



クラウド HPC の最先端： 未来を切り拓く スーパーコンピュータ Microsoft Azure Eagle

日本マイクロソフト株式会社
インフラテクノロジー テクニカル スペシャリスト
松森 健明

2024/6/27

Who am I ?

松森 健明 Takeaki Matsumori

Technical Specialist, Microsoft
(Infrastructure)

技術営業として業種や業界を問わず、すべてのお客様へ
Microsoft Azure をご活用いただくための技術的な計画を支援

Interest:

- Cloud Infrastructure
- Cloud HPC
- Networking

Top500 List : 2023/11 ~

1. Frontier, HPE (1,194.00 PFlop/s)
2. Aurora, Intel (585.34 PFlop/s)
3. Eagle, Microsoft (561.20 PFlop/s)



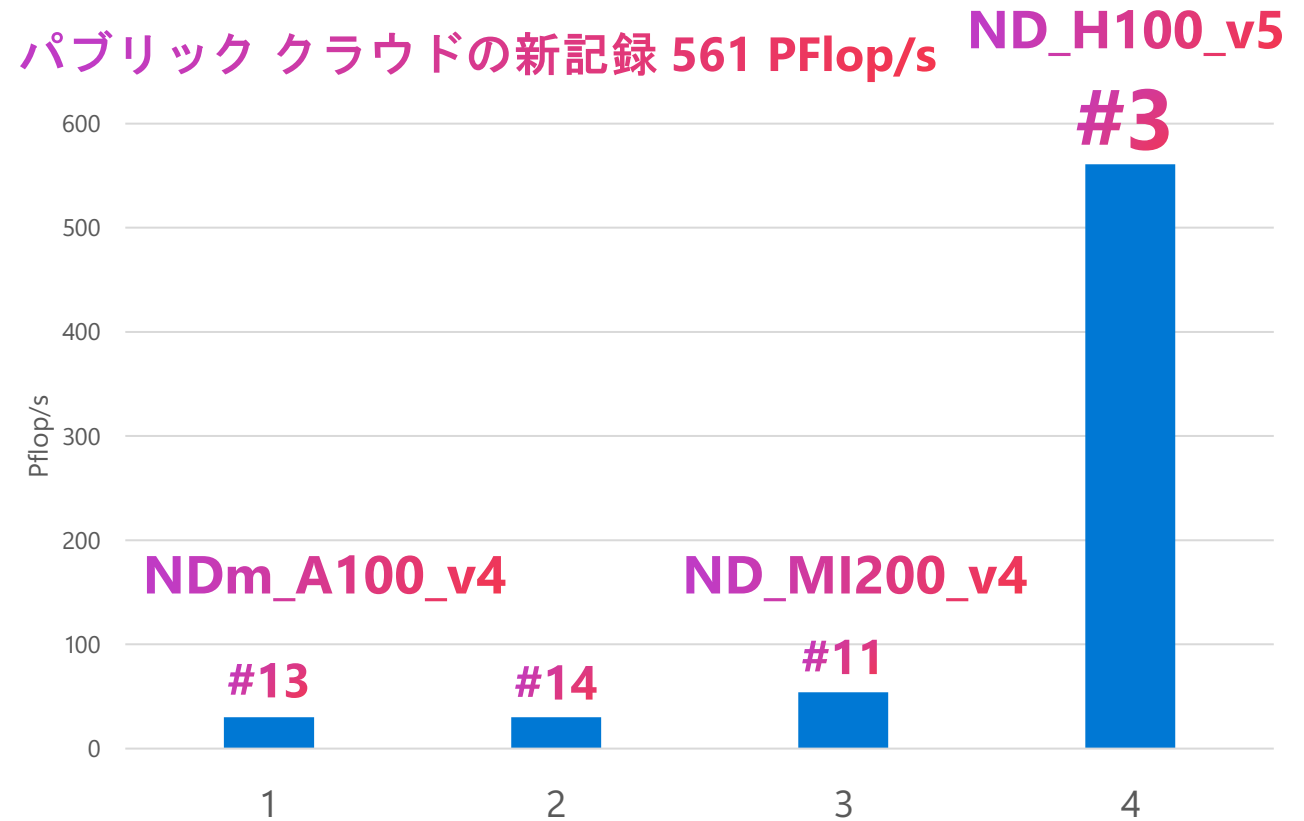
Microsoft Azure の Top500 への取り組み

#3 supercomputer all-up

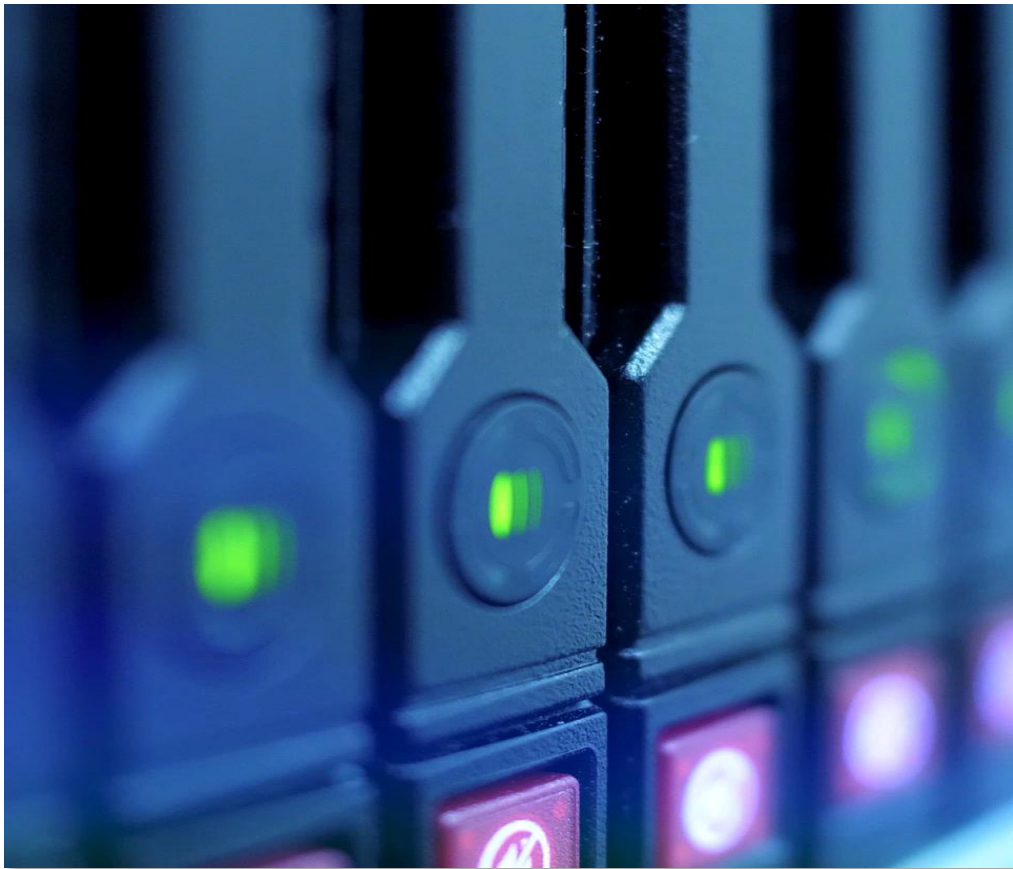
#1 cloud provider

19X performance increase
year-to-year

10X performance increase
list-to-list



High-Performance Linpack (HPL) 561 PFlop/s



- Azure 仮想マシン (ND H100 v5 シリーズ) をベースとしたスーパーコンピュータ
- 14,400 基の NVIDIA H100 GPU
- InfiniBand Quantum 7 NDR 400Gb/s
- Azure HPC/AI のお客様も利用可能な機能を用いて構築

最新の Azure ND H100 v5 VM シリーズ



NVIDIA H100 GPU を 1,000 台規模までスケール可能



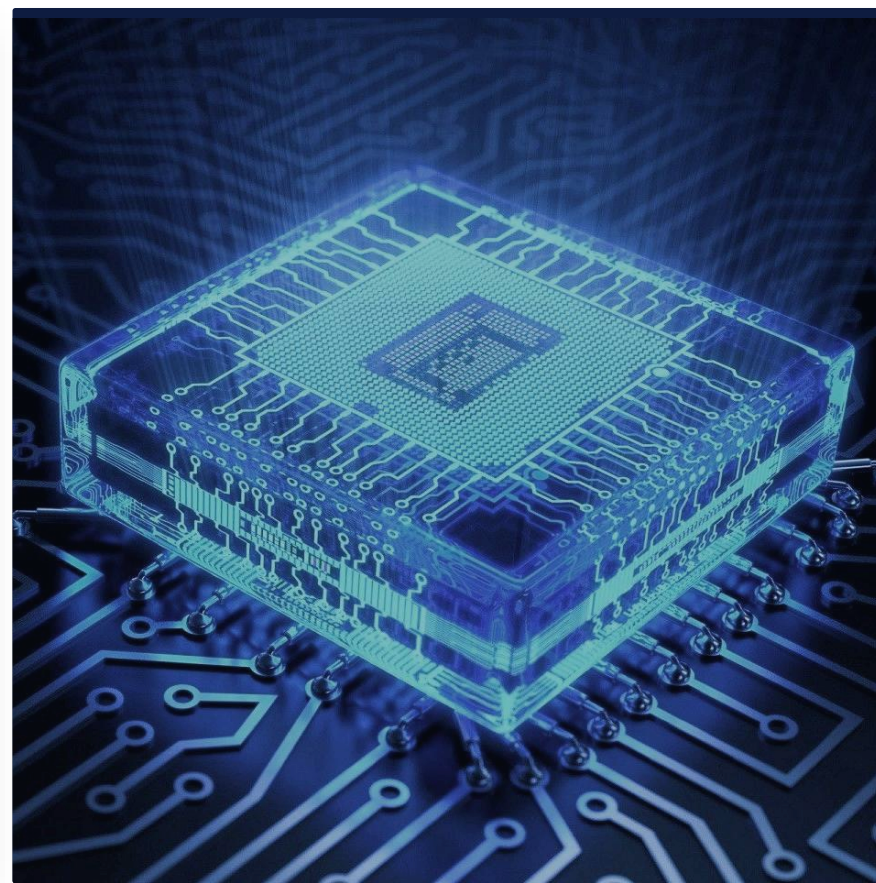
NVIDIA Quantum-2 InfiniBand でノード間接続 (NDR)



生成 AI アプリケーション向けに設計

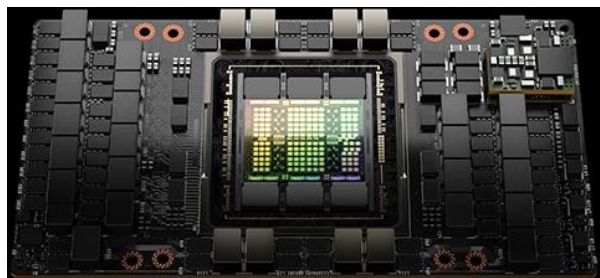


新しいクラスの大規模 AI モデルへ対応



ND H100 v5 > 大規模にスケール可能な AI スーパーコンピュータ

シングル H100 GPU



NCCL
+
NVLink

NVIDIA H100 Tensor Core GPU

- 80 GB of HBM3 Memory
- 2x – 30x A100 performance
- PCIe Gen 5
- 8 GPUs per NDv5 VM

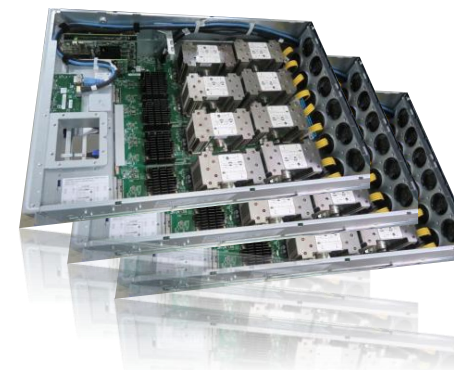
マルチ GPU シングル ND H100 v5 VM 内の NVLink で接続された 8 GPU



NVSwitch + NVLink 4.0

- Connects all 8 GPUs on server
- 3.6 TB/s Bisection BW
- 450 GB/s AllReduce

マルチ GPU VM Quantum-2 InfiniBand で接続された 複数台の ND_H100_v5 VM 千台規模の NDv5, 数千台の H100 GPU までスケール アウト



NCCL
+
NDR

Quantum-2 InfiniBand

- 400 Gigabit dedicated link per GPU (3.2 Terabits/VM)
- Fat-tree topology, connecting O(1000) GPUs
- GPU Direct RDMA

ソフトウェア スタック

OS イメージ

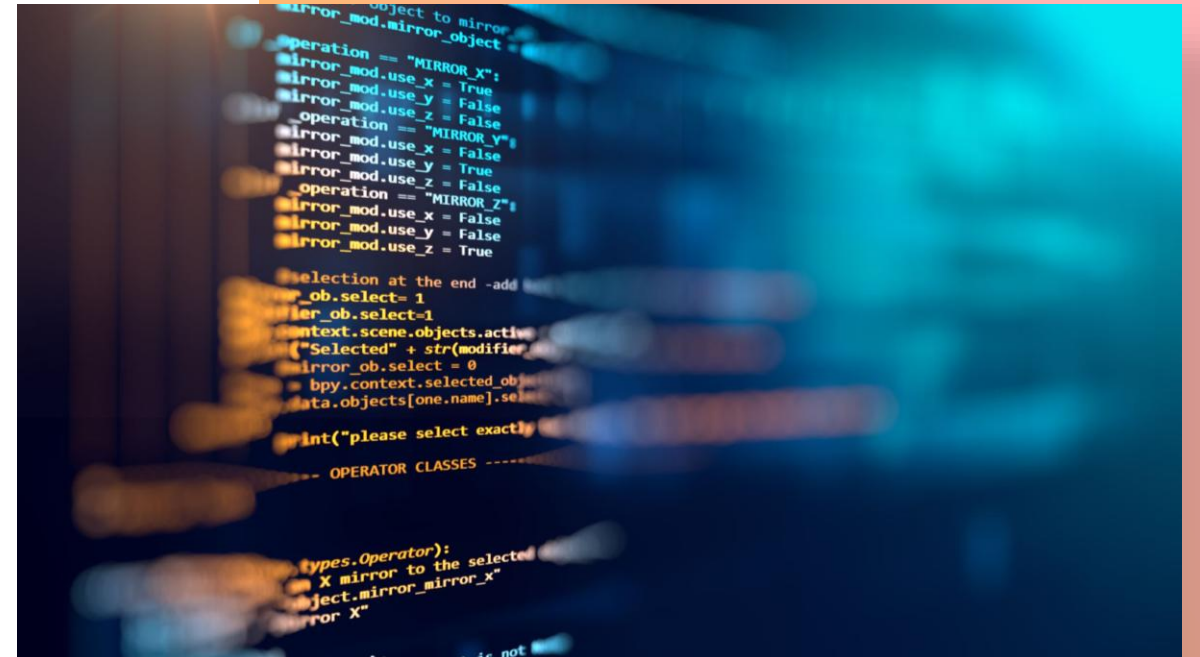
- Azure HPC イメージ
 - Ubuntu-HPC 22.04
- Azure Marketplace からすぐに展開可能

通信ライブラリ

- NVIDIA NCCL 2.19.3
- NVIDIA HPCX 2.16

HPL

- NVIDIA 社ビルドの HPL
- SLURM 22.05.3 経由で実行



Microsoft is powered by Azure AI infrastructure

Copilot for
Security

Microsoft
Copilot in Bing

Copilot for
Microsoft 365

Windows
Copilot

Dynamics 365
Copilot

Azure
OpenAI API

Microsoft のサービスは Azure AI インフラストラクチャ上で実行

Azure AI は Azure HPC インフラストラクチャ上で実行

リアルタイム推論と
低コストコンピューティング



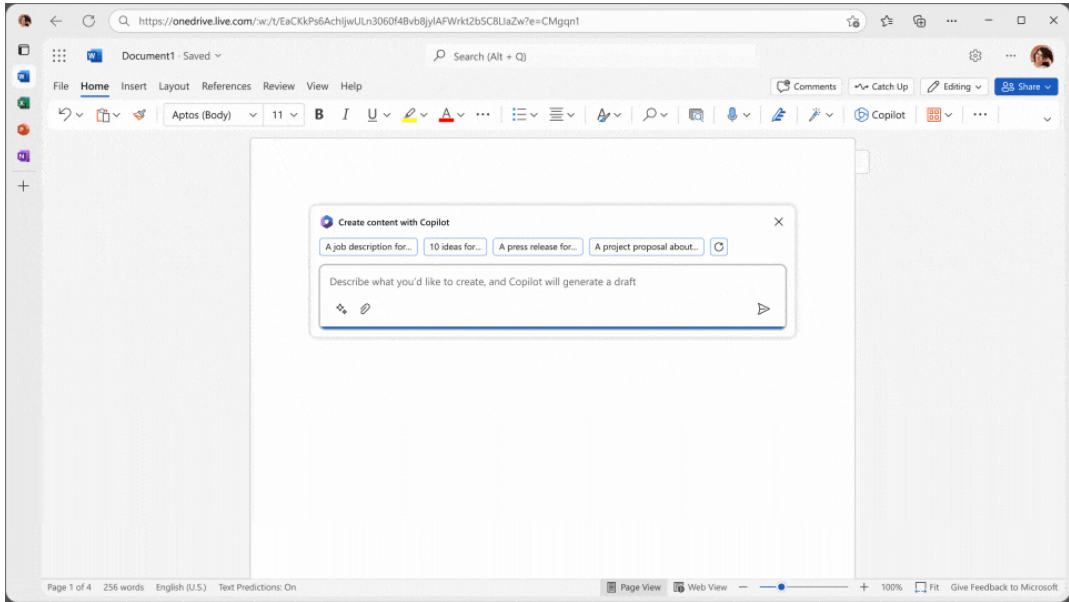
ミッドレンジトレーニングと
高密度推論



分散トレーニングと生成推論

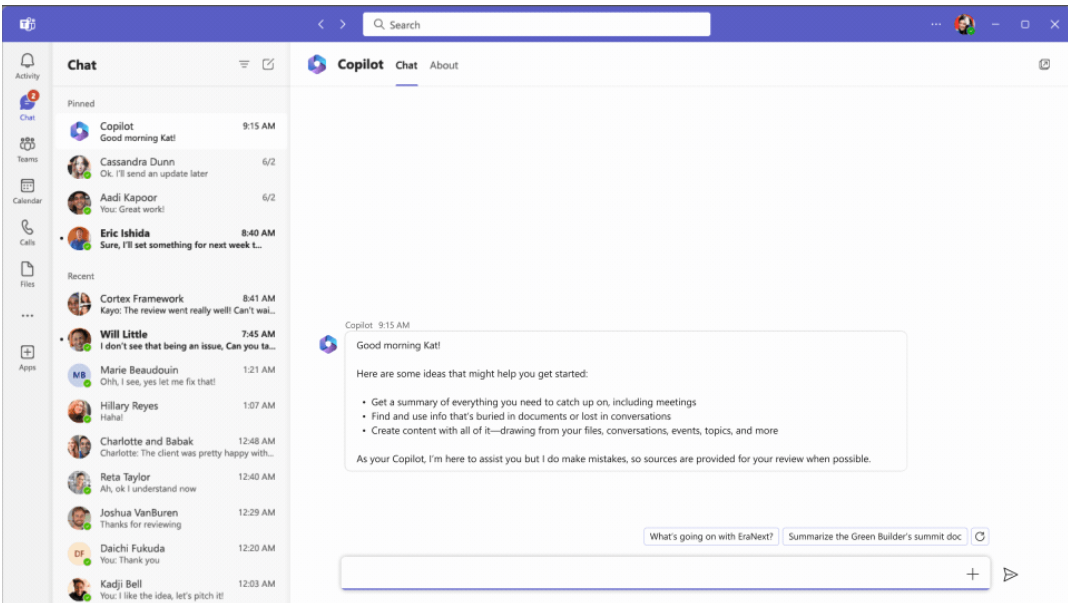
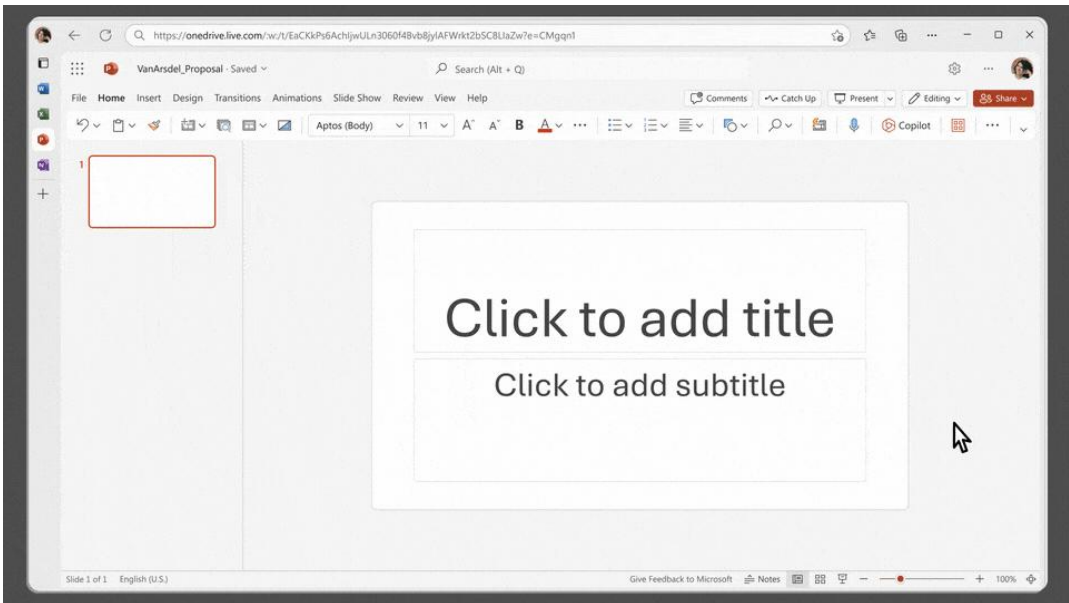


参考：Copilot for Microsoft 365



This screenshot shows the Microsoft Excel application with a data table. The table has columns for Country, Customer, Product, Discount Band, Units Sold, Manufacturing Price, Sale Price, Gross Sales, Discounts, Sales, and COGS. The data is organized by country and product type. A Copilot sidebar is visible on the right side of the application.

Country	Customer	Product	Discount Band	Units Sold	Manufacturing Price	Sale Price	Gross Sales	Discounts	Sales	COGS
France	VanArsdel	Biodegradable Cleaning Products	Low	4345	\$10	\$352	\$1,529,440	\$15,294	\$1,514,146	\$
USA	VanArsdel	Biodegradable Cleaning Products	Low	4245	\$10	\$352	\$1,494,240	\$14,942	\$1,479,298	\$
Australia	VanArsdel	Biodegradable Cleaning Products	None	4124	\$10	\$352	\$1,451,648	\$0	\$1,451,648	\$
Mexico	VanArsdel	Biodegradable Cleaning Products	Medium	3954	\$10	\$352	\$1,391,808	\$83,508	\$1,308,300	\$
France	Contoso Pharmaceuticals	Biodegradable Cleaning Products	Low	4952	\$10	\$127	\$628,904	\$6,289	\$622,615	\$
France	Contoso Pharmaceuticals	Recycled Can	Low	4903	\$5	\$127	\$622,681	\$6,227	\$616,454	\$
Canada	Proseware	Natural Personal Care Products	Low	1954	\$270	\$302	\$590,108	\$5,901	\$584,207	\$
Canada	Proseware	Biodegradable Cleaning Products	High	1990	\$220	\$302	\$600,980	\$84,137	\$516,843	\$
USA	VanArsdel	Compostable Food Packaging	Low	1254	\$3	\$380	\$476,520	\$4,765	\$471,755	\$
USA	Proseware	Natural Personal Care Products	Medium	1465	\$224	\$302	\$442,430	\$26,546	\$415,884	\$
USA	Proseware	Sustainable Snack Packaging	Medium	1410	\$120	\$302	\$425,820	\$25,549	\$400,271	\$
Australia	Proseware	Natural Personal Care Products	Medium	1369	\$224	\$302	\$413,438	\$24,806	\$388,632	\$
USA	VanArsdel	Compostable Food Packaging	Medium	1052	\$3	\$380	\$399,760	\$23,986	\$375,774	\$
Canada	VanArsdel	Compostable Food Packaging	Low	980	\$312	\$380	\$372,400	\$3,724	\$368,676	\$
Australia	Proseware	Sustainable Snack Packaging	Medium	1298	\$120	\$302	\$391,996	\$23,520	\$368,476	\$
Australia	VanArsdel	Compostable Food Packaging	None	954	\$3	\$380	\$362,520	\$0	\$362,520	\$
Canada	Contoso Pharmaceuticals	Biodegradable Cleaning Products	Low	2785	\$110	\$127	\$353,695	\$3,537	\$350,158	\$



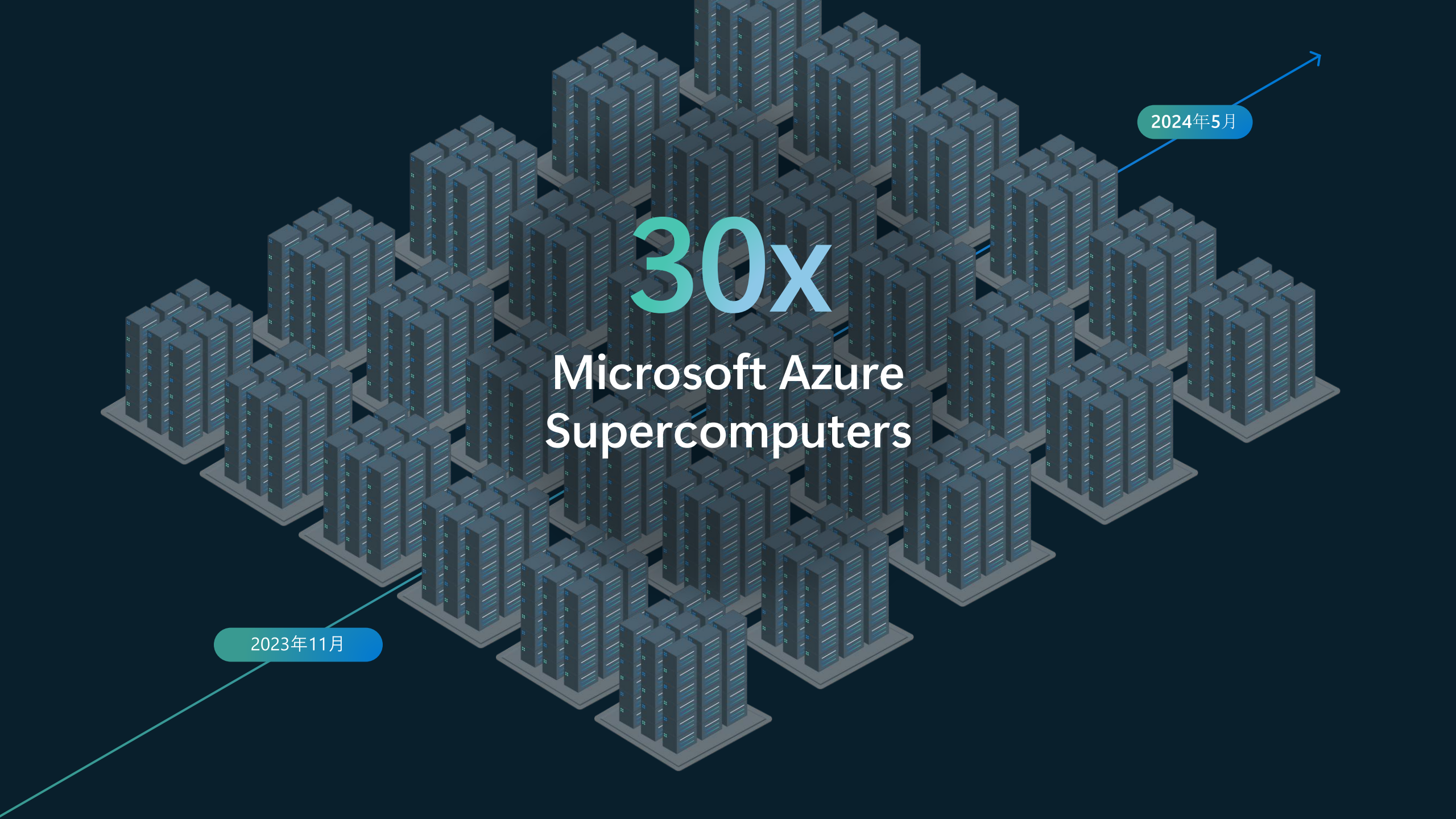
An isometric illustration of several server racks, each composed of multiple horizontal server units. The racks are arranged in a cluster, with some units featuring glowing blue and green lights. The entire scene is set against a dark blue background.

#1 supercomputer for cloud

Microsoft Azure

561.20 PFlop/s

TOP500, 2024 年 5 月



30x

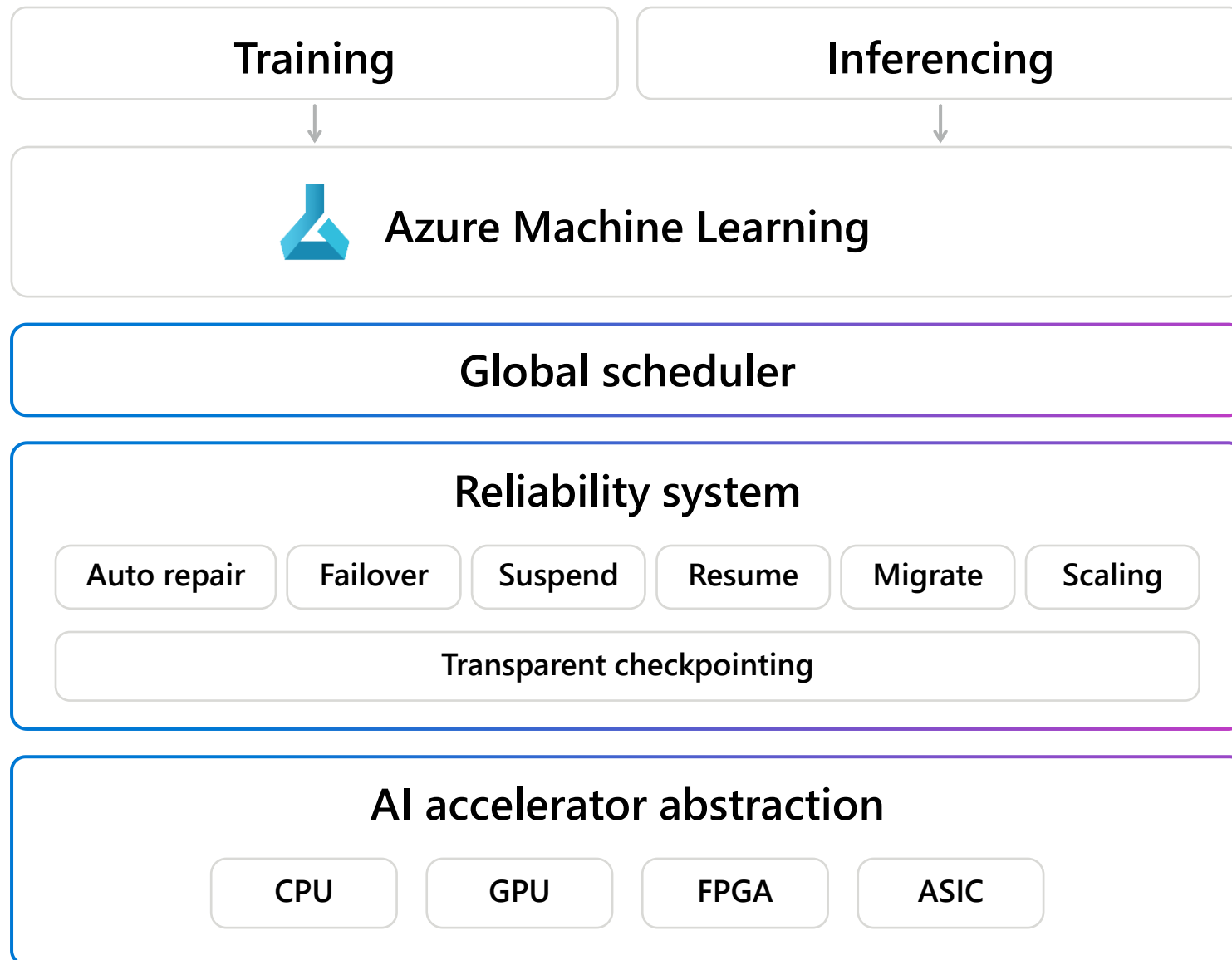
Microsoft Azure
Supercomputers

2024年5月

2023年11月

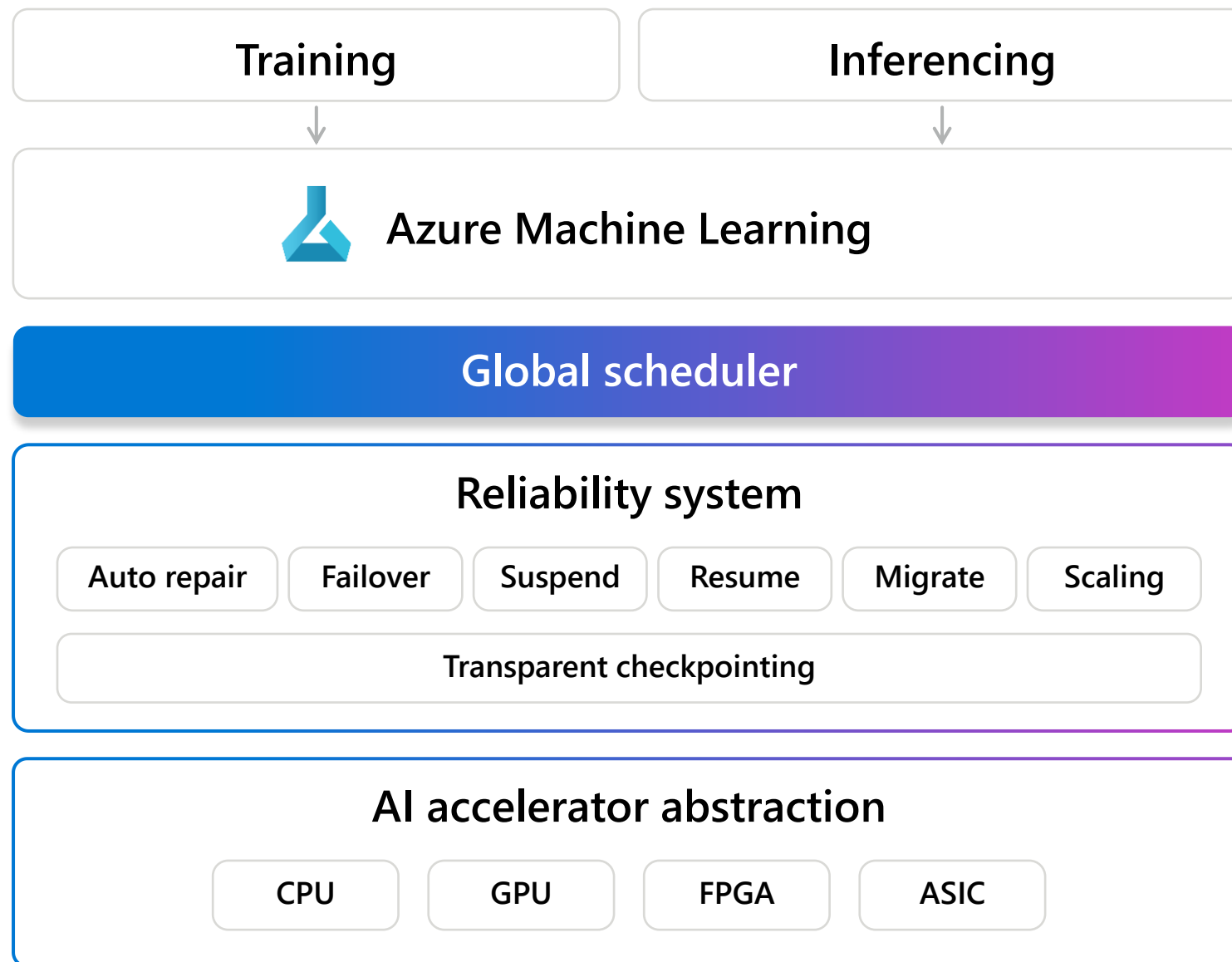
Project Forge

- ✓ リソース使用率を高めるコスト効率の高いインフラストラクチャ
- ✓ 信頼性と耐障害性に優れたAI インフラストラクチャ
- ✓ ワークロードに対しインフラストラクチャと環境を抽象化



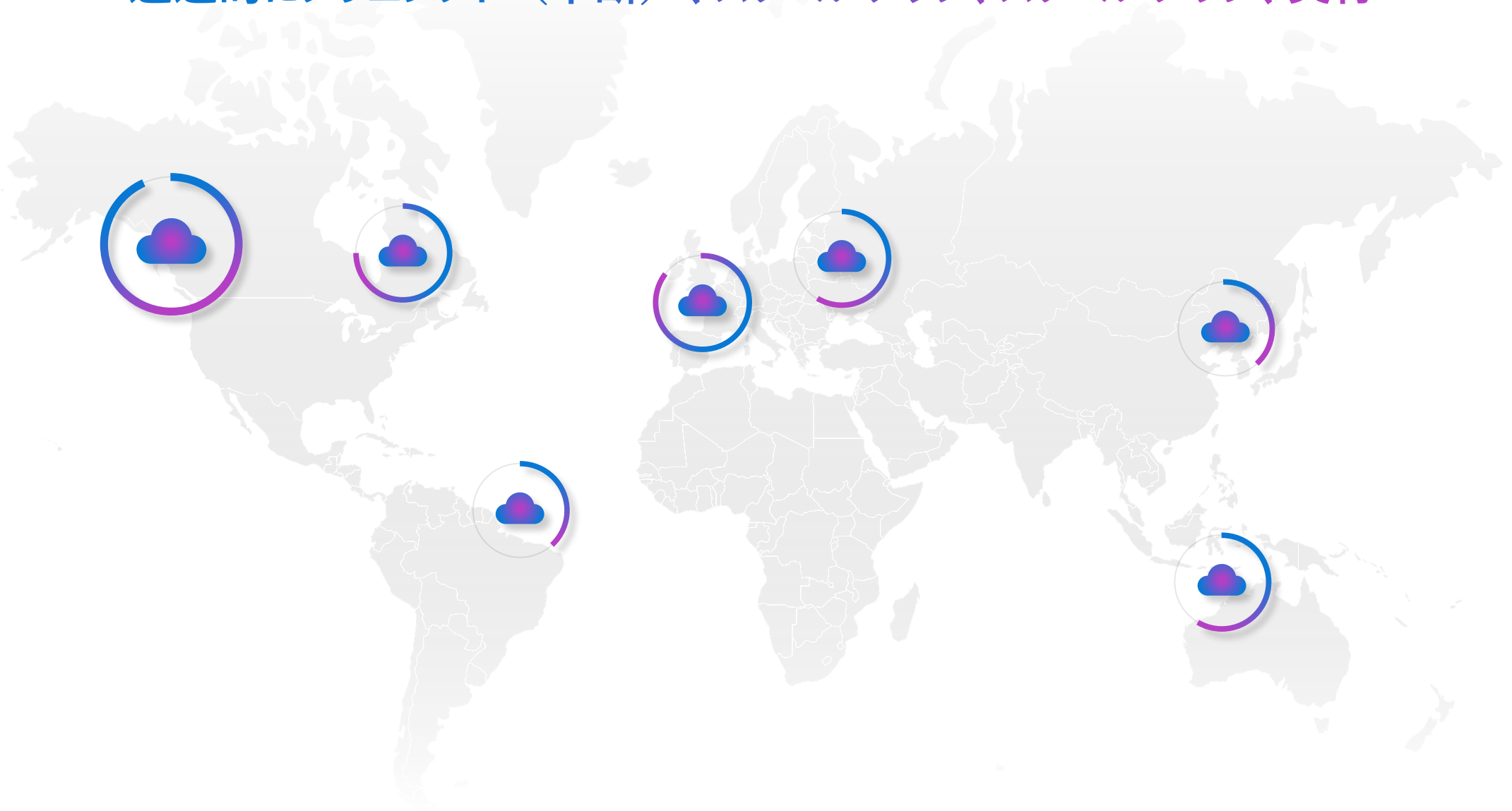
Project Forge

- ✓ リソース使用率を高めるコスト効率の高いインフラストラクチャ
- ✓ 信頼性と耐障害性に優れたAI インフラストラクチャ
- ✓ ワークロードに対しインフラストラクチャと環境を抽象化



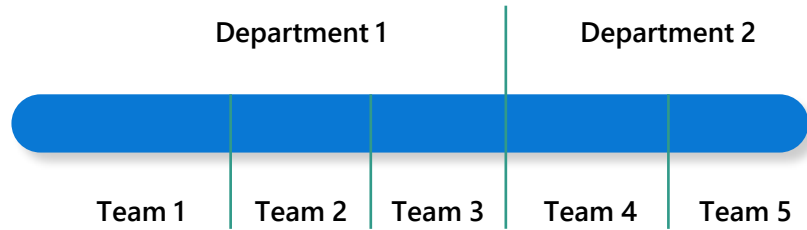
Global scheduler

透過的にプリエンプト（中断）、スケール アップ、スケール ダウン、実行



One Pool

BEFORE



Each team can only use their own capacity



My "Precious"

AFTER

One Pool with reservation – Premium Tier



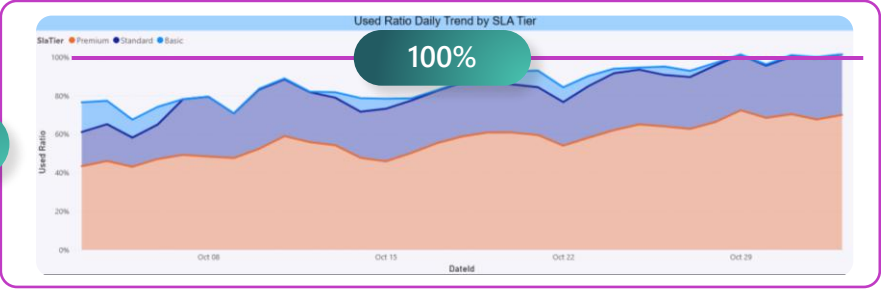
Oversubscription through Standard and Basic Tier



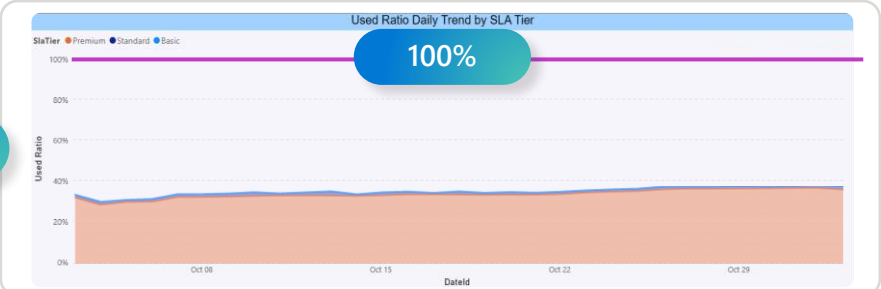
Your GPUs + access to more GPUs
by oversubscription

One Pool

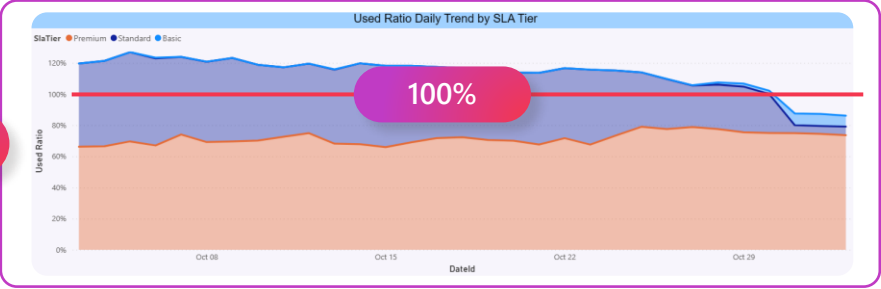
Team-A



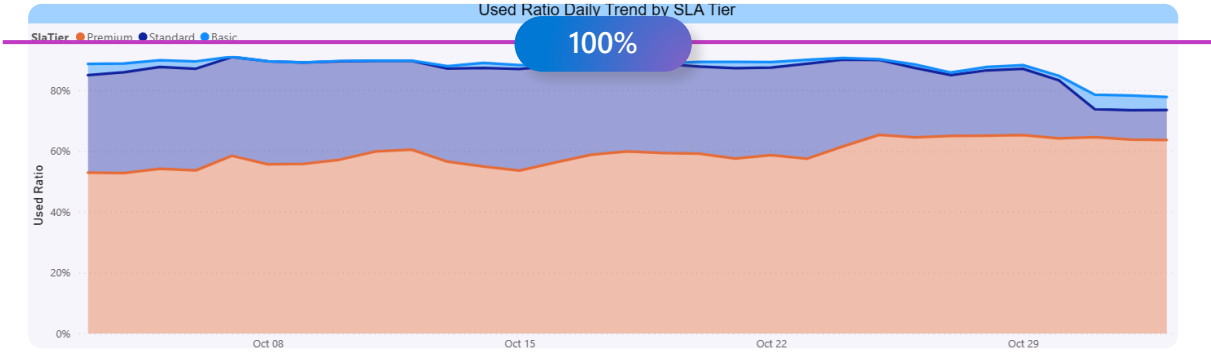
Team-B



Team-C

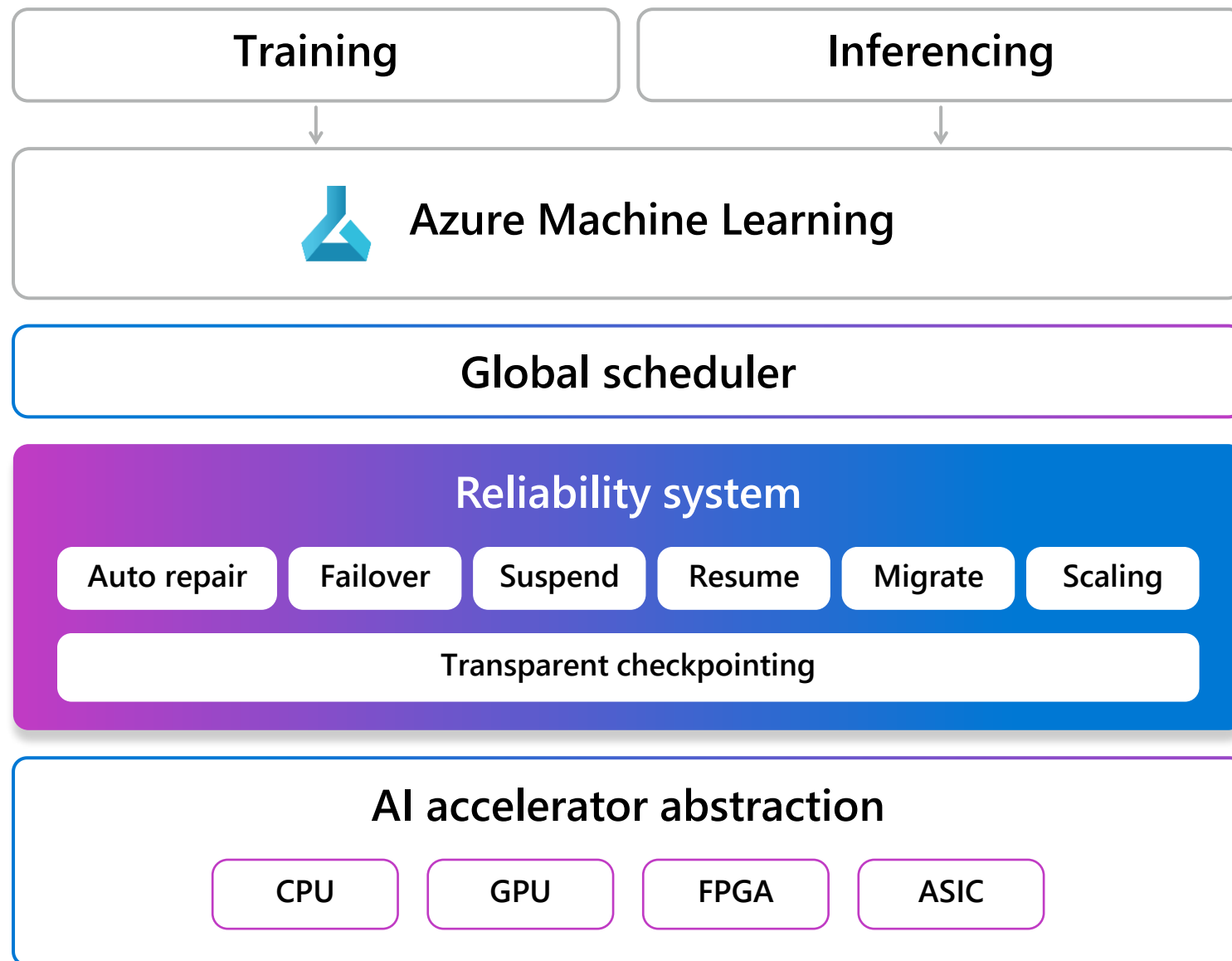


全体の使用率



Project Forge

- ✓ リソース使用率を高めるコスト効率の高いインフラストラクチャ
- ✓ 信頼性と耐障害性に優れたAI インフラストラクチャ
- ✓ ワークロードに対しインフラストラクチャと環境を抽象化



Project Forge 信頼性



自動障害検出



自動再起動



バックグラウンドでの自動修復

Name (run ID)	Experiment	Status	Created on	Duration	GPU count
4c8b86dd-5b57-405d-b438-522004b0...	llmlen-dbc3f56a	✓ Completed	Mar 13, 2024 3:46 PM	6d 7h 33m 20s	256
e664f285-3fa3-482a-b5ad-cf02c5daaa...	Default	✓ Completed	Mar 19, 2024 12:50 PM	5d 20h 37m 54s	1,024
722ffbd8-eccf-4424-9d7e-0562314fcd...	llmlen-dbc3f56a	✓ Completed	May 4, 2024 8:54 PM	5d 6h 32m 21s	256
5f5cdf3-6c9a-4745-a904-00ed03c3b...	phi3_pretraining-327b1213	✓ Completed	Apr 27, 2024 2:58 PM	3d 11h 15m 43s	512
2e20ed7a-90ef-4a85-a1a9-b2b4dfbef...	phi3_pretraining-327b1213	✓ Completed	Apr 27, 2024 3:45 PM	3d 10h 40m 38s	512
ad9cfabc-1eb8-4c6b-9ab9-0fa79b7ce...	phi3_pretraining-327b1213	✓ Completed	Apr 20, 2024 8:37 AM	2d 20h 41m 44s	256
a8c69e69-916b-43bf-8505-26ef4026c...	aims-sing-laptop-a05d4792	✓ Completed	Apr 6, 2024 2:43 AM	3d 6h 58m 16s	512
da4a6ad6-43c3-44d8-8d2a-ebd7fd6e...	phi3_pretraining-327b1213	✓ Completed	Apr 25, 2024 11:17 AM	2d 18h 45m 33s	512
857f67a0-554c-42ff-a2a3-9ee0d531c2...	phi3_pretraining-327b1213	✓ Completed	Apr 19, 2024 2:03 PM	2d 11h 25m 13s	1,024
5f5cdf3-6c9a-4745-a904-00ed03c3b...	phi3_pretraining-327b1213	✓ Completed	Apr 27, 2024 2:58 PM	3d 11h 15m 43s	512

Completed

Apr 19, 2024 2:03 PM

2d 11h 25m 13s

1,024

857f67a0-554c-42ff-a2a3-9ee0d531c2...	phi3_pretraining-327b1213	✓ Completed	Apr 19, 2024 2:03 PM	2d 11h 25m 13s	1,024
ae137e1f-e188-414c-a376-7f228a298a...	lucliu-7d4d42c7	✓ Completed	May 5, 2024 1:49 PM	15h 31m 4s	512

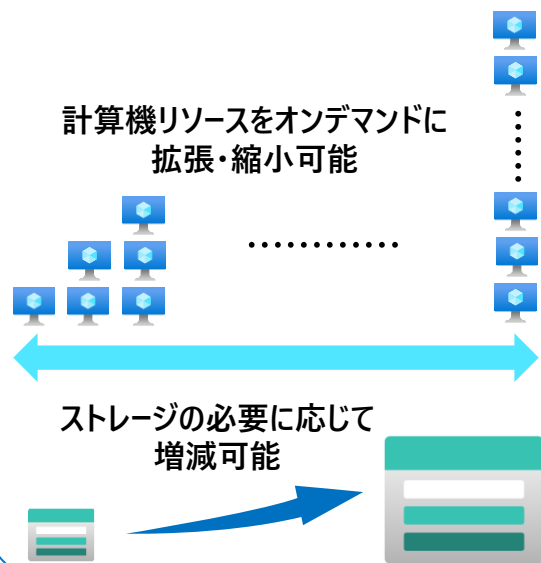
Azure

the world's AI supercomputer

パブリック クラウドのメリット

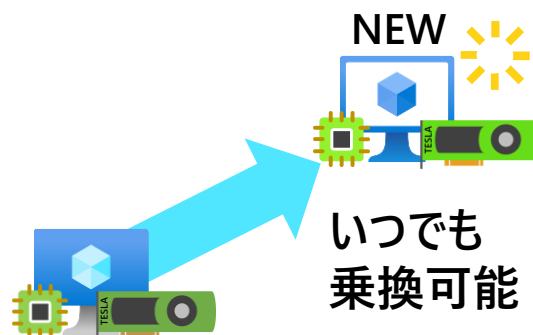
オンデマンド

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

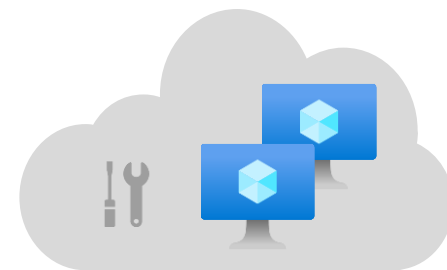
利用に応じて課金される従量課金方式と固定した利用見込みで予約するリザーブドインスタンスで、利用シナリオに応じて従量課金と予約インスタンスを組み合わせることでコスト最適化できます。

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

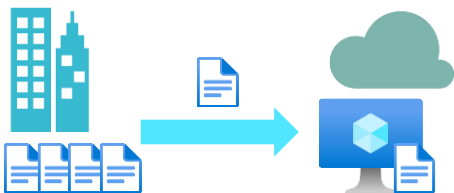
クラウドではハードウェアメンテナンスはクラウド事業者が行うため、ユーザーによるハードウェアメンテナンスは不要です。万が一、ハードウェアの障害が発生しても正常なリソースに移行して運用を継続することができます。



クラウド HPC の利用シナリオ

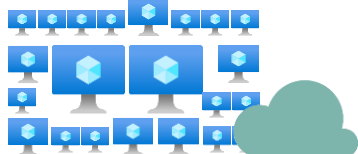
クラウド バースティング

オンプレミス環境の計算リソースが混雑し、ジョブが滞留し長時間の待ちが発生する場合、クラウドの計算リソースにジョブを転送することでジョブの待ち時間を削減することが可能になります。また、クラウド上の計算リソースはジョブが実行されない時は停止し、ジョブが実行されるときのみ起動させることが可能です。



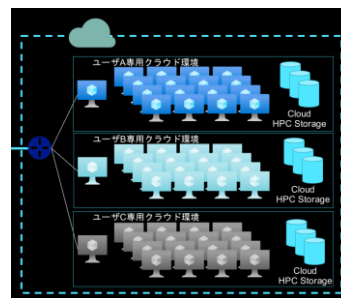
小規模ジョブの クラウド活用

小規模なジョブは、オンプレミス環境ではノードを占有しコアを余らせて実行される場合があります。クラウドの仮想マシンは物理コア数を最小で1コアから利用可能で、ジョブのサイズに応じた仮想マシンを利用することが可能です。また、大量の小規模ジョブに対して、クラウドの計算リソースを一時的に大量に用意することによりスループット性能を向上させることができます。



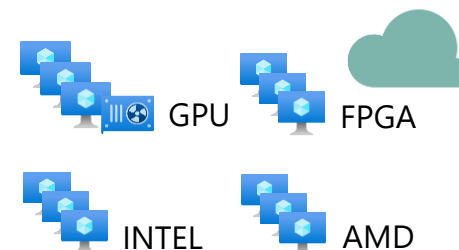
ユーザ毎の 専用クラウド環境

クラウド環境に、ユーザ毎の専用環境を構築することが可能です。ユーザ毎の専用環境を利用することで、個別のカスタマイズが可能になり、他のユーザの影響を受けず計算機環境をご利用いただけます。また、ユーザ毎にサブスクリプションを用意することにより、サブスクリプションごとに課金することが可能になります。



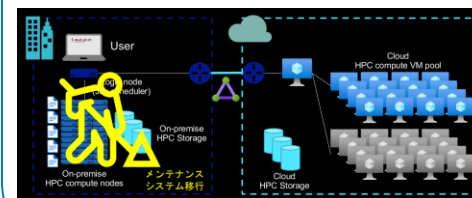
特殊な 計算機環境

オンプレミス環境に通常のCPUで構成されたクラスターを保有している場合、オンプレミス環境にはないGPUやFPGAといった演算加速装置搭載の環境をクラウド上に構成することが可能です。また、オンプレミス環境のCPUと異なるCPUの環境をクラウド上で利用することも可能です。



オンプレ停止時の 計算リソース

オンプレミス環境のメンテナンスやシステム移行時の計算環境としてクラウド環境を使用します。オンプレミス環境が利用できない期間の前後にデータ転送期間を設け、必要なデータのみを転送を行います。クラウド環境もオンプレミス環境と共通のジョブスケジューラーを用いて利用方法をオンプレ環境と共通にすることも可能です。



Azure 上の MPI ワークロードの実行オプション

マネージド サービス

DIY



Batch

- Azure ネイティブ (1P) スケジューラ、フル マネージド サービス
- E2E のライフサイクル管理
- SaaS の活用

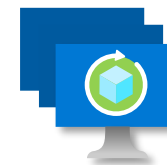
最適化された体験



CycleCloud

- サード パーティ (3P) スケジューラの依存関係
- ジョブのオーケストレーションに対する有限の制御
- ハイブリッド/バースト展開

スケーラビリティ



仮想マシン (IaaS)

- 自組織製スケジューラ
- カスタマイズされたワークフローマネージャー
- 自組織製/カスタム アプリケーション

カスタムメイド

Azure CycleCloud

ユーザー エンパワーメント

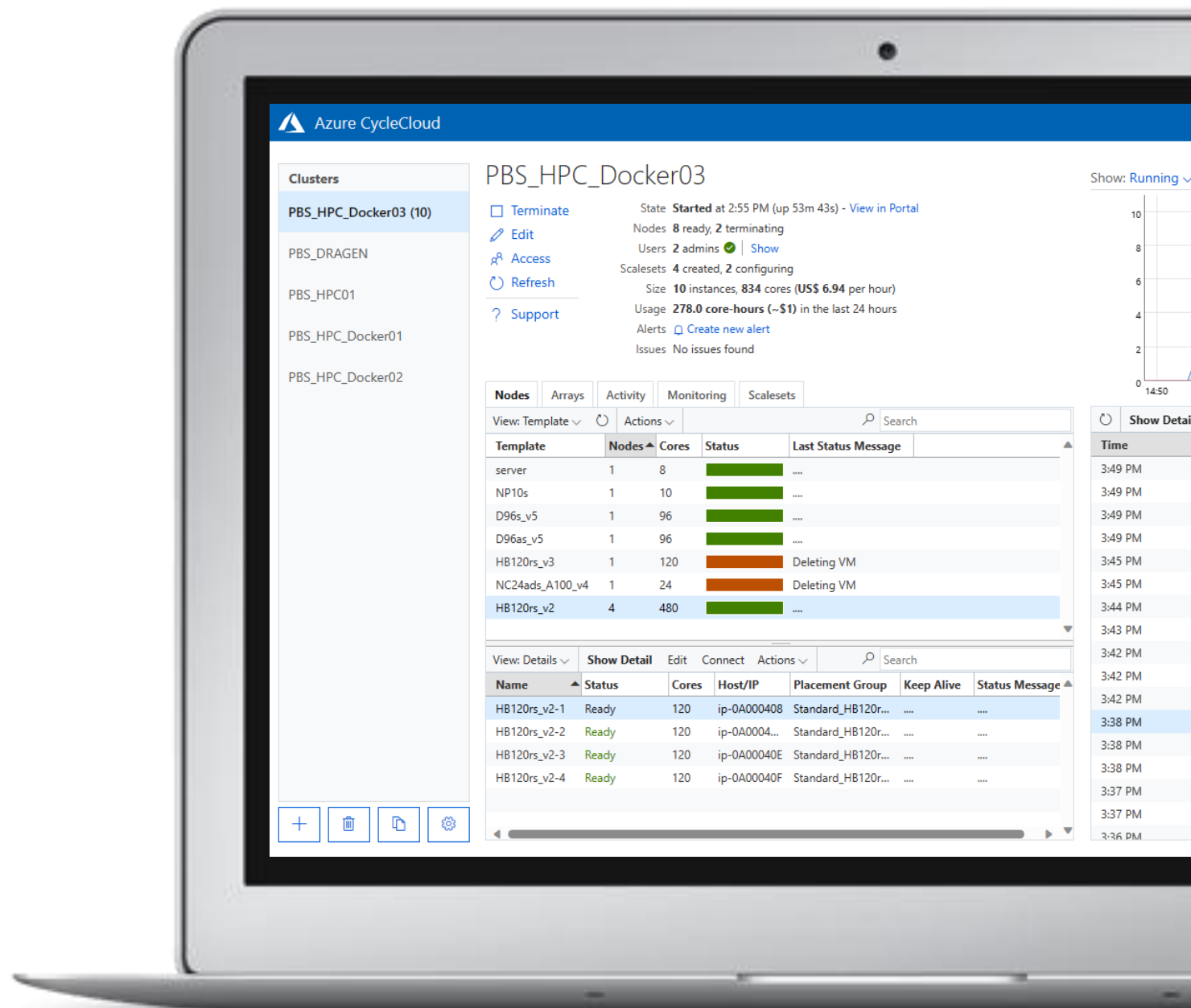
- 既存のワークロードとスケジューラーをクラウド化可能
- リソースへ即時アクセスが可能
- オートスケーリング機能、エラー処理

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 を利用可能
- 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を提供



Azure 仮想マシンの AI ポートフォリオ

Azure インスタンス	NCasT4_v3	NC A100 v4	NC H100 v5	NDm A100 v5	ND H100 v5
コア数	4, 8, 16, 64	24, 48, 96	40, 80	96	96
GPU	1-4x Tesla T4	1-4x A100 GPU, PCIe	1-2x H100 GPU, PCIe	8x A100 GPU, SXM	8x H100 GPU, SXM
メモリー	28/56/110/440 GB	220/440/880 GB	320/640 GB	1900 GB	1900 GB
ローカル ディスク	180/360/2880 GB SSD	1123/2246/4492 GB SSD	4/8 TB SSD	6.4 TB SSD	36 TB SSD
ネットワーク	Azure Network	Azure Network + NVLink GPU Interconnect (pair)	Azure Network + NVLink GPU Interconnect (pair)	Azure Network + InfiniBand EDR + NVLink GPU Interconnect	Azure Network + InfiniBand NDR + NVLink GPU Interconnect
 リアルタイム推論と 低コスト コンピューティング		 ミッドレンジトレーニングと 高密度推論		 分散トレーニングと生成推論	

利用可能な全ての VM シリーズの一覧は以下を参照

[仮想マシンのサイズの概要 - Azure Virtual Machines | Microsoft Learn](#)

Azure HPC ストレージ ソリューション

		Azure Standard Blob	Azure Premium Blob	Azure Premium Files	Azure NetApp Files	Azure Managed Lustre
パフォーマンスと スケール	容量	20+ PiB	20+ PiB	100 TiB	500 TiB	768 TiB
	帯域	15 GB/s (default)	15 GB/s (default)	10 GB/s	4.5 GiB/s	Up to 384 GB/s
	IOPS	20,000 (default)	20,000 (default)	100,000	460,000	>100,000
	遅延	<100 ms	<10 ms	3-5 ms	<1 ms	<2 ms
プロトコル		REST SFTP FUSE HDFS NFSv3 CSI	REST SFTP FUSE HDFS NFSv3 CSI	REST NFSv4.1 (or) SMB CSI	NFSv3 & SMB3* NFSv4.1 & SMB* CSI	Lustre CSI
コスト		\$	\$\$\$	\$\$\$	\$\$\$\$	\$\$\$
		Core			Either	Accelerator

*同一ボリュームへの同時マルチ プロ
トコル アクセス

Azure Managed Lustre File system

Lustre ファイル システムは、HPC 業界で広く利用されている分散ファイルシステムです。高いスループット性能を提供しながら、大規模なストレージ サイズにスケーリングできるオープンソースの並列ファイル システムです。Azure Managed Lustre File system では、**クラウド上に必要なサイズ、必要な性能を設定し、Lustreファイルシステムを素早く展開することができます。**

- Azure Blob Storage との統合
 - Lustre 階層ストレージ管理 (HSM)
 - BLOB コンテナからファイルのインポートすることが可能
 - ジョブが終了したら変更されたデータを Blob にエクスポート
- HPC オークストレーションツールとの連携
 - Azure CycleCloud
 - Azure Batch
 - Azure Kubernetes Service



```
[msadmin@admin ~]$ lfs df -h
```

UUID	bytes	Used	Available	Use%	Mounted on
lustrefs-MDT0000_UUID	598.9G	5.7M	548.3G	1%	/azure-lustre-mount[MDT:0]
lustrefs-OST0000_UUID	15.8T	1.2M	15.0T	1%	/azure-lustre-mount[OST:0]
lustrefs-OST0001_UUID	15.8T	1.2M	15.0T	1%	/azure-lustre-mount[OST:1]
lustrefs-OST0002_UUID	15.8T	1.2M	15.0T	1%	/azure-lustre-mount[OST:2]
lustrefs-OST0003_UUID	15.8T	1.2M	15.0T	1%	/azure-lustre-mount[OST:3]
lustrefs-OST0004_UUID	15.8T	1.2M	15.0T	1%	/azure-lustre-mount[OST:4]
lustrefs-OST0005_UUID	15.8T	1.2M	15.0T	1%	/azure-lustre-mount[OST:5]
lustrefs-OST0006_UUID	15.8T	1.2M	15.0T	1%	/azure-lustre-mount[OST:6]
lustrefs-OST0007_UUID	15.8T	1.2M	15.0T	1%	/azure-lustre-mount[OST:7]
filesystem_summary:	126.7T	9.9M	120.3T	1%	/azure-lustre-mount

クラウド HPC の実現イメージ

大学・研究機関・企業

ユーザ

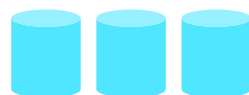


オンプレミスのジョブスケジューラとの連携可能
(ジョブスケジューラにより作りこみが必要な場合があります)

ログインノード

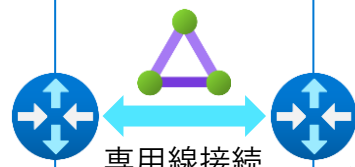


スーパーコンピュータ



ストレージ

Azure ExpressRoute



専用線接続でセキュアに接続
(無制限データプランを選択可能)

Microsoft Azure



Azure CycleCloud

Azure HPC 環境
オーケストレーション

クラウド側の専用ログインノード、ジョブスケジューラの利用可能



Azure NetApp Files



Azure Managed Lustre

クラウド側
共有ファイルシステム

予算アラートによる自動化



ジョブによるリソース要求に合わせてオートスケール

クラウドHPCクラスター



IB マルチノード実行用計算ノード
通常ジョブ(従量課金)

通常ジョブとスポットを選択しコスト削減



IB マルチノード実行用計算ノード
スポットジョブ(スポット料金)



マスターノード



GPU 計算ノード

CPU/GPUなど最新のリソースに乗換や追加が可能



小規模計算用ノード
通常ジョブ(従量課金)



小規模計算用ノード
スポットジョブ(スポット料金)

小規模計算ではジョブが要求したリソース(コア数, メモリ容量)に合わせて
VM のサイズを調整可能

Azure HPC 事例 - GENIAC プロジェクト

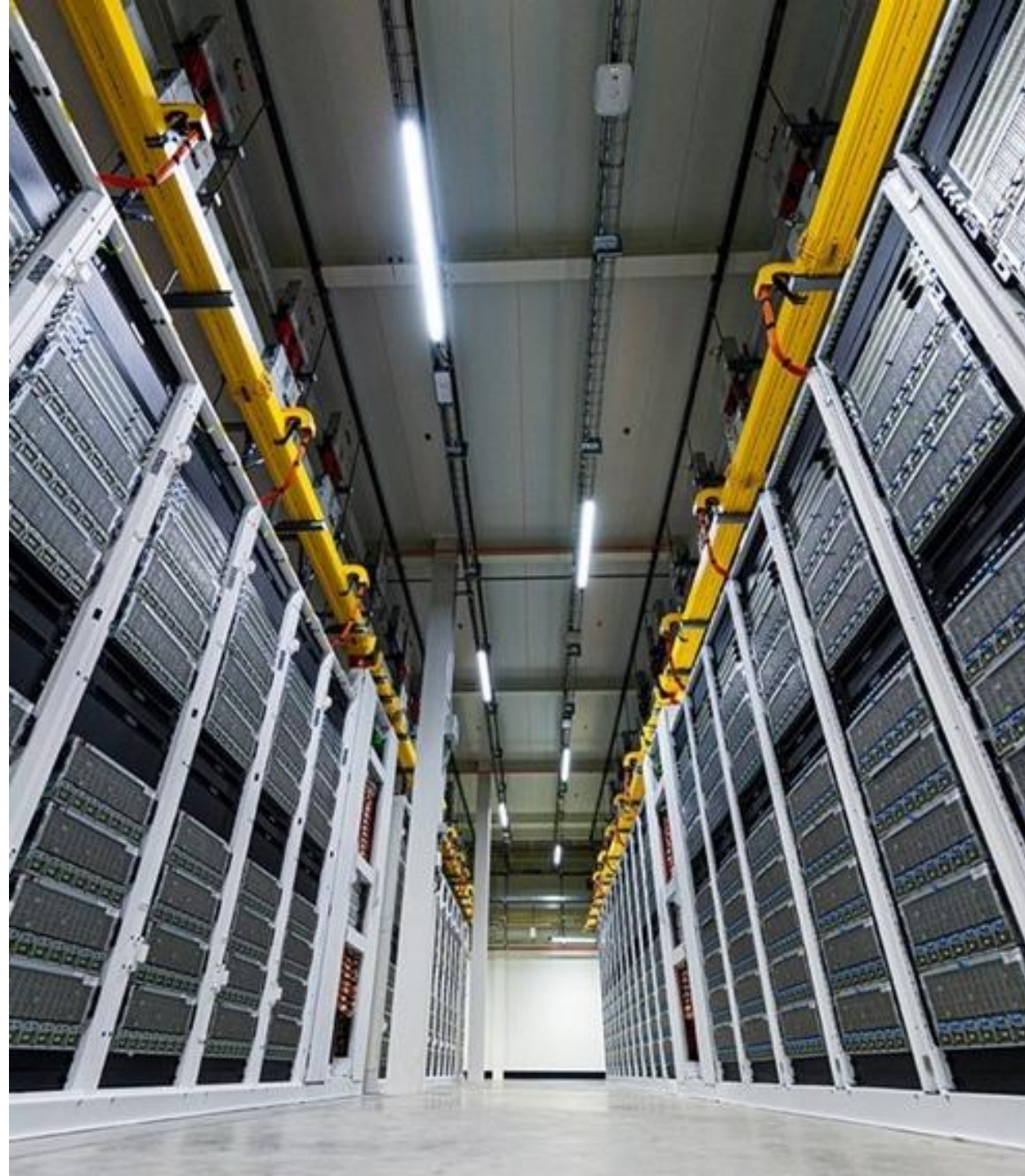
経済産業省及び国立研究開発法人新エネルギー・産業技術総合開発機構は、産業活動や国民生活に大きな影響を与えると考えられる生成 AI の開発力を強化していくため、令和 6 年 2 月 2 日に「GENIAC」プロジェクトを開始しています。

日本マイクロソフトは生成 AI のコア技術である基盤モデルの開発に対する計算資源の提供者として選定され、Microsoft Azure が本プロジェクトにて利用されています。

参考情報

GENIACにおける計算資源の提供支援について、計算資源提供者として Microsoft社を選定し、追加公募を開始しました（METI/経済産業省）

マイクロソフトの計算資源を経済産業省と NEDO が公募する生成 AI 基盤モデル開発プロジェクトに提供 - News Center Japan (microsoft.com)



まとめ

Eagle 一般提供している Azure サービスをベースに構築された世界最速のクラウド型スーパーコンピュータ（の一部）

Azure AI インフラストラクチャー Azure HPC 上で実行される Microsoft が提供する多数の AI サービスの基盤

Azure HPC ワークロードの幅広い実行オプションやストレージオプションを提供し、最新ハードウェアやスケーラビリティ、経済性などのメリットを提供

Azure high-performance computing

HPC と AI のための最先端のクラウドで
イノベーションを解き放つ

Learn more <https://learn.microsoft.com/ja-jp/azure/high-performance-computing/>

お問い合わせはこちら

■営業お問い合わせ先

Tel:0120-167-400 時間 : 9:00 - 17:30 (月～金)

URL: <https://azure.microsoft.com/ja-jp/contact/>