Microsoft Azure

Shaping the Future of Manufacturing through HPC and AIML

Bruce Moxon
Senior Architect and Partner TPM
Azure HPC & AI Customer Solutions and Innovation

brucemoxon@microsoft.com

Top-of-mind Trends

- Engineering Simulation
  - Larger models, higher fidelity simulations
  - New algorithms leveraging high core counts and/or multiple high-end GPUs
  - AI-augmented design space exploration (GNNs, fast physics, materials, supply chain, design for manufacturing)
  - Co-pilot guided/accelerated model development
  - Surrogates -> High Fidelity Simulations -> Prototypes, Digital Twins

- Materials Science and Drug Discovery
  - More complex simulations (molecules, interactions, biologics)
  - AI-augmented property prediction
  - Large scale in silico screening and candidate identification (larger funnel; fail fast)
  - GNN- and LLM-based molecular design

Leaders in Cloud Supercomputing

OpenAI | Microsoft

Today, we are announcing the third phase of our long-term partnership with OpenAI through a multiyear, multibillion dollar investment to accelerate AI breakthroughs to ensure these benefits are broadly shared with the world.

This agreement follows our previous investments in 2019 and 2021. It extends our ongoing collaboration across AI supercomputing and research and enables each of us to independently commercialize the resulting advanced AI technologies.

- **Supercomputing at scale** – Microsoft will increase our investments in the development and deployment of specialized supercomputing systems to accelerate OpenAI's groundbreaking independent AI research. We will also continue to build out Azure's leading AI infrastructure to help customers build and deploy their AI applications on a global scale.
- **New AI-powered experiences** – Microsoft will deploy OpenAI's models across our consumer and

Inflection AI

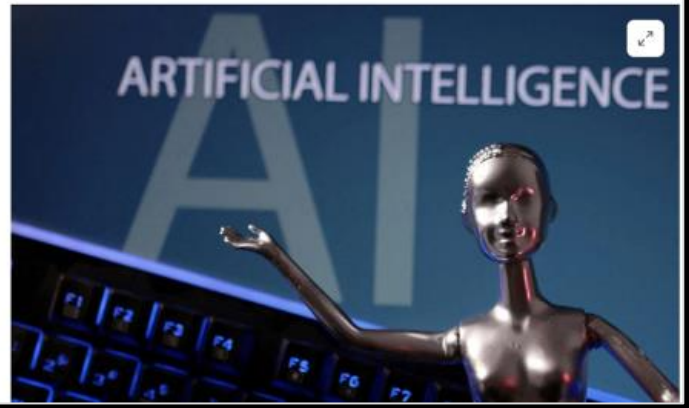
REUTERS® World Business Markets Sustainability More

Disrupted

Microsoft-backed AI startup Inflection raises $1.3 billion from Nvidia and others

By Niket Nishant and Krystal Hu

June 29, 2023 11:57 AM PDT · Updated 3 months ago



April 2021

Microsoft Azure to deploy 4x Supercomputers and Multi-Exabyte Active Data System for UK Met Office

Part of a 10-year managed supercomputing service

Microsoft Azure

November 2021

Azure supercomputer is highest new entry on Top500 at #10 – another cloud first

30 Petaflops Rmax
NVIDIA A100 + AMD EPYC

Microsoft Azure

Azure: the cloud purpose-built for HPC & AI

- ✓ **Genuine HPC approach** platforms, benchmarks, people, and end-to-end experience
- ✓ **Purpose-built platforms** for best performance, and best price-performance, and differentiated solutions
- ✓ **Leading time-to-market for key hardware innovations** to accelerate time-to-solution for customers
- ✓ **Partnering with customers** for the long term to solve HPC and business needs

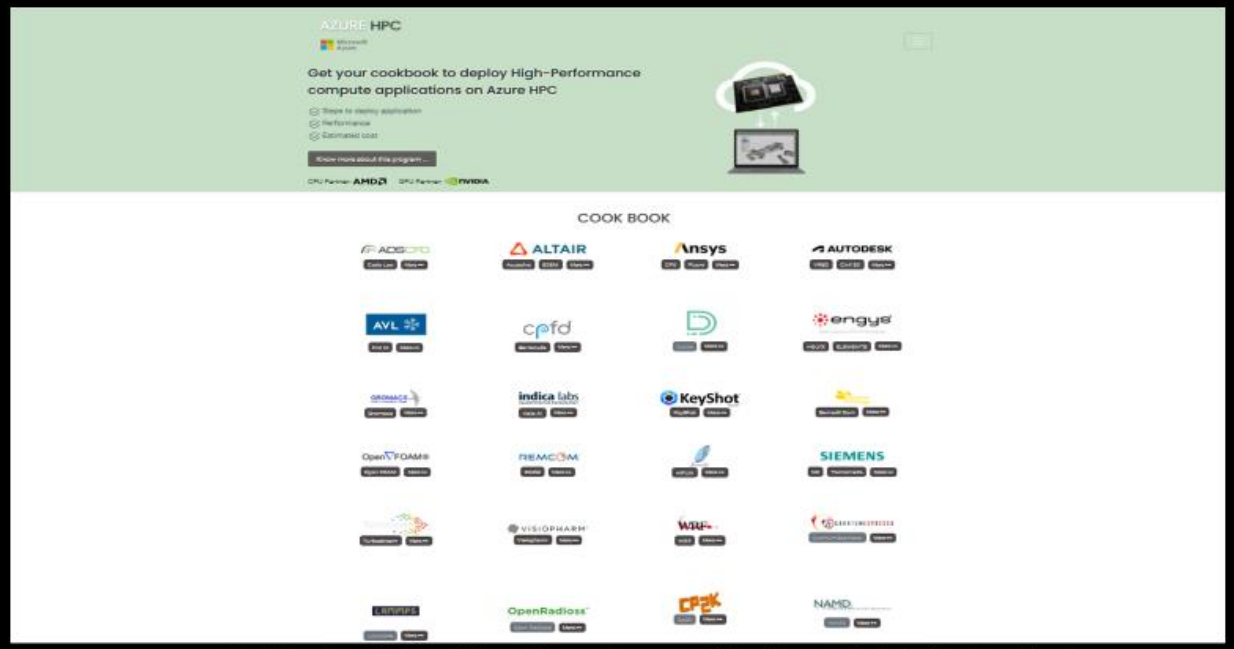
Supercomputing for the most demanding applications	Azure is the only public cloud provider offering the full range of HPC and AI capabilities
InfiniBand HPC & AI clusters for best performance on real workloads	
Compute optimized VMs with "low" latency networks	Compute optimized VMs with "low" latency networks
Azure	Other clouds

Microsoft Azure

Azure HPC & AI breakthroughs

Demonstrating innovation leadership for cloud HPC

2019	20,000 cores MPI job - 1st for cloud
2019	AMD EPYC Rome InfiniBand HPC clusters - 1st for cloud
2019	200 Gb/s HDR InfiniBand with adaptive routing - 1st for cloud
2020	80,000 cores MPI job - 1st for cloud, 12x higher than any other cloud
2020	Top5-scale supercomputer for OpenAI (CPU+GPU) - 1st for cloud
2020	1 TB/sec Parallel File System - 1st for cloud
2021	1.6 Tb/s InfiniBand for NVIDIA A100 clusters - 1st for cloud
2021	HBv3 Milan GA in Azure *at* AMD Launch - 1st for cloud *or* on-prem
2021	10-year Supercomputing-as-a-Service for Met Office - 1st for cloud
2021	#10 supercomputer + 4 more in Top40 - 1st for cloud
2021	#1 cloud on MLPerf benchmark + #2 overall - 1st for cloud
2022	HBv4 Genoa in Azure *at* AMD Launch - 1st for cloud *or* on-prem
2022	400 GB/s NDR InfiniBand proven HPC interconnect - 1st for cloud
2022	*More to come ... and it will probably be 1st for cloud!*



Certification Guide: **https://azurehpc-certification.github.io/index.html**

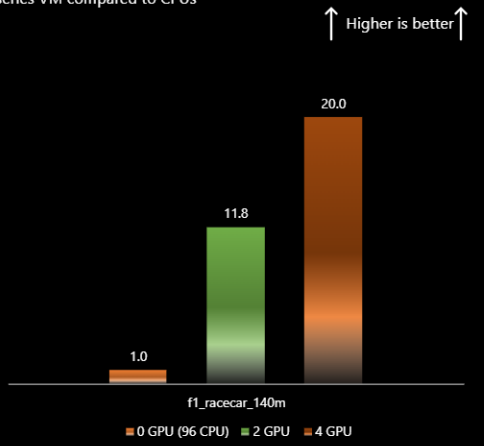
Microsoft

Ansys Fluent

Run your simulations up to 20X Faster on NC A100 v4 series VM compared to CPUs

Ansys Fluent performance on Azure NC A100 v4 series VM powered by 4 NVIDIA Ampere A100 80GB Tensor Core GPUs.

Ansys Fluent is a computational fluid dynamics (CFD) application that's used to model fluid flow, heat and mass transfer, chemical reactions, and more. Ansys Fluent is used in the aerospace, automotive, medical, healthcare, manufacturing, industrial equipment, communication, embedded systems, energy, retail, and consumer goods industries



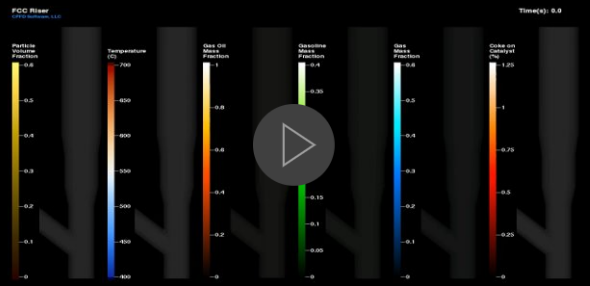
Basis: Pressure based coupled solver, Least Squares cell based, steady for realizable k-epsilon Turbulence solvers containing 140M Hex-core cells shows scalability to higher number of GPUs.

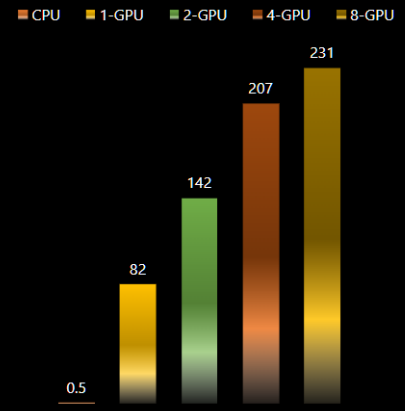
The baseline analysis is performed on 3rd-generation AMD EPYC™ 7V13 (Milan) processors with preview version of Ansys 2023 R2

Microsoft

Barracuda Virtual Reactor

Run your simulations up to 231X Faster on NDm A100 v4 series VM compared to CPUs





Barracuda Virtual Reactor performance on Azure NDm A100 v4 series VM powered by 8 NVIDIA Ampere A100 80GB Tensor Core GPUs.

Virtual Reactor simulates the 3D transient behaviour in fluid-particle systems, including multiphase hydrodynamics, heat balance, and chemical reactions. Uses the Lagrangian formula for the particulate phase, which allows inclusion of discrete particle properties, including the particle size distribution (PSD), composition, temperature, residence time, and history. Provides directional particle filtering through baffles and a GUI.

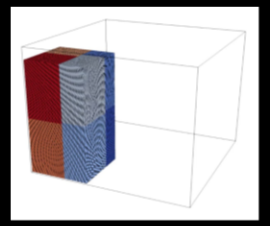
Basis: Particle-based fluid dynamics simulations were run to test the model which contains 40M particles and 0.1M cells

Microsoft

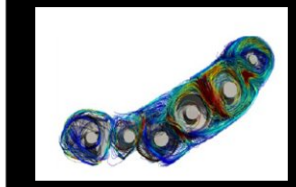
Altair nanoFluidX

Run your simulations up to 6.4X Faster on ND A100 v4 series VM

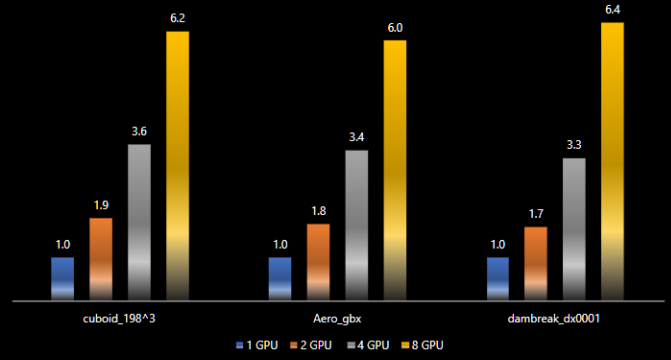
↑ Higher is better ↑



Dam break



Aero gear box



nanoFluidX performance on Azure ND A100 v4 series VM powered by 8 NVIDIA Ampere A100 40GB Tensor Core GPUs.

Automotive, facilities, energy, and environmental industries leverage the GPU scalability of nanoFluidX for predicting the aerodynamic properties of vehicles and facilities

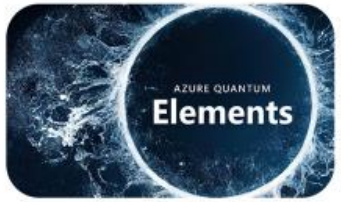
* Basis: GPU-based fluid dynamics simulation using Lagrangian smoothed-particle hydrodynamics (SPH) Solver for the cuboid_198^3 (8M), Aero_gbx (21M) and dambreak_dx0001 (54M) particles.

Microsoft | Azure Quantum Our story Try Azure Quantum Explore Get started today All Microsoft

Azure Quantum Elements

Purpose-built to accelerate scientific discovery

Sign up to learn more about private preview >



Read the customer stories

See how customers are innovating today with Azure Quantum Elements.

Learn more >

Accelerate scientific discovery with Azure Quantum Elements

Sign-up to learn more about the private preview and get the latest updates from the Azure Quantum team

Azure Quantum Elements is a system that boosts productivity for chemistry and materials science R&D. Researchers and product developers can screen candidates, study mechanisms, and design molecules and materials through state-of-the art computing capabilities and enterprise-grade services. Azure Quantum Elements includes simulation workflows optimized for scaling on Azure HPC clusters, AI-accelerated computing, augmented reasoning using AI, integration with quantum tools to start experimenting with existing quantum hardware, and access in the future to Microsoft's quantum supercomputer. With Azure Quantum Elements customers will be able to:

- Accelerate time to impact, with some customers seeing a six-month to one-week speedup from project kick-off to solution.[1]
- Explore more materials, with the potential to scale from thousands of candidates to tens of millions.[2]
- Speed up certain chemistry simulations by 500,000 times, effectively compressing nearly one year of research into one minute.[3]
- Improve productivity with Copilot in Azure Quantum Elements to query and visualize data, write code, and initiate simulations.
- Get ready for quantum computing by addressing quantum chemistry problems today with AI & HPC, while experimenting with existing quantum hardware and getting priority access to Microsoft's quantum supercomputer in the future.
- Save time and money by accelerating R&D pipeline and bringing innovative products to market more quickly.

Sign up

Fill out this form to learn more about the Private Preview

First Name *

Last Name *

Company/Organization Email *

example@domain.com

Company/Organization Name *

Job Title *

DOE TEC^4 PNNL Collaboration

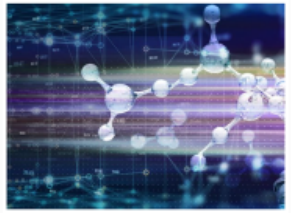
HPCwire

Search... Go

PNNL Collaborates with Microsoft, Micron to Bring Computational Chemistry to the Masses

September 21, 2023

RICHLAND, Wash., Sept. 21, 2023 — Pacific Northwest National Laboratory is collaborating with leading technology companies Microsoft Corp. and Micron Technology to make computational chemistry—a challenging subject but one with far-reaching significance for our lives—broadly available to applied researchers and industrial users.



Computational chemistry problems are incredibly complex but the payoffs for energy research and other applications are enormous. Image credit: Nathan Johnson for PNNL.

The project, known as TEC^4 (Transferring Exascale Computational Chemistry to Cloud Computing Environment and Emerging Hardware Technologies), is part of a broad effort announced by the Department of Energy to quicken the transfer of technology from fundamental research to innovation that can be scaled into products and capabilities that support the economic health and security of the nation.

Road to Quantum

Accelerating scientific discovery with Azure Quantum

Jun 21, 2023 | Jason Zander, Executive Vice President, Strategic Missions and Technologies, Microsoft

Today, Microsoft is announcing new advances to Azure Quantum aimed at accelerating scientific discovery.

Resources

ISV and Open Source Validations and Characterizations

- ISV and Open Source Validations and Characterizations (Cookbook)
- Azure Architecture Center - Azure Architecture Center | Microsoft Learn

Azure Quantum Elements

- Accelerating scientific discovery with Azure Quantum - The Official Microsoft Blog
- Microsoft Azure Quantum Blog | Research, Development & Insights
- Azure Quantum Elements aims to compress 250 years of chemistry into the next 25 (microsoft.com)
- Unlocking the power of Azure for Molecular Dynamics - Microsoft Azure Quantum Blog
- Azure Quantum Elements Private Preview (microsoft.com)
- Azure Quantum | Elements demo (microsoft.com)

HPC Resource Stack on Azure

Transformative Services

Azure Machine Learning | Azure Data Lake | Azure ML Compute

Workload Orchestration

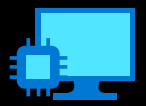
VM Scale Sets | Azure Batch | Azure CycleCloud | Azure Kubernetes Service

Fast, Secure Networking

ExpressRoute | InfiniBand

High Performing Storage

Azure NetApp Files | Azure Managed Lustre | Scalable Object Storage | File and Object Caching



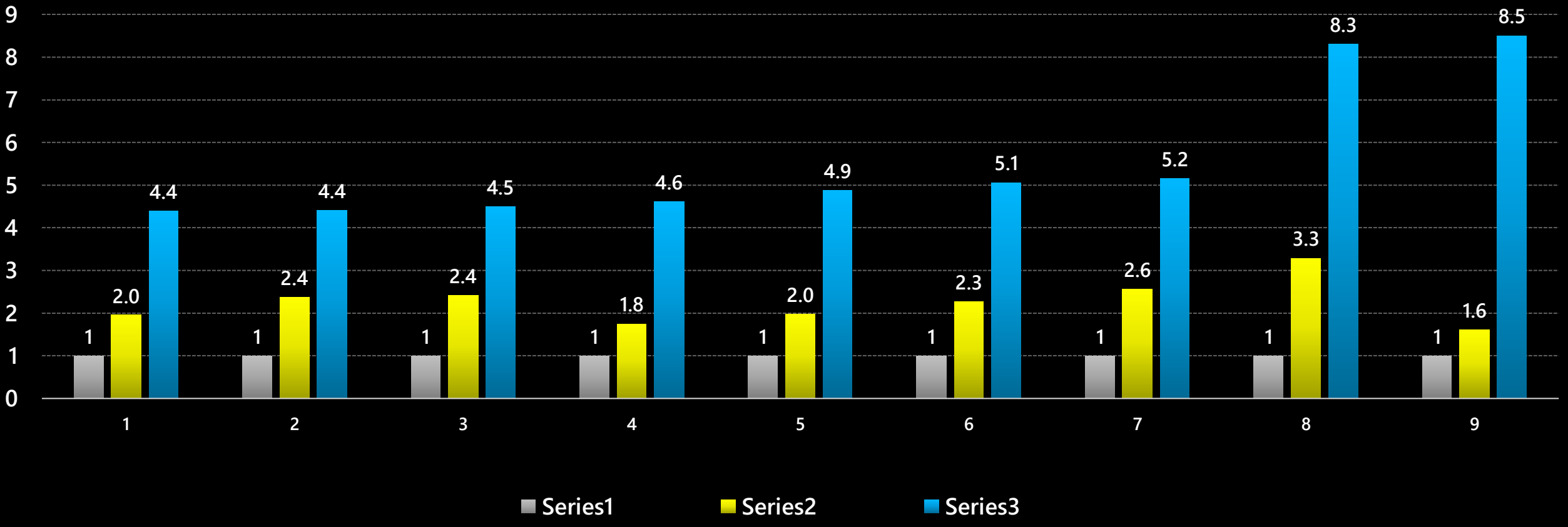
Optimized Compute

H-Series | N-Series | Dedicated Supercomputing

Microsoft Azure

Relative HPC Performance

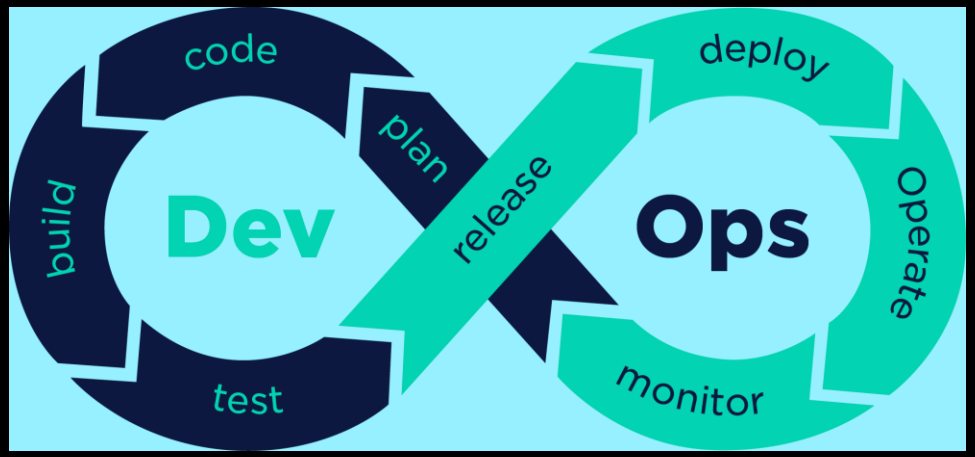
HBv4/HX v. HBv3 v. 4 yr old HPC server, 1 VM (Server), Higher = Better



For a deep dive into the data that went into this analysis, check out the HPC Tech Community Blog: aka.ms/HX-HBv4/TechPreviewBlog

** 4 year old HPC server is represented here by Azure HC-series with Intel "Skylake" 8168 processors. Azure is using HC-series VMs as a reasonable proxy for common, on-premises bare metal HPC server performance from 2018/2019*

Bringing DevOps to HPC/HTC



https://www.360logica.com/blog/agile-to-devops/

Traditional / on-prem HPC/HTC and AIML

- Multiple constituencies
- Collective requirements, architecture
- Benchmarks and POCs
- Formal procurement
- 3-5 year lifetime; rolling technology refresh

On-demand / Cloud

- Project-centric, optimized deployments
- Rapid, Independent Technology Adoption
- Per-project rqmts, dev-ops-maint-retire
- Campaign compute and storage
- On-demand deployment and scaling (up/down) matched to project phasing (dev-ops-maint)

https://insidehpc.com/2021/04/hpc-devops-powered-by-the-cloud/

Microsoft Azure

Azure H-series for HPC

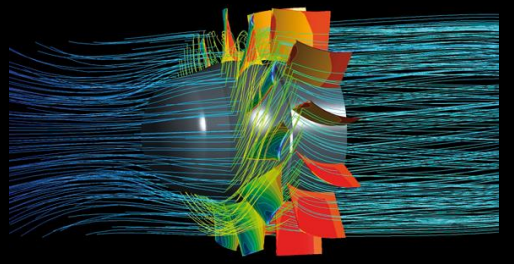
	HX	HBv4	HBv3	HBv2	HC
Processor	176 cores AMD Genoa (Preview) AMD Genoa-X (GA)	176 cores AMD Genoa AMD Genoa-X	120 cores AMD Milan-X	120 cores AMD Rome	44 cores Intel Skylake
Memory	1.4 TB	688 GB DDR5	448 GB DDR4	448 GB DDR4	352 GB DDR4
DRAM Bandwidth	750 GB/s	750 GB/s	350 GB/s	350 GB/s	190 GB/s
InfiniBand	400 Gb/s	400 Gb/s	200 Gb/s	200 Gb/s	100 Gb/s
SSD	3.6 TB NVMe	3.6 TB NVMe	1.8 TB NVMe	900 GB NVMe	700 GB SSD
Availability	Available	Available	Available	Available	Available

GPU Computing

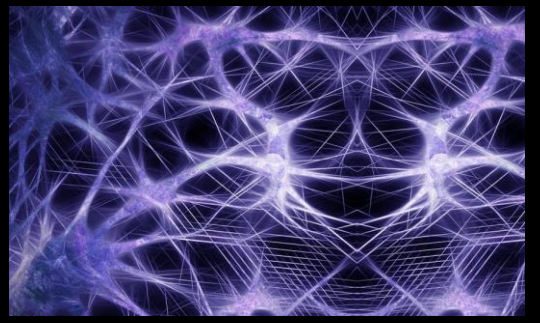
Visualization

Rendering

HPC/Simulation



Deep-Learning/AI



NV

Workload centric PCs & Workstations

Variable configuration for GPU workstations for content consumption

Modern workspace for interactive collaboration

NC

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

ND

N-series GPU VMs on Azure

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)			Azure Network	Azure Network + NVLink GPU Interconnect

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			

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