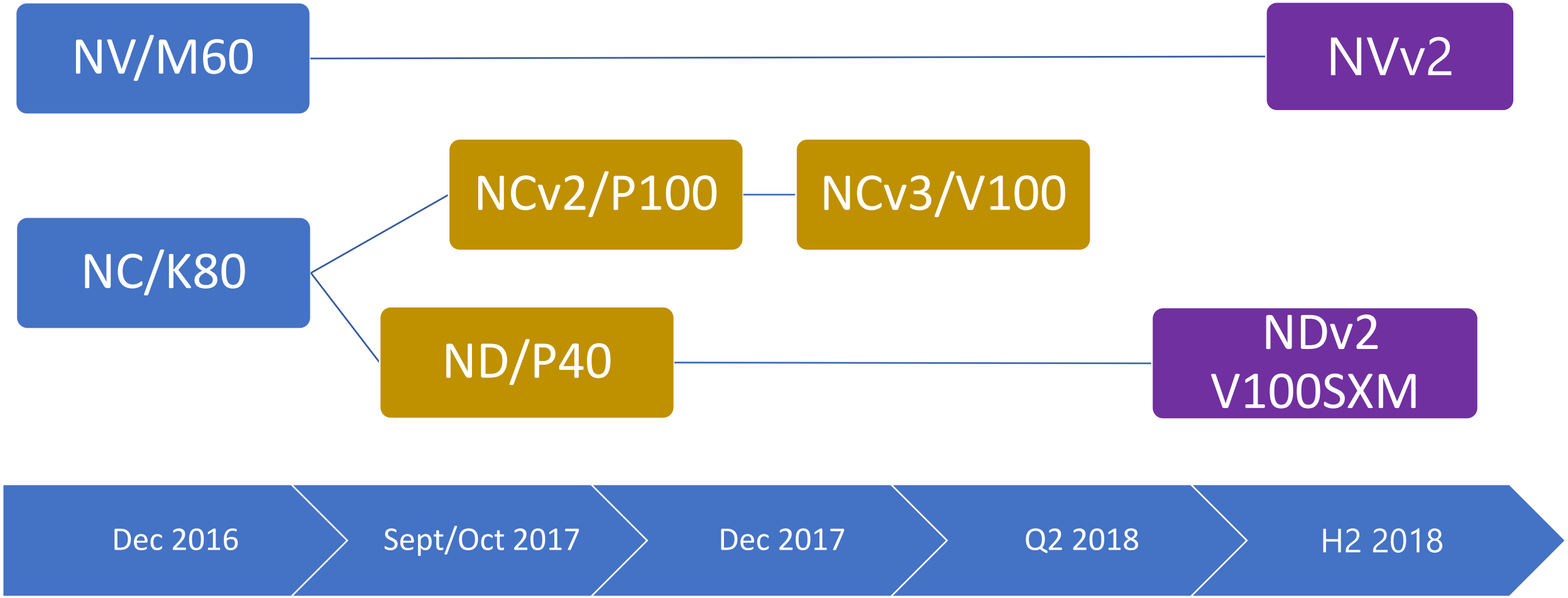




https://azure.microsoft.com/en-us/blog/availability-of-star-ccm-on-microsoft-azure/


Microsoft Azure Roadmap
Microsoft

GPU SKU Roadmap Across Generation



NV_v2 – Updated GPU Visualization Platform

- Visualization optimized GPU instances featuring NVIDIA Tesla M60 GPUs
- Broadwell based CPU processor with doubled memory from previous generation (up to 448 GB)
- Premium storage support (SSD backed)
- Get faster results for the your graphic intensive 2D and 3D applications
- Grid license included with each GPU instance
- Specs:
  - 2048 NVIDIA CUDA cores per GPU
  - 36 H.264 1080p30 streams
  - GPU Memory 8 GB/GPU

NVIDIA Quadro Virtual Workstation Driver

Azure NV/NVIDIA Tesla M60 GPUs

Azure Virtual Machines

	NV6s_v2	NV12s_v2	NV24s_v2
Cores	6	12	24
GPU	1 x M60	2 x M60	4 x M60
Memory	112 GB	224 GB	448 GB
Local Disk	~700 GB SSD	~1.4 TB SSD	~3 TB SSD
Network	Azure Network	Azure Network	Azure Network
GRID Licenses	1	2	4

Compute Virtual Machines : NC_v2

	NC6s_v2	NC12s_v2	NC24s_v2	NC24rs_v2
Cores	6	12	24	24
GPU	1 x P100 GPU	2 x P100 GPU	4 x P100 GPU	4 x P100 GPU
Memory	112 GB	224 GB	448 GB	448 GB
Disk	~700 GB SSD	~1.4 TB SSD	~3 TB SSD	~3 TB SSD
Network	Azure Network	Azure Network	Azure Network	InfiniBand

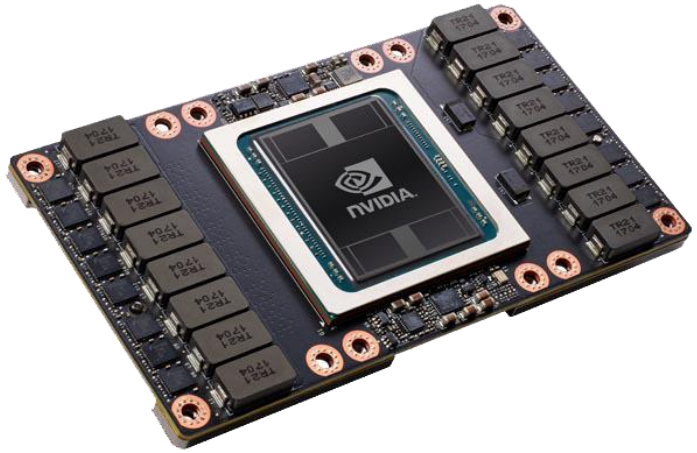
Next-Gen GPU Compute VM: NC_v3

	NC6s_v3	NC12s_v3	NC24s_v3	NC24rs_v3
Cores	6	12	24	24
GPU	1 x V100 GPU	2 x V100 GPU	4 x V100 GPU	4 x V100 GPU
Memory	112 GB	224 GB	448 GB	448 GB
Disk	~700 GB SSD	~1.4 TB SSD	~3 TB SSD	~3 TB SSD
Network	Azure Network	Azure Network	Azure Network	InfiniBand

Coming Soon
ND_v2 – Volta Generation GPU Compute

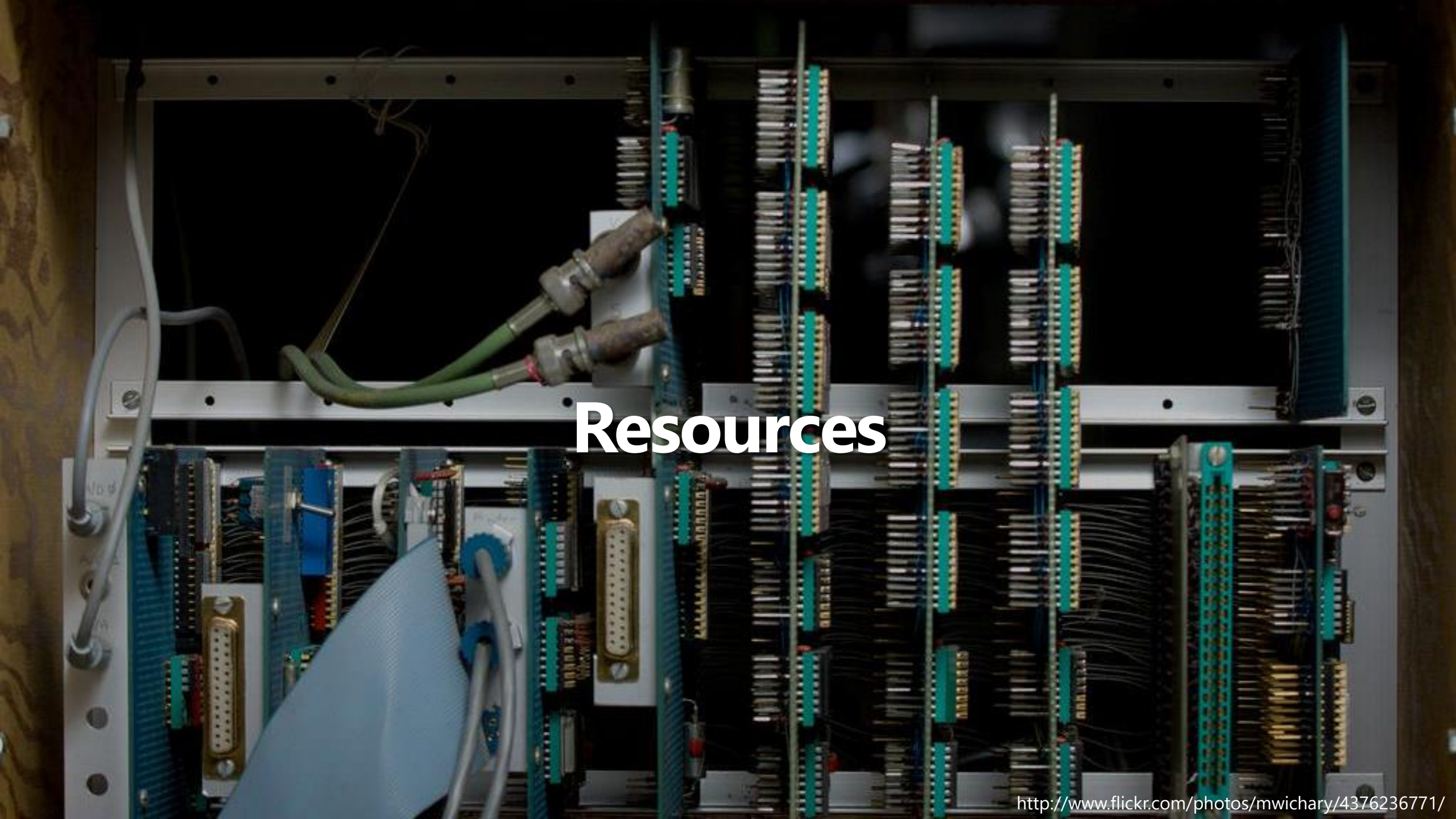
- Volta SXM GPU instances - NVIDIA V100 GPUs
- 8X NVIDIA V100 GPUs interconnected with NVLink mesh
- Tensor Core technology to deliver over 100 Teraflops per second (TFLOPS) of deep learning performance
- Excellent for accelerating machine training jobs and HPC
- Skylake based processor with premium storage support (SSD backed)
- Specs:
  - 640 NVIDIA Tensor Core
  - FP64 - 7.8 TFLOPS of double precision floating point performance
  - FP32 – 15.7 TFLOPS of single precision performance
  - GPU Memory 16 GB
  - 300 GB/s GPU interconnect through NVLink

	ND40s_v3
Cores	40 cores
GPU	8 x V100 SXM
Memory	768 GB
Local Disk	~1.3 TB SSD
Network	Azure Network + NVLink GPU interconnect



GPU Deep Learning VM: ND

	ND6s	ND12s	ND24s	ND24rs
Cores	6	12	24	24
GPU	1 x P40	2 x P40	4 x P40	4 x P40
Memory	112 GB	224 GB	448 GB	448 GB
Disk	~700 GB SSD	~1.4 TB SSD	~3 TB SSD	~3 TB SSD
Network	Azure Network	Azure Network	Azure Network	InfiniBand



Resources

http://www.flickr.com/photos/mwichary/4376236771/

Big Compute Industry Team

Scaling big compute through partners, evangelism, training and industry insight.

Large Deal Engagement

Direct customer engagement with industry experts and aligned partners. End to end major deal management

Team driving in top 20 deals

Field Evangelism & Readiness

More than 10 formalized field readiness programs. Correct lack of awareness by PDU and Generalists. Industry/SKU deep dives.

Partner/ISV Recruiting and Pairing

HPC partner recruiting and management by industry. Technical enablement and seller exposure

10 Partners Recruited 40 WW Partners targeted

Industry Hunting Scenarios/Engineering Planning

Sellers to drive pipeline through industry specific sales scenarios. Provide engineering guidance on product efficacy and needs.

GBB Partnership

Tight GBB coordination to engineering and industry teams. Proactive partner pairing with GBB team

Strategic Account Alignment

Strategic account planning by industry with Global Account team to proactively drive scenarios. Penetrate top 130 MS accounts

Next steps

Learn more about Azure for Big Compute.
https://azure.microsoft.com/solutions/high-performance-computing/

Explore Azure solutions for your industry.
https://azure.microsoft.com/solutions/

Engage a partner to help implement your high-performance solution.
https://azure.microsoft.com/partners/directory/?solution=high-performance-computing