Microsoft

PCCC22 (第22回PCクラスタシンポジウム)

Microsoft AzureのクラウドHPCを支える基盤技術

日本マイクロソフト株式会社
クラウドインフラアーキテクト本部
クラウドソリューションアーキテクト
五十木 秀一 (Shuichi Gojuki) (in shugo)
2022/12/05

自己紹介

五十木 秀一 SHUICHI GOJUKI (Ph.D in Physics)

所属：
日本マイクロソフト株式会社
エンタープライズアーキテクト統括本部 クラウドインフラアーキテクト本部
クラウド ソリューション アーキテクト
業種や業界は問わず、すべてのHPCなお客様へクラウドをご利用いただくための技術支援

特技：
プログラムチューニング、ベンチマーク、コンピュータアーキテクチャ
専門分野：
理論物理学、ハイパフォーマンスコンピューティング

Microsoft Azure

アジェンダ

Microsoft AzureのクラウドHPCを支える基盤技術

- 仮想マシン
  - 仮想マシンアップデート
  - InfiniBand
- ネットワーク
  - 仮想ネットワーク
  - 高速ネットワーク
  - 近接通信配置グループ
- ストレージ
- オーケストレーション
  - Azure CycleCloud

AzureにおけるHPCリソース

Compute

GPU/FPGA (N-Series)

HPC w/ **InfiniBand** (H-Series)

汎用/計算最適化 (D/F-Series)

ストレージ/メモリ 最適化 (M/L-Series)

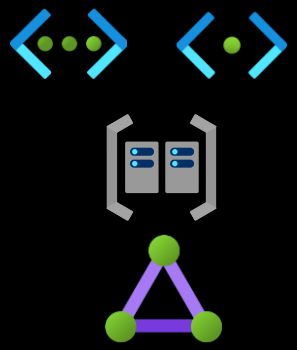
Storage

NFS (Azure NetAppFiles)

NFS Cache (Azure HPC Cache)

オブジェクト (Azure Blob)

Network



仮想ネットワーク/サブネット

高速ネットワーク

近接通信配置グループ

ExpressRoute (専用回線)

Security

Azure Firewall

Network Security Group

Azure Defender

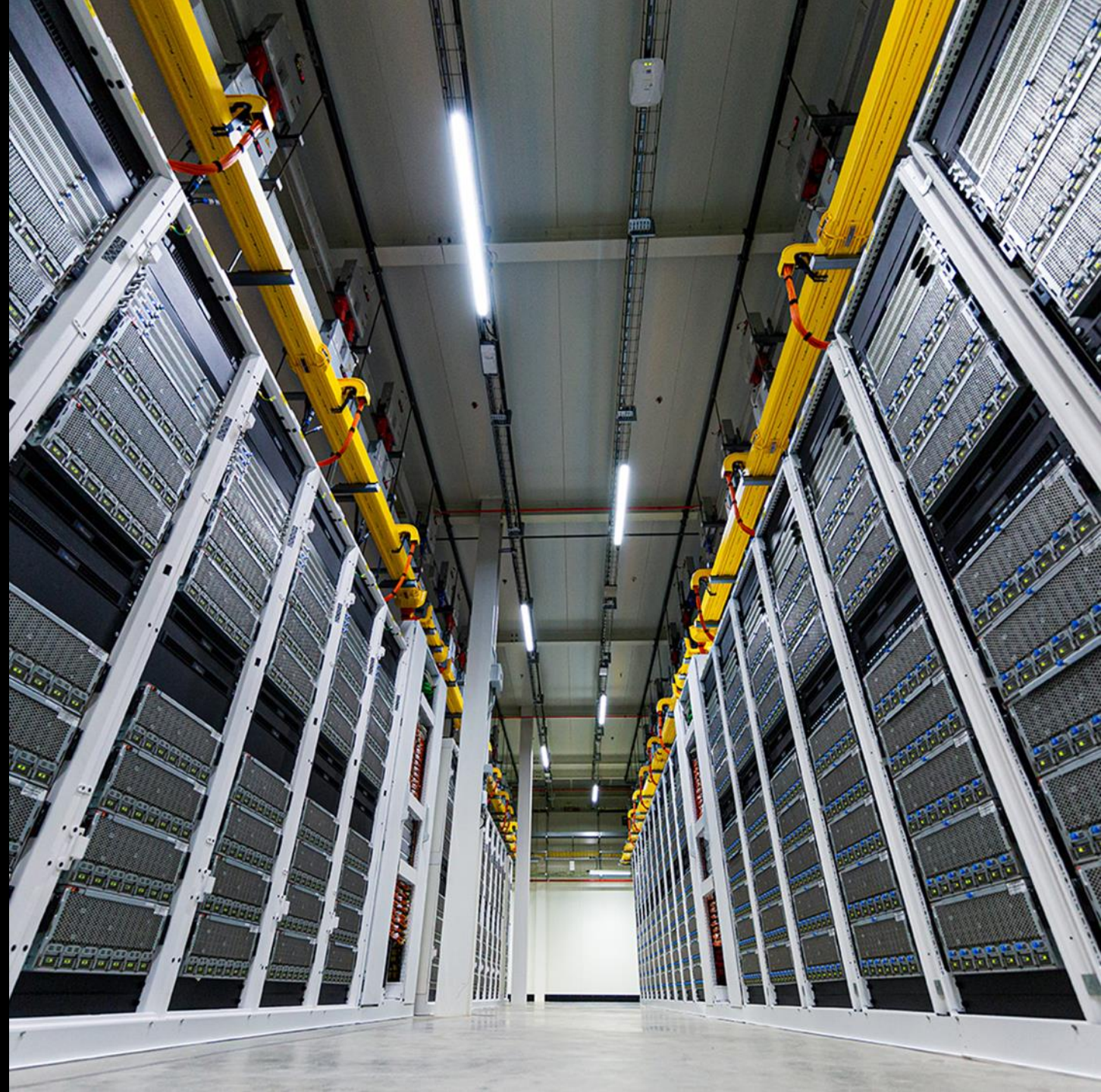
Workload Orchestration

VM Scale Sets

Azure CycleCloud

Azure Batch

仮想マシン

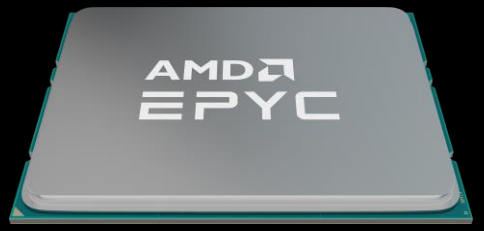


HBv4/HX (AMD 4th Gen EYPC Genoa)

preview

HBv4/HX シリーズ VM は、計算流体力学、有限要素解析、フロントエンドおよびバックエンドEDA、レンダリング、分子動力学、計算地球科学、気象シミュレーション、金融リスク分析などのさまざまなHPCワークロード向けに最適化されています。

- ✓ AMD 第4世代 EPYC™ 7004シリーズを搭載
- ✓ 最大176コア、マルチスレッドなし
- ✓ 688GB(HBv4)/1408GB(HX)メモリ搭載
- ✓ ノード当たりのDDR5メモリ帯域幅 800MB/sを提供
- ✓ 768MB L3キャッシュ
- ✓ 400Gbps NDR InfiniBand搭載
- ✓ 1.8TB NVMe x 2を搭載

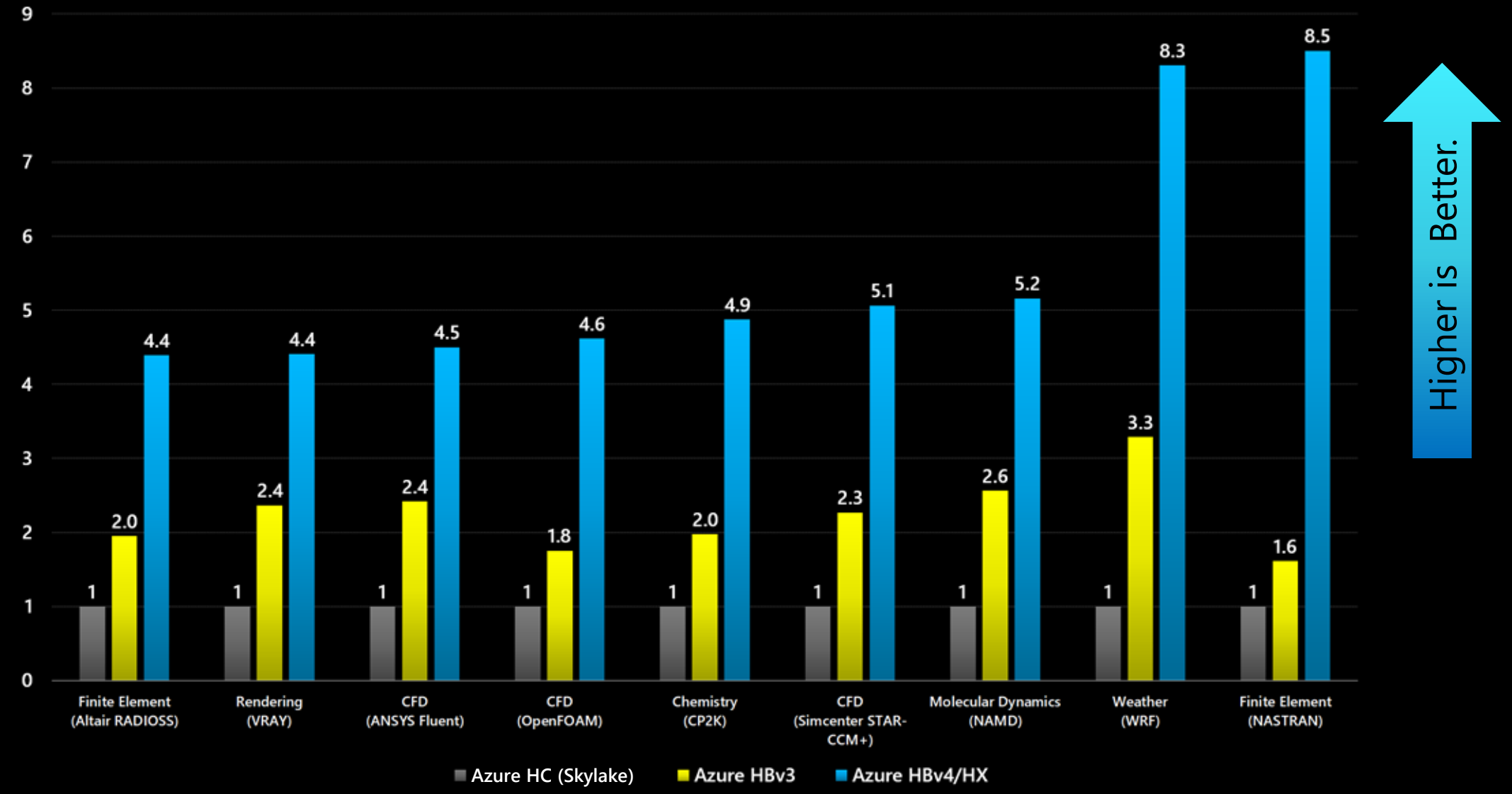


Microsoft Azure

	HBv4	HX
CPU周波数	AMD EPYC 7004-series Genoa 2.4 GHz (max. single core: 3.7 GHz)	
VMあたりのコア数	176(HB176rs_v4) 144(HB176-144rs_v4) 96(HB176-96rs_v4) 48(HB176-48rs_v4) 24(HB176-24rs_v4)	176(HX176rs) 144(HX176-144rs) 96(HX176-96rs) 48(HX176-48rs) 24(HX176-24rs)
メモリバンド幅	800 GB/s	
搭載メモリ量	3.9 GiB – 28.7GiB /core, 688GiB	8 GiB – 59GiB /core, 1408GiB
ローカルディスク	1.8 TB NVMe x 2	
InfiniBand	400 Gbps NDR InfiniBand	
接続ネットワーク	80 Gbps	

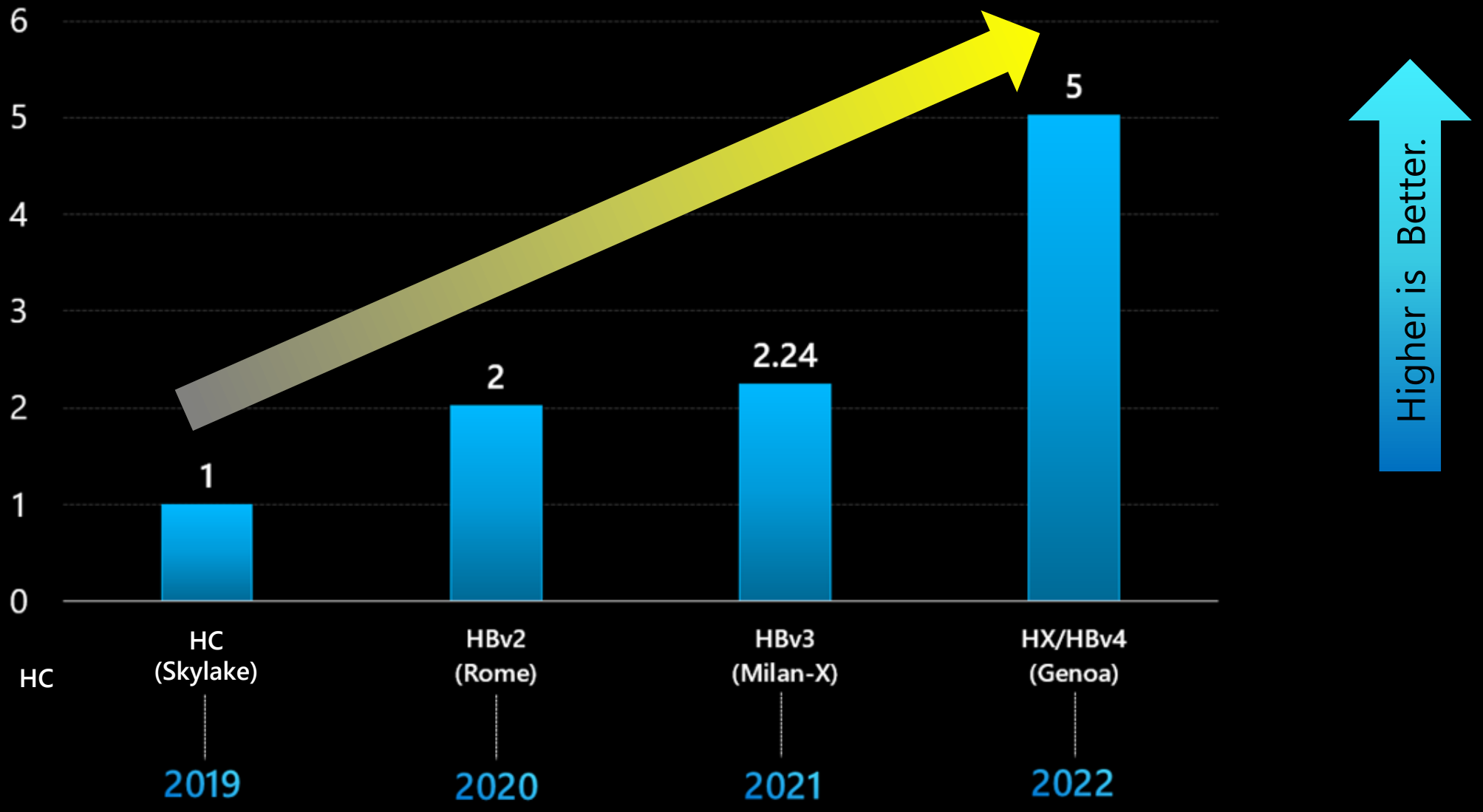
プレビュー期間はGenoaで利用、GA時はGenoa-Xにアップデート予定

Azure HBv4/HX-series v. HBv3-series v. HC (Skylake)



Constant Azure HPC Improvement

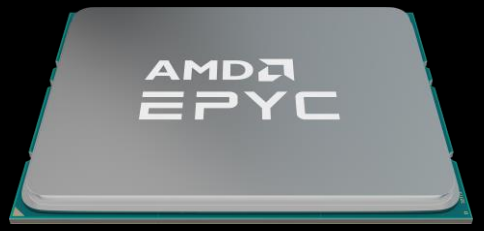
Siemens Simcenter STAR-CCM+ (Civil)



HBv3 (AMD 3rd Gen EYPC w/ 3D V-Cache Milan-X)

HBv3 シリーズ VM は、流体力学、明示的および暗黙的な有限要素分析、気象モデリング、地震処理、貯水池シミュレーション、RTL シミュレーションなど、HPC アプリケーションのために最適化されています。

- ✓ AMD 第三世代 EPYC™ 7V73Xを搭載
- ✓ あらゆるHPCワークロードに対応するためノード当たりのコア数を選択可能
- ✓ ノード当たりのメモリ帯域幅 350MB/sを提供
- ✓ 200Gbps HDR InfiniBand搭載
- ✓ 960GiB NVMe x 2を搭載



Microsoft Azure

	HBv3
CPU周波数	AMD EPYC 7V73X – Milan-X 1.9 GHz (max. single core: 3.5 GHz all cores turbo: 3.0 GHz)
VMあたりのコア数	120(HB120rs_v3) 96(HB120-96rs_v3) 64(HB120-64rs_v3) 32(HB120-32rs_v3) 16(HB120-16rs_v3)
メモリバンド幅	350 GB/s
搭載メモリ量	3.75 GiB - 28GiB /core, 448GiB
ローカルディスク	960 GiB NVMe x 2
InfiniBand	200 Gbps HDR InfiniBand
接続ネットワーク	50 Gbps (40 Gbps利用可能)

東日本リージョンでご利用いただけるようになりました

NC_A100_v4

NC_A100_v4 シリーズの仮想マシンは、NVIDIA A100 PCIe GPUと第3世代 AMD EPYC 7V13(Milan)プロセッサが搭載されています。このVMには、最大4個のNVIDIA A100 PCIe GPU(それぞれに80GBのメモリを装備)、最大96個の非マルチスレッドAMD EPYC Milanプロセッサコア、880GiBのシステムメモリが踏査入れています。これらのVMは、次のような実際のApplied AIワークロードに最適です。

- ✓ GPUで高速化された分析とデータベース
- ✓ 大量の前処理と後処理があるバッチ推論
- ✓ 自律運転の強化学習
- ✓ 石油とガスの貯留層シミュレーション
- ✓ 機械学習開発
- ✓ ビデオ処理
- ✓ AI/ML Webサービス

	NC24ads_A100_v4	NC48ads_A100_v4	NC96ads_A100_v4
CPU	AMD EPYC 7V13 – Milan (2.45GHz)		
コア数	24	48	96
GPU	NVIDIA A100 (80GB) x 1	NVIDIA A100 (80GB) x 2	NVIDIA A100 (80GB) x 4
メモリ容量	220GiB	440GiB	880GiB
ローカルディスク	1123 GiB NVMe	2246 GiB NVMe	4492 GiB NVMe
Network bandwidth (Mbps)	20,000	40,000	80,000

Microsoft Azure

東日本リージョンでご利用いただけるようになりました

NVadsA10 v5

NVadsA10v5 シリーズ仮想マシンには、NVIDIA A10 GPU と AMD EPYC 74F3V(Milan) CPU が搭載され、基本周波数は 3.2 GHz、全コアのピーク周波数は 4.0 GHz です。NVadsA10v5 シリーズでは、部分的な NVIDIA GPU を備えた仮想マシンを導入しています。24 GiB フレーム バッファーを備えた完全な A10 GPU に対して 4 GiB フレーム バッファーを備えた 6 分の 1 の GPU を下限として、GPU で強化されたグラフィックス アプリケーションと仮想デスクトップに対して適正なサイズの仮想マシンを選択します。

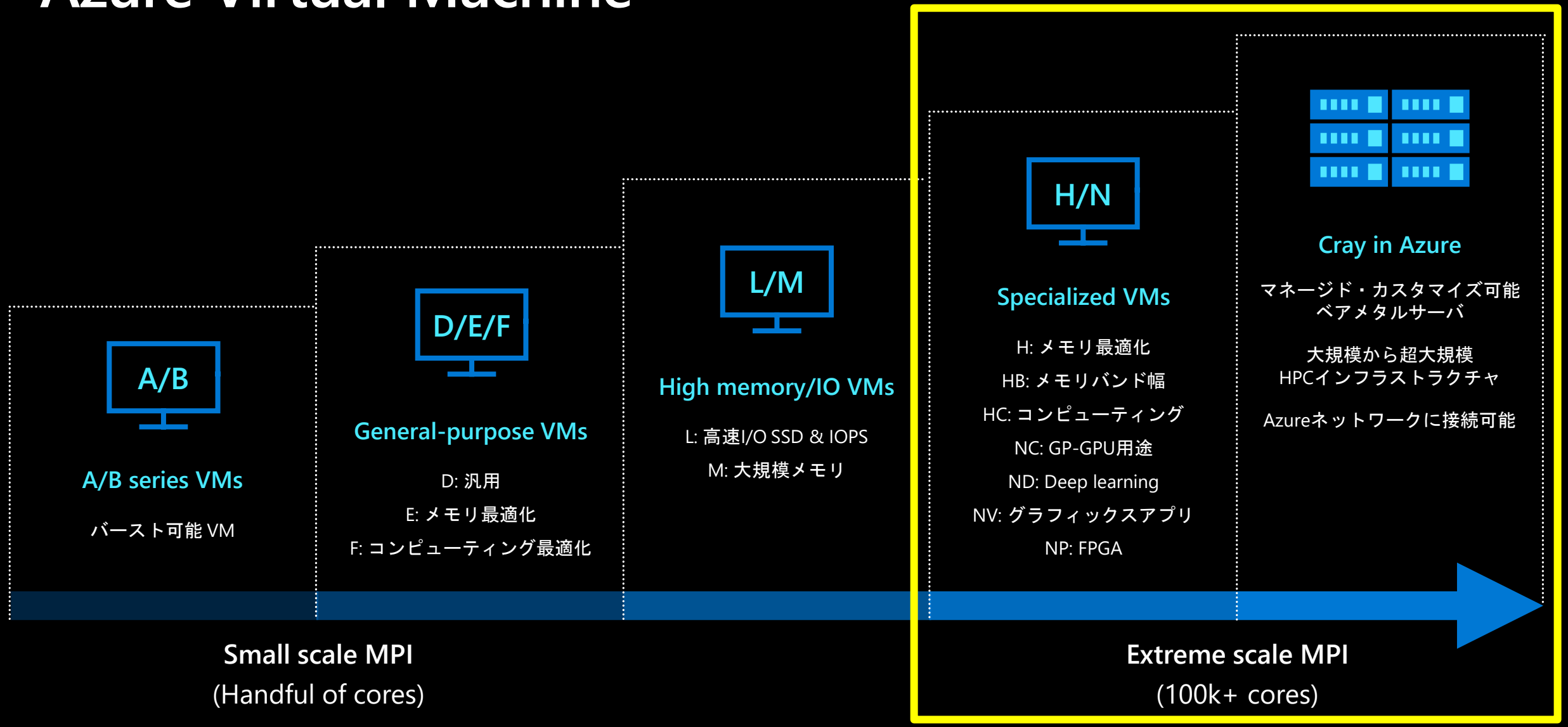
	NV6ads_A10_v5	NV12ads_A10_v5	NV18ads_A10_v5	NV36ads_A10_v5	NV36adms_A10_v5	NV72ads_A10_v5
CPU	AMD EPYC 74F3V – Milan (3.2GHz)					
コア数	6	12	18	36	36	72
GPU	NVIDIA A10 (4GB) x 1/6	NVIDIA A10 (8GB) x 1/3	NVIDIA A10 (12GB) x 1/2	NVIDIA A10 (24GB) x 1	NVIDIA A10 (24GB) x 1	NVIDIA A10 (24GB) x 2
メモリ容量	55GiB	110GiB	220GiB	440GiB	880GiB	880GiB
ローカルディスク	180 GiB SSD	360 GiB SSD	720 GiB SSD	720 GiB SSD	720 GiB SSD	1400 GiB SSD
Network bandwidth (Mbps)	5,000	10,000	20,000	40,000	80,000	80,000

Microsoft Azure

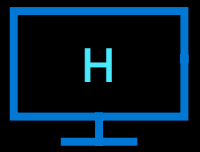
東日本リージョンでご利用いただけるようになりました

Azure Virtual Machine

InfiniBand

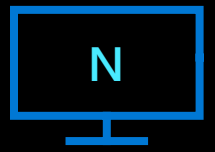


InfiniBand搭載 Virtual Machine



Hシリーズ (InfiniBand)

- H16r (FDR)
- HB60rs (EDR)
- HC44rs (EDR)
- HB120rs_v2 (HDR)
- HB120rs_v3 (HDR)
- HB176rs_v4 (NDR) *preview
- HX176rs (NDR) *preview



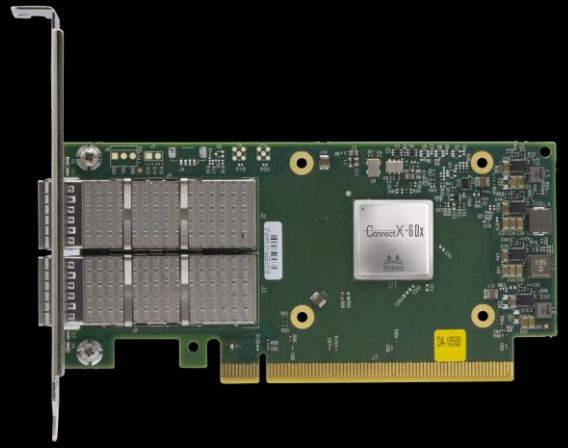
Nシリーズ (GPU + InfiniBand)

- NC24r (2x NVIDIA K80 + FDR)
- NC24rs_v2 (4 x NVIDIA P100 + FDR)
- NC24rs_v3 (4 x NVIDIA V100 + FDR)
- ND24rs (4 x NVIDIA P40 + FDR)
- ND40rs_v2 (8 x NVIDIA V100 + EDR)
- ND96rs_v4 (8 x NVIDIA A100 + 8 x HDR)

InfiniBand (Hardware)

Connect-X 5 Adapter

- Powering HB/HC/NDv2 series VMs
- EDR 100Gb/s InfiniBand
- Up to 200M messages/second



Connect-X 6 Adapter

- HBv2/HBv3/NDv4 VM
- HDR 200Gb/s InfiniBand
- Up to 215M messages/second
- PCI Gen4

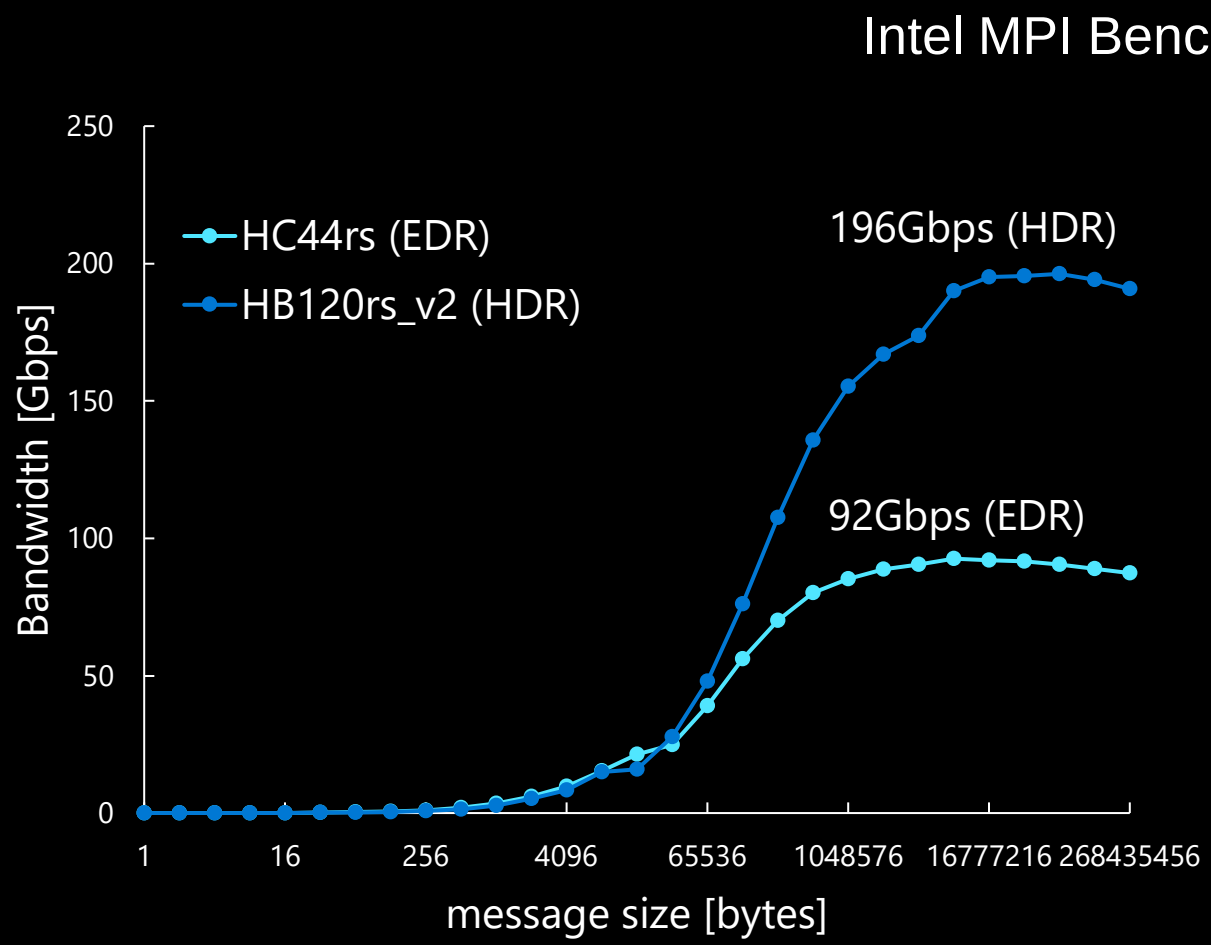
- 動的接続転送 (DCT)
  - Reliable and scalable transport
  - Lesser Memory footprint
- MPIコレクティブのオフロード (hcoll)
  - Collective offload framework
  - Asynchronous execution
  - Supports blocking/non-blocking collective
- UD multicast (MCAST)
  - Unreliable datagram (UD) based multicast
  - Create a mcast group and broadcast
- アダプティブルーティング

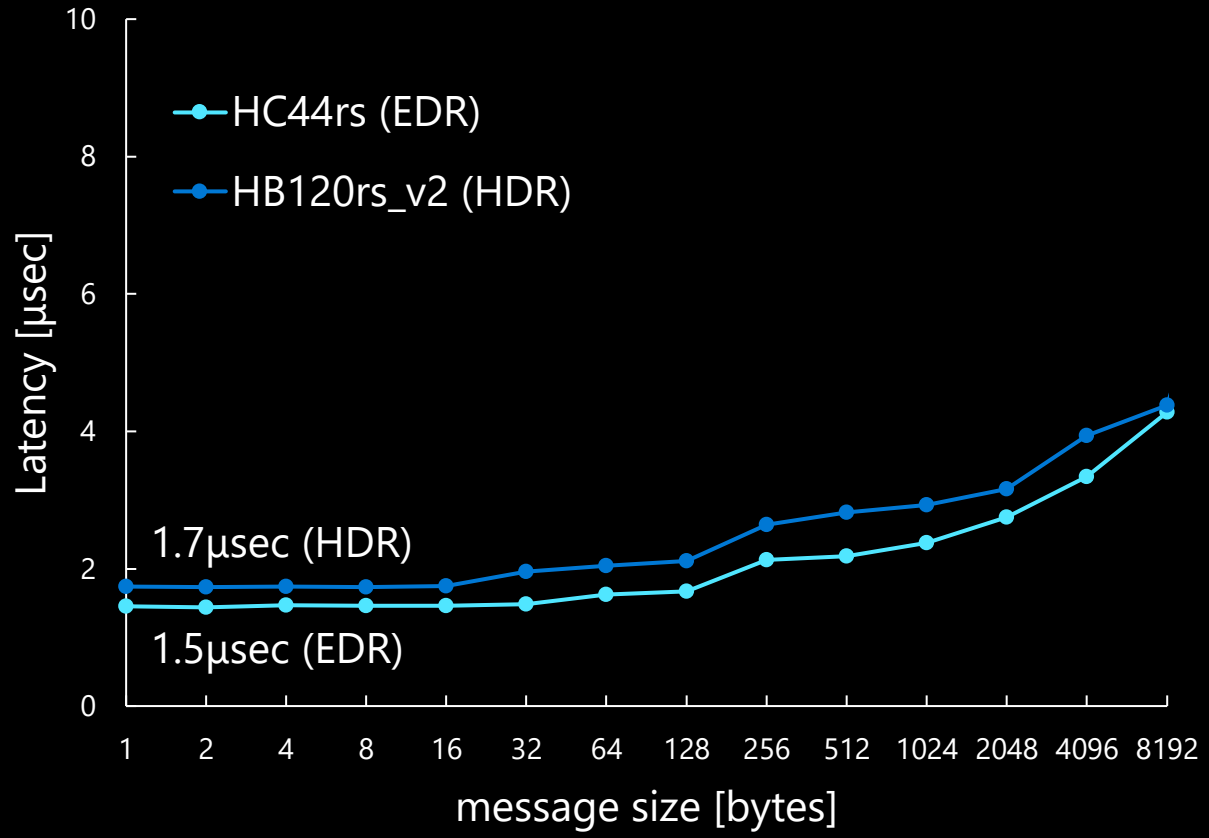
Adaptive Routing on Reliable Transport | Enhanced vSwitch / vRouter Offloads

InfiniBand (Software)

MPI Support	HPC-X, Intel MPI, OpenMPI, MVAPICH2, MPICH
Additional Frameworks	UCX, libfabric, PGAS
OS Support	CentOS 7.6以降, RHEL 7.6以降, Ubuntu 18.04以降, SLES 12 SP4以降, WindowsServer 2016以降 (CentOS 8.1、Windows Server 2019以降 推奨)
Orchestrator Support	Azure CycleCloud, Azure Batch, Azure Kubernetes Service, Virual Machine Scale Sets, Microsoft HPC Pack

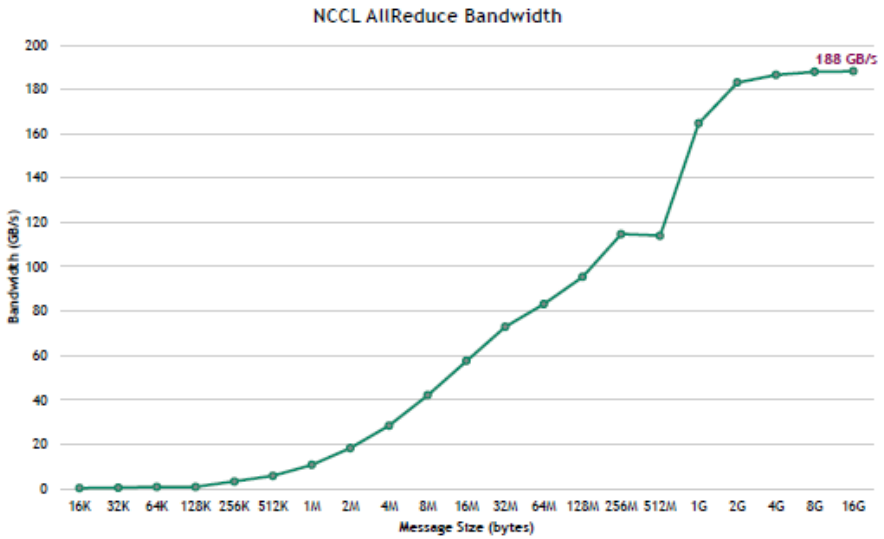
InfiniBand Performance (MPI Bandwidth/Latency)

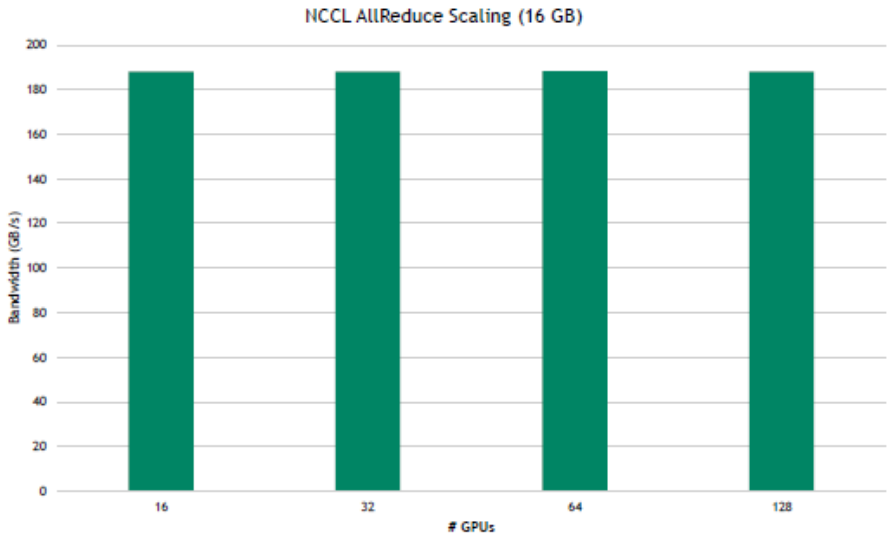




InfiniBand Performance (NCCL Performance)

NCCL Performance Results - NDv4





- **NCCL Collective Benchmarks**
  - NCCL 2.8.3
  - 16 x NDv4 VM Instances
- Theoretical Peak = 200 GB/s (w/o SHARP)
- Consistent bandwidth of ~188 GB/s up to 16 VMs

GTC, April 2021, "Microsoft Azure InfiniBand HPC Cloud User Experience and Best Practices"

ネットワーク

仮想マシンのネットワーク帯域幅

Azure の VM には多様なサイズと種類があり、パフォーマンス機能の組み合わせもそれぞれ異なっています。機能の 1 つがネットワーク スループット (帯域幅) で、メガビット/秒 (Mbps) で測定されます。 仮想マシンは共有ハードウェアでホストされているため、同じハードウェアを共有する仮想マシン間でネットワーク容量が公平に分配される必要があります。 大きな仮想マシンには、小さい仮想マシンよりも相対的に多くの帯域幅が割り当てられます。

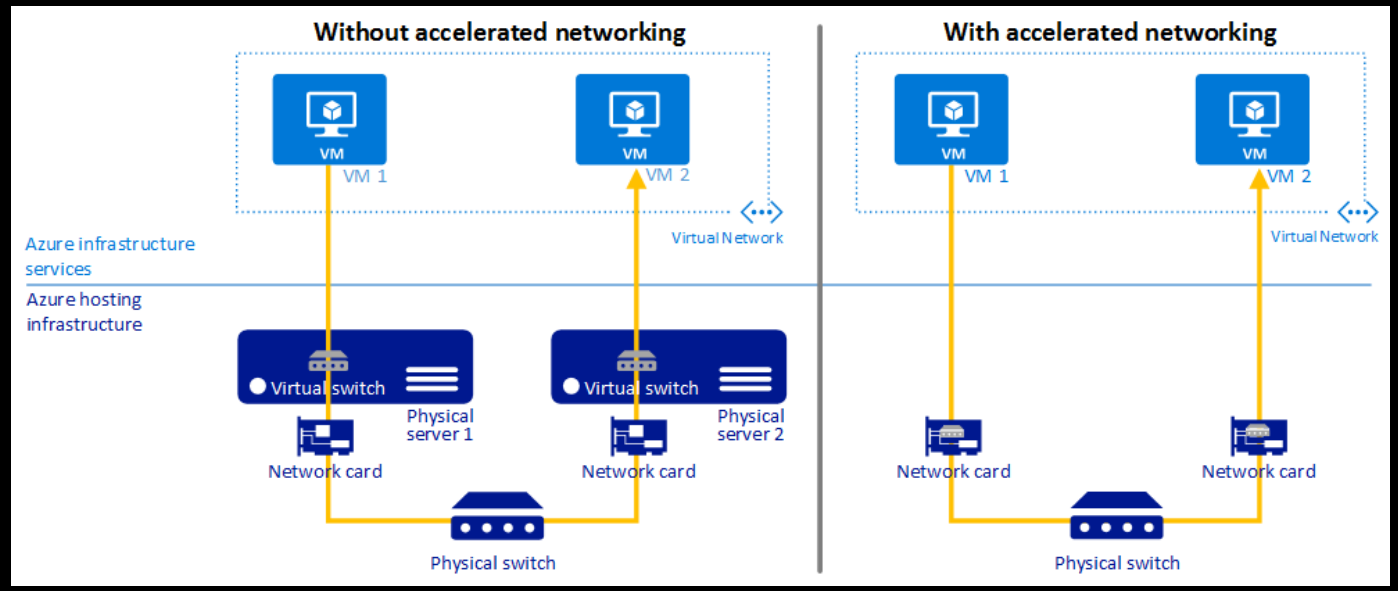
Ds_v4シリーズの例：

サイズ	vCPU	メモリ:GiB	最大 NIC 数	必要なネットワーク帯域幅 (Mbps)
Standard_D2s_v4	2	8	2	5000
Standard_D4s_v4	4	16	2	10000
Standard_D8s_v4	8	32	4	12500
Standard_D16s_v4	16	64	8	12500
Standard_D32s_v4	32	128	8	16000
Standard_D48s_v4	48	192	8	24000
Standard_D64s_v4	64	256	8	30000

高速ネットワーク (Accelerated Networking)

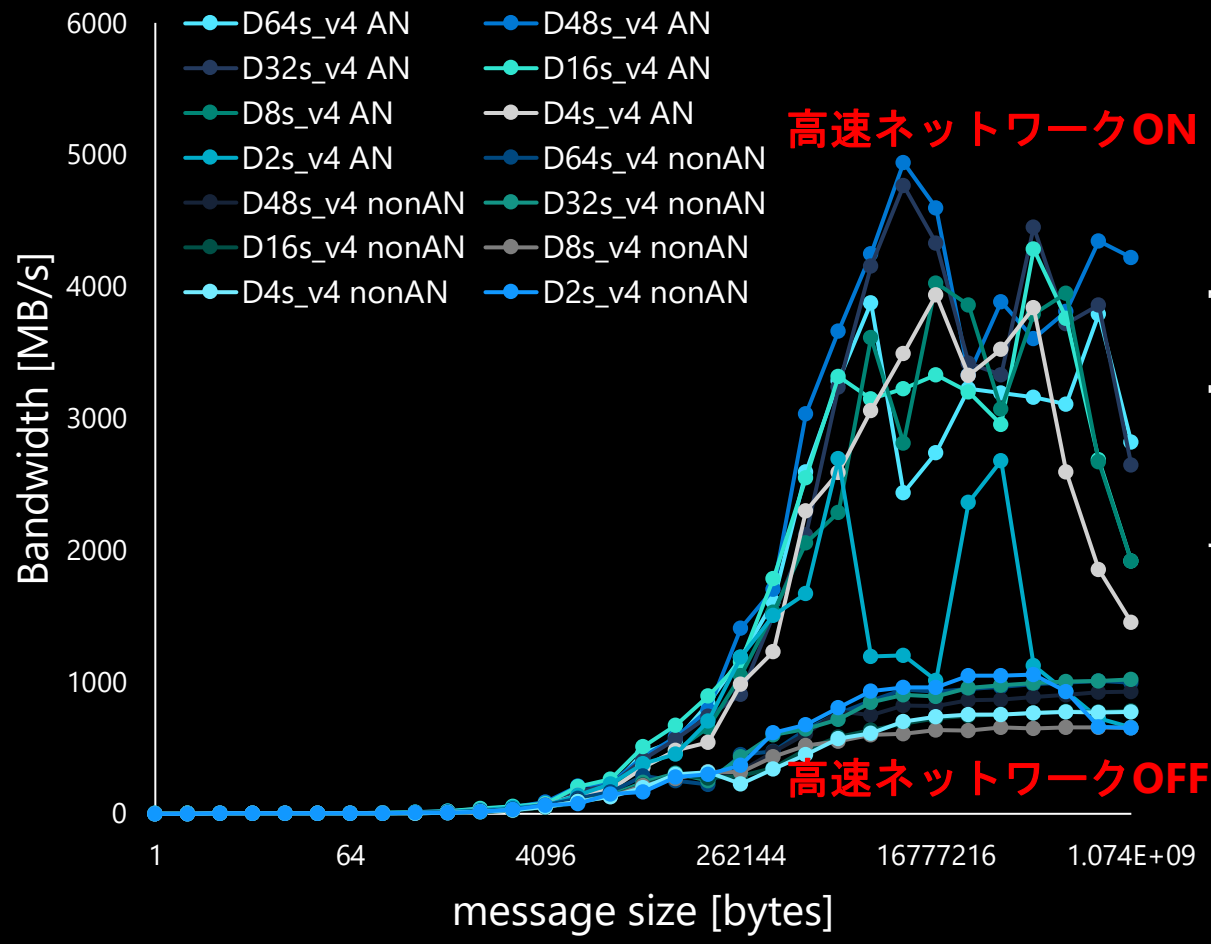
- 高速ネットワークを使用しない：
  - VMに出入りするすべてのネットワークがホストと仮想スイッチをスキャンする
- 高速ネットワークを使用する：
  - ネットワークトラフィックはVMのNICに到達した後、直接VMに転送される

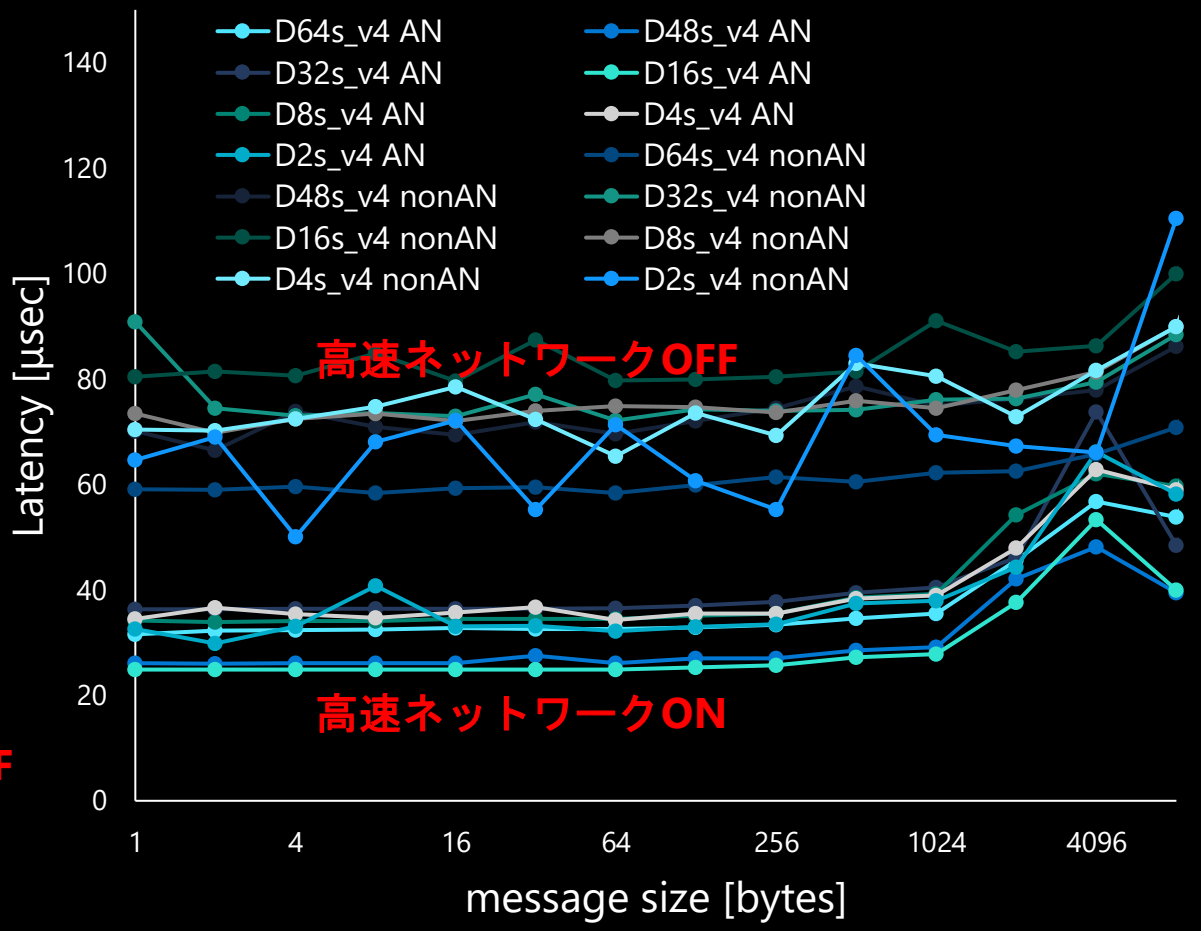
- **待ち時間の短縮/1秒当たりのパケット数(pps)の向上**
  データ パスから仮想スイッチがなくなるので、パケットからホストでポリシー処理に消費される時間がなくなります。 また、VM 内で処理できるパケットの数も増えます。
- **ジッターの削減**
  仮想スイッチの処理は、適用する必要があるポリシーの量によって異なります。 また、処理を行っている CPU のワークロードにも依存します。 ポリシーの適用をハードウェアにオフロードすると、パケットが VM に直接配信されるので、その変動が解消されます。 オフロードを行うと、ホストから VM への通信、すべてのソフトウェア割り込み、すべてのコンテキスト切り替えもなくなります。
- **CPU使用率の削減**
  ホストの仮想スイッチをバイパスすることによって、ネットワーク トラフィックを処理するための CPU 使用率を削減できます。



高速ネットワーク (Accelerated Networking: AN)

Ds_v4シリーズの例：



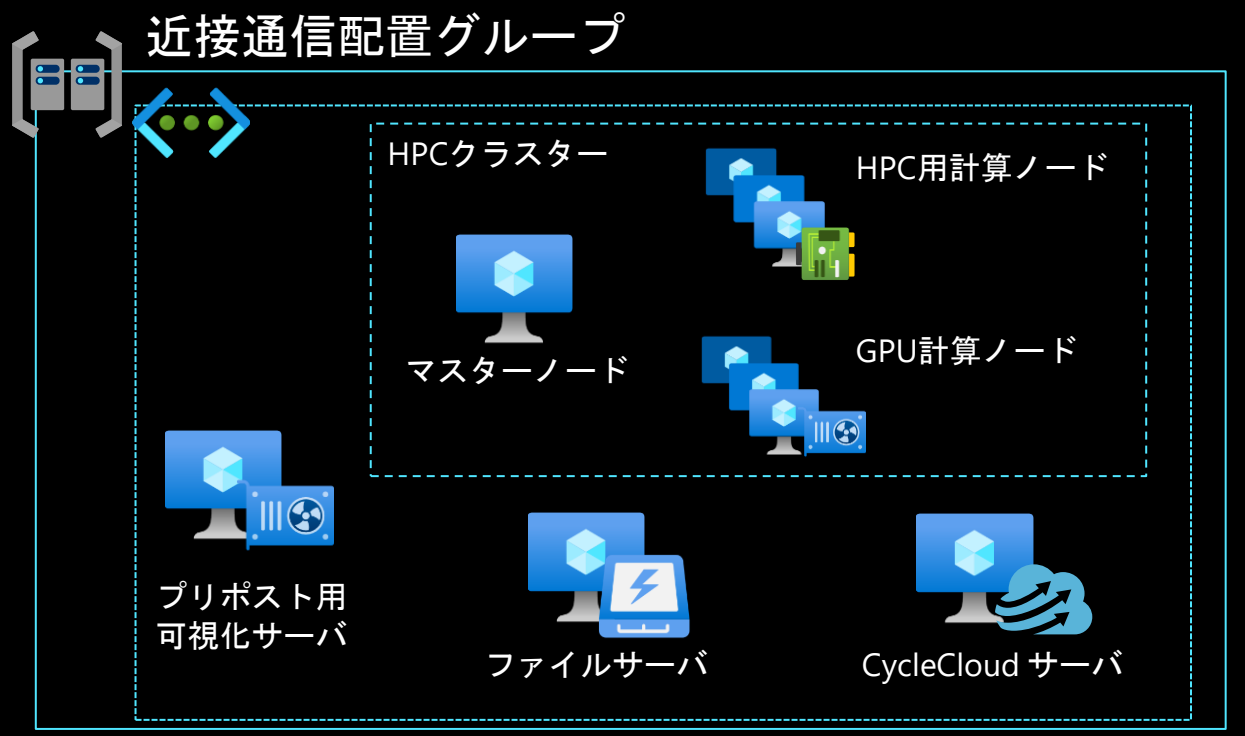


Intel MPI Benchmark pingpong: 近接通信配置部ループ設定, 高速ネットワーク有効(AN), 高速ネットワーク無効(nonAN)

近接通信配置グループ (Proximity Placement Group)

近接通信配置グループは、Azure コンピューティング リソースが互いに物理的に近くに配置されるようにするために使用される論理的なグループ化です。 近接通信配置グループは、短い待ち時間が要件であるワークロードに役立ちます。

- スタンドアロン VM 間の短い待ち時間。
- 1 つの可用性セットまたは仮想マシン スケール セット内の VM 間の短い待ち時間。
- スタンドアロン VM、複数の可用性セット内の VM、または複数のスケール セット内の VM の間の短い待ち時間。 1 つの配置グループ内に複数のコンピューティング リソースを配置して多層アプリケーションを構成できます。
- 異なるハードウェアの種類を使用した複数のアプリケーション層間の短い待ち時間。 たとえば、1 つの近接通信配置グループ内での、可用性セット内の M シリーズを使用したバックエンドとスケール セット内の D シリーズ インスタンス上のフロント エンドの実行。



ファイルサーバとHPCクラスター、可視化サーバなどはファイルI/Oを必要とするため物理的に近くに配置される必要があります。
同一の近接通信配置グループ内に設定し、高速ネットワークと一緒に設定することによりレイテンシを軽減することができます。

ストレージ

Azure Storage

ローカルストレージ	Local Disk：SSD, NVMe
	Managed Disk : Premium SSD, Ultra SSD
Azure Blob ストレージ	Blob
	Azure Data Lake Storage Gen2
NAS ソリューション	NFS Server (IaaS)
	Azure NetApp Files
	Azure HPC Cache
分散ファイルシステム	Lustre, BeeGFS
	Cray Clusterstor











利用に応じて最適なストレージを選択

START
Which do I use?

- Premium SSD
- Azure NetApp Files
- Ultra SSD
- Azure HPC Cache
- Blob
- Lustre/Bee GFS
- ADLS Gen2
- Cray Clusterstor
- Roll-your-own NFS Server

Shared or Local/Stageable?

Local/Stageable → **Azure Local Storage**

Low latency, dataset is small/portable and can be recopied across VM farm if desired

Persistent or Scratch?

Scratch → Ephemeral Disks

Persistent → Latency or Cost-sensitive?

Cost-Sensitive → SSD

Latency-sensitive → SSD

Shared → Object/HDFS or File?

Object → **Azure Blob Storage**

Unstructured Data at Exabyte Scale.

Blob (High Throughput/Premium)

Latency is not a factor, not File-dependent, Large Files, app can be re-written

ADLS Gen2

Large-scale analytics, Hadoop-compatible, ABFS driver acceptable

File/NAS → NFS or Parallel FS?

NFS → **File/NAS: Small/Medium Scale HPC Environment**

NAS exists (NFS) in current environment, simplicity trumps performance requirement

Cheap, scratch, small # of clients

Write-many (unique); write once, read once; home dirs; enterprise apps

Read-heavy portion of dataset; write some, read many

Read-heavy (75%+); write once, read many; hide latency

Write-heavy — Balanced — Read-heavy

Parallel FS → **File/NAS: High Scale HPC Environment**

Current environment is using Parallel Filesystem, heavy concurrent file access required

Managed or Orchestrated?

Orchestrated → Lustre / BeeGFS

Managed → Cray

MB/sec –> Low Single-digit GB/sec throughput

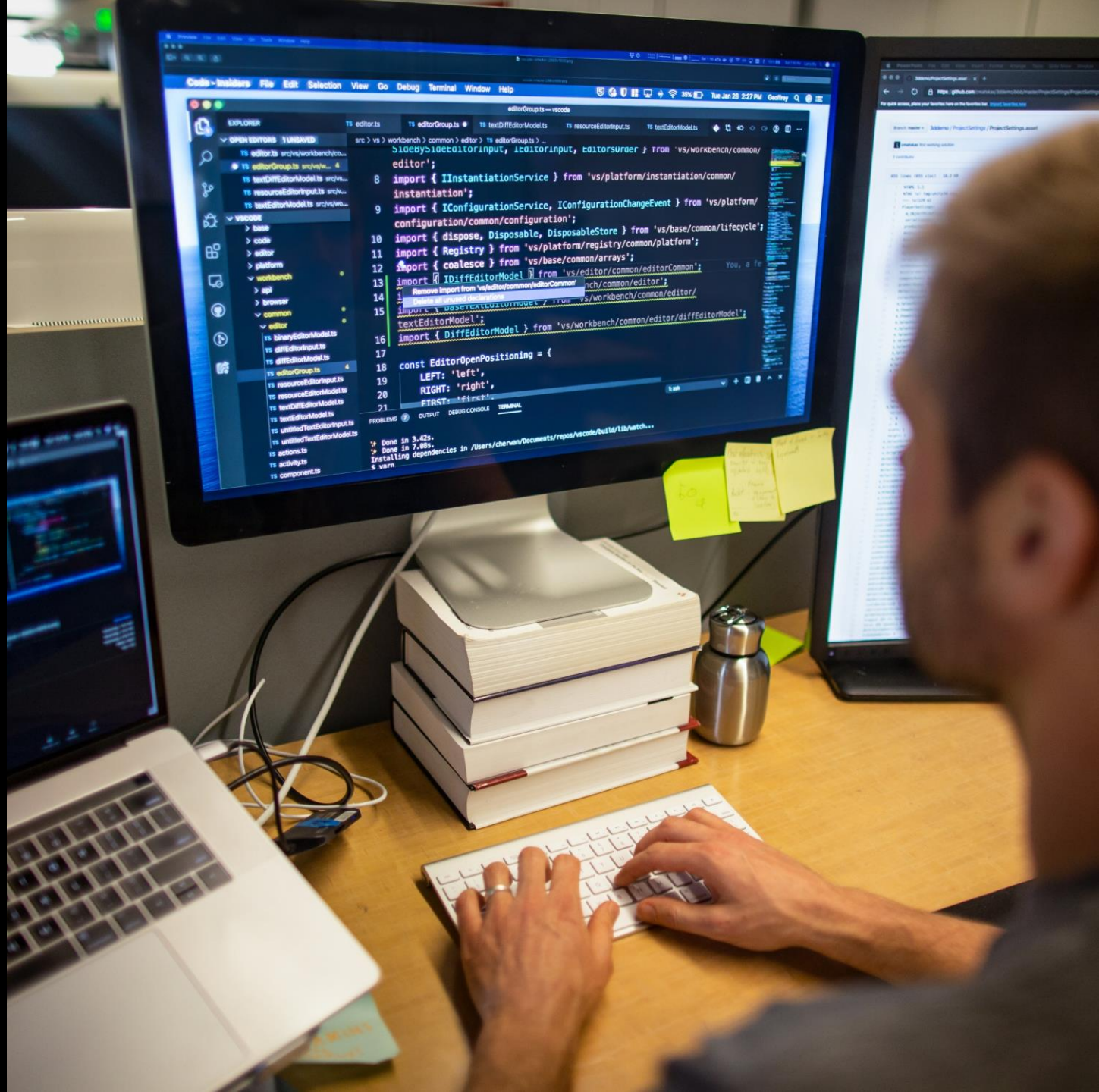
Single-digit GB/sec –> Double-digit GB/sec throughput

Double-digit GB/sec throughput

Baremetal HPC

Increased performance, complexity, parallel IO requirements

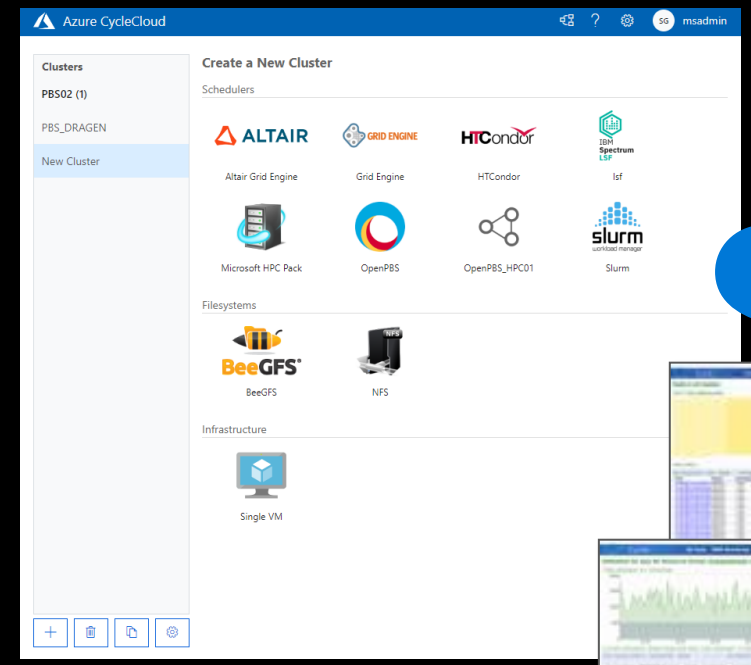
オーケストレーション



Azure CYCLE CLOUD

あらゆる規模のHPCクラスターやビッグコンピューティング クラスターを作成、管理、運用、最適化

- ✓ HPCクラスターを簡単に作成および管理
  - CycleCloudのウェブアプリケーションからHPCクラスターの構築や管理が可能
- ✓ あらゆるジョブ スケジューラまたはソフトウェアスタックを使用
  - Slurm, Grid Engine, HPC Pack, HTCondor, LSF, PBS Pro, Symphonyなど様々なジョブスケジューラーを選択可能
  - HPCアプリケーションやディープラーニングのフレームワークをインストール済みのイメージーを利用可能
  - カスタムイメージでアプリケーションやジョブスケジューラのカスタマイズ可能
- ✓ HPCクラスターを任意のサイズに自動スケーリング
  - リソースの受領に合わせて、スケジューラ対応の自動スケーリング機能
- ✓ クラスターを制御および監視
  - Active DirectoryやLDAPサーバと統合しロールベースのアクセス制御を提供
  - コストの通知および制御
  - パフォーマンスの監視
- ✓ クラスターをカスタマイズ
  - テンプレートにより、目的に合わせたクラスタ構成やアプリケーションをカスタマイズ可能
- ✓ ハイブリッドHPCを実現
  - Avere, Microsoft HPC Pack, 組み込みのデータ転送ツールのサポートにより、バーストおよびオンプレミス環境とのハイブリッドHPC環境を実現



Azure CycleCloud

ユーザ エンパワーメント

- 既存のワークロードとスケジューラをクラウド化可能
- リソースへ即時アクセスが可能
- オートスケーリング機能、エラー処理

IT マネージメント

- 内部および外部クラウドのワークフローを連携
- 認証・認可にActive Directoryの利用可能
- 安全で一貫性のあるアクセスを提供

ビジネス マネジメント

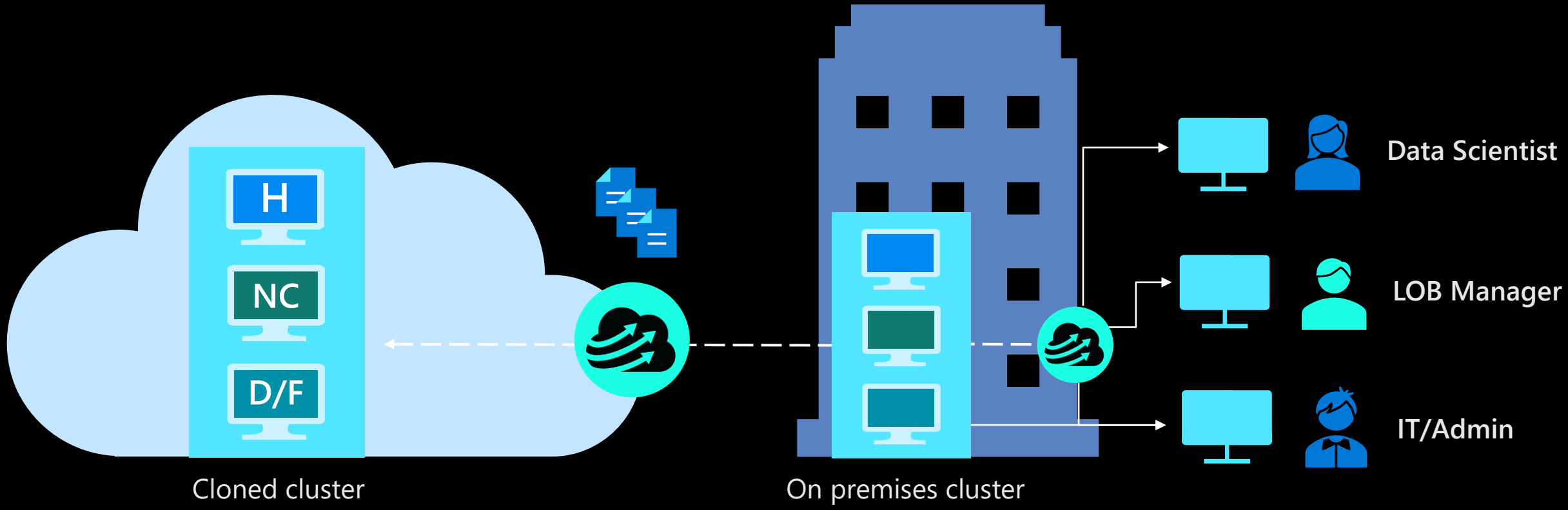
- コスト管理（使用量とコスト）
- コストを管理・制御するためのツールの提供

Azure CycleCloud シナリオ

戦略的なHPCアプリケーション環境をAzureに移行する際のギャップを埋める

オンプレミス環境とミラーリングするためのテンプレートを使って、CycleCloudはアプリケーションを書き換えることなくロードできる同一の環境を提供します。

CycleCloudは、アクセス、認証、コスト管理、コンプライアンス監査報告のためのポリシーベースのシステムを提供することで、ガバナンスの問題を解決します。



まとめ

Microsoft Azure

Microsoft AzureのクラウドHPCを支える基盤技術

- ✓仮想マシン
  - ✓仮想マシンアップデート
  - ✓ **InfiniBand**
- ✓ネットワーク
  - ✓仮想ネットワーク
  - ✓ **高速ネットワーク**
  - ✓ **近接通信配置グループ**
- ✓ストレージ
- ✓オーケストレーション
  - ✓ **Azure CycleCloud**

Microsoft Azureで
クラウドでも本物のHPC環境を!

お問い合わせはこちら

日本マイクロソフト文教営業本部
HPCお問い合わせ窓口
MSKK_HPCinquiry4Edu@microsoft.com

Microsoft