Microsoft

Exascale in the Era of Hyperscale

The many flavors of Azure's unique journey to 10^{18}

Evan Burness

Microsoft Azure (2017-2019)
Principal Program Manager for H-series VMs for high-performance computing

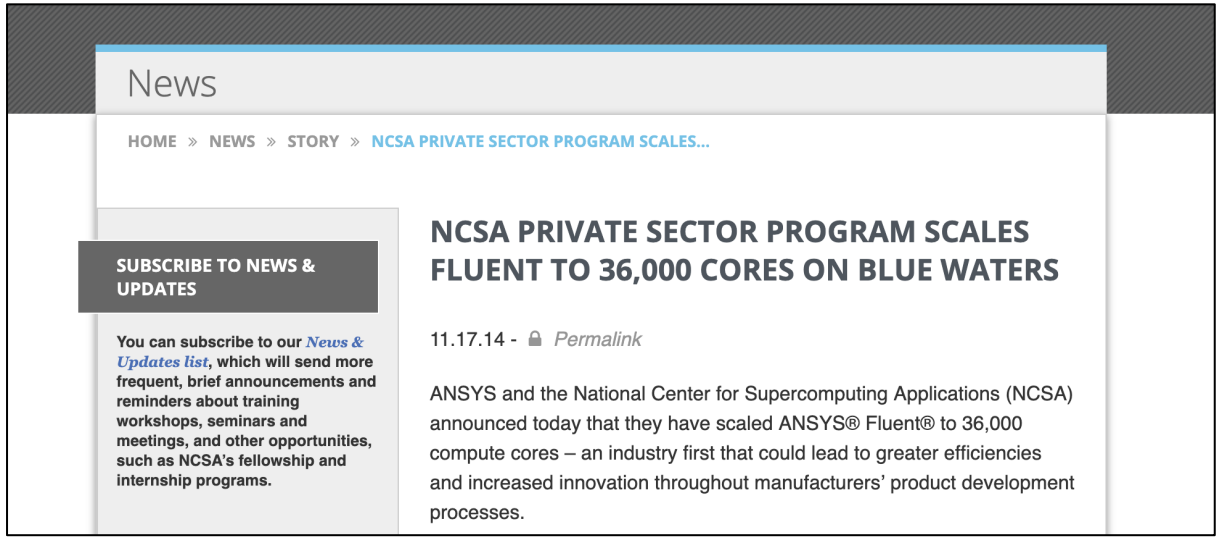
Cycle Computing (2016-2017)
Director for High-Performance Computing

National Center for Supercomputing Applications (Univ. Illinois) (2009-2016)
Program Manager, Private Sector Program

My Early Experiences with Really Large Scale

System Overview

- System: Cray XE/XK hybrid supercomputer
- Cabinets: 288
- Peak Performance: 13 PF
- System Memory: 1.5 PB
- XE Compute Nodes: 22,640 AMD 6276 "Interlagos" processors
- XK Compute Nodes: 4,224 NVIDIA® GK110 "Kepler" GPU accelerators
- Interconnect: Gemini



News

HOME » NEWS » STORY » NCSA PRIVATE SECTOR PROGRAM SCALES...

SUBSCRIBE TO NEWS & UPDATES

You can subscribe to our *News & Updates list*, which will send more frequent, brief announcements and reminders about training workshops, seminars and meetings, and other opportunities, such as NCSA's fellowship and internship programs.

NCSA PRIVATE SECTOR PROGRAM SCALES FLUENT TO 36,000 CORES ON BLUE WATERS

11.17.14 - *Permalink*

ANSYS and the National Center for Supercomputing Applications (NCSA) announced today that they have scaled ANSYS® Fluent® to 36,000 compute cores – an industry first that could lead to greater efficiencies and increased innovation throughout manufacturers' product development processes.

54 regions worldwide **140** available in 140 countries

* Two Azure Government Secret region locations undisclosed

Inside Azure Datacenters

Amsterdam, NL

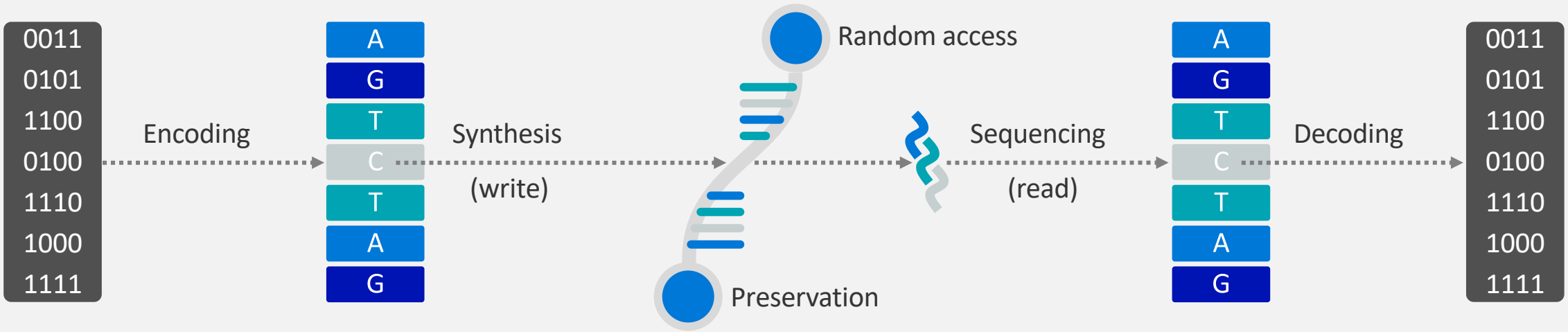
Cheyenne, WY

DNA – Project Palix

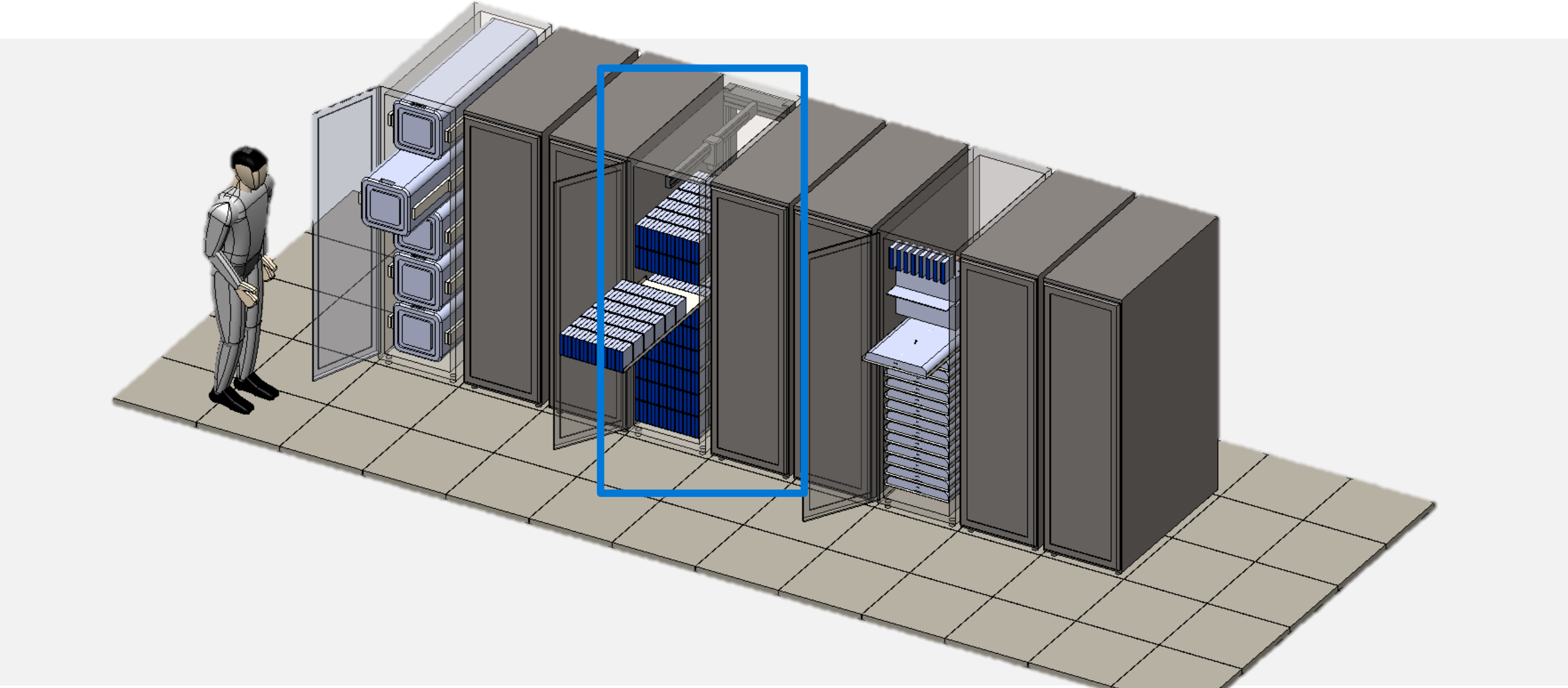
Joint project between Microsoft and the University of Washington

>1,000x denser than tape

2,000 years in ~10˚C



Storing more than 1ZB in a datacenter rack



Azure is Exascale

000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000,
000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000,
000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000,
000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000,
000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000,
000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000,
000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000,
000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000,
000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000,
000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000,
000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000

Azure is Exascale

(With one minor exception ☺)

Specialized Compute Fleet

Entry Level VMs

Dev/Test Workolads

General Purpose VMs

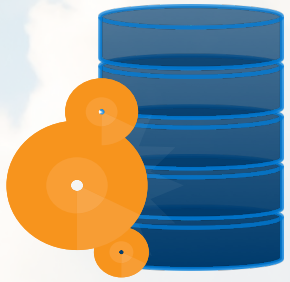
Common Applications, Web servers etc

Compute Optimized VMs

Gaming, Analytics

Large Memory VMs

Large Databases



>80,000 IOPs Premium Storage

Low latency, high throughput apps

Storage optimized VMs

No SQL Databases (Cassandra, MongoDB), Data warehousing

InfiniBand

CPU VMs

Batch processing, fluid dynamics, monte carlo simulation

GPU VMs

NV -Graphic based applications

NC2 – Advanced Sim (P100)

ND1 – AI Inferencing (P40)

ND2* – AI Training (V100/V100 SXM)

FPGA*

Virtual Machines – HPC

FPGA Microservices – AI/Edge

Cray Services in Azure

IB Connected CPU/GPU/Storage available in cloud

High-Performance Computing VMs (H)

Powered by **AMD EPYC** and **Intel Xeon Platinum**

	HB	HC
CPU	AMD EPYC	Intel Xeon Platinum
Cores/VM	60	44
Clock Speed*	2.55 GHz	3.4 GHz
Memory Bandwidth	263 GB/sec	191 GB/sec
Memory	4 GB/core 240 total	8 GB/core 352 GB
Local Disk	700 GB NVMe	
InfiniBand	100 Gb EDR w/ ConnectX-5 (SRIOV)	
Network**	Gen 2 SmartNIC (Arria10 FPGA accelerated)	

*All cores, non-AVX, peak Boost/Turbo frequencies

**More on Azure SmartNIC at https://bit.ly/2OAainE

High-Performance Computing VMs (H)

Powered by **AMD EPYC** and **Intel Xeon Platinum**

	HB	HC
Targets	CFD, Seismic, Weather,	FEA, MD, Chemistry
Workload Driver	Memory Bandwidth	Raw Compute
Max MPI job per VMSS (300 VMs)	18,000 cores	13,200 cores
RDMA Verbs	All	All
MPI Support	MVAPICH2, OpenMPI, MPICH, Platform MPI, Intel MPI, HPC-X	
Additional Frameworks	Unified Communication X libfabric, PGAS	
Azure Storage Support	Premium	Premium
OS Support for RDMA	CentOS/RHEL 7.5+ SLES 12 SP4+ WinServer 2016+	CentOS/RHEL 7.5+ SLES 12 SP4+ WinServer 2016+

Azure H-series
HW + SW Stack

Project Olympus

Flexible and Modular design to handle wide variety of public cloud workloads

Open Compute Project open source design

Compute

Intel, AMD, ARM64 CPUs

High density GPU expansion for HPC/AI

NVM (DRAM+battery) and 3DXP for low-latency

Storage

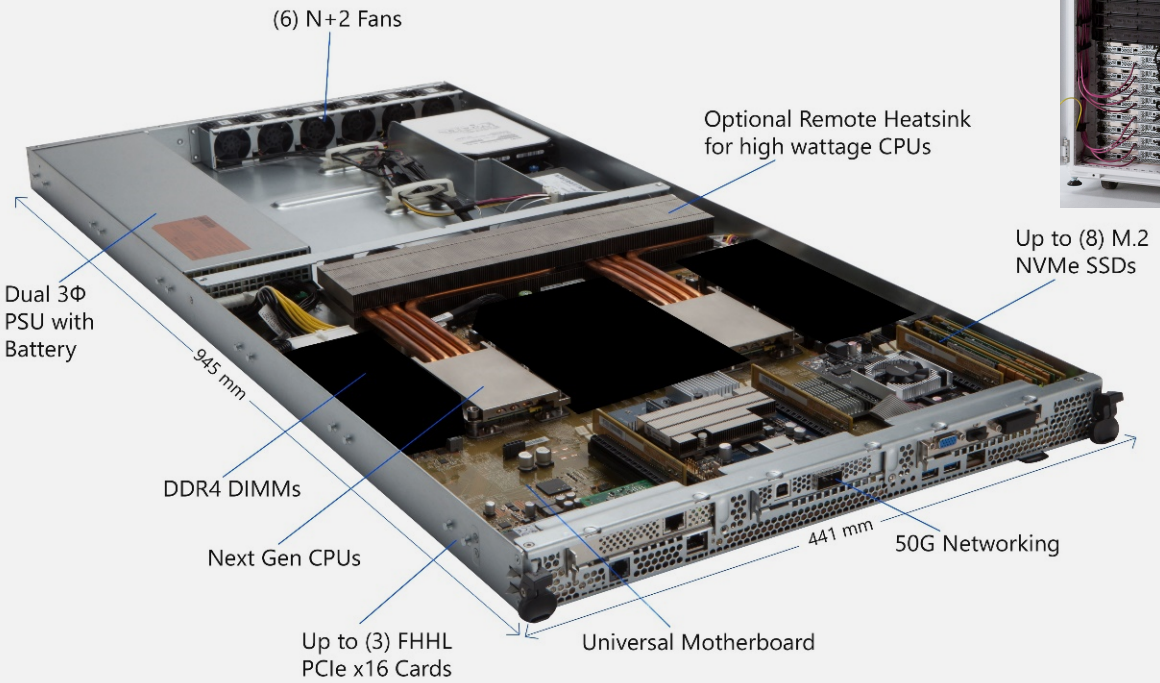
High density HDD and Flash expansion

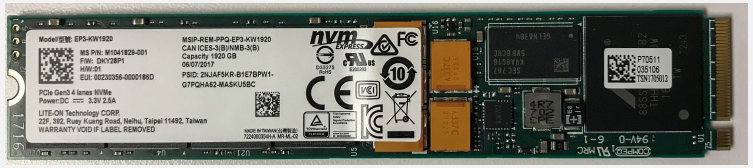
Microsoft custom designed SSDs

Networking

50 Gbps networking

Accelerated VMs using FPGAs





Microsoft SSD

The Azure OS

Minimal version of Windows Server Core 2016 running Hyper-V

Only native x64 code – no managed code on the host

Remove all language packages except English

Disable unused drivers and features such as print spooler, etc.

Minimal set of server roles

How we shape the VM

Managing CPU Groups

CPU groups are managed through the Hyper-V Host Compute Service, or HCS. A great description of the HCS, its genesis, links to the HCS APIs, and more is available on the Microsoft Virtualization team's blog in the posting Introducing the Host Compute Service (HCS).

> **Note**
>
> Only the HCS may be used to create and manage CPU groups; the Hyper-V Manager applet, WMI and PowerShel management interfaces don't support CPU groups.

Microsoft provides a command line utility, cpugroups.exe, on the Microsoft Download Center which uses the HCS interface to manage CPU groups. This utility can also display the CPU topology of a host.

How CPU Groups Work

Allocation of host compute resources across CPU groups is enforced by the Hyper-V hypervisor, using a computed

https://bit.ly/2Y0nIz1

How we shape the host

Hyper-V Host CPU Resource Management

12/15/2017 • 4 minutes to read • Contributors

Hyper-V host CPU resource controls introduced in Windows Server 2016 or later allow Hyper-V administrators to better manage and allocate host server CPU resources between the "root", or management partition, and guest VMs. Using these controls, administrators can dedicate a subset of the processors of a host system to the root partition. This can segregate the work done in a Hyper-V host from the workloads running in guest virtual machines by running them on separate subsets of the system processors.

For details about hardware for Hyper-V hosts, see Windows 10 Hyper-V System Requirements.

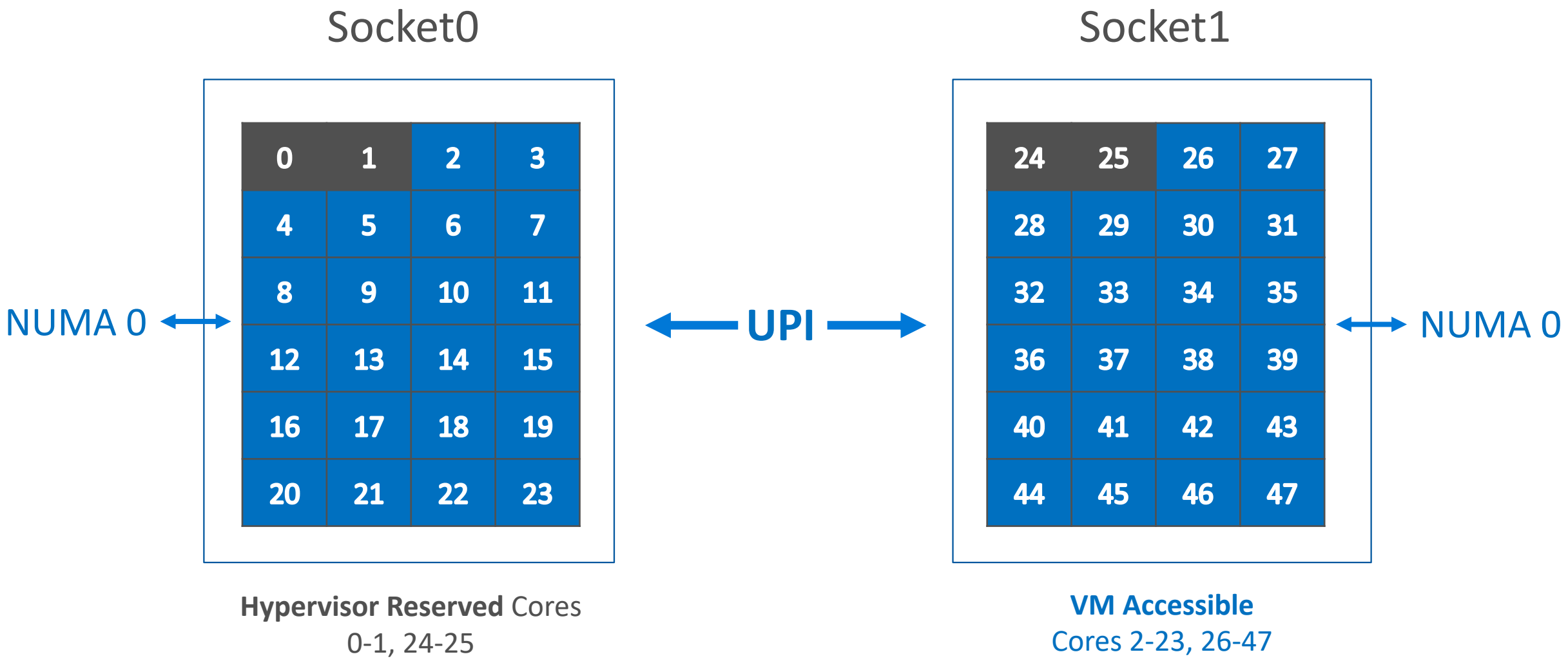
Background

Before setting controls for Hyper-V host CPU resources, it's helpful to review the basics of the Hyper-V architecture. You can find a general summary in the Hyper-V Architecture section. These are important concepts for this article:

- Hyper-V creates and manages virtual machine partitions, across which compute resources are allocated and

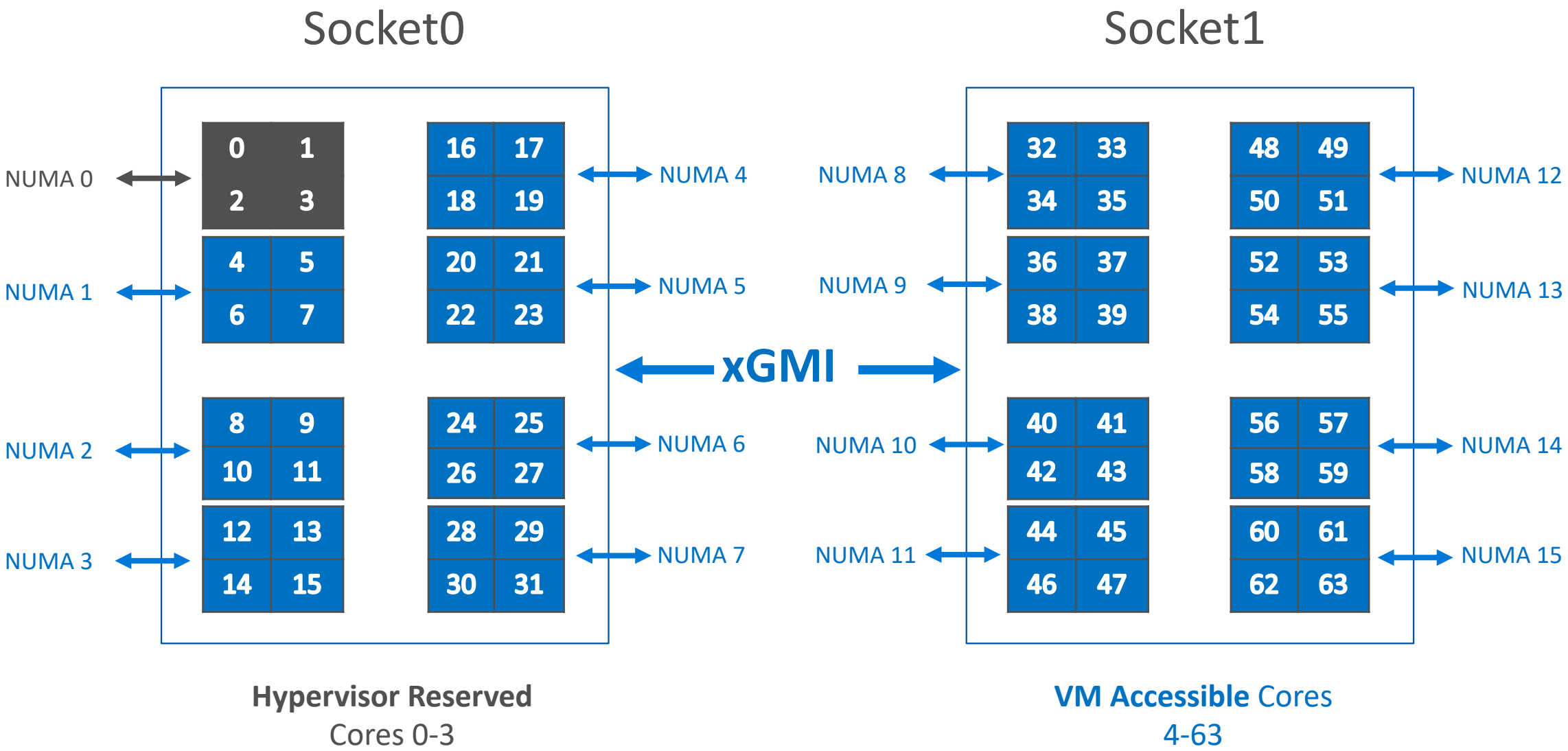
https://bit.ly/2IqTKyS

Azure HC44rs VMs



**VM accessible cores enumerate as 0-43 within the VM*

Azure HB60rs VMs



**VM accessible cores enumerate as 0-59 within the VM*

MPI Bandwidth

OSU MPI Bandwidth Test v5.4.3

CLI: ./mvapich2-2.3.install/bin/mpirun_rsh -np 2 -hostfile ~/hostfile MV2_CPU_MAPPING=36 ./mvapich2-2.3/osu_benchmarks/mpi/pt2pt/osu_bw

#bytes	Bandwidth (MB/s)	Bandwidth (Gb/s)
2	6.18	0.04944
4	13.27	0.10616
8	26.58	0.21264
16	53.51	0.42808
32	106.81	0.85448
64	211.24	1.68992
128	386.98	3.09584
256	756.32	6.05056
512	1434.2	11.4736
1024	2663.8	21.3104
2048	4396.99	35.17592
4096	6365.86	50.92688
8192	8137.9	65.1032
16384	9218.29	73.74632
32768	10564.61	84.51688
65536	11275.6	90.2048
131072	11633.7	93.0696
262144	11856.27	94.85016
524288	11962.69	95.70152
1048576	12025.43	96.20344
2097152	12038.7	96.3096
4194304	11290.92	90.32736

MPI_send_recv

OSU MPI Latency Test v5.4.3

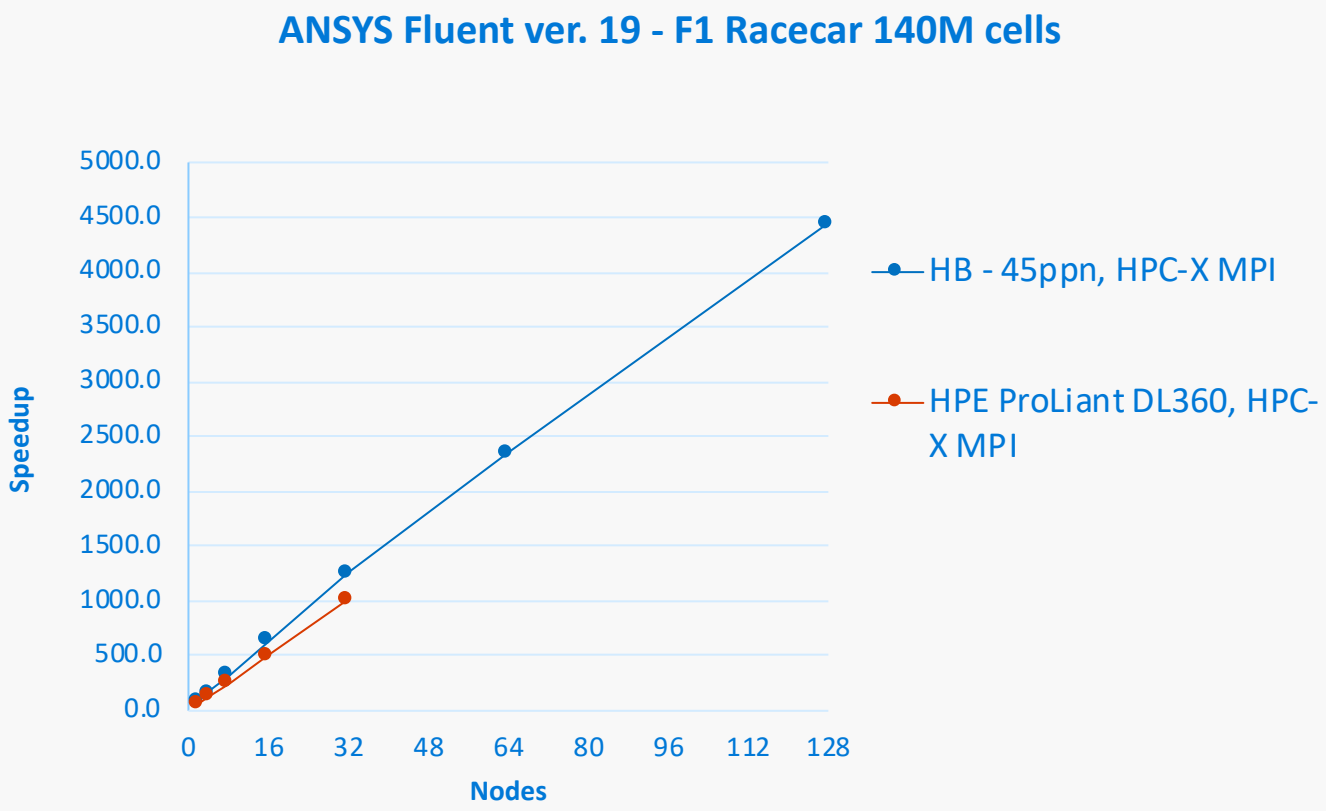
CLI: ./bin/mpirun_rsh -np 2 -hostfile ~/hostfile MV2_CPU_MAPPING=[INSERT CORE #] ./osu_latency

#bytes	Latency [usec] (MPICH 3.3 + CH4)	Latency [usec] (OpenMPI 4.0.0)	Latency [usec] (MVAPICH2 2.3)
2	1.84	1.78	2.08
4	1.84	1.79	2.08
8	1.85	1.79	2.05
16	1.85	1.79	2.1
32	1.87	1.82	2.12
64	2	1.95	2.13
128	2.05	2	2.18
256	2.48	2.44	2.75
512	2.57	2.52	2.81
1024	2.76	2.71	2.97
2048	3.09	3.11	3.34
4096	3.72	3.91	4.44

Azure H-series
Benchmarks & Applications

ANSYS Fluent on Azure HB

24% faster than bare-metal with identical InfiniBand stack



Nodes	Instance	Version	PPN	Cores	MPI	Solver Rating
2	HB60rs	v192	45	90	HPCX	81.9
4	HB60rs	v192	45	180	HPCX	159.7
8	HB60rs	v192	45	360	HPCX	324.8
16	HB60rs	v192	45	720	HPCX	638.6
32	HB60rs	v192	45	1440	HPCX	1260.9
64	HB60rs	v192	45	2880	HPCX	2351.0
128	HB60rs	v192	45	5760	HPCX	4447.9

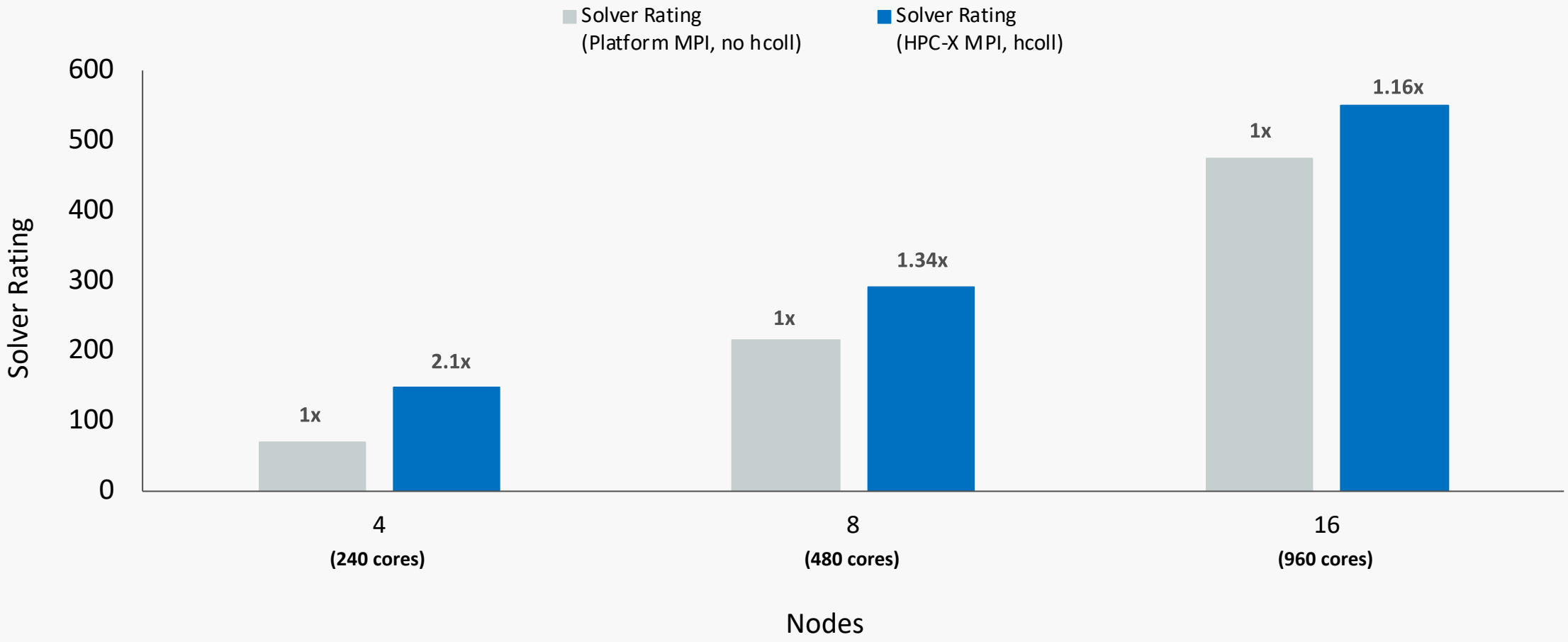
Nodes	Instance	Version	PPN	Cores	MPI	Solver Rating
2	HPE DL360	v19	32	64	HPCX	63.6
4	HPE DL360	v19	32	128	HPCX	128.79
8	HPE DL360	v19	32	256	HPCX	255.3
16	HPE DL360	v19	32	512	HPCX	509.4
32	HPE DL360	v19	32	1024	HPCX	1017.36

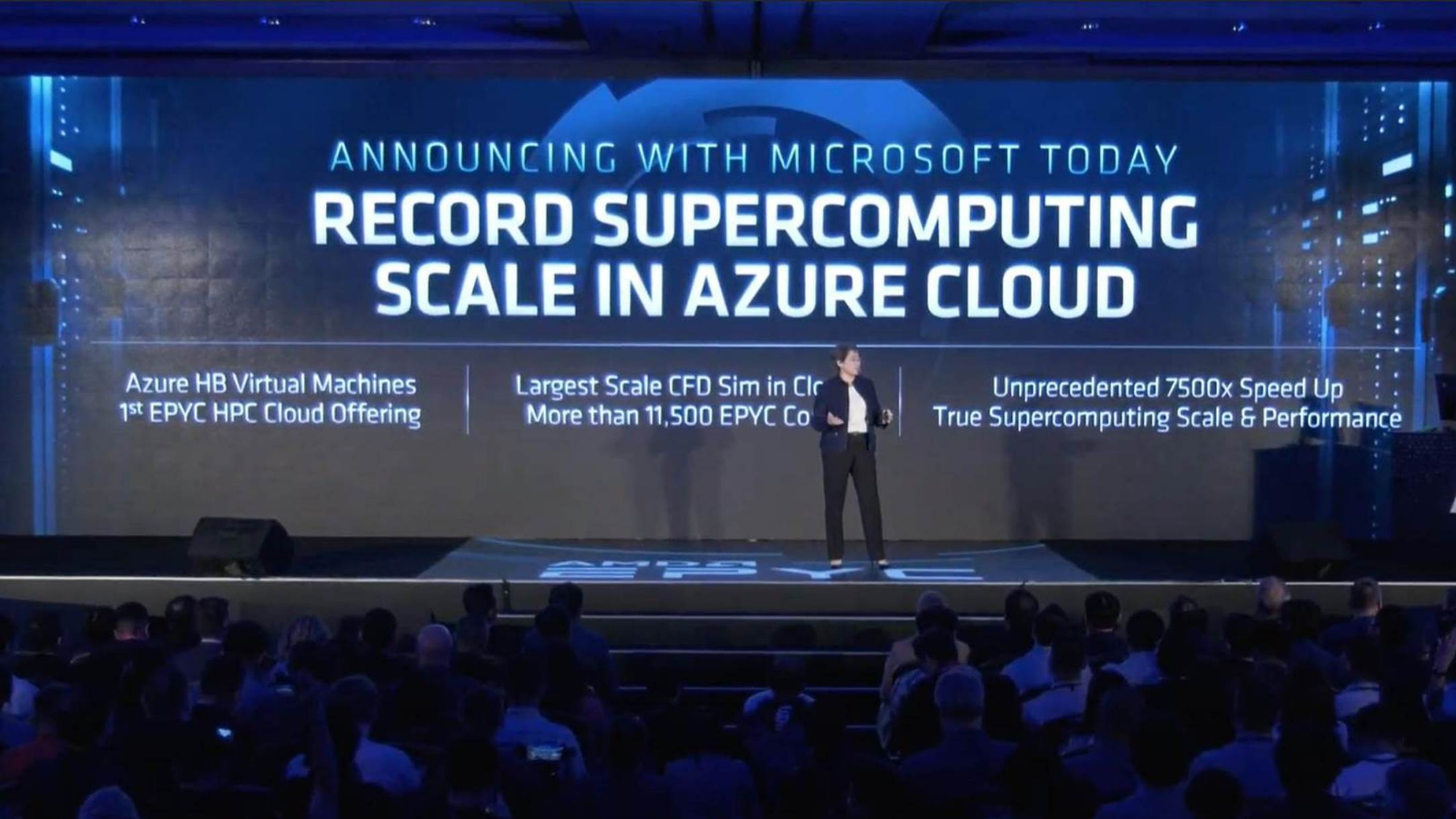
HB60rs uses Mellanox Connect-X5 EDR InfiniBand
HPE ProLiant DL360 Gen9 with Xeon E5 2697A v4 at 3.1 GHz with MellanoxConnectX-5 EDR InfiniBand
Details at: https://bit.ly/30wO7WX

Azure Experiments with Hardware Offload for MPI

HPC-X MPI on Mellanox ConnectX-5 showing up to 2.1x perf boost on Azure HB

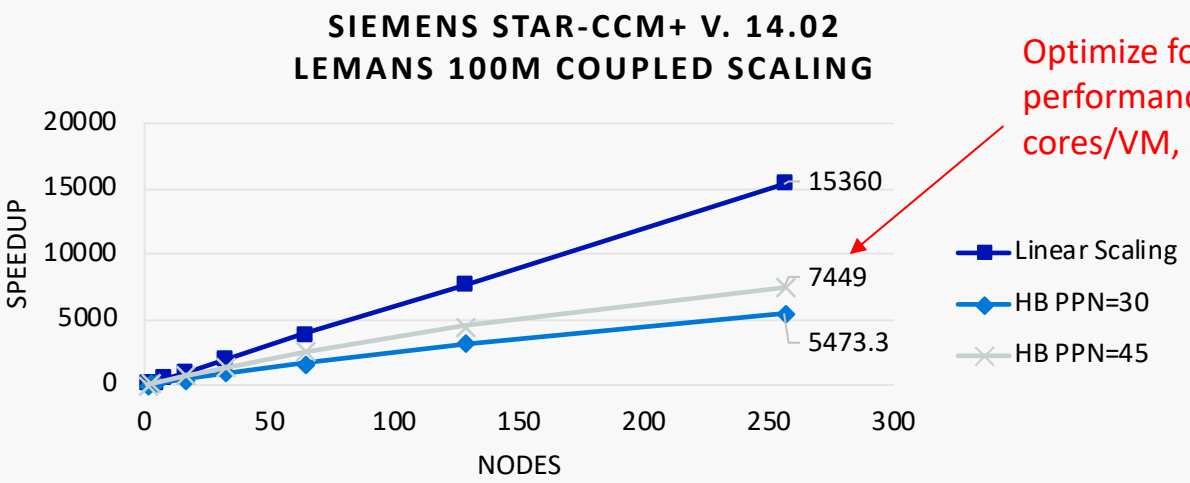
Azure HB with ANSYS Fluent ver. 18; F1 racecar 280m cells; Solver Rating; Higher = Better



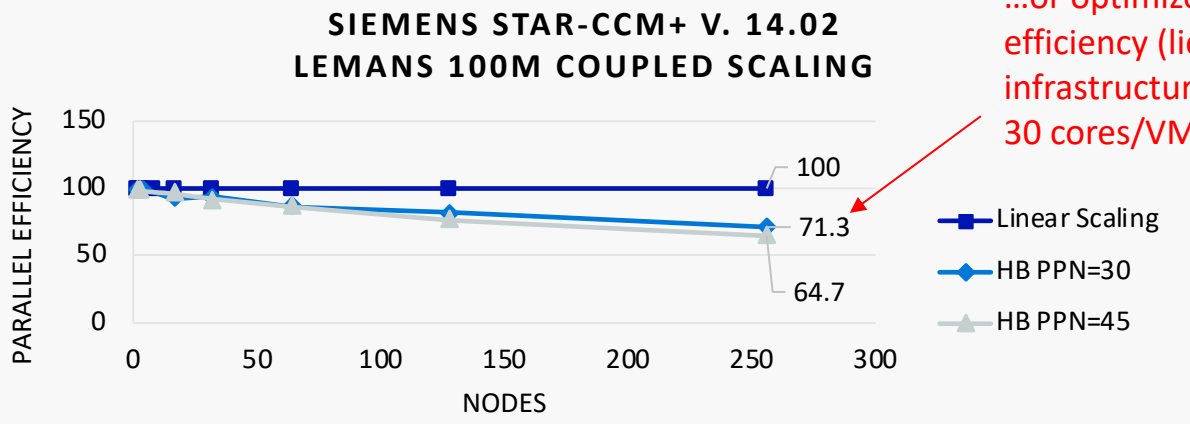

ANNOUNCING WITH MICROSOFT TODAY
RECORD SUPERCOMPUTING SCALE IN AZURE CLOUD
Azure HB Virtual Machines
1st EPYC HPC Cloud Offering
Largest Scale CFD Sim in Clo
More than 11,500 EPYC Co
Unprecedented 7500x Speed Up
True Supercomputing Scale & Performance

Siemens Star-CCM+ on Azure HB

Scaling beyond 11,500 cores on the Public Cloud



Optimize for raw performance with 45 cores/VM, or….



…or optimize for scaling efficiency (licensing, infrastructure, etc.) with 30 cores/VM

Hosts	Cores	PPN	Sample Elapsed Time	Speedup (Node based)	SpeedUp (Core based)	ParallelEff
1	30	30	632.5	1.00	30	100.0
2	60	30	314.2	2.00	59.9	99.9
4	120	30	158.4	3.97	119.2	99.3
16	480	30	42.58	14.89	446.6	93.0
32	960	30	21.14	29.86	895.7	93.3
64	1920	30	11.34	55.49	1664.8	86.7
128	3840	30	6.12	104.60	3138.1	81.7
256	7680	30	3.46	182.44	5473.3	71.3

Hosts	Cores	PPN	Sample Elapsed Time	Speedup (Node based)	SpeedUp (Core based)	ParallelEff
1	45.0	45	496.6	1.00	45	100.0
2	90.0	45	251.7	1.98	89	99.1
16	720.0	45	32.16	15.48	697	96.7
32	1440.0	45	17.0	29.28	1318	91.5
64	2880	45	9.13	55.67	2505	87.0
128	5760	45	5.2	98.92	4451	77.3
256	11520	45	3.1	165.53	7449	64.7

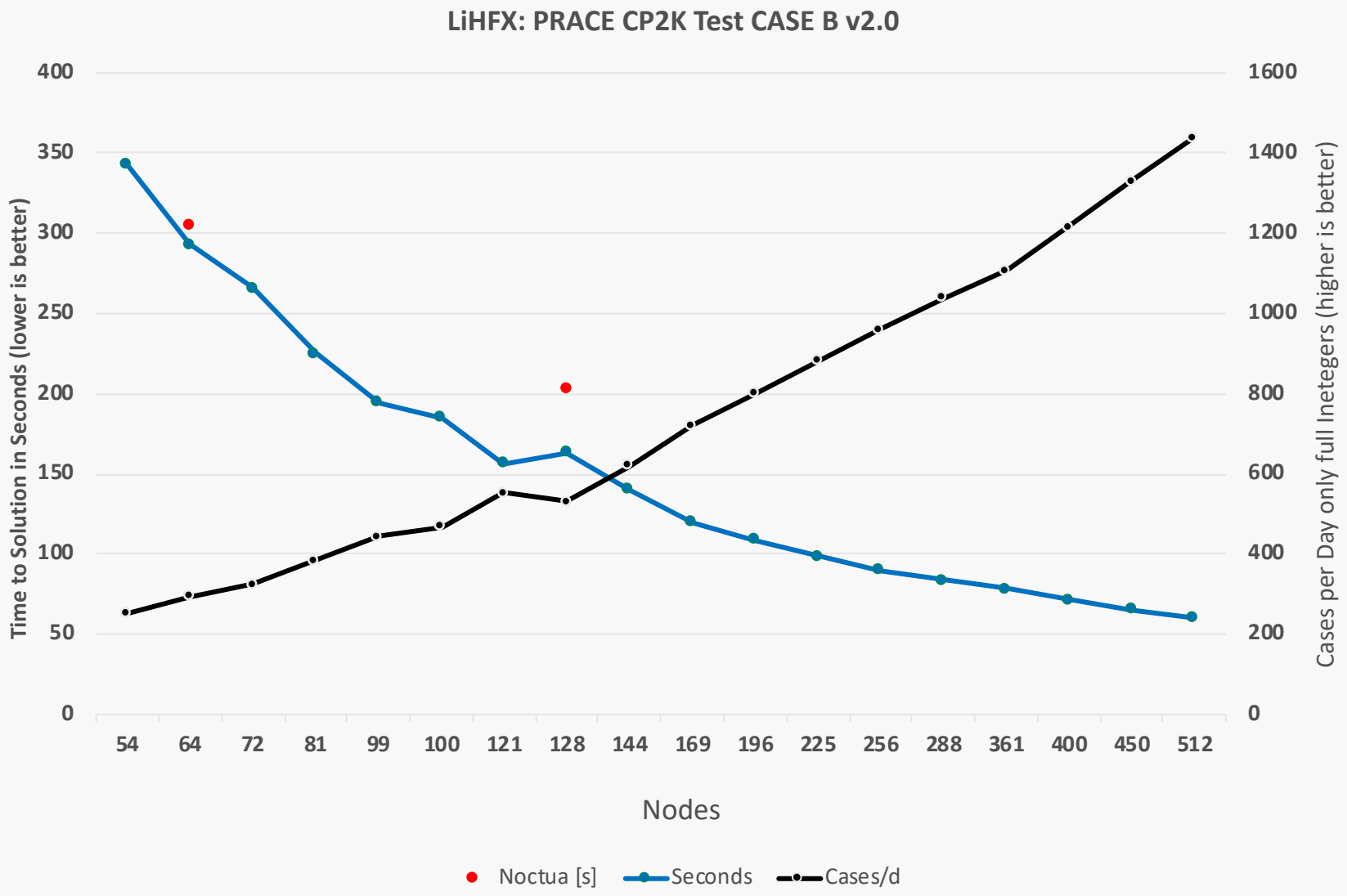
ANSYS Fluent on Azure HB-series

88% efficiency at 128 VMs, within 1% of XC50

83% Efficiency at 288 VMs (12,960 MPI ranks)

Very useful learnings about scalability of EPYC

CP2K on Azure HC-series



4-25% more performant at identical node counts v. Noctua (Xeon Gold 6148 w/ Omnipath)

Scaling to 512 VMs (22,528 MPI ranks)

Lessons So Far

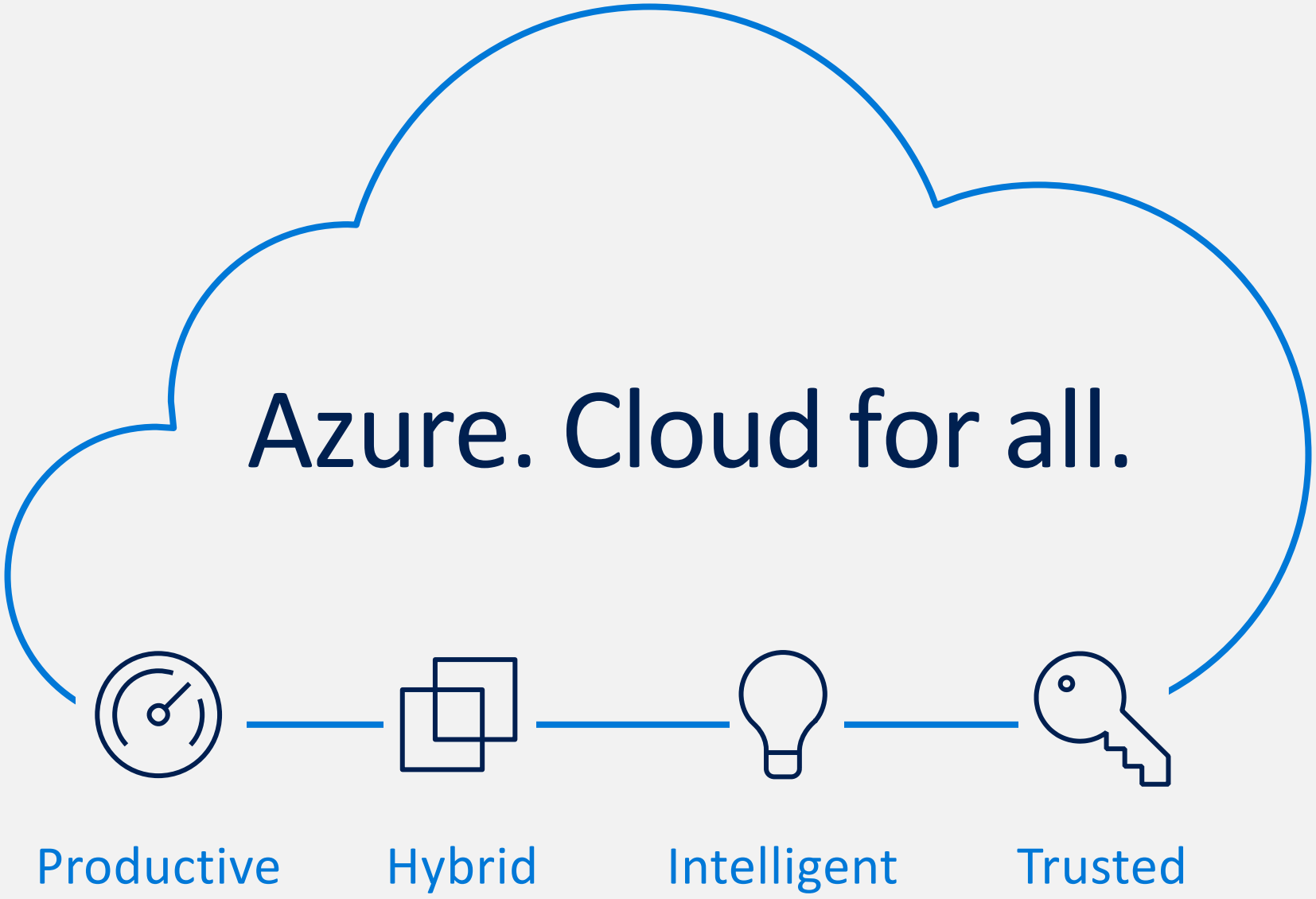
Azure VM-based HPC is becoming very capable and cost-effective for large (petascale) real-world, tightly-coupled problems (provided you use the HPC-optimized kit!)

10s of petaFLOPS for a single Azure-based CPU HPC environment is within reach (GPUs higher of course)

Things that are different: essential to get hypervisor config right

Things that are the same: memory bandwidth really matters for real-world apps, hardware offload for collectives helps, diverse workloads on the same fabric can impair performance

Bottom line: Azure is in the HPC space because it’s core to our mission to empower every individual and organization to achieve more


Azure. Cloud for all.
Productive
Hybrid
Intelligent
Trusted