



PCクラスターワークショップ in 神戸 2022

新しいクラウドHPCと Microsoft Azureの最新情報のご紹介

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](#))
2022/06/23



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



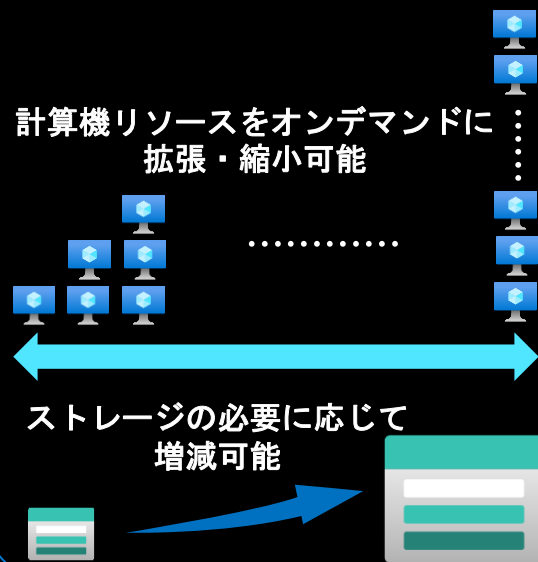
クラウドならではの HPCシナリオ



クラウドのメリット

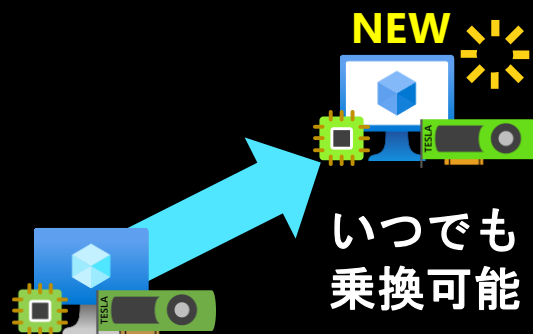
オンデマンド

必要な時に、必要なリソースを選択し、必要な量のリソースを利用することができます。また、不要になったら、停止することでコスト削減が可能です。



乗換可能

新しいCPU, GPU, ネットワークインターコネクトがリリースされると、Azureでも最新のCPU, GPU, ネットワークインターコネクトが搭載された仮想マシンが利用可能になります。いつでも最新のリソースに乗換可能なので、常に最新の計算機環境をご利用いただけます。



従量課金と予約

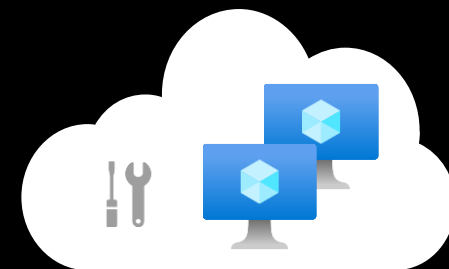
利用に応じて課金される従量課金方式と固定した利用見込みで予約するリザーブドインスタンスで、利用シナリオに応じて従量課金と予約インスタンスを組み合わせてコスト最適化できます。

固定利用を超える部分は従量課金で利用分のみコスト



メンテナンス不要

クラウドではハードウェアメンテナンスはクラウド事業者が行うため、ユーザによるハードウェアメンテナンスは不要です。万が一、ハードウェアの障害が発生しても正常なリソースに移行して運用を継続することができます。



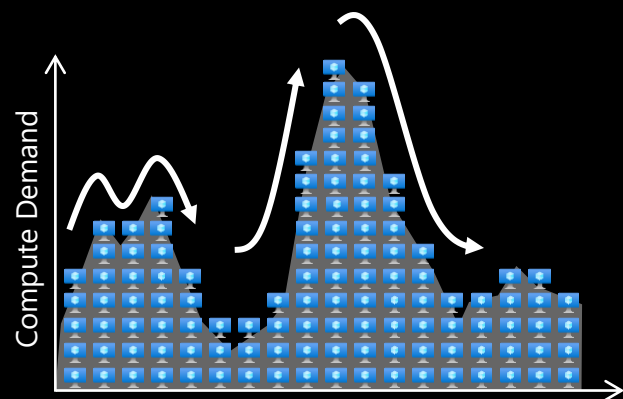
2つのクラウド利用モデル

ハイブリッドモデル：オンプレミスと連携

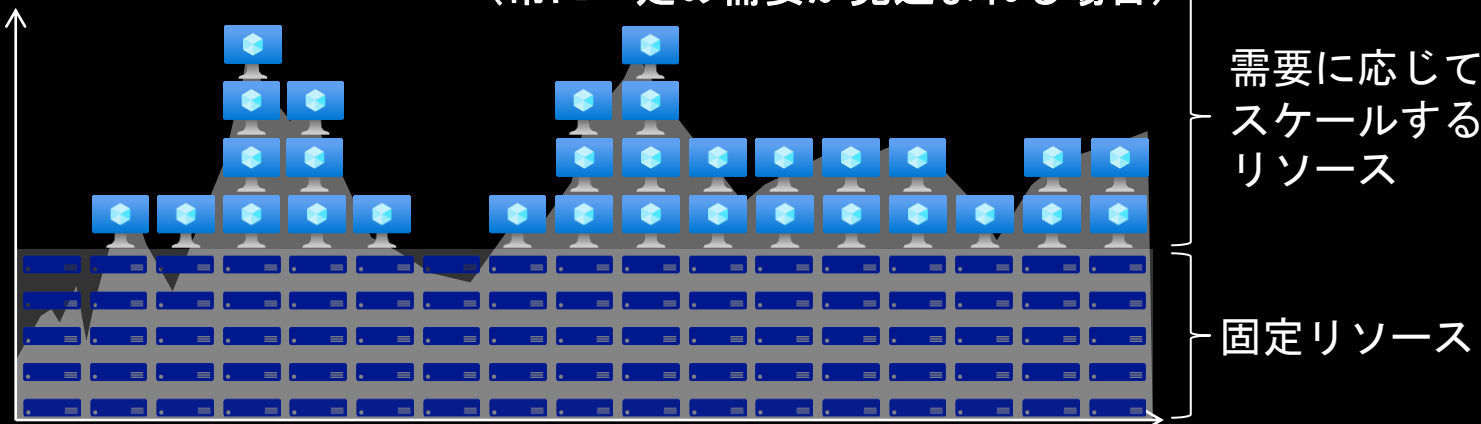
フルクラウドモデル：固定リソースもクラウド

(常に一定の需要が見込まれる場合)

固定リソースあり



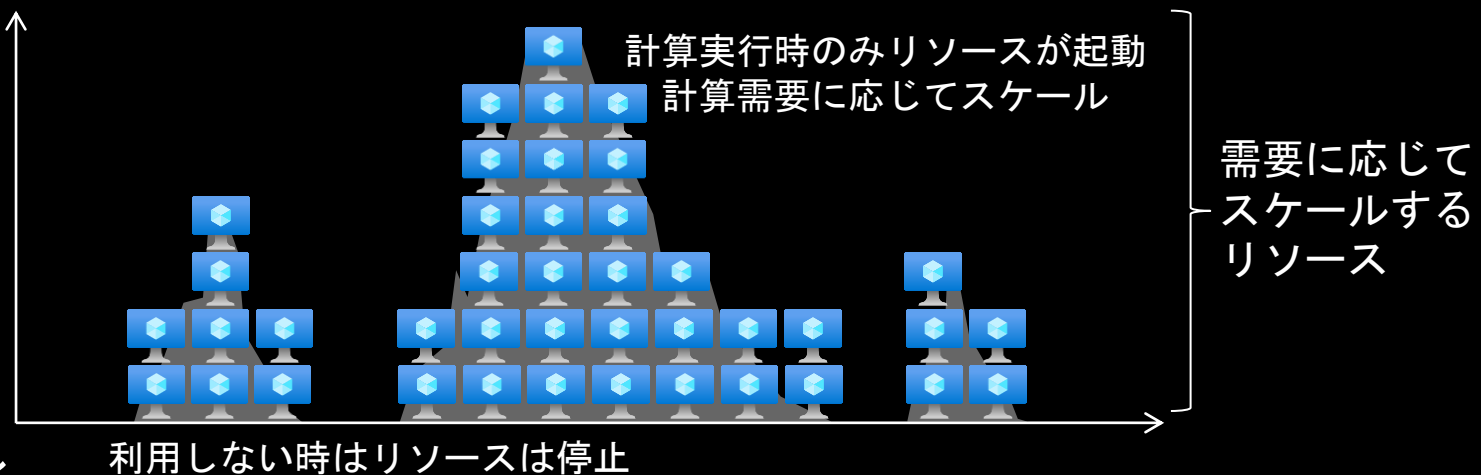
- ✓ 需要に応じてリソースを増減
- ✓ 利用した分だけ課金（従量課金）
- ✓ 固定した需要に対する固定リソース（リザーブドインスタンス）



フルクラウドモデル

計算実行時のみリソースが起動
計算需要に応じてスケール

固定リソースなし



スケールアップとスケールアウト

1コアVM  ×1



64コアVM  ×64

シングルノードジョブで
利用するコア数に応じて
VMサイズを選択

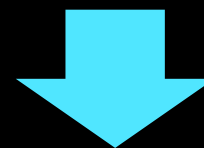
IB搭載VM
4ノード MPI-ジョブ



IB搭載VM
Nノード MPI-ジョブ

大規模マルチノードジョブに対応可能

限られた計算ノードに複数のジョブを実行



需要に応じて計算ノードを
スケールアウト

様々なリソースに対応



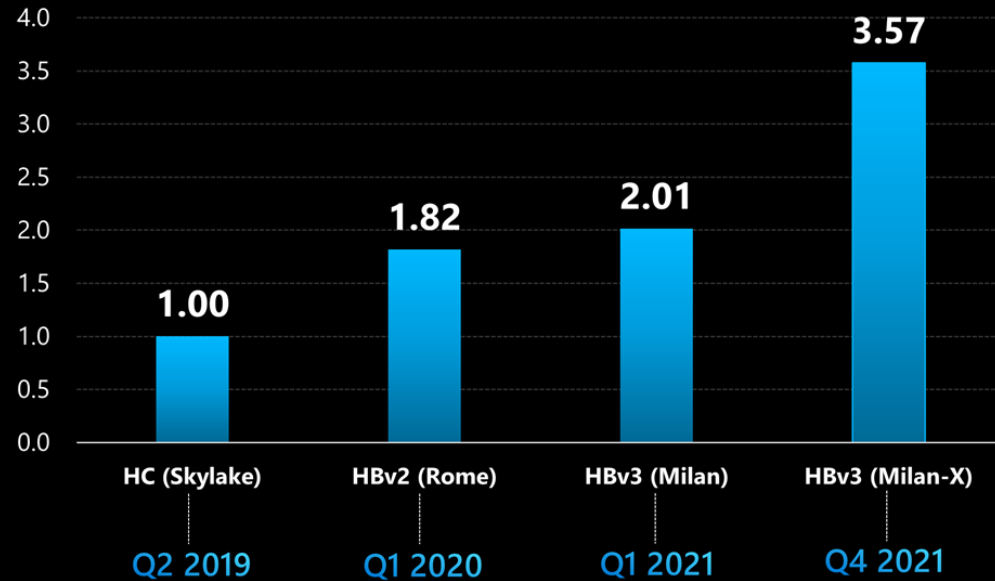
一つのクラスターに様々なリソースを準備可能。

計算ノードは計算実行時に起動し、利用しない時は停止させることで課金されない。

常に最新のリソースを利用可能

Constant Azure HPC Improvement

Relative Performance, Ansys Fluent 2021 R1, F1_racecar_140m



最新のCPUを搭載したAzure VMに乗換可能



Q2 2019: HC (Intel Skylake)

× 1.82



Q1 2020: HBv2 (AMD EPYC Rome)

× 2.01



Q1 2021: HBv3 (AMD EPYC Milan)

× 3.57



Q4 2021: HBv3 (AMD EPYC Milan-X)

最新のCPUを搭載したAzure VMを利用することで、
最高のパフォーマンスの計算機環境を利用し続けることが可能

Microsoft Azureにおける HPCのご紹介





Global



60 Azure regions

Largest geographical footprint of any cloud provider with more than **60** Azure regions



Secure



Microsoft Cyber Defense Operations Center

>3,500 full-time security professionals

6.5 trillion global signals daily

\$1 billion annual cybersecurity investment



Compliant

91 Compliance offerings

GLOBAL

- ISO 27001:2013
- ISO 27017:2015
- ISO 27018:2014
- ISO 22301:2012
- ISO 9001:2015
- ISO 20000-1:2011
- SOC 1 Type 2
- SOC 2 Type 2
- SOC 3
- CIS Benchmark
- CSA STAR Certification
- CSA STAR Attestation
- CSA STAR Self-Assessment
- WCAG 2.0 (ISO 40500:2012)

U.S. GOVT

- FedRAMP High
- FedRAMP Moderate
- EAR
- ITAR
- DoD DISA SRG Level 5
- DoD DISA SRG Level 4
- DoD DISA SRG Level 2
- DFARS
- DoE 10 CFR Part 810
- NIST SP 800-171
- NIST CSF
- Section 508 VPATs
- FIPS 140-2
- CJIS
- IRS 1075
- CNSSI 1253

INDUSTRY

- PCI DSS Level 1
- GLBA (US)
- FFIEC (US)
- Shared Assessments (US)
- SEC 17a-4 (US)
- CFTC 1.31 (US)
- FINRA 4511 (US)
- SOX (US)
- 23 NYCRR 500 (US)
- OSFI (Canada)
- FCA + PRA (UK)
- APRA (Australia)
- FINMA (Switzerland)
- FSA (Denmark)
- RBI + IRDAI (India)
- MAS + ABS (Singapore)
- NBB + FSMA (Belgium)
- AFM + DNB (Netherlands)
- AMF + ACPR (France)
- KNF (Poland)
- European Banking Authority (EBA)
- FISC (Japan)
- HIPAA BAA (US)
- HITRUST Certification
- GxP (FDA 21 CFR Part 11)
- MARS-E (US)
- NHS IG Toolkit (UK)
- NEN 7510:2011 (Netherlands)
- FERPA (US)
- CDSA
- MPAA (US)
- FACT (UK)
- DPP (UK)

REGIONAL

- Argentina PDPA
- Australia IRAP Unclassified
- Australia IRAP PROTECTED
- Canada Privacy Laws
- China GB 18030:2005
- China DJCP (MLPS) Level 3
- China TRUCS / CCCPPF
- EU EN 301 549
- EU ENISA IAF
- EU Model Clauses
- EU – US Privacy Shield
- GDPR
- Germany CS
- Germany IT-Grundschutz workbook
- India MeitY
- Japan CS Mark Gold
- Japan My Number Act
- Netherlands BIR 2012
- New Zealand Gov CIO Framework
- Singapore MTCS Level 3
- Spain ENS High
- Spain DPA
- UK Cyber Essentials Plus
- UK G-Cloud
- UK PASF

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



Fast, Secure Networking

ExpressRoute

InfiniBand

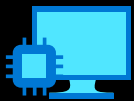


High Performing Storage

Azure HPC Cache

Azure NetApp Files

Clusterstore



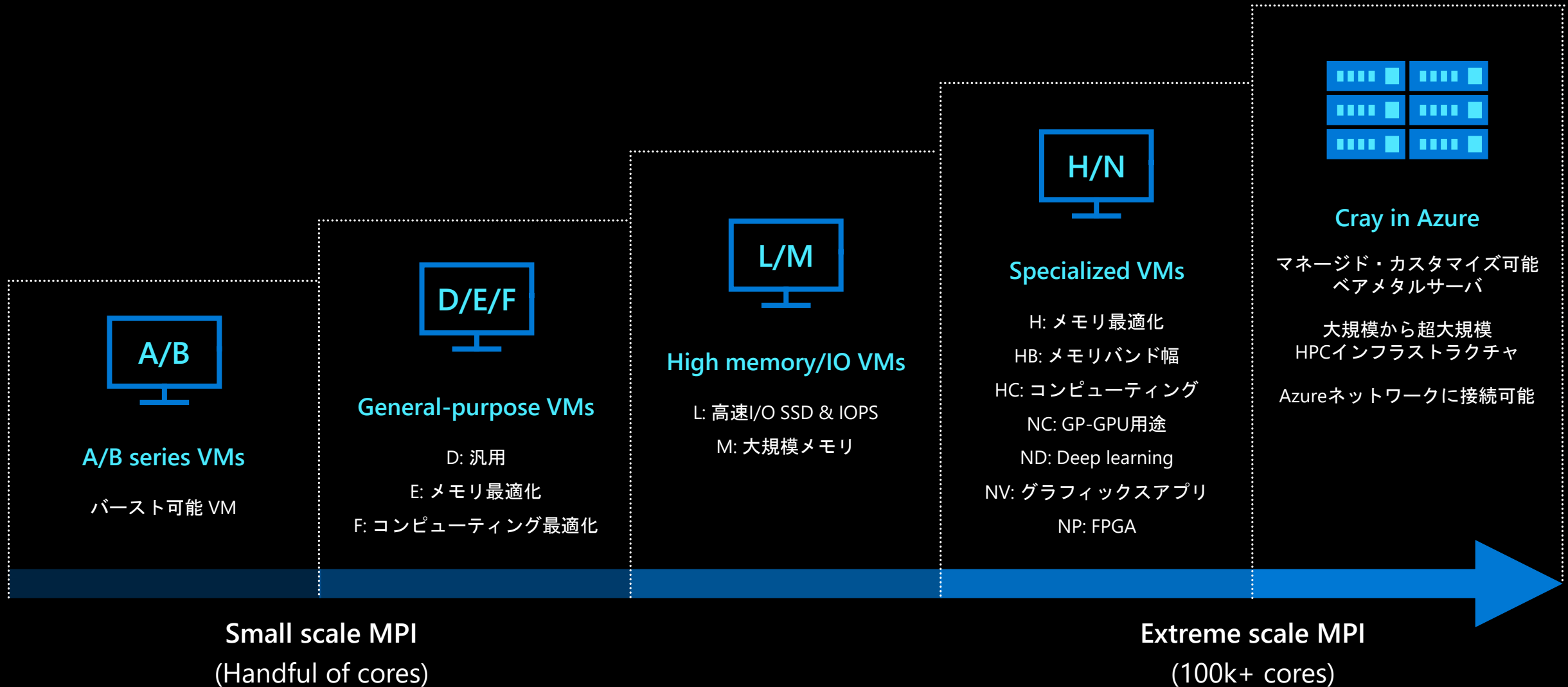
Optimized Compute

H-Series

N-Series

Cray

Solve any HPC, AI workload—at any scale



HPC用途 VM (H-Series)

	 HBv3	 HBv2	 HB	 HC	 H
Workload Optimized	メモリバンド幅	メモリバンド幅	メモリバンド幅	高密度コンピューティング	大規模メモリHPC
CPU	AMD EPYC 3 rd Gen w/ 3D V-Cache "Milan-X"	AMD EPYC 2 nd Gen "Rome"	AMD EPYC 1 st Gen "Naples"	Intel Xeon Platinum 1 st Gen "Skylake"	Intel Xeon E5 v3 "Haswell"
Cores/VM	120	120	60	44	16
TeraFLOPS/VM (FP64)	4 TF	4 TF	1.92 TF	2.6 TF	.7 TF
Memory Bandwidth	350 GB/s	350 GB/s	263 GB/s	191 GB/s	80 GB/s
Memory	448 GB total	4 GB/core, 480 total	4 GB/core, 240 total	8 GB/core, 352 GB	14 GB/core, 224 GB
Local Disk	2 * 960 GB NVMe	960 GB NVMe	700 GB NVMe		2 TB SATA
InfiniBand	200 Gb HDR	200 Gb HDR	100 Gb EDR		56 Gb FDR
Network	50 GbE	50 GbE	50 GbE		16 GbE

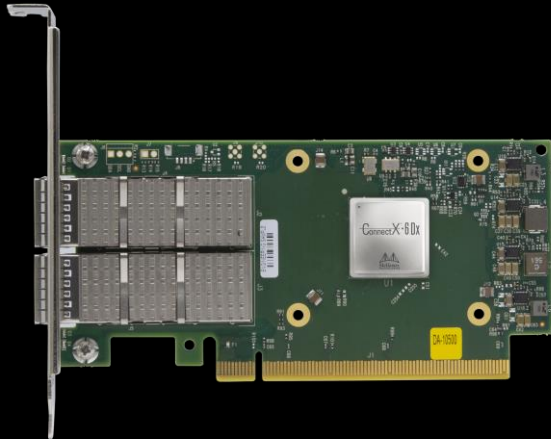
**All cores, non-AVX, peak Boost/Turbo frequencies*

ハードウェア インターコネクト



Connect-X 5 Adapter

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



Connect-X 6 Adapter

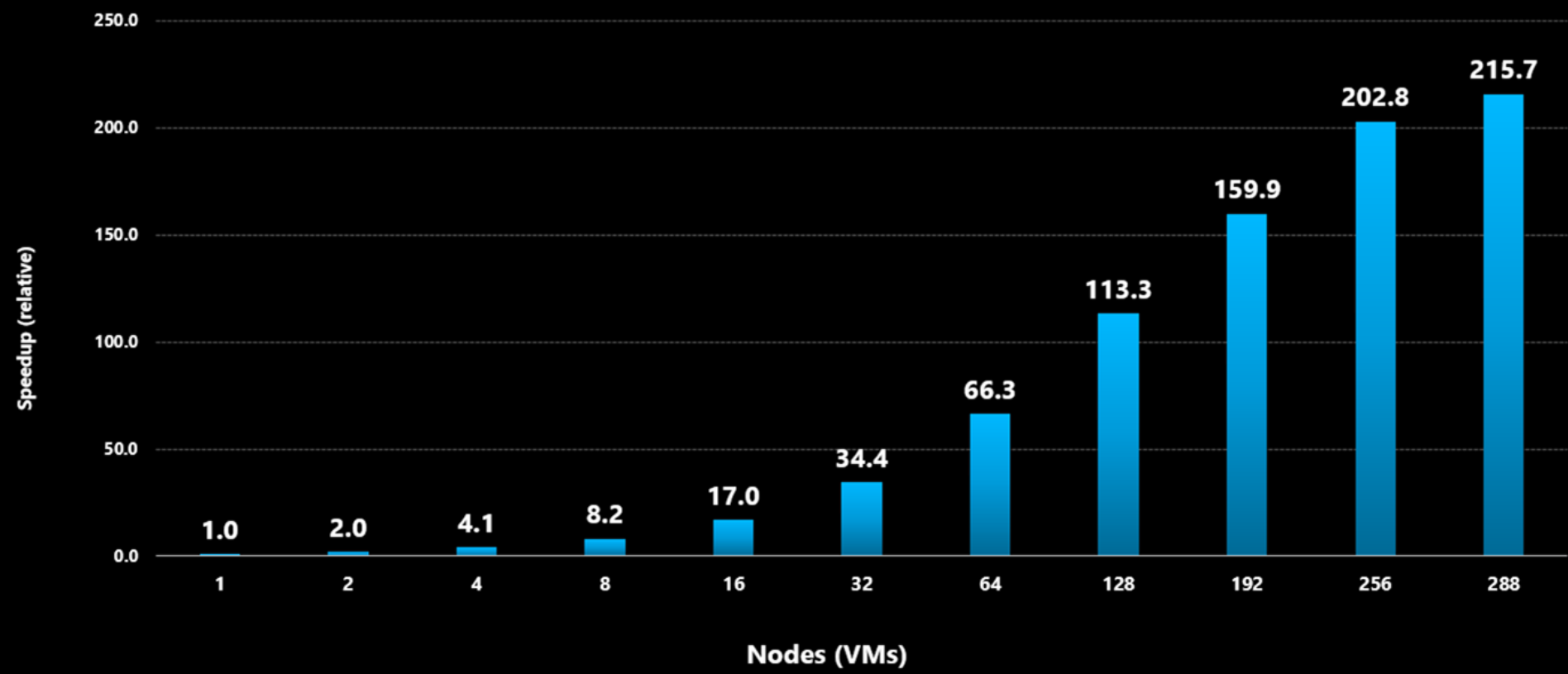
- HBv3/HBv2 VM
- HDR 200Gb/s InfiniBand
- Up to 215M messages/second
- PCI Gen4

Adaptive Routing on Reliable Transport | Enhanced vSwitch / vRouter Offloads

HBv3 Performance Benchmark

HPC at Extreme Scale on Azure

HBv3 (120 EPYC 7V13 cores); Scaling Curve up to 33,408 CPU cores
Siemens Star-CCM+ 15.04.088 – Le Mans 100m Coupled Solver Model – 20 iterations



GPU コンピューティング

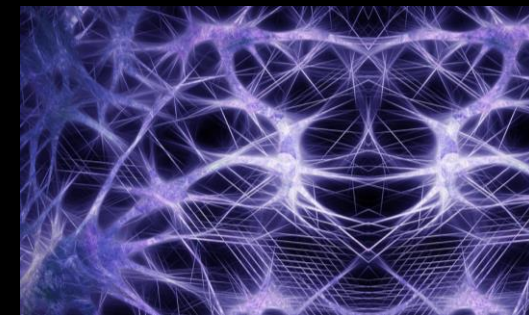
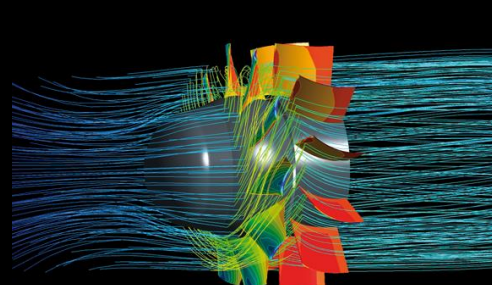
可視化



レンダリング



HPC/シミュレーション ディープラーニング/AI



PC/ワークステーションのワークロード

利用に応じて構成変更可能な
GPUワークステーション

インタラクティブなコラボレーションのための
モダン・ワークスペース

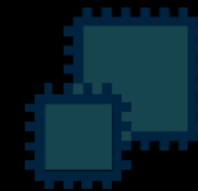


MPIを用いたマルチノードHPC/MLワークロードをIBを用いてスケールアップ

高密度シングルノードでのトレーニングやHPCワークロードを高速なNVLINK
で接続されたマルチGPU VMのスケールアップで実現



コンピューティング用途 GPU VM (NC) & 可視化用途 GPU VM(NV)

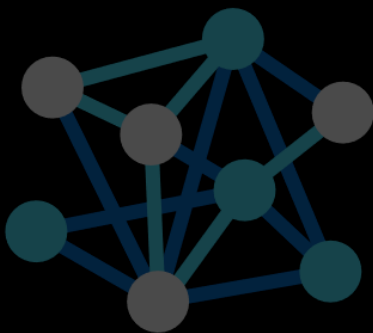


	 NC	 NCv2	 NCv3	 NCasT4_v3
Cores	6, 12, 24	6, 12, 24	6, 12, 24	4, 8, 16, 64
GPU	Tesla K80	Tesla P100	Tesla V100	Tesla T4
Memory	56/112/224 GB	112/224/448 GB	112/224/448 GB	28/56/110/440 GB
Local Disk	340/680/1440 GB SSD	736/1474/2948 GB SSD	736/1474/2948 GB SSD	180/360/2880 GB SSD
Network	Azure Network + InfiniBand (largest size only)			

	 NV	 NVv3	 NVv4
Cores	6, 12, 24	12, 24, 48	4, 8, 16, 32
GPU	Tesla M60	Tesla M60	Radeon Instinct MI25
Memory	56,112,224 GB	112/224/448 GB	14/28/56/112 GB
Local Disk	340/680/1440 TB SSD	320/640/1280 GB SSD	88/176/352/704 GB
Network	Azure Network		



ディープラーニング向け GPU VM (ND Series)



	 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

deep learning

Azure CycleCloud



ユーザエンパワーメント

- 既存のワークフローと使い慣れたスケジューラーをクラウドで利用可
- 計算機環境へ簡単にアクセス可能
- オートスケール機能、エラー処理を提供



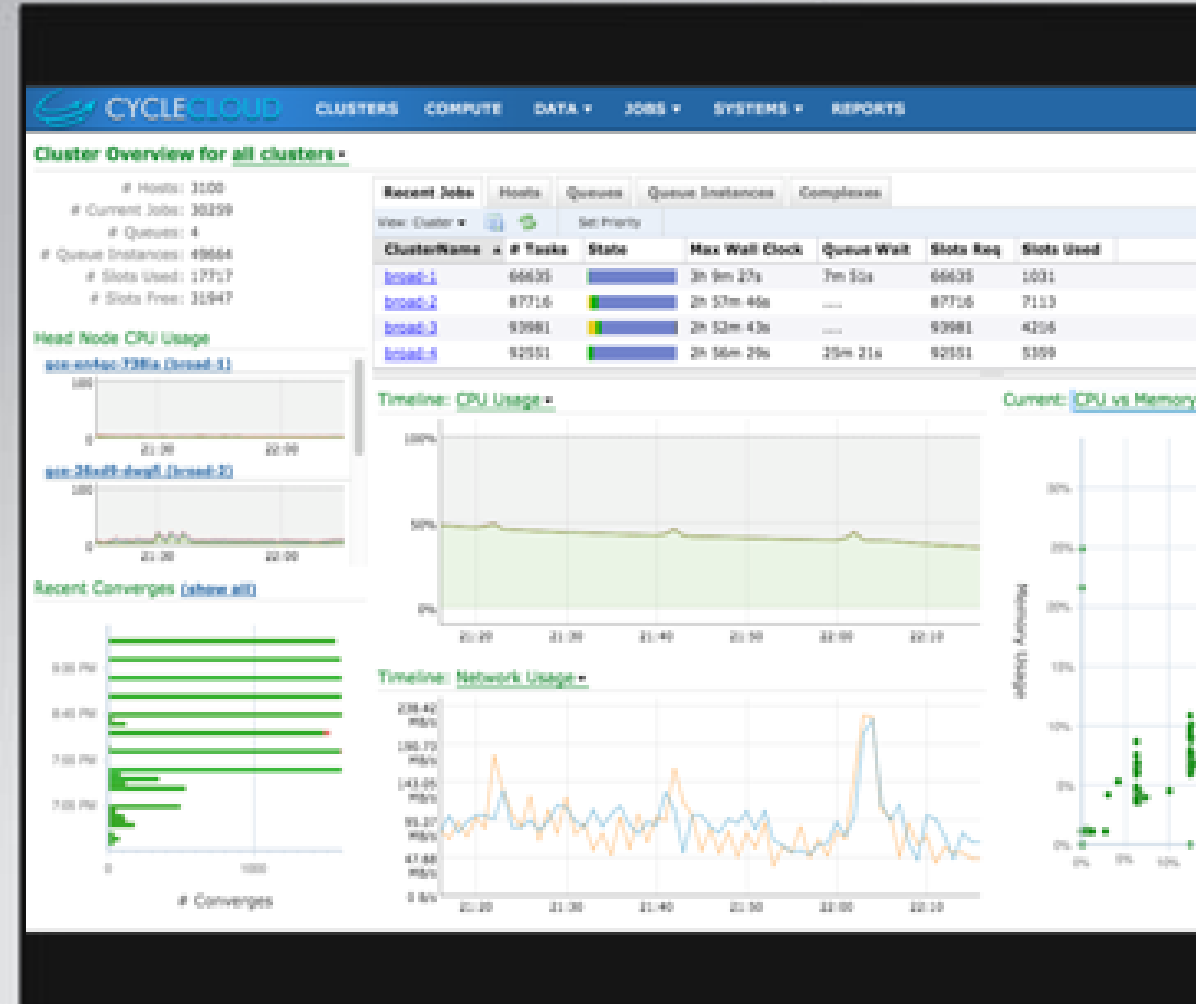
IT マネジメント

- オンプレミスと外部クラウドに対応したワークフローをリンク可能
- 認証と承認にActive Directoryを使用可能
- 一貫性のあるアクセスを提供



コスト管理

- 使用量と支出をリンク可能
- コストを管理・制御するためのツールを提供

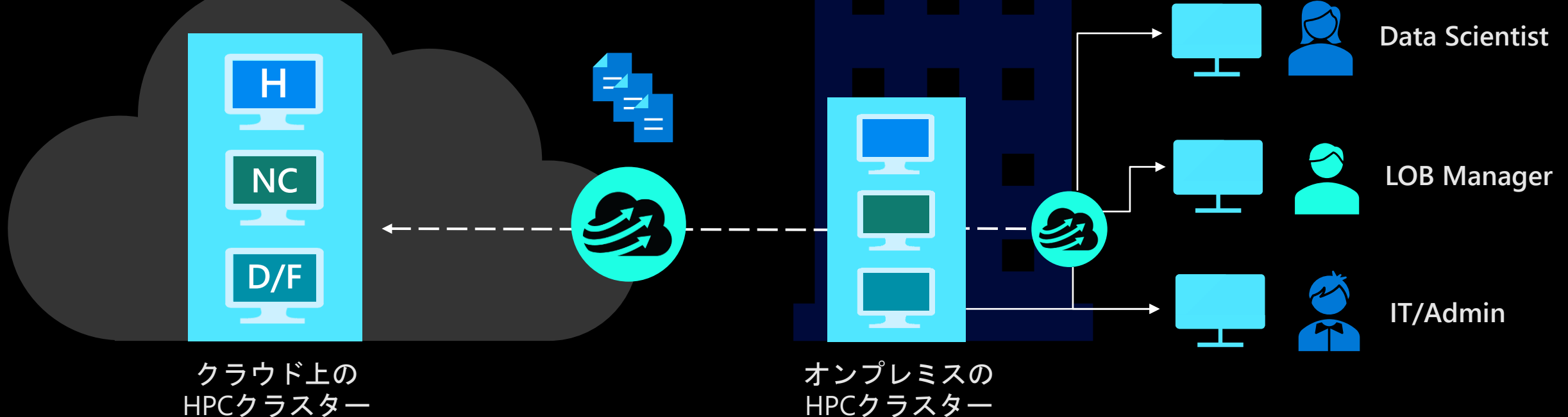


Azure CycleCloud 利用イメージ

オンプレミスと使い慣れたジョブスケジューラーと利用方法をクラウド環境で提供

- テンプレートを用いてクラウド上のHPC環境をカスタマイズ可能
- 計算リソースはジョブ実行時に自動起動、ジョブが終了すると自動停止
- ジョブの本数、サイズに応じて計算機リソースをオートスケール（上限の設定が可能）
- 様々な計算機リソースをあらかじめ準備しておくことで、用途に応じてリソースの選択が可能

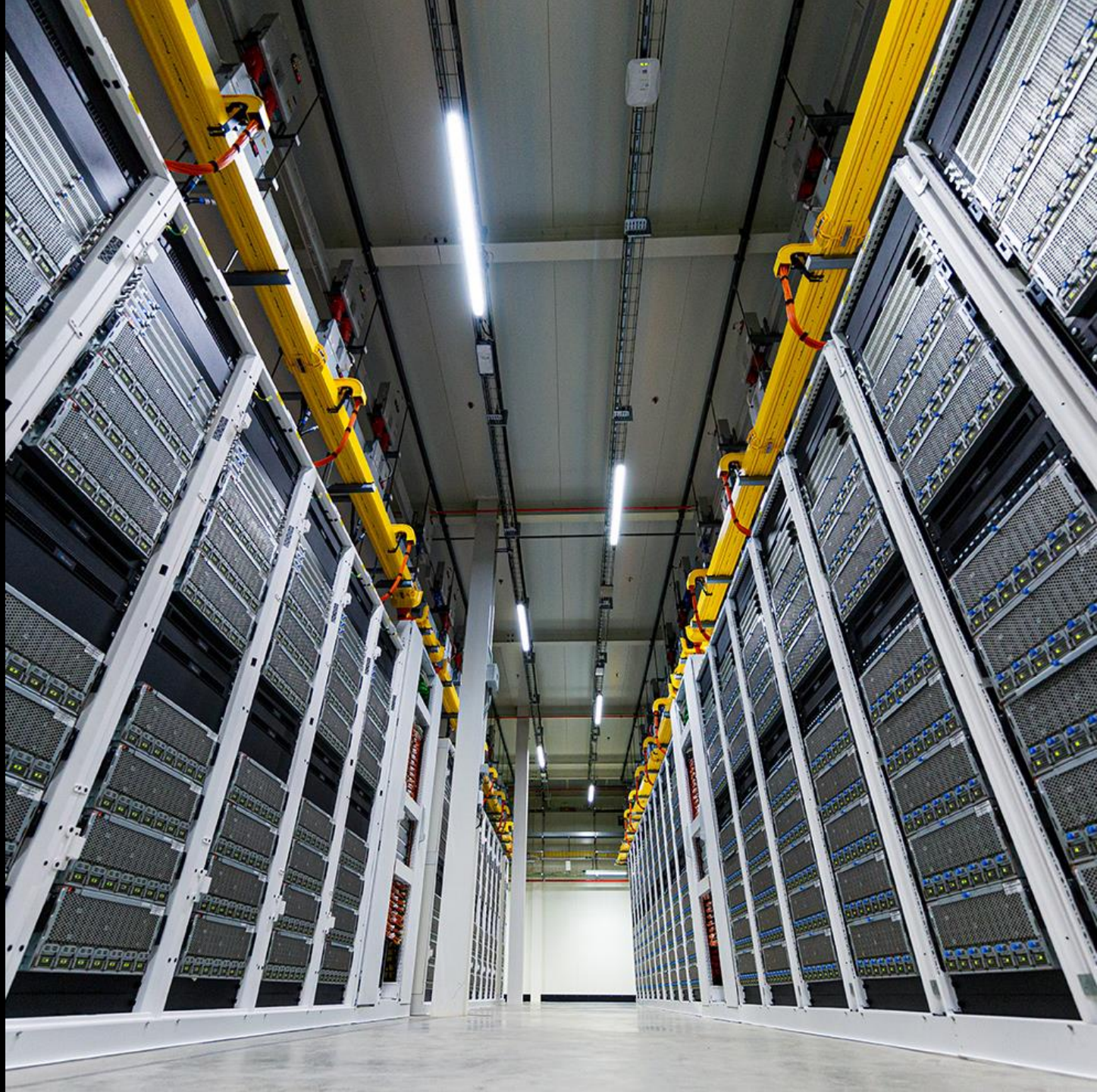
- クラウドの計算機リソースへジョブ投入
- クラウドとオンプレミスのクラスターを選択可能
- 管理者はCycleCloudサーバのウェブインターフェイスでクラウド上のHPC環境を管理・運用・デプロイ



Azure CycleCloud がサポートするジョブスケジューラー



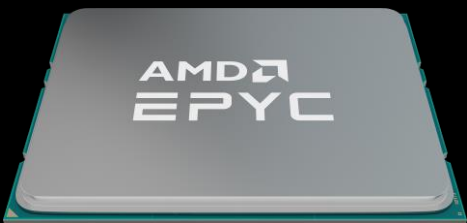
Azure Update



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



	
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)

Azure HBv3 (Milan-X)

VMサイズそのままCPUをAMD EPYC MilanからMilan-Xにアップデート



HBv3	Milan-X	Milan
CPU	AMD EPYC 7V73X	AMD EPYC 7V13
CPU Frequency (base/peak)	1.9GHz/3.5GHz	2.45GHz/3.675GHz
L1d/L1i cache	32KB/32KB	
L2 cache	512KB	
L3 cache (per CCD)	96MB (3D V-Cache)	32MB
Cores(sockets)/node	120(2)	
メモリバンド幅	350GB/s	
搭載メモリ量	448GiB	
ローカルディスク	960 GiB NVMe x2	
InfiniBand	200Gbps HDR InfiniBand	
接続ネットワーク	50 Gbps	

HBv3 update with "Milan-X"

Azure HBv3仮想マシンを第3世代 AMD EPYC Milan-X にアップデートします。メモリ性能に依存するさまざまなワークロードのパフォーマンス、スケーリング効率、コスト効率を大幅に改善します。

CFD: 80%性能向上

EDA RTL: 60%性能向上

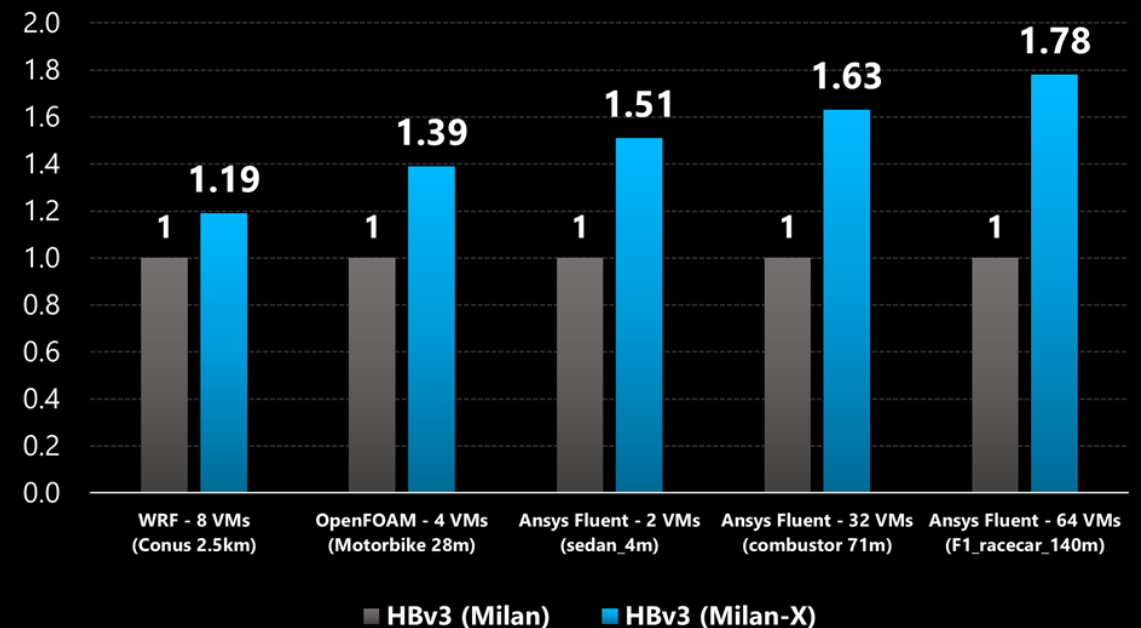
FEA(陽解法): 50%性能向上

気象シミュレーション: 19%性能向上



Relative HPC Performance

HBv3 (Milan-X) v. HBv3 (Std. Milan)



HBv3 update with "Milan-X"

スケール性能により、モデル規模が大きくなればなるほどよりコストメリットが高くなる

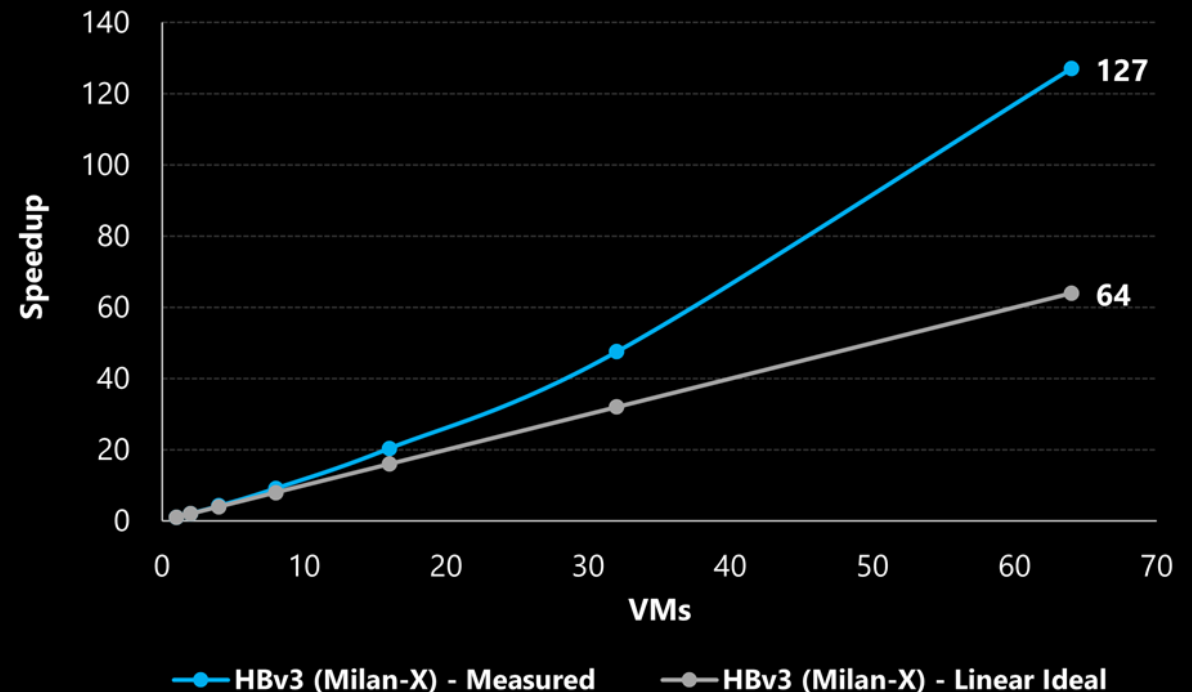


200% Scaling Efficiency

Ansys Fluent 2021 R1, F1_racecar_140m
Speedup (Solver Rating) v. Linear Ideal

NVIDIA Quantum InfiniBand
大規模キャッシュ効果

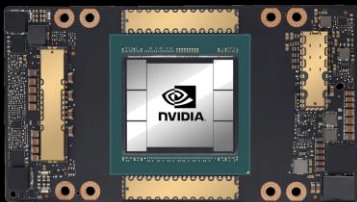
線形以上のスケーリング効率を実現することにより、結果を得るまでの時間を劇的に短縮することができるだけでなく、Azureのコンピューْتُコストを削減することができます。



NDm A100 v4

NDm A100 v4 シリーズの仮想マシンは、ハイエンドのディープラーニングトレーニングと密結合のスケールアップおよびスケールアウトの HPC ワークロード向けに設計された、Azure GPU ファミリに新たに追加された主力製品です。NVIDIA A100 80GB × 8 + AMD EPYC 7V12 + Mellanox HDR InfiniBand × 8でさまざまなAI、MLのフレームワークを活用して優れたパフォーマンスを発揮します。また、業界標準のNCCL2通信ライブラリに対応し、AIおよびHPCツールによってInfiniBand相互接続で優れたスケーラビリティを実現します。

- ✓ NVIDIA A100 80GB × 8 (NVLINK) 搭載
- ✓ AMD 第2世代 EPYC 7V12 (96コア/ノード) 搭載
- ✓ 200Gbps HDR InfiniBand × 8 で ノードあたり1.6Tb/sの相互接続帯域幅を提供



	
CPU	AMD EPYC 7V12 – Rome (2.45GHz)
コア数	96
GPU	NVIDIA A100 (80GB) x 8 with NVLINK 3.0
メモリ容量	1900GiB
ローカルディスク	6.4 TiB SSD
InfiniBand	200 Gbps HDR InfiniBand x 8

Top500 (SC21) に Azure NDmv4クラスターが10位にランクイン

2021年11月に開催されたSC21で発表されたTop500に、Azure NDmv4クラスターが10位にランクイン。浮動小数点演算性能30.05PFLOPS(実行効率76%)を記録している。ISC2021にランクインした4システム(32～35位)と合わせて5システムがTop500にランクイン



NOVEMBER 2021

The 58th annual edition of the TOP500 saw little change in the Top10. The Microsoft Azure system called Voyager-EUS2 was the only machine to shake up the top spots, claiming No. 10. Based on an AMD EPYC processor with 48 cores and 2.45GHz working together with an NVIDIA A100 GPU and 80 GB of memory, Voyager-EUS2 also utilizes a Mellanox HDR Infiniband for data transfer.

8	JUWELS Booster Module - Bull Sequana XH2000 , AMD EPYC 7402 24C 2.8GHz, NVIDIA A100, Mellanox HDR InfiniBand/ParTec ParaStation ClusterSuite, Atos Forschungszentrum Juelich (FZJ) Germany	449,280	44,120.0	70,980.0	1,764
9	HPC5 - PowerEdge C4140, Xeon Gold 6252 24C 2.1GHz, NVIDIA Tesla V100, Mellanox HDR Infiniband, DELL EMC Eni S.p.A. Italy	669,760	35,450.0	51,720.8	2,252
10	Voyager-EUS2 - ND96amsr_A100_v4, AMD EPYC 7V12 48C 2.45GHz, NVIDIA A100 80GB, Mellanox HDR Infiniband, Microsoft Azure Azure East US 2 United States	253,440	30,050.0	39,531.2	



NC_A100_v4

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

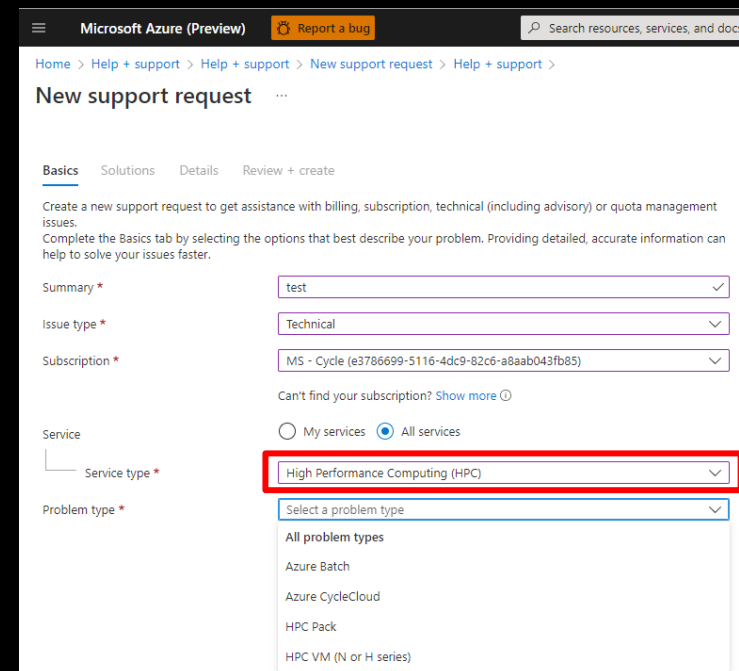


	NC24ads_A100_v4	NC48ads_A100_v4	NC96ads_A100_v4
CPU	AMD EPYC 7V13 – Milan (2.45GHz)		
コア数	24	48	96
GPU	NVIDIA A100 (80GB) x 1	NVIDIA A100 (80GB) x 2	NVIDIA A100 (80GB) x 4
メモリ容量	220GiB	440GiB	880GiB
ローカル ディスク	1123 GiB NVMe	2246 GiB NVMe	4492 GiB NVMe
Network bandwidth (Mbps)	20,000	40,000	80,000

Azure HPC Support

HPC専門のサポートチーム

- ✓ HPCの専門知識のあるテクニカルアドバイザーが対応
- ✓ Azure Portalから問い合わせ可能
- ✓ メールおよび電話での対応
- ✓ US, EMEA, APACにサポートチームあり
- ✓ 24x7対応



Microsoft Azure (Preview) Report a bug Search resources, services, and docs

Home > Help + support > Help + support > New support request > Help + support >

New support request

Basics Solutions Details Review + create

Create a new support request to get assistance with billing, subscription, technical (including advisory) or quota management issues. Complete the Basics tab by selecting the options that best describe your problem. Providing detailed, accurate information can help to solve your issues faster.

Summary * test

Issue type * Technical

Subscription * MS - Cycle (e3786699-5116-4dc9-82c6-a8aab043fb85)

Can't find your subscription? Show more

Service

Service type * High Performance Computing (HPC)

Problem type * Select a problem type

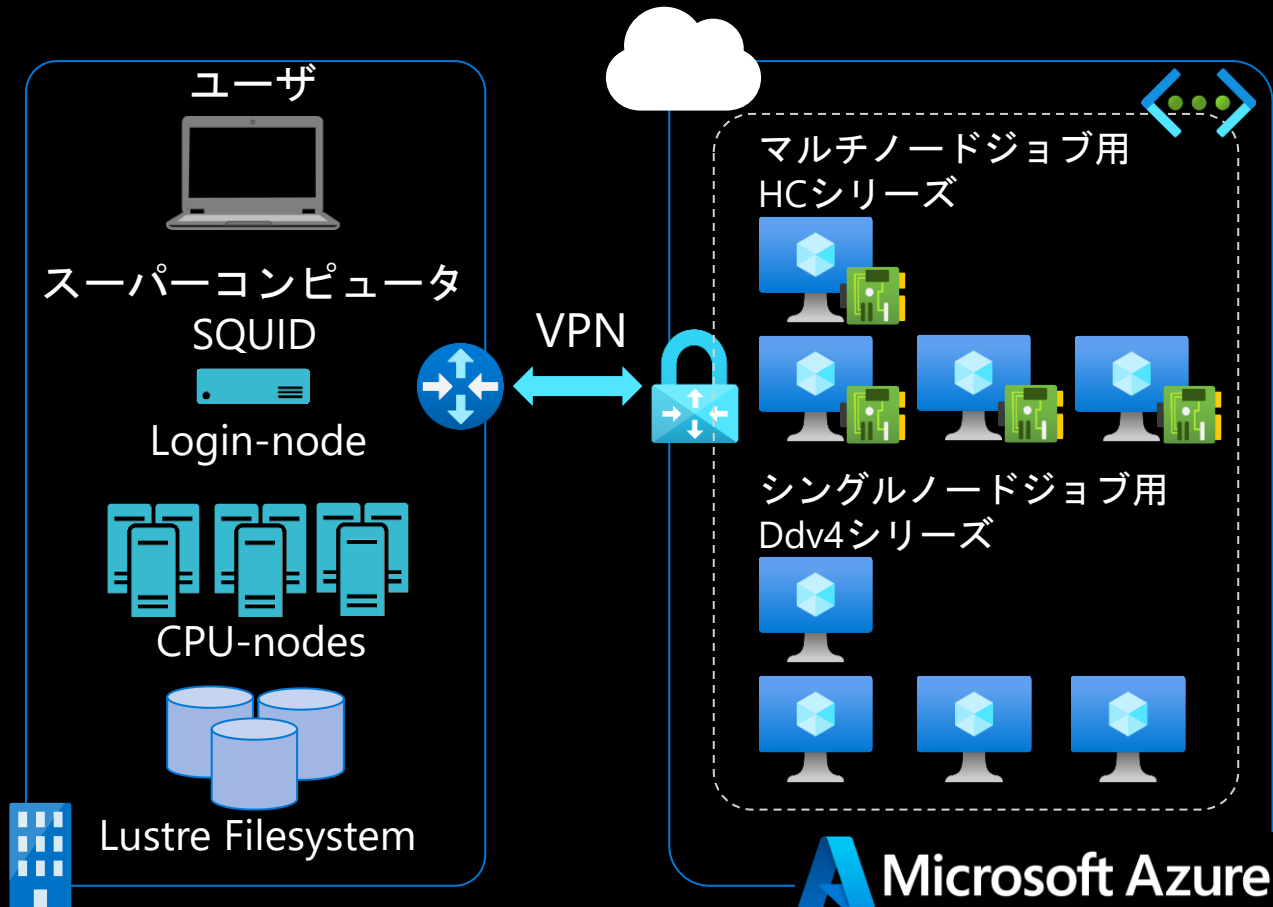
- All problem types
- Azure Batch
- Azure CycleCloud
- HPC Pack
- HPC VM (N or H series)

			
HPC Platform	Azure CycleCloud	Microsoft HPC Pack	Cray ClusterStor in Azure
• ハイパフォーマンス コンピューティング VM – Hシリーズ(H, HB, & HC) • GPU VM – N シリーズ (N, NV, NC & ND) • FPGA VM – NP シリーズ • InfiniBand, GPU ドライバー • GPU パフォーマンス etc	• Cyclecloud サポート • クラスター 構成 • クラスター 関連 • スケジューラー関連問題 • クラスター デプロイ問題	• Windows ベース HPC ソリューション • HPC pack サポート	• Lustre parallel Filesystem

クラウドHPC 事例紹介



事例：大阪大学サイバーメディアセンター様 スーパーコンピュータSQUID+クラウドバーステイング

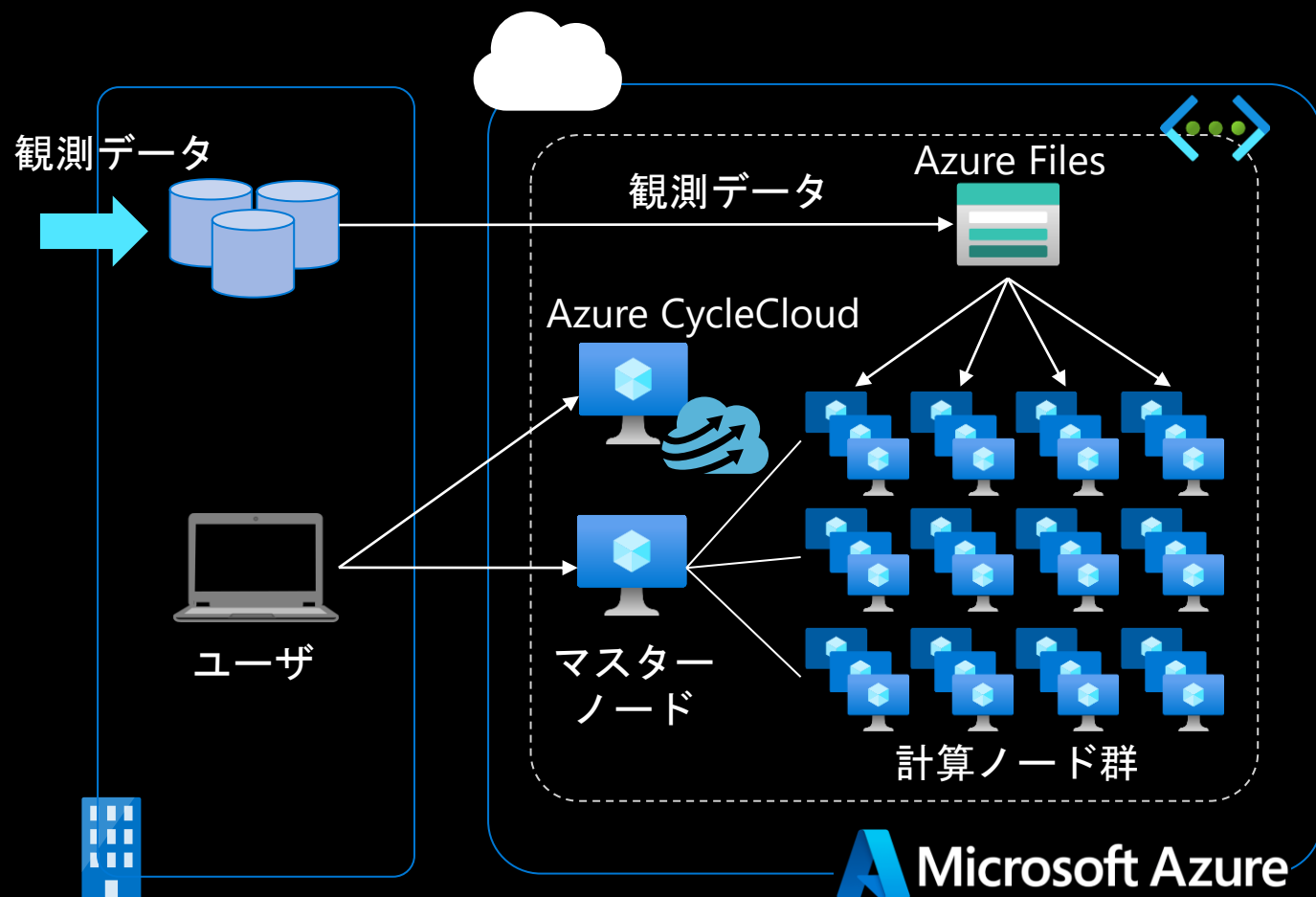


スーパーコンピュータが混雑時に、クラウドの計算リソースで処理

- ✓ ジョブの実行待ちを減らせる
- ✓ マルチノードジョブ用にIB搭載のVMでスケールアップ&スケールアウト
- ✓ シングルノードジョブ用にはIB非搭載のVMでスケールアウト
- ✓ オンプレミスストレージ環境をクラウド計算リソースにマウント
- ✓ 高速I/Oが必要なジョブではクラウド計算リソースのローカルディスクヘステーシング

ハイブリッドモデル

事例：京都大学防災研究所様 CycleCloudを用いたリアルタイム洪水予測システム



台風による河川のリアルタイム洪水予測

- ✓ Azure CycleCloudを用いることにより、計算が必要な時のみ計算ノードを起動
- ✓ 必要な計算機リソースをオンデマンドにクラウドに展開
- ✓ 計算が行われないときは計算リソースは停止

オンプレミスのスーパーコンピュータではできなかったリアルタイム洪水予測ができるようになった

フルクラウドモデル

まとめ



Microsoft AzureのHPCリソース

- ✓ 仮想マシン
- ✓ ネットワーク
- ✓ ストレージ
- ✓ HPC専門サポートチーム

クラウドならではのHPCシナリオ

- ✓ オンデマンド
- ✓ 最新のリソースに乗り換え可能
- ✓ 従量課金と予約インスタンス

Microsoft Azureでクラウドならではの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