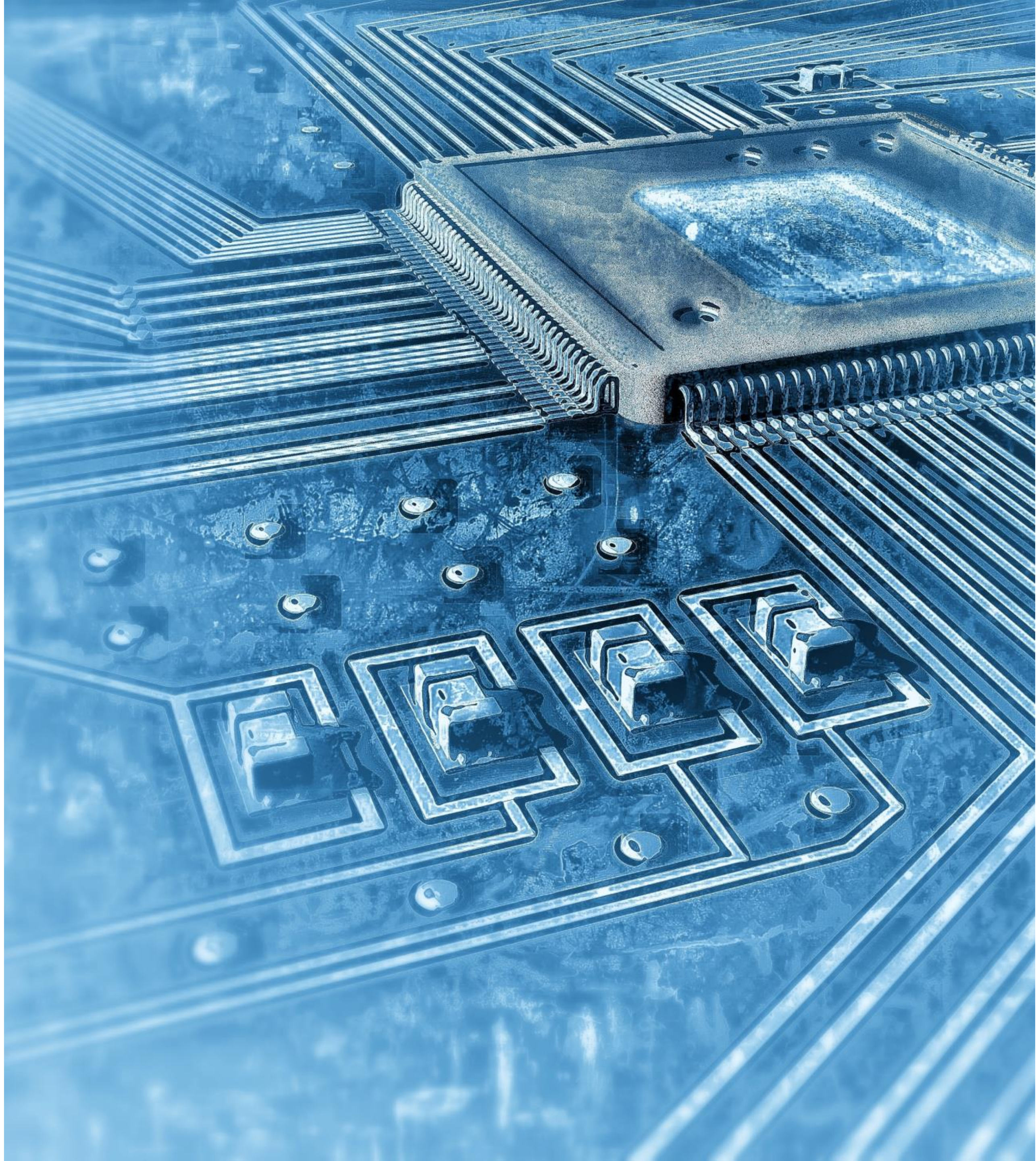
Microsoft

SC21 vSCC
Azure Webinar

August 23, 2021



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/webinars/tutorials with more details about Azure and the cloud component of the competition as they become available

Andy Howard

Azure HPC

Agenda

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

Microsoft

HPC on Azure

Accelerate | Connect | Excite

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 inter-connects as well as edge-to-cloud connectivity

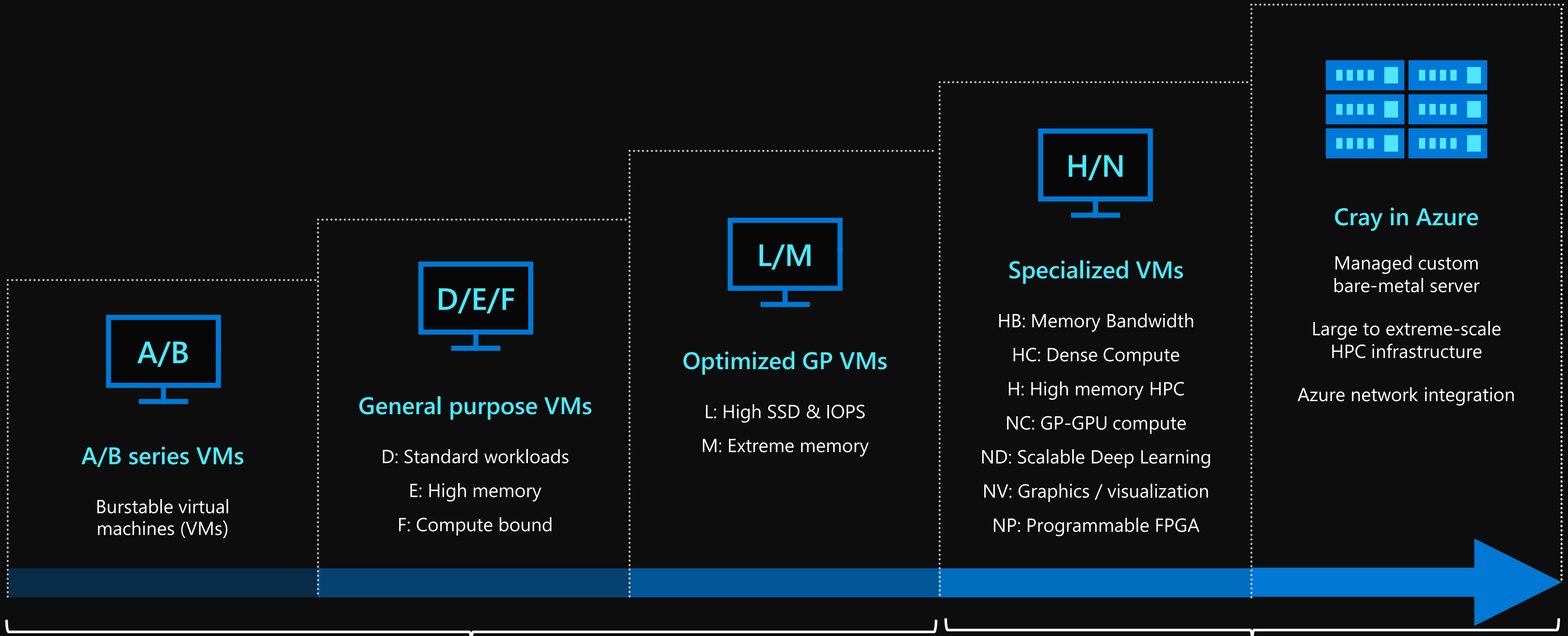
High Performing Storage

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

Workload Orchestration

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

Solve any HPC, AI workload — at any scale



A/B series VMs

Burstable virtual
machines (VMs)

General purpose VMs

D: Standard workloads

E: High memory

F: Compute bound

Optimized GP VMs

L: High SSD & IOPS

M: Extreme memory

Specialized VMs

HB: Memory Bandwidth

HC: Dense Compute

H: High memory HPC

NC: GP-GPU compute

ND: Scalable Deep Learning

NV: Graphics / visualization

NP: Programmable FPGA

Cray in Azure

Managed custom
bare-metal server

Large to extreme-scale
HPC infrastructure

Azure network integration

Small scale MPI
(Handful of cores)

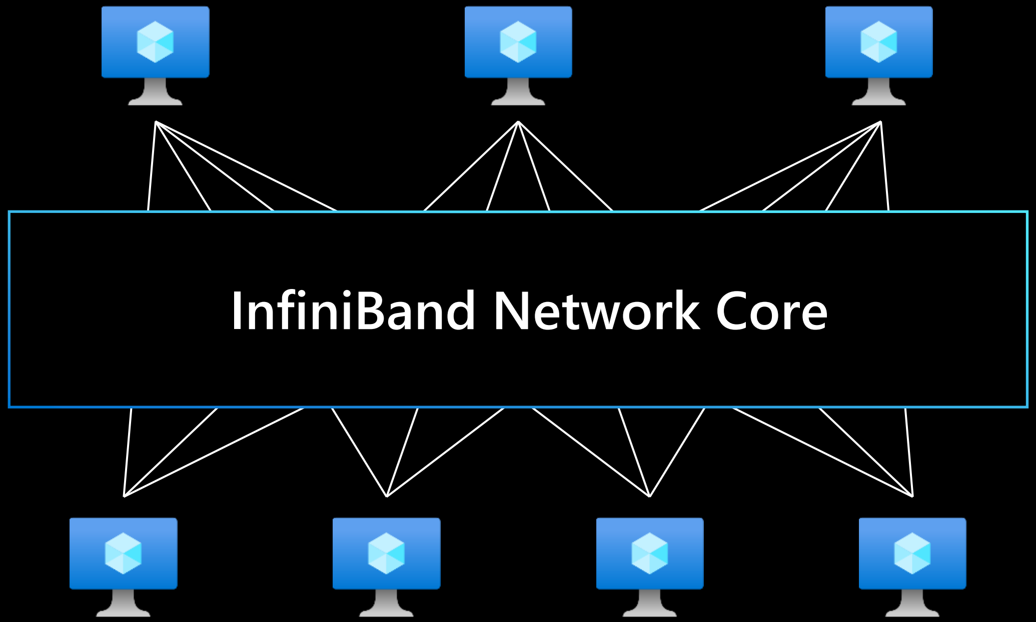
Extreme scale MPI
(100k+ cores)

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

Dynamic Connected Transport

Bare-metal passthrough

Intelligent Adaptive Routing

H

CPU VMs with InfiniBand

HB – Scalable AMD HPC
HC – Scalable Intel HPC

HB-series

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

HC-series

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

Azure HPC v. NSF Track 1

Supercomputing at Scale

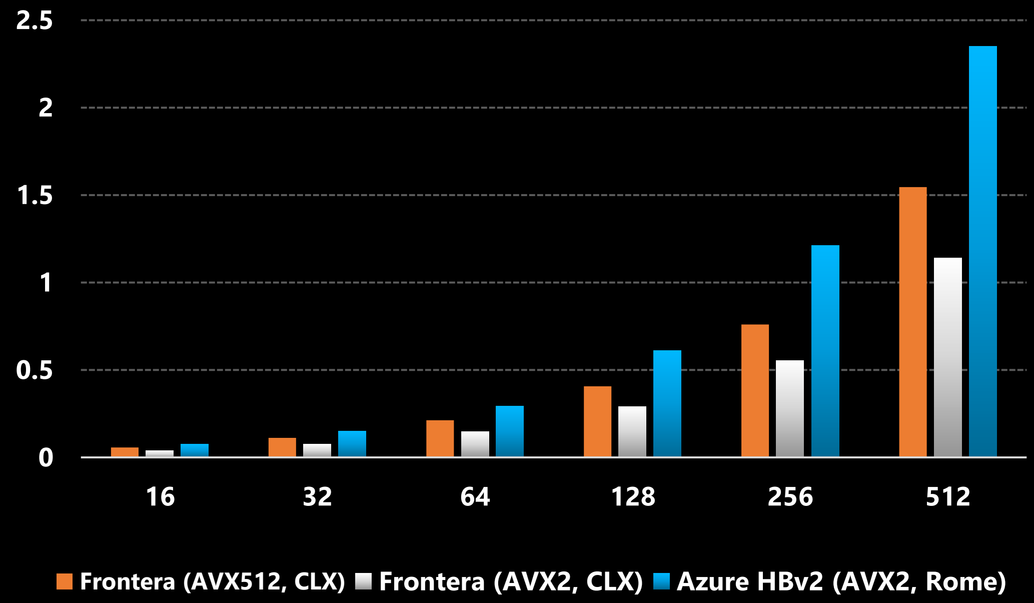
Azure HBv2 outperforms TACC “Frontera” by 40-90% on equivalent test (NAMD 2.14)

Azure AI for Health working with Beckman Institute at Univ Illinois on COVID19 modeling

Azure is putting “NSF Track1” supercomputing capabilities all over the planet

Azure v. a TOP10 Supercomputer

NAMD, nanoseconds/day, higher = better



Azure HPC v. NCAR “Cheyenne”

Supercomputing at Scale

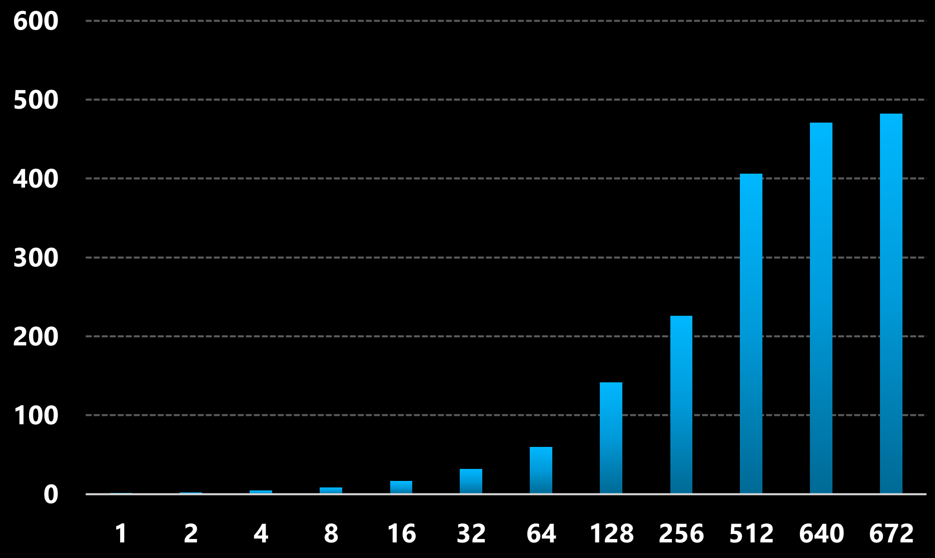
Azure HBv2 outperforms NCAR “Cheyenne” by 2.2x (672 Azure VMs v. 1,024 Cheyenne nodes)

Demonstrates Azure’s ability to run large-scale, and highly impactful weather simulations, and value of Azure’s continuous injection of HPC technology rather than acquisition of static hardware config

80,640 MPI ranks (Feb 2020 Cloud HPC Record)
110% scaling @128 VMs
72% scaling at 672 VMs (Model not big enough)

Simulation Speedup

WRF v. 4.1.3, OpenMPI 4.02, Azure HPC CentOS 7.7
Hurricane Maria, 371m gridpoints, 1km, 1 hr, 3s time-step



HBv2

HBv3 – The New Cloud HPC Flagship

Highest performance, most cost-effective CPU for HPC

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

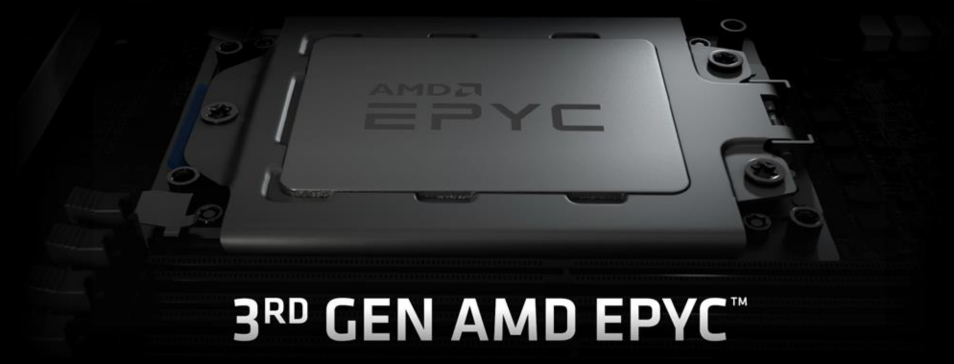
Range of sizes to fit greater range of customer needs

+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)				

Highest Perf per VM

Highest Perf per Core

*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.

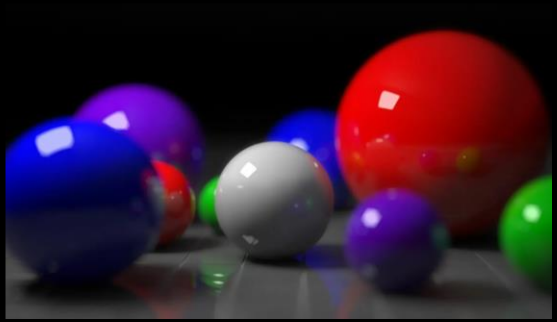
High-Performance Computing VMs (H)

	Available Now	Available Now	Available Now	Available Now
	HBv2	**HB**	**HC**	**H**
Workload Optimized	Memory Bandwidth	Memory Bandwidth	Dense Compute	Large-Memory HPC
CPU	AMD EPYC 2nd Gen “Rome”	AMD EPYC 1st Gen “Naples”	Intel Xeon Platinum 1st Gen “Skylake”	Intel Xeon E5 v3 “Haswell”
Cores/VM	120	60	44	16
TeraFLOPS/VM (FP64)	4 TF	0.9 TF	2.6 TF	0.7 TF
Memory Bandwidth	353 GB/s	263 GB/sec	191 GB/sec	82 GB/s
Memory	4 GB/core, 480 total	4 GB/core, 240 total	8 GB/core, 352 GB	14 GB/core, 224 GB
Local Disk	900 GB NVMe	700 GB NVMe		2 TB SATA
InfiniBand	200 Gb HDR	100 Gb EDR		56 Gb FDR
Network	32 GbE	32 GbE		16 GbE

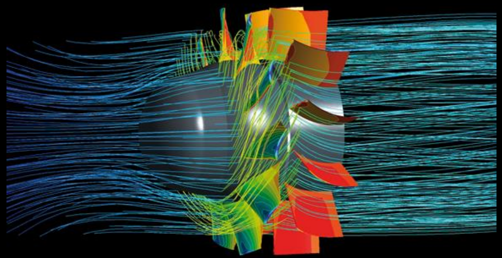
GPU Products in Azure

Visualization

Rendering



HPC/Simulation



Deep-Learning/AI

NV

Graphics Applications

Virtual Desktops & Workstations: Turnkey, Deskless, Cloud-Native

NC

GP-GPU Compute

Flexible sizes with broad global footprint GPU VMs for lightweight and midrange AI, analytics, simulation, and rendering.

ND

Scalable Deep Learning

Scale-up & out for dense AI and HPC with multi-GPU VMs featuring NVLINK interconnect, and InfiniBand

NVv4

Flexible AMD GPU VDI platform

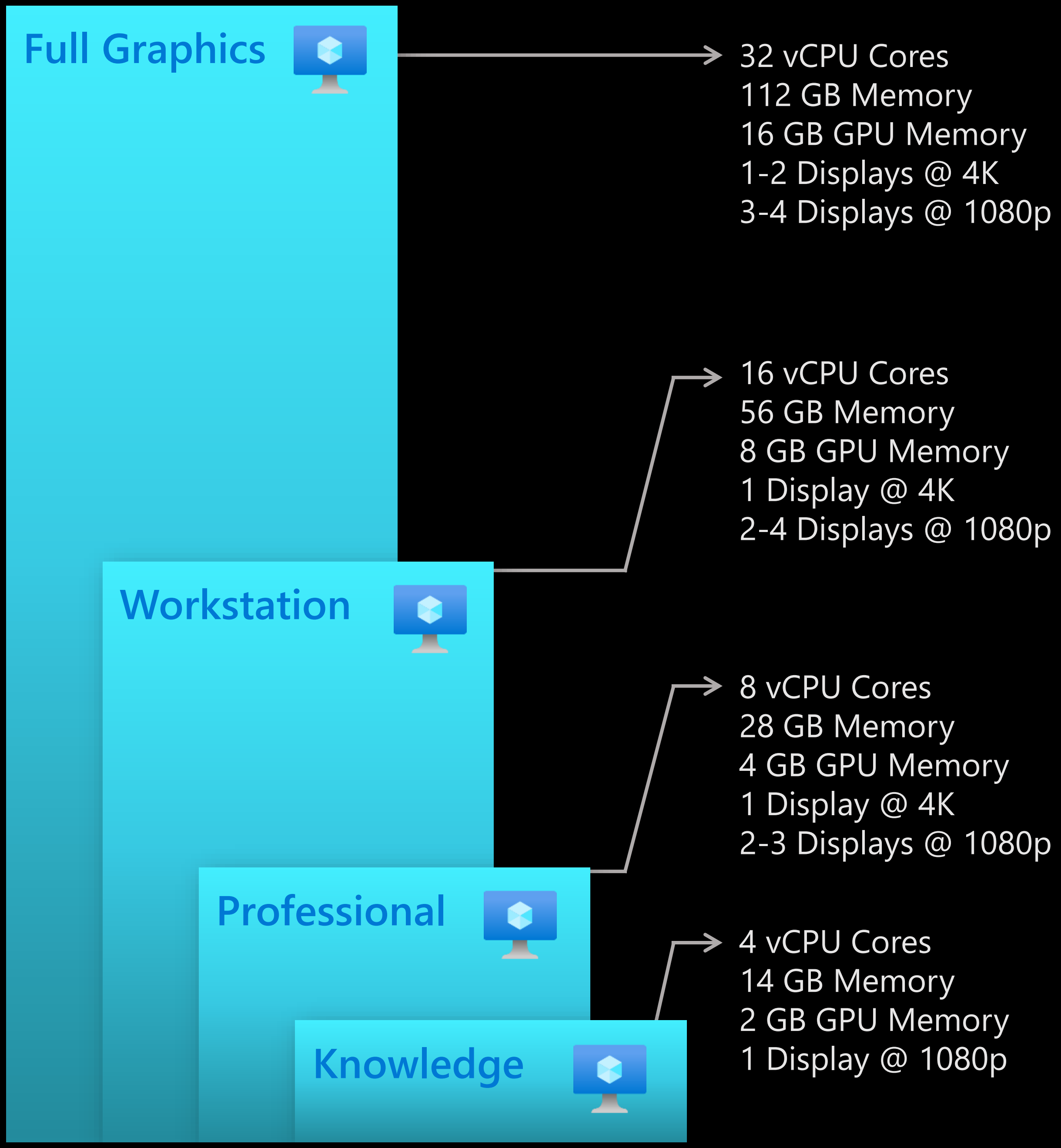
AMD Rome EPYC CPU + Radeon Instinct MI25 GPU

Whole or fractional dedicated GPU acceleration

Right size your workload from 2GB to 16GB
of dedicated HBM2 GPU memory

Most price competitive GPU SKU for VDI: $.10/hour

Continued updates coming soon: Linux guest support,
Hardware encoding, additional Windows Guest OS ver



Full Graphics:
32 vCPU Cores
112 GB Memory
16 GB GPU Memory
1-2 Displays @ 4K
3-4 Displays @ 1080p

Workstation:
16 vCPU Cores
56 GB Memory
8 GB GPU Memory
1 Display @ 4K
2-4 Displays @ 1080p

Professional:
8 vCPU Cores
28 GB Memory
4 GB GPU Memory
1 Display @ 4K
2-3 Displays @ 1080p

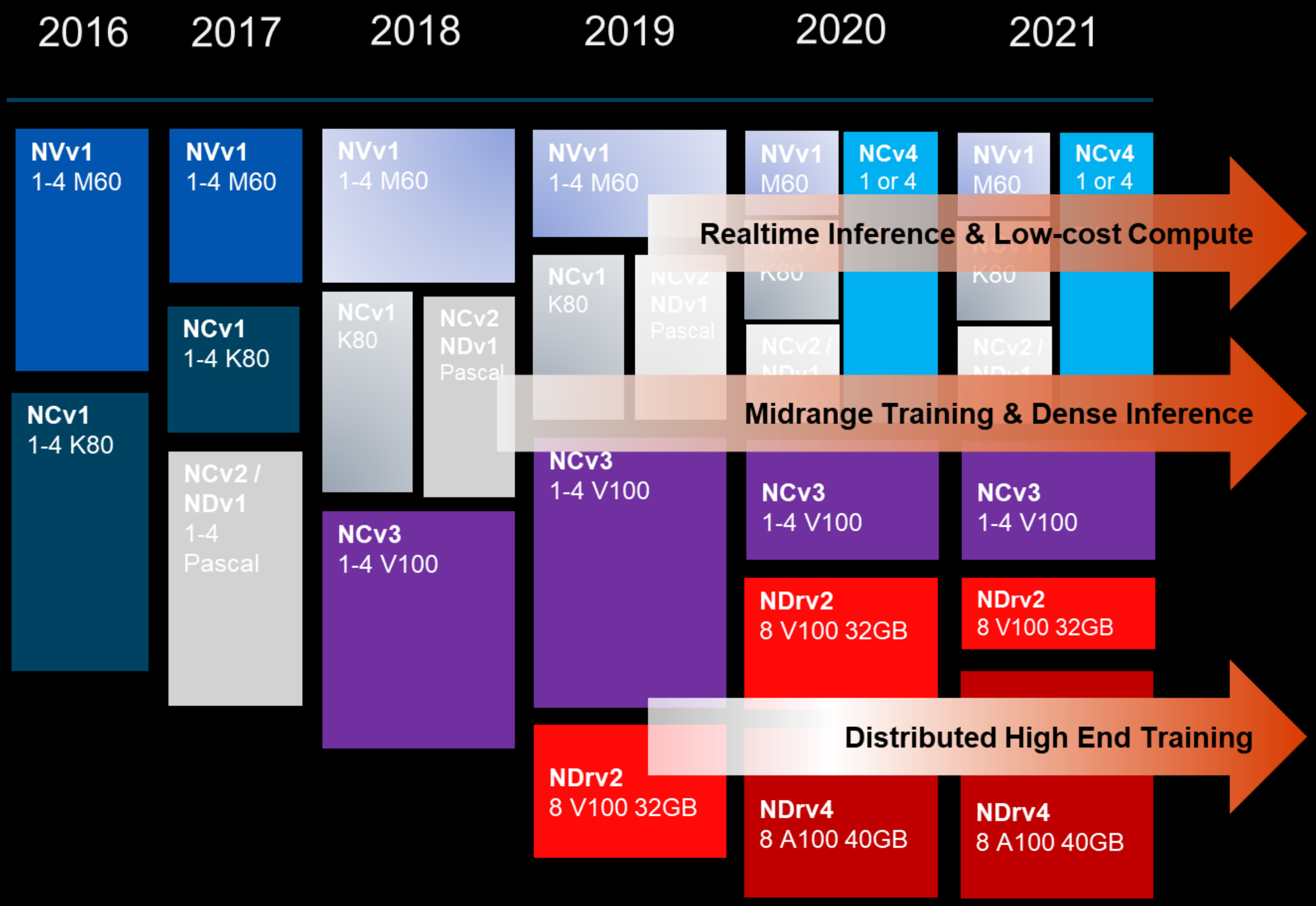
Knowledge:
4 vCPU Cores
14 GB Memory
2 GB GPU Memory
1 Display @ 1080p

NC/ND

GPU-enabled VMs

NC – GP-GPU Compute
ND – Scalable Deep Learning

AI Workloads Are Continuously Evolving



2016 2017 2018 2019 2020 2021

NVv1 1-4 M60

NCv1 1-4 K80

NVv1 1-4 M60

NCv1 1-4 K80

NCv2 / NDv1 1-4 Pascal

NVv1 1-4 M60

NCv1 K80

NCv2 NDv1 Pascal

NCv3 1-4 V100

NVv1 1-4 M60

NCv1 K80

NCv2 NDv1 Pascal

NCv3 1-4 V100

NDrv2 8 V100 32GB

NVv1 M60

NCv4 1 or 4

K80

NCv2 / NDv1

NCv3 1-4 V100

NDrv2 8 V100 32GB

NDrv4 8 A100 40GB

NVv1 M60

NCv4 1 or 4

K80

NCv2 / NDv1

NCv3 1-4 V100

NDrv2 8 V100 32GB

NDrv4 8 A100 40GB

Realtime Inference & Low-cost Compute

Midrange Training & Dense Inference

Distributed High End Training

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)

NCT4_v3

NVIDIA T4 universal deep learning accelerator

AMD Rome EPYC CPU + NVIDIA T4 GPU

High core count per T4 ratio: up to 16 CPUs (no HT) per T4

GPU Memory 16 GB DDR6 300 GB/sec

2560 CUDA Cores / 3RAM20 NVIDIA Tensor Cores per T4

Broad regional rollout with multi-zonal availability

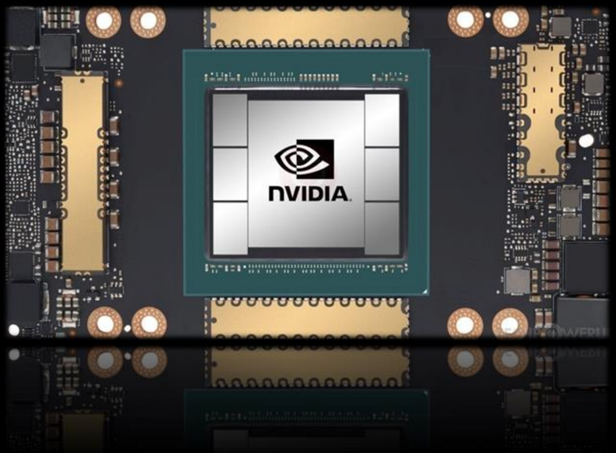
AccelNet enabled for low-latency, consistent networking

Ideal for inferencing, video encoding and lighter GPU compute scenarios

	NCas4_T4_v3	NC8as_T4_v3	NC16as_T4_v3	NC64as_T4_v3
Cores	4	8	16	64
GPU	1xT4	1xT4	1xT4	4 x T4
RAM	28 GB	56 GB	112 GB	432 GB

NDr_v4

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

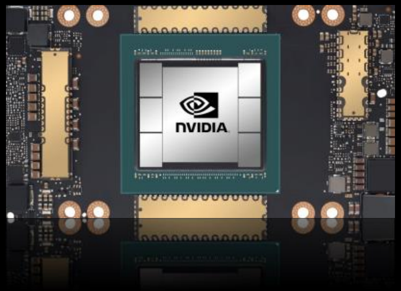
Custom, ground-up platform with PCIe Gen 4-based connectivity for optimal system level performance

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

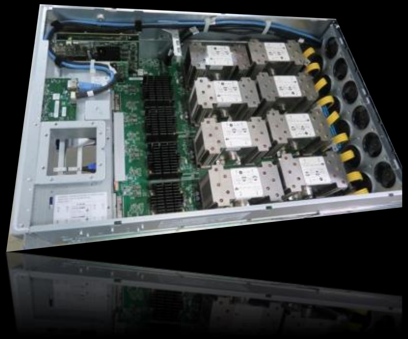
ND A100 v4: Massively Scalable AI Supercomputer

Single A100 GPU



→ NCCL+NVLink

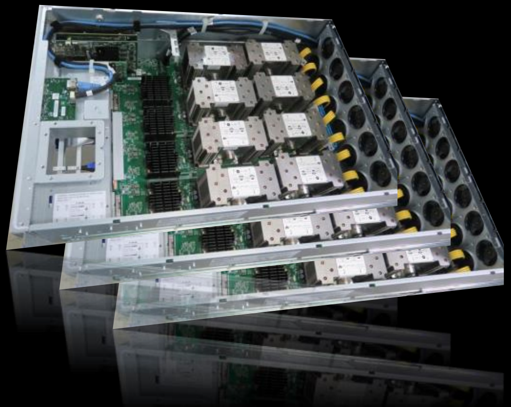
Multi-GPU with NVLINK
1 NDv4 VM, 8 A100s



→ 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

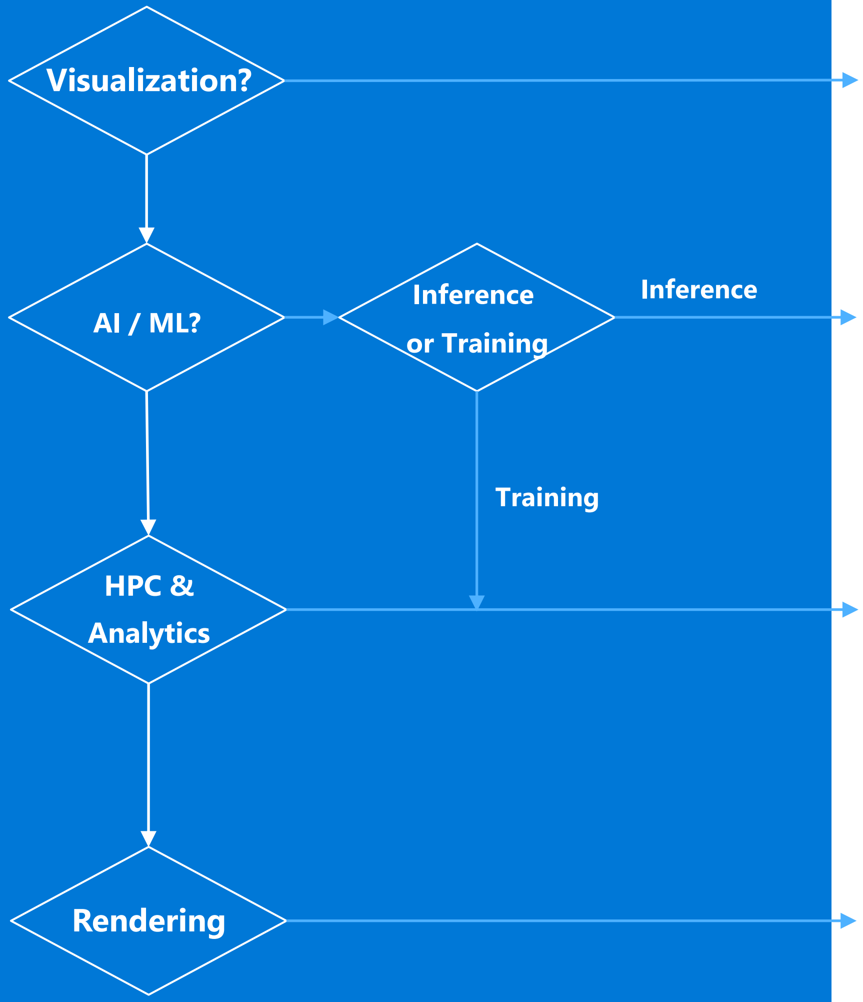
- Between the 8 GPUs local GPUs within each VM
- 2.4 Terabits full-duplex, non-blocking

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

GPU VM Triage



Visualization?

AI / ML?

Inference or Training

Inference

Training

HPC & Analytics

Rendering

Start at top and work down to find a GPU VM Solution

M60

Workload Fit:	VM / GPU Solution:	
Large Dataset (CFD / FEA):	NV_v3	M60
Conventional CAD / Modeling:	NV	M60

V100 P100 P40 K80

Workload Fit:	VM / GPU Solution:	
Large Model:	NC_v3	V100 PCIe
Large Batch Size:	ND	P40
General Purpose:	NC_v2	P100
Simple Models:	NC	K80

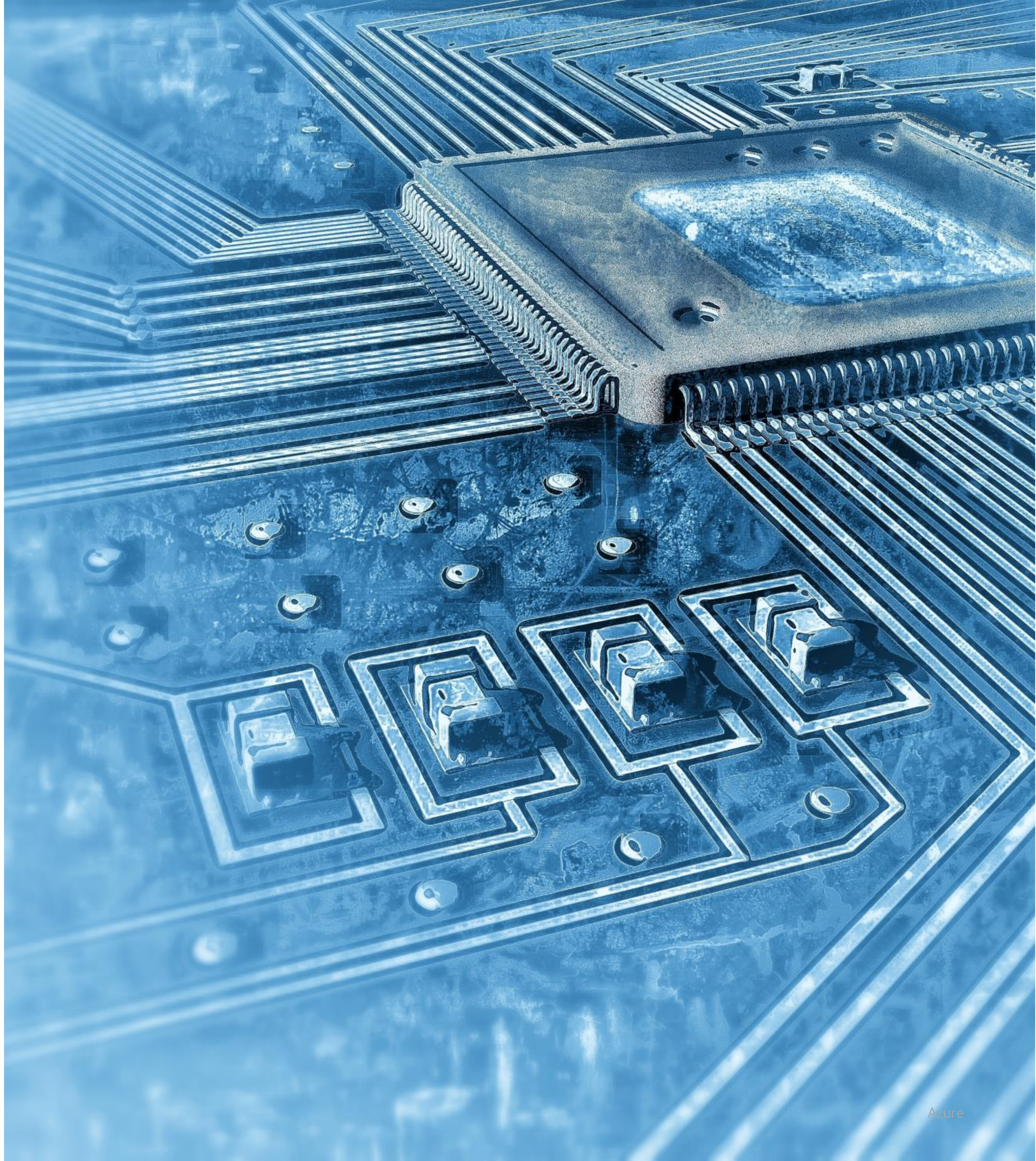
V100 SXM V100 PCIe P100 K80

Workload Fit:	VM / GPU Solution:	
Cost-effective development VM:	NC_v2	P100
Cost-effective deployment:	NC_v3	V100 PCIe
Multi-GPU optimized (6-8 GPUs):	NDr_v2	V100 SXM + EDR
Large jobs (8-500 GPUs):	NCr_v3	V100 PCIe + FDR
Exploration & Education:	NC	K80

P100 P40

Workload Fit:	VM / GPU Solution:	
General Purpose:	NC_v2	P100
Large Textures & High Resolution:	ND	P40

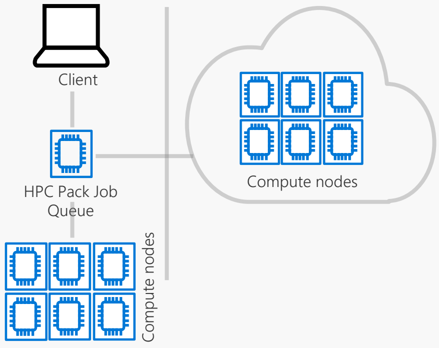
HPC Software Platform



Services for Workload Management

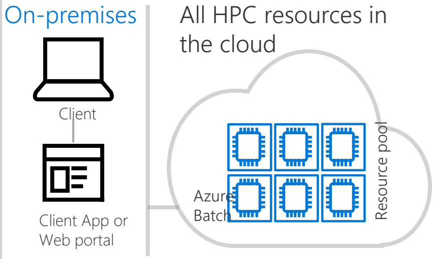
HPC Pack

HPC Scheduler



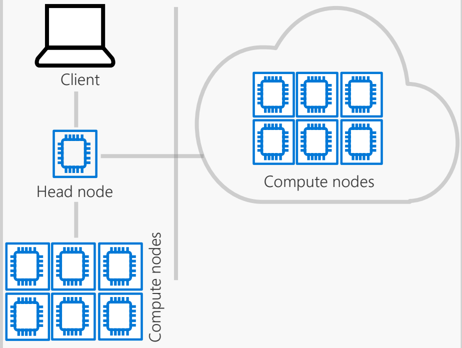
Azure Batch

Batch Jobs



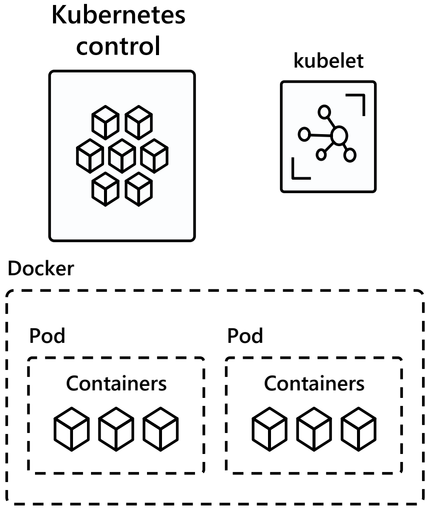
Azure CycleCloud

HPC Clusters

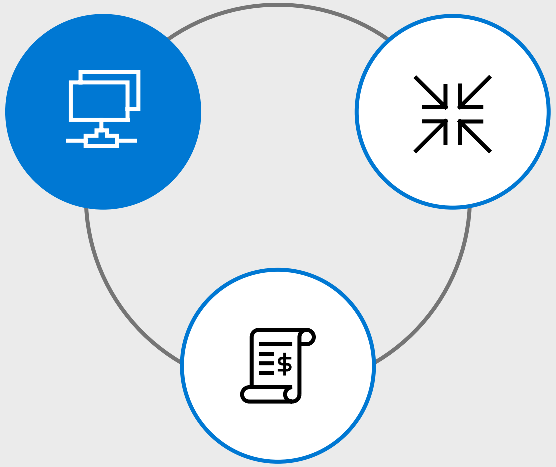


Azure Kubernetes Service

Containers



Azure CycleCloud



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

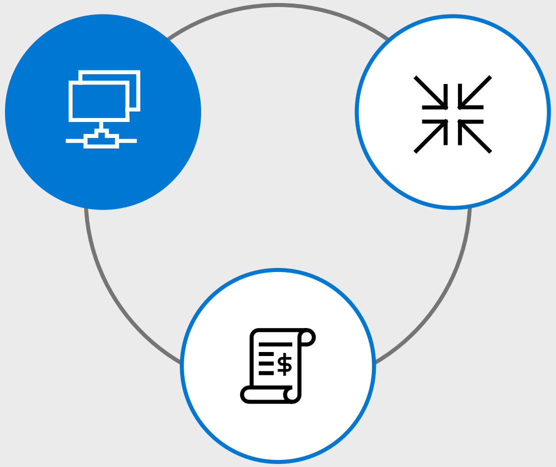
Provide secure and consistent access

Business management

Able to link usage to spend

Provide tools to manage and control costs

Azure CycleCloud



Traditional Scheduler Orchestration

Scheduler Support

Provides autoscaling and orchestration for:

Slurm
OpenPBS
IBM Spectrum LSF
IBM Spectrum Symphony
Grid Engine
+ others

Microsoft

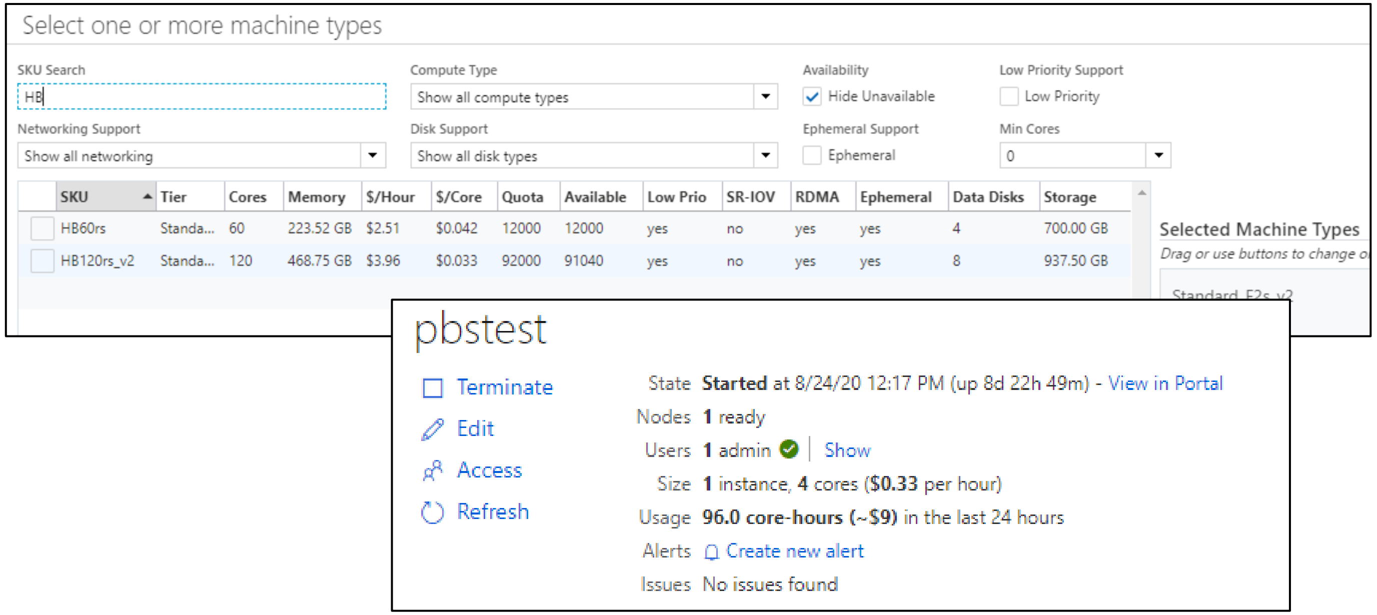
Competition budgets & access

Accelerate | Connect | Excite

Competition Budgets

- **Different testing budgets for each month leading up to the competition**
- **Will not be as large as the competition budget**
- **May go up and down depending on the month**
- **Monthly 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: HB
Compute Type: Show all compute types
Availability: ☑ Hide Unavailable
Low Priority Support: □ Low Priority
Networking Support: Show all networking
Disk Support: Show all disk types
Ephemeral Support: □ Ephemeral
Min Cores: 0

	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_F2s_v2

pbstest

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 will only be increased if teams ask for them!
  - During testing, some reasonable quotas will be set, but will likely be lower than during the actual competition
- **Team advisors will get login information by mid-September**

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

Microsoft

Q&A

Accelerate | Connect | Excite

Microsoft