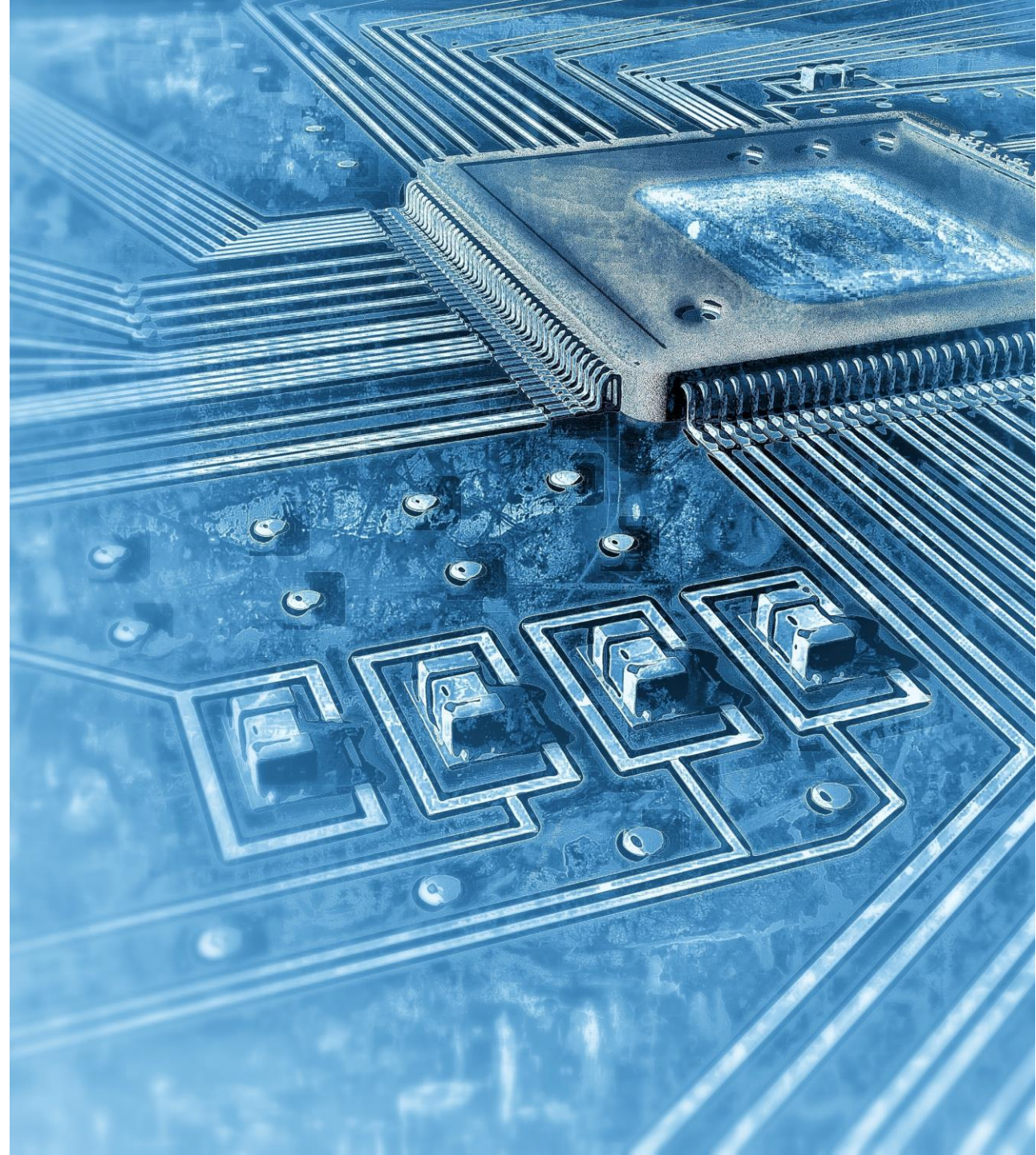




SC22 Student Cluster Competition Azure Webinar

October 11, 2022



Welcome and Introduction

- This will be a short(ish) presentation, followed by a longer Q&A session – please put questions in chat
- We're recording this session, the recording and these slides will be posted on the webinars page
- There will be follow-up conversations with more details about Azure and the cloud component of the competition as they become available

Welcome and Introduction



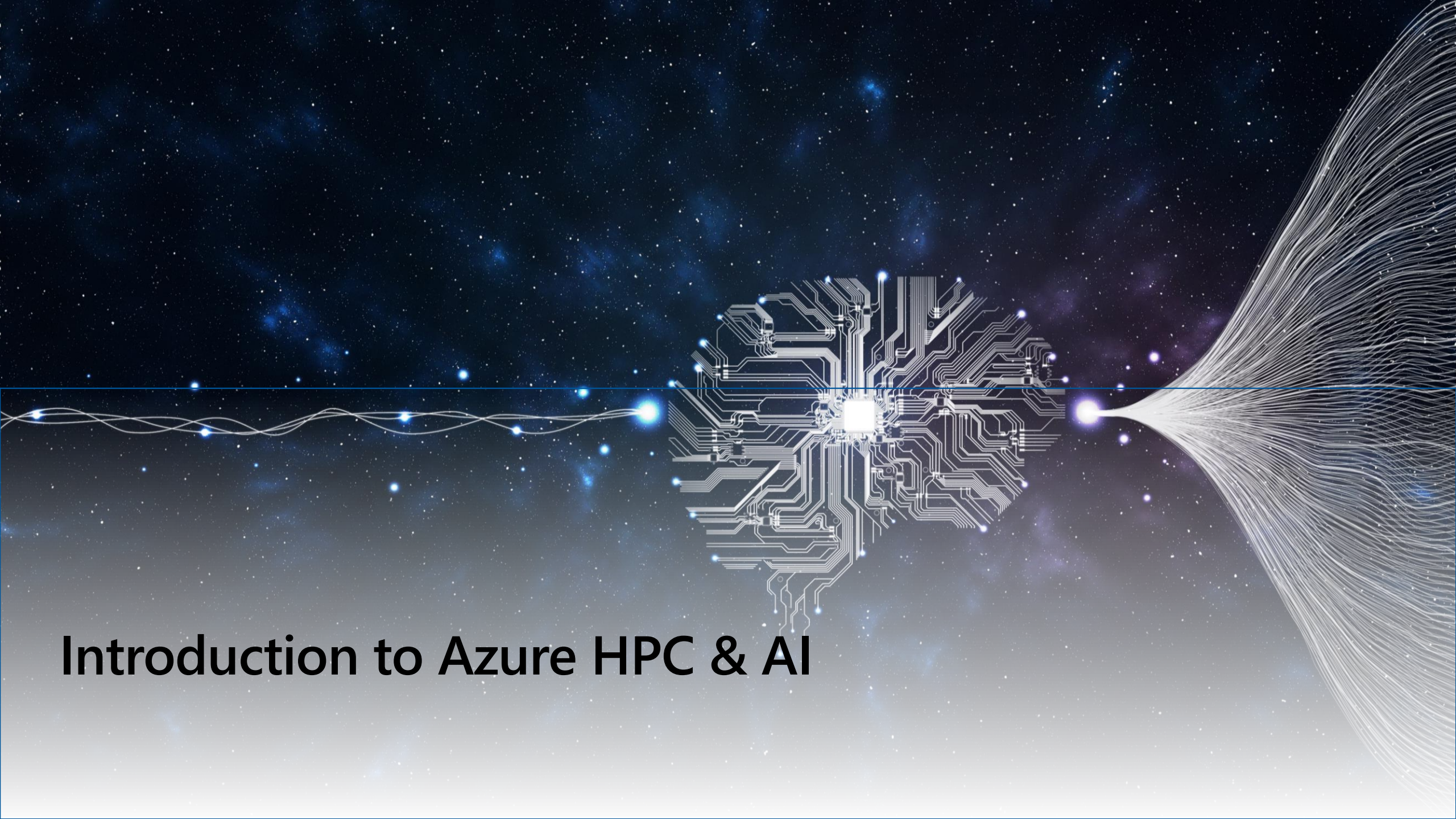
Andy Howard
Azure HPC/AI Software &
Services - CycleCloud



Laura Parry
Sr. Specialist GBB - EMEA

Agenda

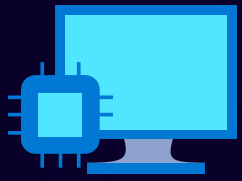
- Overview of HPC on Azure
- Testing vs Competition budgets
- Access and Quotas
- Q&A



Introduction to Azure HPC & AI

~~HPC in the Cloud~~

A cloud built for HPC



Purpose-built HPC

A full range of CPU and GPU capabilities that help applications scale to 80K+ cores



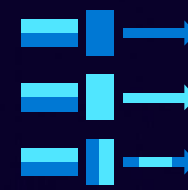
Fast, Secure Networking

Fast InfiniBand interconnects as well as edge-to-cloud connectivity



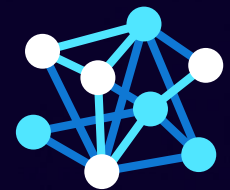
High Performing Storage

A range of storage capabilities to support simple-to-complex storage needs



Workload Orchestration

End-to-end workflow agility using known, familiar tools & processes



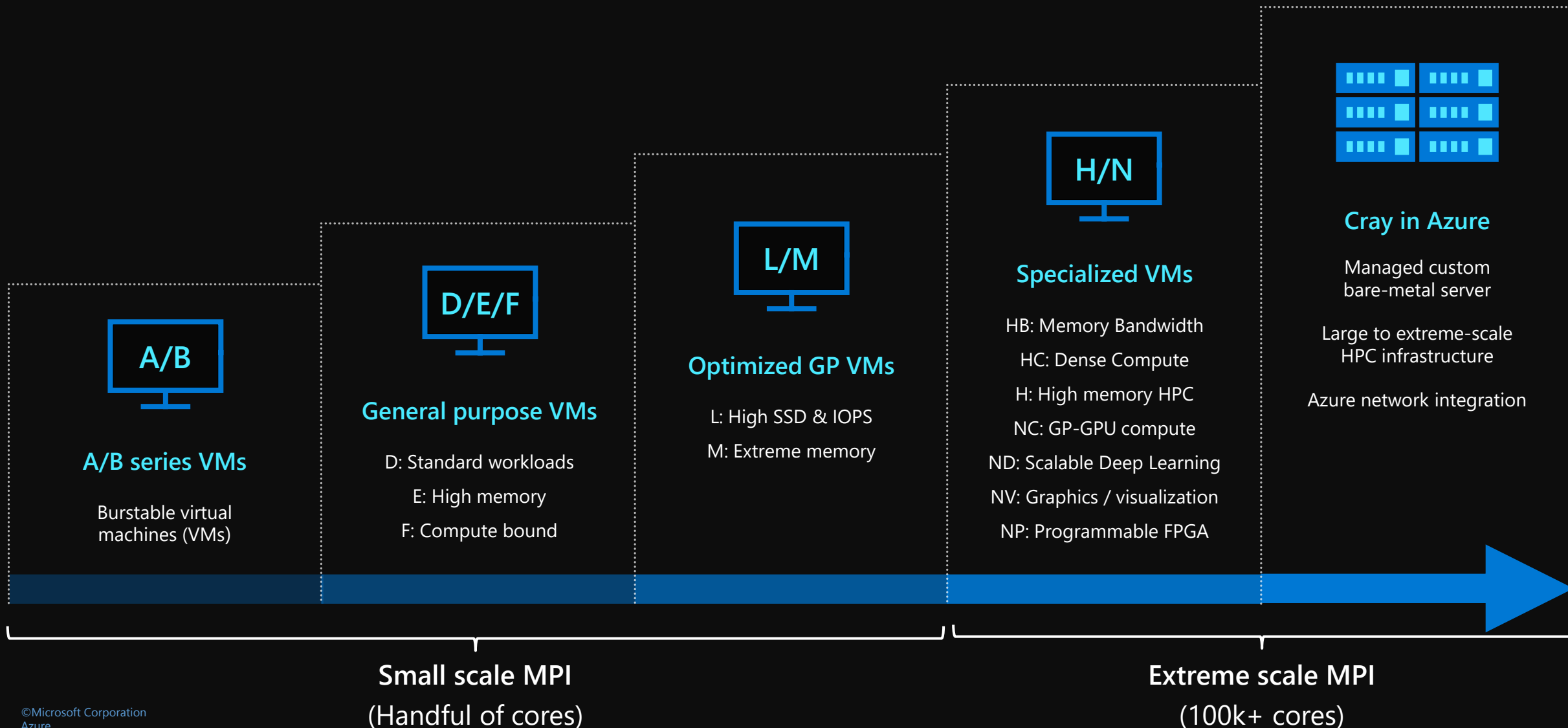
Intelligence Services

AI, machine learning, and deep learning at supercomputer scale

Azure HPC & AI breakthroughs

- 2019 • 20,000 core MPI job
1st for cloud
- 2019 • AMD EPYC Rome HPC IB clusters
1st for cloud
- 2019 • 200 Gb/s HDR with adaptive routing
1st for cloud
- 2020 • 80,000 core MPI job
1st for cloud, 12x better than any other cloud
- 2020 • Top5-scale supercomputer for OpenAI
1st for cloud
- 2020 • 1 TB/sec Parallel File System
1st for cloud
- 2021 • 1.6 Tb/s InfiniBand for NVIDIA A100
1st for cloud
- 2021 • HBv3 Milan GA in Azure *at* AMD Launch
1st for cloud *or* on-prem
- 2021 • 10-year Supercomputing-aaS for Met Office
1st for cloud
- 2021 • 10th Fastest Supercomputer is an Azure Cluster
Azure is the only public cloud with 4 entries in Top500 list

Solve any HPC, AI workload — at any scale



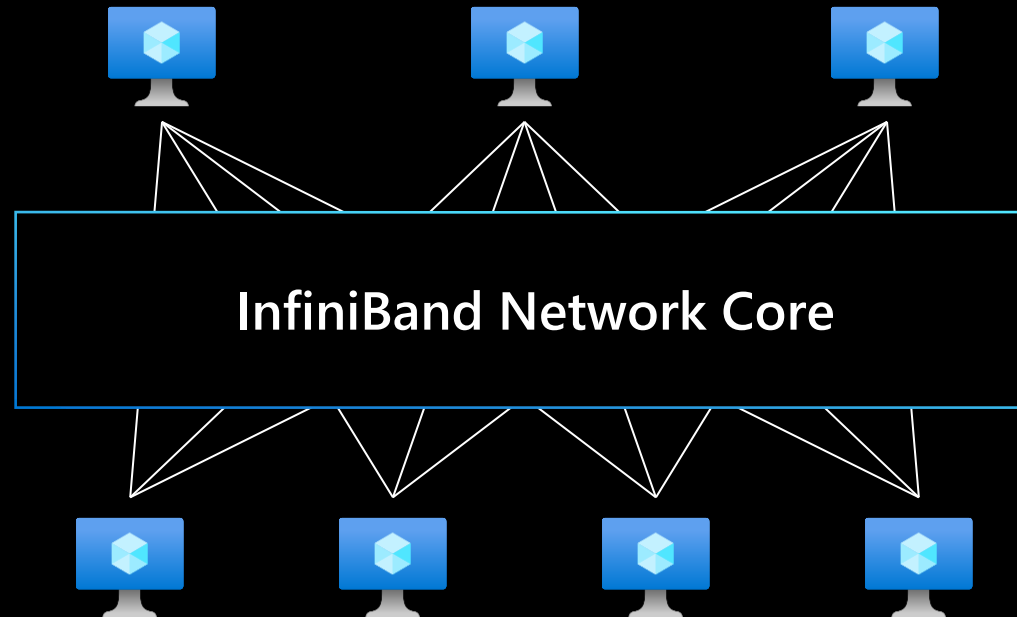
Non-blocking Fat Tree topology

Hardware offload of MPI collectives

Full MPI & NCCL Integration

< 1.5 microsecond latencies

Up to 1.6 Tb/s per VM



Bare-metal passthrough

Dynamic Connected Transport

Intelligent Adaptive Routing



CPU VMs with InfiniBand

HB – Scalable AMD HPC

HC – Scalable Intel HPC





Scalable AMD HPC

AMD EPYC 2nd and 3rd Gen Processors

4 TeraFLOPS FP64 / 8 TeraFLOPS FP32

350 GB/S memory bandwidth

200 GB HDR InfiniBand

MPI Scaling to > 80,000 Cores

0.9 – 1.8 TB NVMe SSD + Azure Premium Storage



Scalable Intel HPC

Intel Xeon Platinum 1st Gen Processors

2.7 TeraFLOPS FP64 / 5.4 TeraFLOPS FP32

190 GB/S memory bandwidth

100 GB EDR InfiniBand

MPI Scaling to > 20,000 Cores

700 GB SSD + Azure Premium Storage

HBv3 – The New Cloud HPC Flagship

Highest performance, most cost-effective CPU for HPC

Performance leadership both **per VM** or **per core**

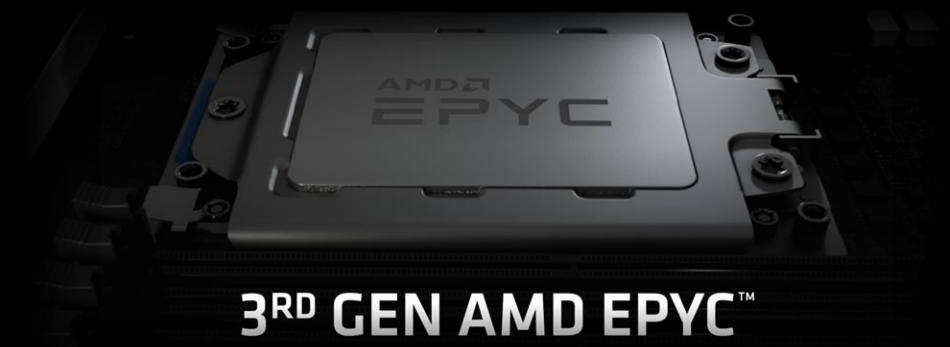
Range of sizes to fit greater range of customer needs

+19% IPC from Zen3 core v. Zen2 core, Up to 32 MB L3/core

Simpler NUMA topology (4 NUMA domains per VM)

Large SSD gains* - 2x size, 4.7x IOPS, 3.6x bandwidth

200 Gb HDR InfiniBand, MPI jobs up to 80,000 cores



- **Available now** in East US, South Central US, and West Europe
- **Q4 2021** expansion to West US 3 and Singapore

Azure HBv3 VM Sizes

VM Size	120 CPU cores	96 CPU cores	64 CPU cores	32 CPU cores	16 CPU cores
VM Name	standard_HB120rs_v3	standard_HB120-96rs_v3	standard_HB120-64rs_v3	standard_HB120-32rs_v3	standard_HB120-16rs_v3
Similar to...	EPYC 7713	EPYC 7643	EPYC 7543	EPYC 7313	EPYC 72F3
InfiniBand	200 Gb	200 Gb	200 Gb	200 Gb	200 Gb
Peak CPU Frequency*	3.675 GHz	3.675 GHz	3.675 GHz	3.675 GHz	3.675 GHz
RAM per VM	448 GB				
RAM per core	3.75 GB	4.67 GB	7 GB	14 GB	28 GB
Memory B/W per VM	350 GB/s				
Memory B/W per core	2.91 GB/s	3.65 GB/s	5.46 GB/s	10.9 GB/s	21.9 GB/s
L3 Cache per VM	480 MB				
L3 Cache per core	4 MB	5 MB	7.5 MB	15 MB	30 MB
SSD Perf per VM	2 * 960 GB NVMe – 6.9 GB/s (Read) / 2.9 GB/s (Write), 200k IOPS (Read) / 190k IOPS (Write)				



*Clock frequencies are based on non-AVX workload scenarios and are based on measured frequency delivery for workloads as captured by the Azure HPC team with AMD EPYC 7003-series processors and corresponding system firmware as of January 2021. Experienced clock frequency by a customer is a function of a variety of factors, including the coding and usage of a given application. Frequencies indicated above are not necessarily indicative of final clock frequencies for EPYC 7003-series processors.

High-Performance Computing VMs (H)

	Available Now	Available Now	Available Now
	HBv2	HB	HC
Workload Optimized	Memory Bandwidth	Memory Bandwidth	Dense Compute
CPU	AMD EPYC 2 nd Gen "Rome"	AMD EPYC 1 st Gen "Naples"	Intel Xeon Platinum 1 st Gen "Skylake"
Cores/VM	120	60	44
TeraFLOPS/VM (FP64)	4 TF	0.9 TF	2.6 TF
Memory Bandwidth	353 GB/s	263 GB/sec	191 GB/sec
Memory	4 GB/core, 480 total	4 GB/core, 240 total	8 GB/core, 352 GB
Local Disk	900 GB NVMe	700 GB NVMe	
InfiniBand	200 Gb HDR	100 Gb EDR	
Network	32 GbE	32 GbE	

GPU Computing

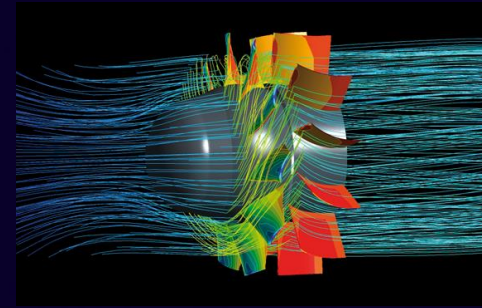
Visualization



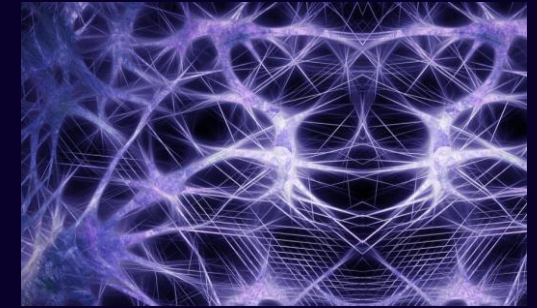
Rendering



HPC/Simulation



Deep-Learning/AI



Workload centric PCs & Workstations

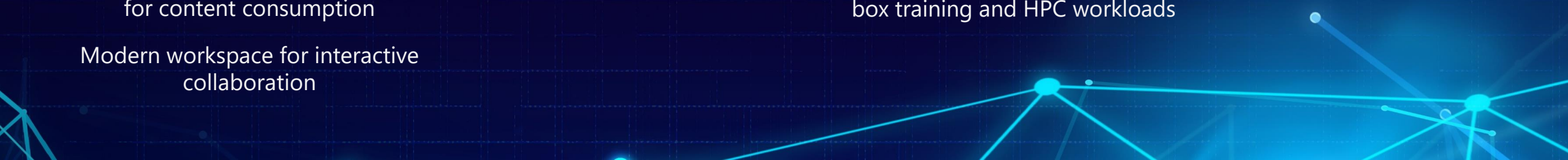
Variable configuration for GPU workstations
for content consumption

Modern workspace for interactive
collaboration

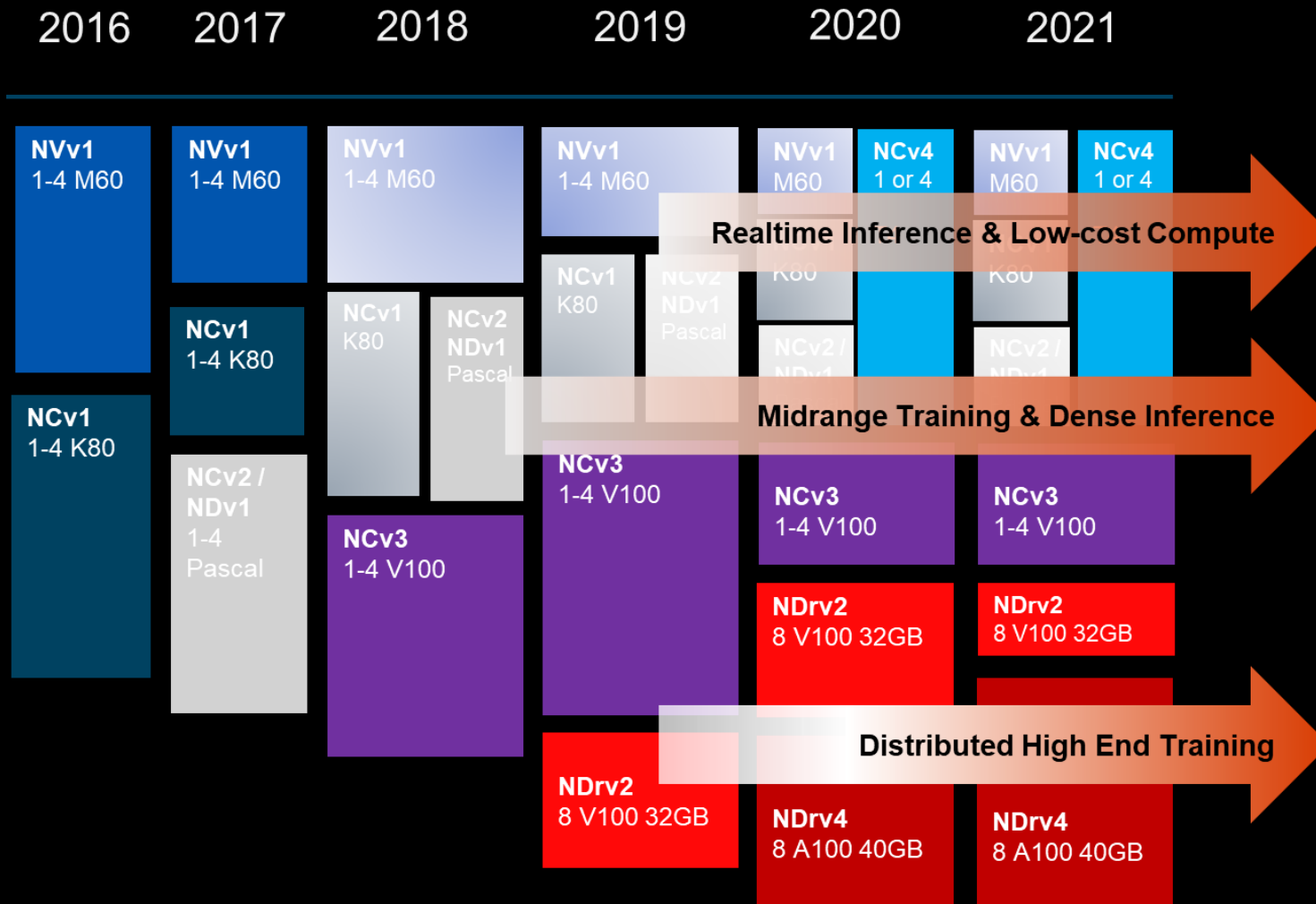


Scale out using IB for multimode HPC and ML workloads on any MPI stack

Scale-up multi-GPU VMs with fast NVLINK interconnect for high-density single
box training and HPC workloads



AI Workloads Are Continuously Evolving



Workload optimized & evolving

- 3 mature and distinct AI workloads

Showcase the best of AI Hardware on Azure

- Purpose built with latest hardware and hypervisor features
- Unique capacity strategies for each segment

One Azure HW platform, multiple-consumption models

- Azure IaaS VMs & AML developer platform
- AI developer services (e.g., Cognitive & scenario AI services) and MSFT AI research
- Solutions (e.g., Office 365/Power BI/Bing)

GPUs for Compute (NC) & Visualization (NV)

compute

Azure Instance →	NC	NCv2	NCv3	NCasT4_v3	NC A100 v4
Cores	6, 12, 24	6, 12, 24	6, 12, 24	4, 8, 16, 64	24, 48, 96
GPU	Tesla K80	Tesla P100	Tesla V100	Tesla T4	A100 Tensor Core
Memory	56/112/224 GB	112/224/448 GB	112/224/448 GB	28/56/110/440 GB	220/440/880 GB
Local Disk	340/680/1440 GB SSD	736/1474/2948 GB SSD	736/1474/2948 GB SSD	180/360/2880 GB SSD	1123/2246/4492 GB
Network	Azure Network + InfiniBand (largest size only)				

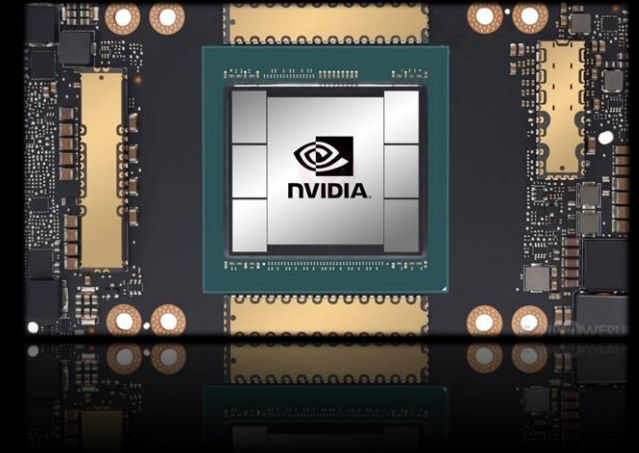
visualize

Azure Instance →	NV	NVv3	NVv4	NVads A10 v5
Cores	6, 12, 24	12, 24, 48	4, 8, 16, 32	6, 12, 18, 36, 72
GPU	Tesla M60	Tesla M60	Radeon Instinct MI25	A10 Tensor Core
Memory	56,112,224 GB	112/224/448 GB	14/28/56/112 GB	55/110/220/440/880 GB
Local Disk	340/680/1440 TB SSD	320/640/1280 GB SSD	88/176/352/704 GB	180/320/720/1400 GB
Network	Azure Network			

GPUs for Deep Learning (ND series)

Deep learning

Azure instance →	ND	NDv2	ND A100 v4	NDm A100 v4
CPU Cores	6,12,24	40	96	96
GPU	1x, 2x, or 4x P40 GPUs	8x V100 32 GB (NVLink) GPUs	8x A100 40 GB GPUs	8x A100 80 GB GPUs
Memory	12/224/448 GB	672 GB	900 GB	1900 GB
Local Disk	736/1474/2948 GiB SSD	2948 GiB SSD	6 TB SSD	6.4 TB SSD
Network	Azure Network + InfiniBand EDR	Azure Network + InfiniBand EDR + NVLink GPU Interconnect	Azure Network + InfiniBand EDR + NVLink GPU Interconnect	Azure Network + InfiniBand EDR + NVLink GPU Interconnect



Flagship Nvidia offering for tightly-coupled GPU workloads at scale: Model-Parallel Training and HPC

Ampere SXM GPU instances: 8X NVIDIA A100 GPUs interconnected with NVLink + NVSwitch

One 200 Gigabit InfiniBand HDR link per GPU with full NCCL2 support and GPUDirect RDMA

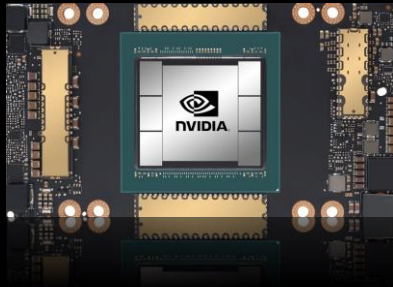
Custom, ground-up platform with PCIe Gen 4-based connectivity for optimal system level performance

AI supercomputer cluster with thousands of tightly-coupled GPUs

Per NDRv4 VM	Configuration
Physical CPU Cores	96 AMD EPYC 2ND GEN Cores
A100 GPUs	40 GB x 8 (with NVLink)
RAM	896 GB
NVMe Local Disk	7 TB
IB Connectivity	8 x 200 Gigabit HDR + GPUDirect RDMA

ND A100 v4: Massively Scalable AI Supercomputer

Single A100 GPU



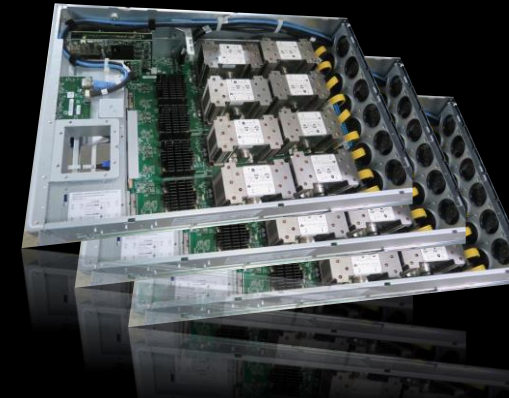
NCCL+NVLink

Multi-GPU with NVLINK
1 NDv4 VM, 8 A100s



NCCL+HDR

Multi-GPU with HDR InfiniBand
Up to hundreds of NDv4 VMs, thousands of A100s



NVIDIA A100 Tensor Core GPU

- 40 GB of HBM2 Memory
- 2x – 20x V100 performance
- PCIe Gen 4, AMD Rome host
- 8 per VM

NVSwitch + NVLink 3.0

- Between the 8 GPUs local GPUs within each VM
- 2.4 Terabits full-duplex, non-blocking

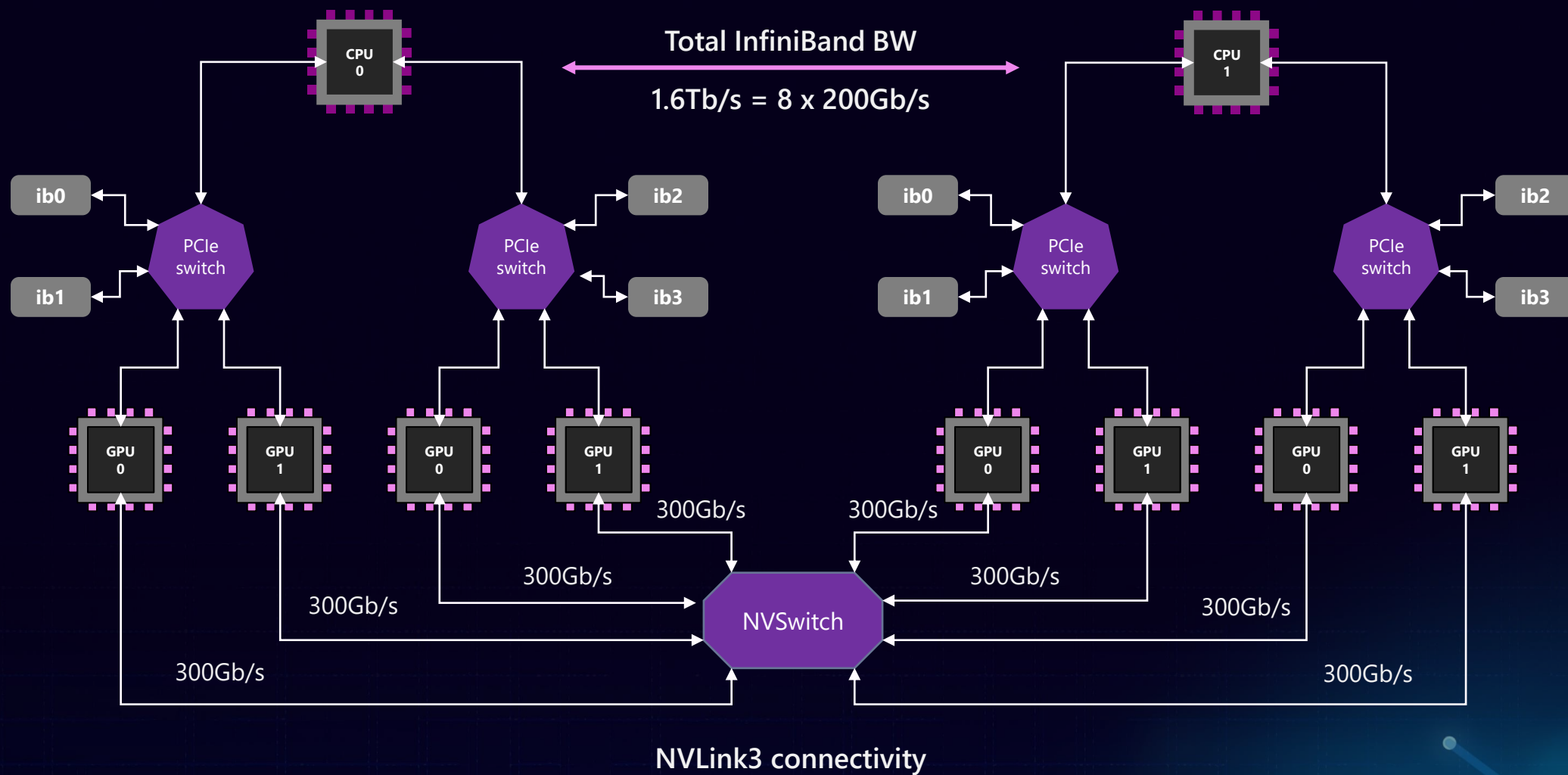
Mellanox InfiniBand HDR Fabric

- 200 Gigabit dedicated link per GPU (1.6 Terabits/VM)
- Topology agnostic fat-tree
- Any to any, all to all, fully subscribed up to thousands of GPUs
- Dynamically provisioned via VMSS
- GPUDirect RDMA

Learn more at: <https://aka.ms/AISCforYou>

Preview sign-up: <http://aka.ms/AzureA100SignUpForm>

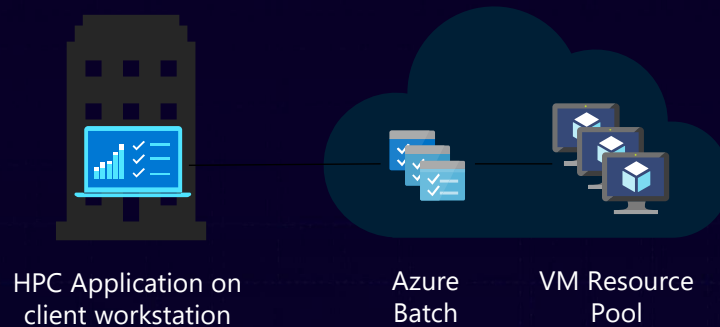
ND_v4 Node Topology



Services for HPC Workload Management

Azure Batch

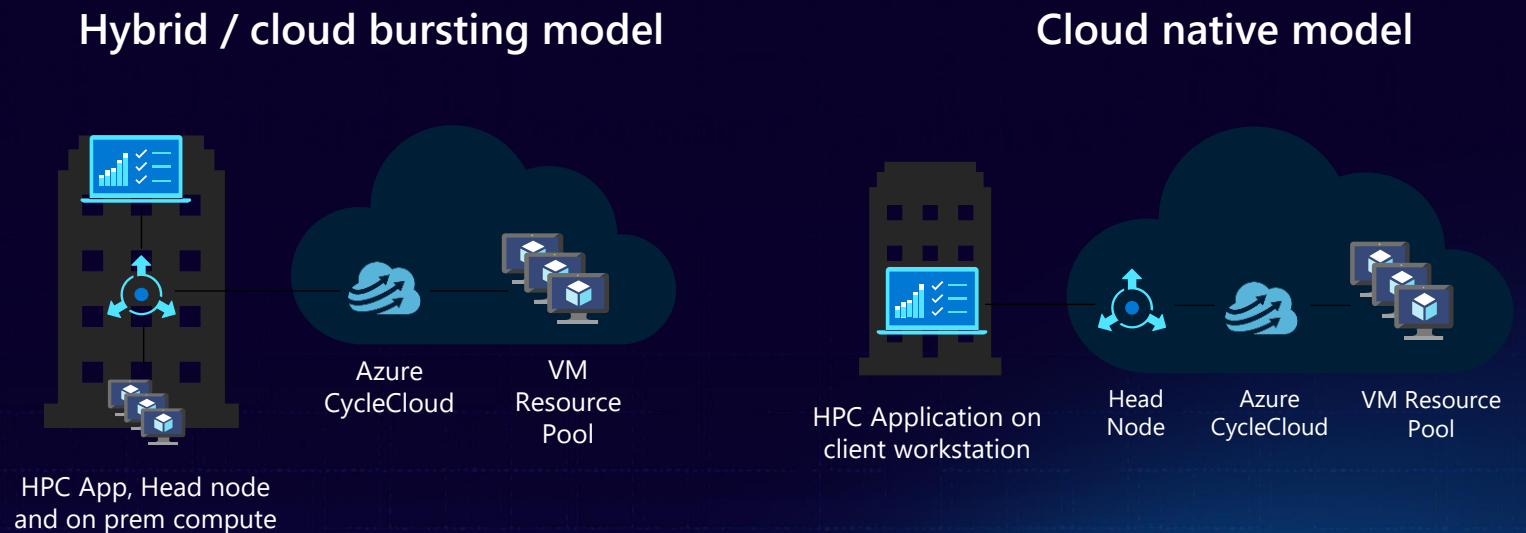
Cloud-native job scheduling



- HPC-as-a-Service Model
- All HPC resources are cloud-based

Azure CycleCloud

Traditional cluster scheduler orchestration



- Support for third party schedulers
- Traditional HPC scaling methodology, but using Azure

Azure Cycle Cloud

User empowerment

- Able to cloud-enable existing workflows and schedulers
- Enable instant access to resources
- Provide auto-scaling, error handling

IT management

- Link workflows for internal and external clouds
- Use Active Directory for authentication and authorization
- Provide secure and consistent access

Business management

- Able to link usage to spend
- Provide tools to manage and control costs



Support for Traditional Job Schedulers





Competition budgets & access

Accelerate | Connect | Excite



Competition Budgets

- Teams will have access to a test window before the competition
- Will not be as large as the competition budget
- Testing budgets will be announced later once the committee has finalized them
- Two great ways to approximate price of a cluster:
 - Azure CycleCloud pricing information
 - Azure Pricing Calculator

<https://azure.microsoft.com/en-us/pricing/calculator>

Select one or more machine types

SKU Search:

Compute Type:

Availability: ☒ Hide Unavailable

Low Priority Support: ☐ Low Priority

Networking Support:

Disk Support:

Ephemeral Support: ☐ Ephemeral

Min Cores:

	SKU	Tier	Cores	Memory	\$/Hour	\$/Core	Quota	Available	Low Prio	SR-IOV	RDMA	Ephemeral	Data Disks	Storage
☐	HB60rs	Standa...	60	223.52 GB	\$2.51	\$0.042	12000	12000	yes	no	yes	yes	4	700.00 GB
☐	HB120rs_v2	Standa...	120	468.75 GB	\$3.96	\$0.033	92000	91040	yes	no	yes	yes	8	937.50 GB

Selected Machine Types

Drag or use buttons to change o

Standard E2s v2

pbstest

☐ Terminate

Edit

Access

Refresh

State **Started** at 8/24/20 12:17 PM (up 8d 22h 49m) - [View in Portal](#)

Nodes **1** ready

Users **1** admin | [Show](#)

Size **1** instance, **4** cores (**\$0.33** per hour)

Usage **96.0 core-hours (~\$9)** in the last 24 hours

Alerts [Create new alert](#)

Issues No issues found

Access to Azure

- Each team will receive login information for a dedicated CycleCloud install/bastion VM
 - ***Note: It is highly recommended that you don't enable public IPs or password logins on your clusters!!!***
- Access will be restricted to a single Resource Group in Azure and dedicated VNETs/Subnets
- VM Family quotas will be set ahead of the competition to ensure fair access to resources
 - Quotas for HPC VM types are pre-determined this year
 - If you want to make an argument for a different VM type, please reach out via Discourse
- Team advisors will get login information by mid-October
- Testing window is currently scheduled for October 24-30

Why CycleCloud?

- Easier for committee to setup and manage environments
- Easier for teams to get started without having to learn intricacies of Azure
- Out-of-the-box autoscaling capabilities to keep costs down
- Realtime cost reporting across clusters managed by CycleCloud down to the minute
 - A special plugin will be installed to allow teams to query their Azure spend, both the total for each month and the current hourly and minute burn rates
 - Getting started resources for CycleCloud are available on the Microsoft Docs site: <https://docs.microsoft.com/en-us/azure/cyclecloud/?view=cyclecloud-8>
 - Azure HPC modules on MS Learn: <https://aka.ms/hpclearn>
 - Azure CycleCloud walkthrough webinar from last year: <https://www.youtube.com/watch?v=oa02aV5MaBU>



Q&A

Accelerate | Connect | Excite



