

PCクラウドコンソーシアム HPCクラウド部会 第3回ワークショップ

生成AIを活用して加速するHPCクラウド戦略

佐々木 啓

アマゾン ウェブ サービス ジャパン合同会社

パブリックセクター 教育・研究技術本部

シニアソリューションアーキテクト

2026/2/25



自己紹介

佐々木 啓 / Sasaki Kei

アマゾン ウェブ サービス ジャパン 合同会社
パブリックセクター技術統括本部 教育・研究技術本部
シニアソリューションアーキテクト



- 大学や研究機関のお客様に対して研究活動、教育現場、経営事務のクラウド化を検討するための技術支援を担当
- お客様が最新のクラウド技術を効果的に活用するための専門的なアドバイスやソリューションの選定をサポート



好きなAWSサービス : AWS ParallelCluster、Kiro

クラウドの最近



HPC on Cloud の急成長

- HPC on Cloud は、過去成長率年間約 7-8 %
だったが、2024 年成長率は **23.5 %**
- AI が成長の主な原因ではあるが、
従来の HPC ワークロードも成長率 **8.4 %**

今後もこのトレンドは拡大予定

HPC on Cloud は
急成長し続ける領域



HPC/AI Market Trends: size and growth

23.5%

2024 Growth

Fastest growth in 20+ Years!
Historically, HPC/AI grew at ~7-8% per year*

~\$60B

2024 Total HPC/AI
Spending

~\$10 billion on cloud HPC *



HPC without newer AI workloads

- Projected to ~\$98B by 2033
- ~8.4% CAGR over forecast period

AI infra is a huge, fast-growing pool

- Global AI infra spending to reach ~\$758B by 2029
- Q2'25 AI hardware spend +166% YoY to \$82B
- >90% of AI server spend already on accelerated systems

**HPC and AI
are growing
individually
and together**



米国政府にAIとHPCインフラに最大 500 億ドルの投資を発表

10 年間かけて最大 500 億ドルの投資を行い、政府向けのAI/HPC 関連インフラを構築する計画

Amazon to invest up to \$50 billion to expand AI and supercomputing infrastructure for US government agencies

Company news Artificial Intelligence AWS Amazon Data Centers

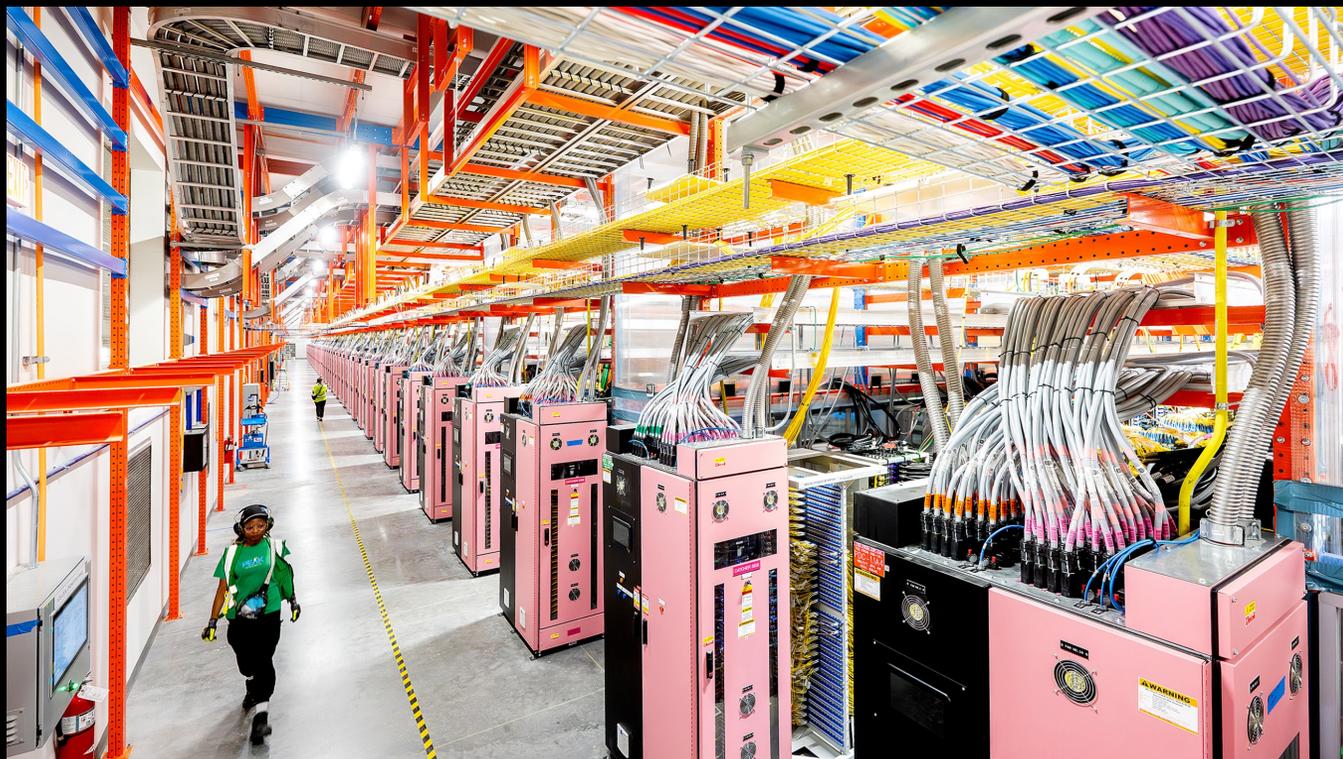
Share



- 2026年着工、約1.3GWのAI・スパコン容量をAWS Top Secret/Secret / GovCloud (US) リージョンに追加
- 提供サービス
 - Amazon SageMaker AI (モデル訓練・カスタマイズ)
 - Amazon Bedrock (モデル・エージェントデプロイ)
 - Amazon Nova、Anthropic Claude、生成AIモデル
 - AWS Trainium AIチップ、NVIDIA AIインフラ
- 国家安全保障、科学研究・イノベーション、自律システム開発
- サイバーセキュリティ、エネルギーイノベーション、ヘルスケア
- 期待される効果
 - シミュレーション×AIの統合で、数週間～数ヶ月かかっていた分析を数時間に短縮
 - 従来のHPCワークフローからAI加速型ディスクバリエーションへの転換

Project Rainier : 世界最大級のAIコンピューティングクラスターが稼働開始(2025/10)

記録的な速さで約50万個のTrainium2チップが供給され、Anthropic社は2025年末までに100万個以上のチップにまで拡大

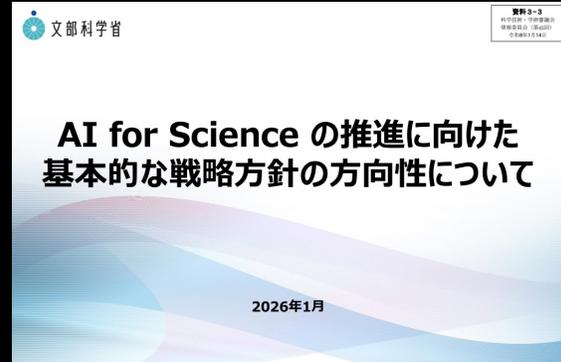


<https://www.aboutamazon.com/news/aws/aws-project-rainier-ai-trainium-chips-compute-cluster>



国内: 基盤AIの学習環境 / AIサービスの利活用

国内の取り組み



研究へのAI利活用、国際連携



GENIAC: 基盤モデル開発支援事業

AWSの支援活動



AI JAMへのAI企業として参加、ハッカソン支援(一例)



支援プログラムやAI活用ロードマップ実行の技術支援

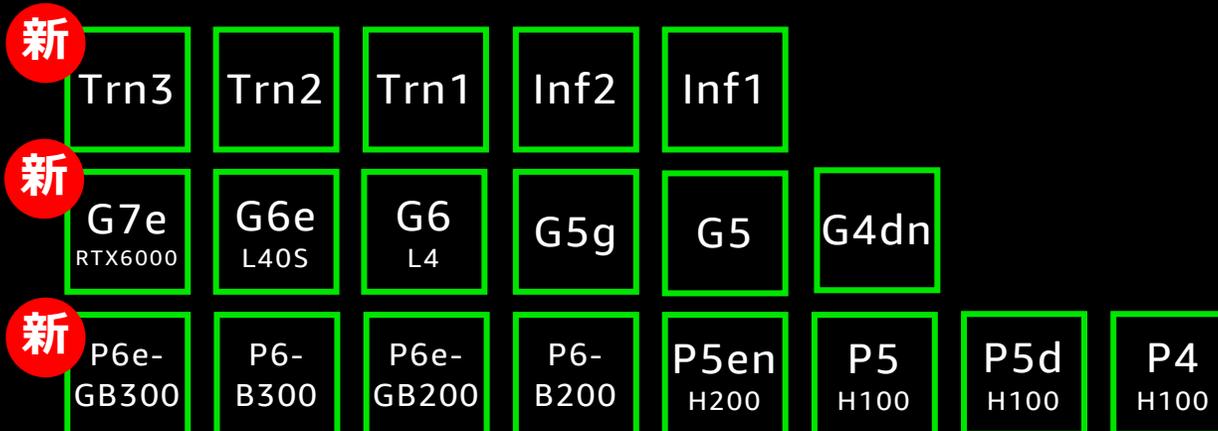


最新のサービスアップデート

HPC/AI/ML向けの EC2 インスタンス

1. アクセラレータ搭載 (高速化コンピューティング)

GPU、AWS ML アクセラレータ、
FPGA 搭載インスタンス



2. HPC 最適化

HPC クラスタや分子動力学解析など、
ノード間の高速ネットワークを要す
るワークロードに最適



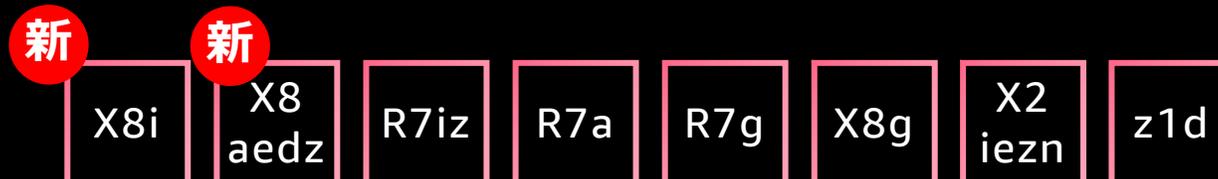
3. コンピュート最適化

小中規模の科学計算など、
単一インスタンスでの高CPUタスク



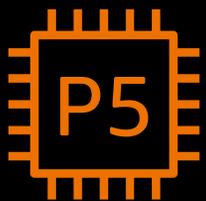
4. メモリ最適化

半導体シミュレーションなど

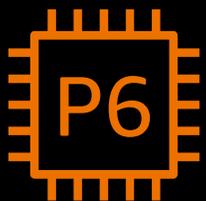


機械学習トレーニング向けインスタンス

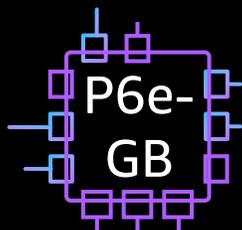
Nvidia GPU 搭載の、機械学習のトレーニングに特化した最上位 P5, P6, P6e-GB インスタンスファミリー



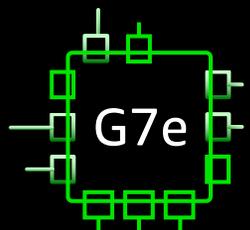
- NVIDIA H100/H200 を最大 8 基搭載
- H100は **単一 (シングル) GPU 検証**も可能



- NVIDIA Blackwell B200/B300 を 8 基搭載
- P5 からの既存の分散学習基盤をそのままBlackwell 世代にリフレッシュしやすい
- Capacity Blocks for ML / HyperPodにて利用が可能



- GB200/GB300 NVL72 (Grace Blackwell) 搭載のUltraServers (1 ラック = 1 インスタンス)
- 兆パラメータ級の大規模な一括 LLM 学習・推論に最適化された設計
- P6e-GB300はGPUメモリ(288 GB HBM3e/1枚)、P6e-GB200比で1.5倍のFP4演算能力(スパース性利用なし)

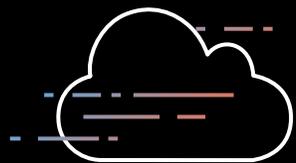


- **GTX Pro 6000 Brackwell**を **1/2/4/8 基**搭載したGPUモデル
- GPUメモリ (**96GB/GPU**)、単一GPUで 最大70Bパラメータ モデルをFP8で実行可能
- ネットワーク帯域幅 最大1,600Gbps / EFA経由 GPUDirect RDMA 対応

Cloud Continuum: 必要な場所にどこでもAWSを

← 一貫したエクスペリエンスを実現する同じインフラストラクチャ、サービス、API、ツール →

REGIONS



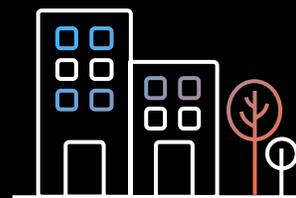
AWS Regions

METRO AREAS & TELCO NETWORKS



Amazon CloudFront
AWS Local Zones

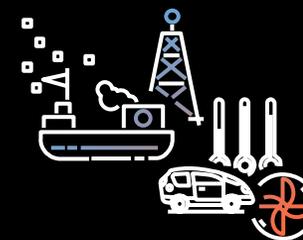
ON-PREMISES



AWS Outposts
AWS Dedicated Local Zones
Amazon EKS Hybrid Nodes

New **AWS AI Factories**

FAR EDGE



Amazon EKS Anywhere
Amazon ECS Anywhere
AWS IoT

AWS AI Factory は、お客様のデータセンターと電力を活用しながら、最新のAIチップ・高速ネットワーク・高性能ストレージをフルマネージドで提供し、BedrockやSageMakerを含むAWSのAIサービスをオンプレミスでもクラウドと同じセキュリティと運用体験で利用可能にする、専用AIインフラストラクチャです。

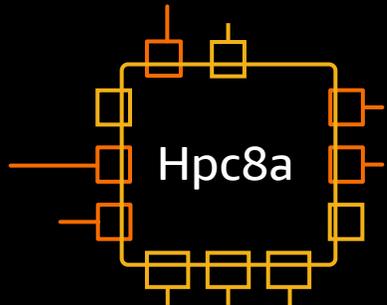


© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved.

<https://www.aboutamazon.com/news/aws/aws-data-centers-ai-factories>

Amazon EC2 Hpc8a instances

前世代 (Hpc7a) 比で性能 最大40%向上、メモリ帯域幅 42%増加、コストパフォーマンスで最大25%向上



- AMD EPYC 第5世代 "Turin" プロセッサ搭載、最大周波数 4.5GHz、SMT無効
- 1インスタンスあたり 192物理コア、768 GiB メモリ
- 最大300 Gbps Elastic Fabric Adapter (EFA) 帯域
- 第6世代 AWS Nitro System 搭載
- 96xlargeの1サイズ構成、ワークロードに応じてコア数を調整しコアあたりメモリ比を最適化可能
- 対象ワークロード：計算流体力学 (CFD)、高解像度気象モデリング、衝突シミュレーション、創薬 等
- 提供リージョン：US East (Ohio)、Europe (Stockholm), etc

Amazon EC2 X8i Instance

Intel Xeon6プロセッサ搭載、メモリ最適化インスタンス

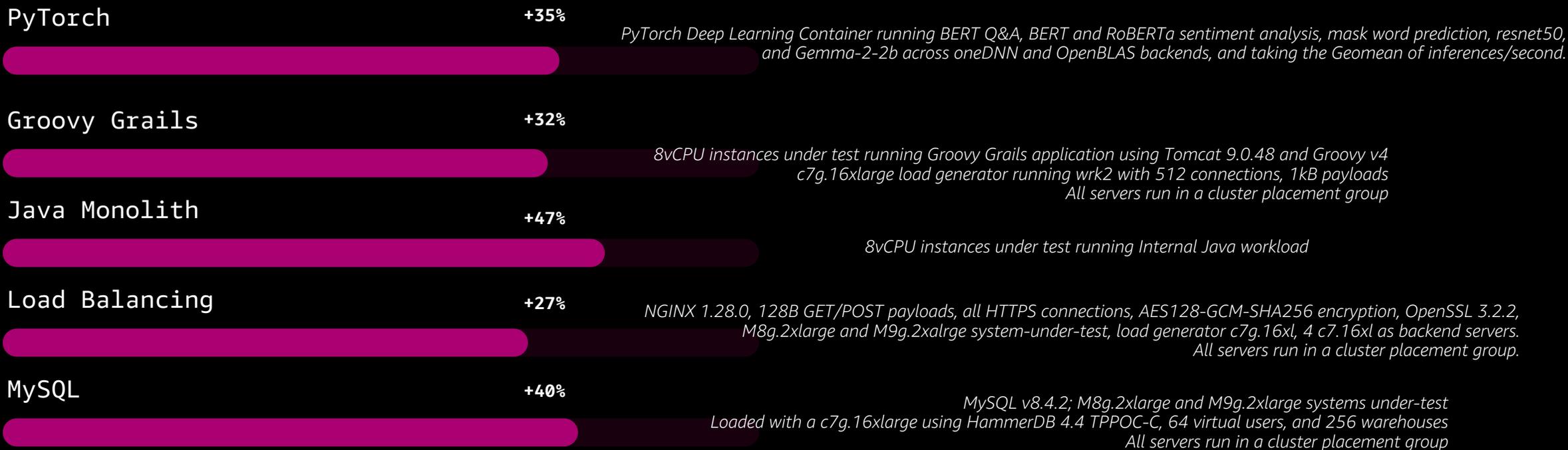
- カスタム版 Intel Xeon 6 プロセッサ搭載、同等 Intel プロセッサ比で**クラウド最高のパフォーマンスと最速メモリアクセス**を提供
- 旧世代 X2i 比でメモリ容量 最大1.5倍（最大6TB）、メモリ帯域幅 最大3.4倍
- X2i 比で全体性能 最大43%向上、SAP認定 SAPS 性能値 最大50%向上
- PostgreSQL 最大47%、Memcached 最大88%、AI推論 最大46% 向上
- SAP HANA、インメモリデータベース、大規模データベース、EDA（半導体設計）に最適
- 提供リージョン：バージニア北部、オハイオ、オレゴン、フランクフルト

<https://aws.amazon.com/jp/blogs/news/amazon-ec2-x8i-instances-powered-by-custom-intel-xeon-6-processors-are-generally-available-for-memory-intensive-workloads/>

AWS Graviton5

第5世代カスタムプロセッサ

- [M9g](#) がプレビュー中 (C9g・R9g は 2026 年予定))
- M9g インスタンスで Graviton4比 25% 高速、DB30%・Web35%・ML35% の性能向上を実現
- シングルソケット192 コア/チップ、ネットワーク・EBS帯域幅も大幅向上
- **Nitro Isolation Engine** で数学的に証明された分離セキュリティ、

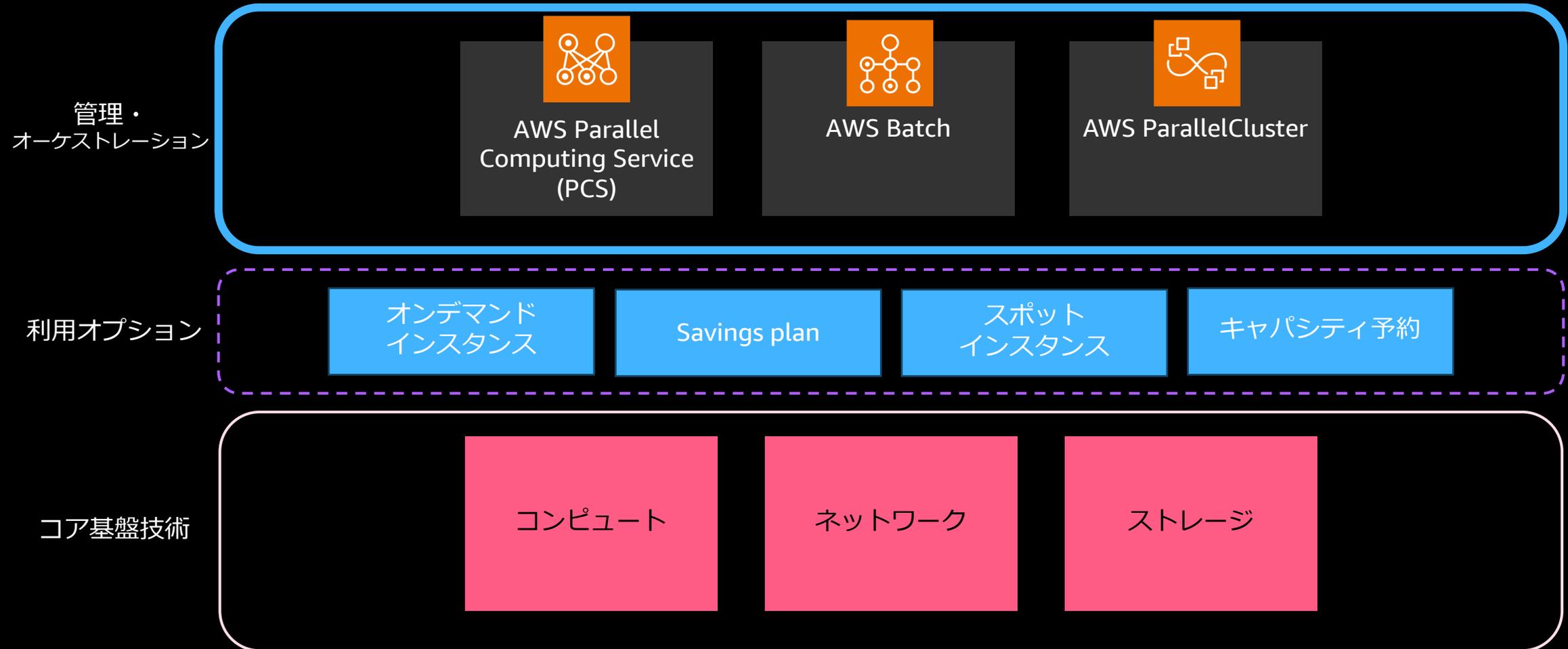


※ M8g と M9g の比較



オーケストレーションサービス

HPC on AWS におけるビルディングブロック



オーケストレーション関連ツールの位置付け

HPC ジョブスケジューラー

	アンマネージド (ツール)	マネージド
独自スケジューラー		 AWS Batch コンテナ化されたワークロードの計画、スケジューリング、実行を行うフルマネージドのバッチコンピューティングサービス
HPC スケジューラー (Slurm)	 AWS ParallelCluster CLI/API/GUI から Slurm 環境を自動デプロイできるオープンソースのクラスター管理ツール	 AWS PCS クラウド専門知識なしでHPCクラターを迅速構築・運用可能なフルマネージドSlurmサービス

R&D 向け Webポータル構築ツール



RES

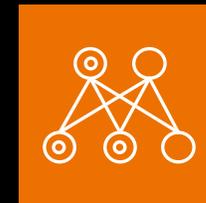
R&Dチームがクラウドの専門知識なしで、エンジニアリング用デスクトップ環境を作成・管理可能なオープンソースのウェブポータル構築ツール

ML 向け



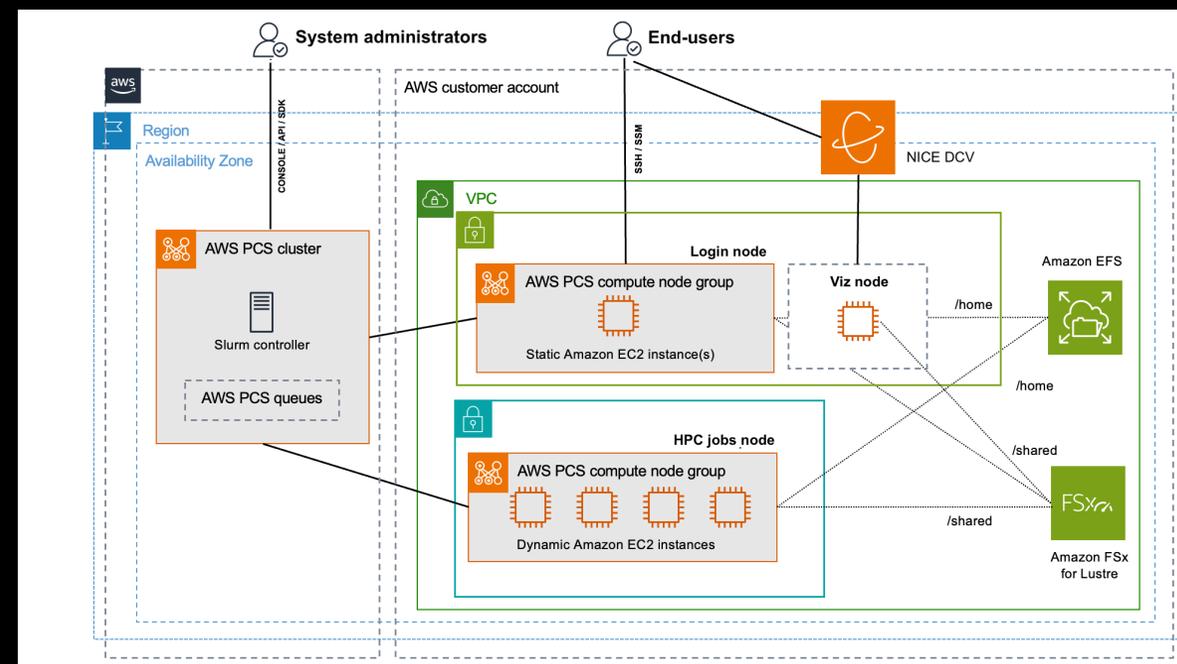
Amazon SageMaker
HyperPod

AWS Parallel Computing Service (AWS PCS)



HPC クラスターの設定と管理を支援する新しいマネージドサービス

- **Slurmスケジューラーを使用** アカウンティング, カスタム設定(Fair-share/QoS/プリエンブション)
- **AWSマネジメントコンソール、SDK、CLIを通じてアクセス可能**
- ノード数が負荷に応じて**自動的にスケール**
- 各ノードのインスタンスタイプとソフトウェアは**自由にカスタマイズ可能**
- **各種ファイルストレージ** (Amazon FSx for Lustre等) への接続
- **閉域網**で利用でき、AWS PrivateLink によるAPI呼び出しにも対応、IPv6対応
- AWS の標準的な**ロギングサービス** (Amazon CloudWatch, AWS CloudTrail) に対応
- **HIPPA適格** (ヘルスケア・ライフサイエンス領域にも対応)
- AWS CloudFormationやTerraform を使用することでクラスター管理タスクを自動化

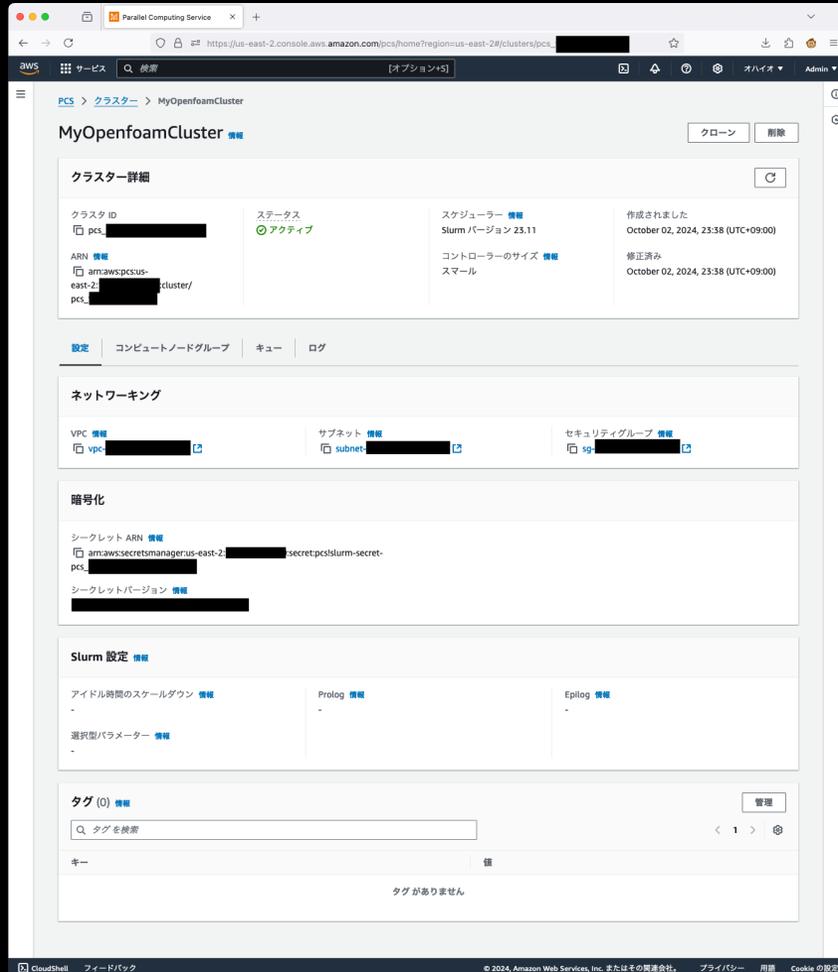


<https://aws.amazon.com/blogs/aws/announcing-aws-parallel-computing-service-to-run-hpc-workloads-at-virtually-any-scale/>

AWS PCS の操作方法



マネジメントコンソールによる操作



AWS CLI による操作

```
$ aws pcs create-cluster --cluster-name "MyOpenfoamCluster" ...
```

```
$ aws pcs create-compute-node-group --cluster-identifier pcs_ xxxxxxxxxx --compute-node-group-name 'login' --scaling-configuration minInstanceCount=1,maxInstanceCount=1 --instance-configs instanceType=c6i.xlarge ...
```

...

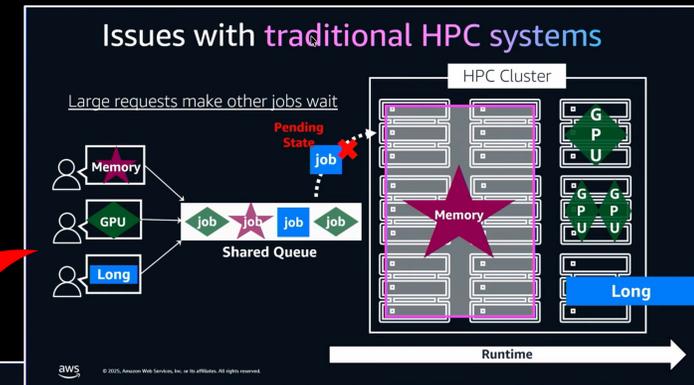


AWS SDK により
各種プログラミング言語での
操作にも対応

豊田中央研究所: ParallelCluster から PCS へ移行し様々な計算ニーズへ対応

- 様々な HPC ワークロードのニーズがある豊田中央研究所が、シンプルな EC2 の利用から開始して、ParallelCluster から PCS へと移行した経緯を紹介。
- PCS では研究者の計算実行準備にかかる手間が大幅に削減され (30 分で準備完了)、HPC on AWS が社内で浸透した。

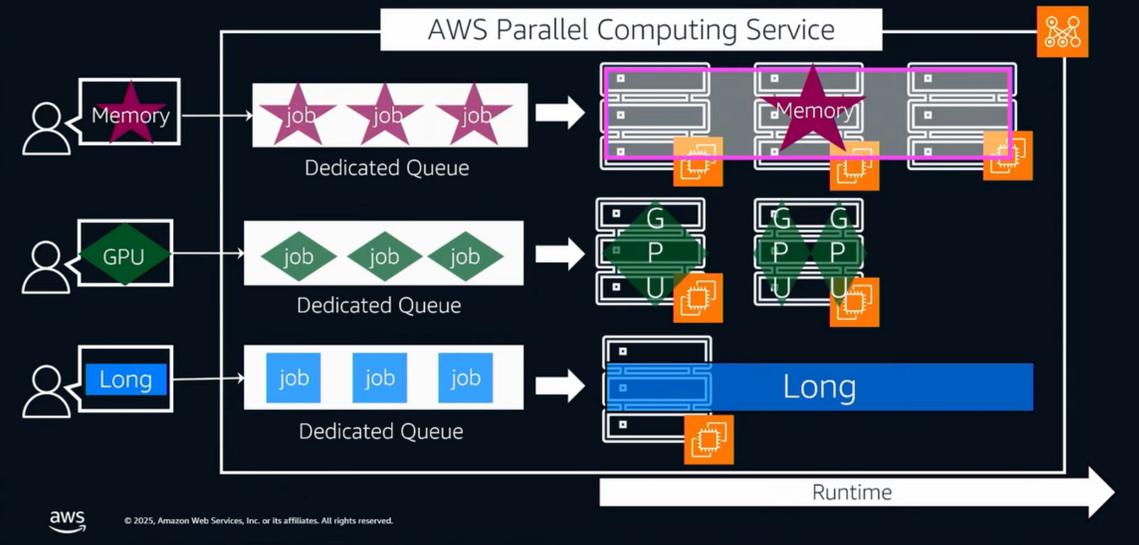
オンプレミスDCにおける課題



限られたリソース
で大規模実行時は
長い待ちが発生

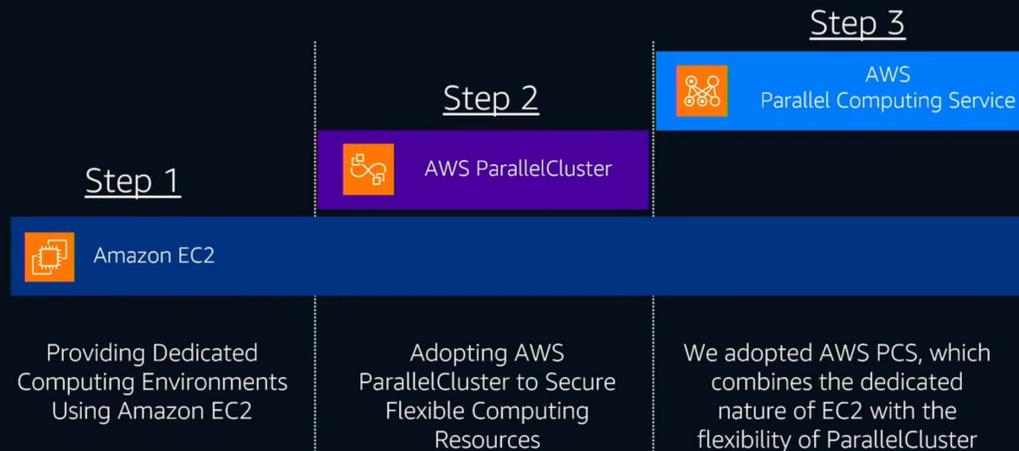
PCS利用時のアーキテクチャ

Tailored Computing for Researchers



ニーズごとにキューを分けて計算。
運用の手間も減り、研究に集中できる環境

Pursuing a Flexible Environment that Meets Diverse Needs



AWS Blog: 第一三共株式会社 創薬研究クラウドプラットフォームにおけるモダナイゼーションの取り組み

- ゲノム解析・構造予測など大規模並列計算の需要が急増
- 1論文で数百人規模のDNAシーケンサーデータを扱う時代、マルチオミクスデータも複雑化
- ParallelCluster運用ではHPC管理が属人化し、ノウハウ継承・スケーラビリティに課題



- AWS PCSへモダナイゼーションし、コンソール中心の操作で引き継ぎを容易に
- Step Functions・Systems Managerでユーザー追加を自動化、Slurmアップグレードも自動化し保守負担を軽減
- データ解析用に大容量・高速ストレージ環境をデプロイし、解析後に停止するコスト最適化サイクルを実現

Amazon Web Services ブログ

第一三共株式会社：創薬研究クラウドプラットフォームにおけるモダナイゼーションの取り組み

by Takehiro Nakajima | on 23 6月 2025 | in Amazon EC2, Amazon FSx for Lustre, Amazon Simple Storage Service (S3), AWS Health, AWS Parallel Computing Service, AWS ParallelCluster, AWS Step Functions, AWS Systems Manager, Best Practices, General, Healthcare, High Performance Computing, Life Sciences | Permalink | Share

このブログは、第一三共株式会社 研究統括部 研究イノベーション企画部およびモダリティ第一研究所と、アマゾン ウェブ サービス ジャパン合同会社 シニア ソリューション アーキテクト 中島丈博による共著です。

はじめに

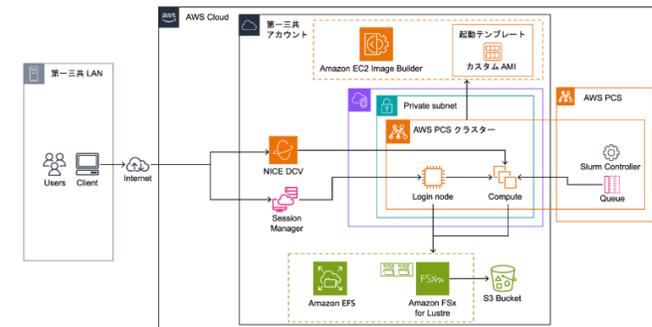
近年、バイオインフォマティクスの分野では、ゲノム解析や構造予測など、大規模な並列計算を必要とする業務が急速に増加しています。第一三共株式会社では、これまでも [AWS ParallelCluster](#) を活用し、ハイパフォーマンスコンピューティング (HPC) 環境を構築・運用し、研究環境の効率化とスピードアップを実現してきました。(参考: [AWS Summit Japan 2024 第一三共株式会社における創薬研究クラウドプラットフォーム](#)) 一方で、HPC 環境の継続的な利用においては障害対応や構成管理などの保守作業があり、これらの担当者にはスキルやノウハウが必要となることから対応が属人的になる場合もありました。このような体制は、ノウハウの継承やチーム全体での利活用の妨げとなり、将来的なスケーラビリティの確保にも課題を残します。こうした背景から、より安定かつ柔軟に運用でき、管理者やユーザーが気軽に利用可能なプラットフォームを目指して、AWS が提供する新しいマネージドサービス [AWS Parallel Computing Service \(AWS PCS\)](#) の試験構築に着手しました。本記事では、今回採用したアーキテクチャやモダナイゼーションの取り組みを通じて得られた、AWS PCS の使用感や運用面の知見をご紹介します。

AWS アーキテクチャのご紹介

今回採用した、創薬研究クラウドプラットフォームにおける AWS のモダナイゼーションアーキテクチャについてご紹介します。

HPC モダンアーキテクチャ

データサイエンティストの研究環境である HPC クラスター構成は、AWS PCS とクラスター運用を効率化するための複数 AWS サービスで構築されています。



お役立ちリンク

開始方法リソースセンター
AWS の最新情報
AWS イベントスケジュール
builders.flash - AWS 公式ウェブマガジン
日本国内のお客様事例

フォローをお願いいたします

Twitter
Facebook
LinkedIn
Twitch
最新情報メール
RSS フィード



<https://aws.amazon.com/jp/blogs/news/daiichisankyo-drug-discovery-research-cloud-platform-modernization/>

© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved.

データやAI との融合

AWSのストレージサービス

データの格納 (ストレージリソースの提供)



Amazon S3 and
Amazon S3 Glacier

Object (API)



Amazon EBS

Block



Amazon EFS

NFS



Amazon FSx for
Windows File Server

SMB



Amazon FSx for
NetApp ONTAP

SMB / NFS / iSCSI



Amazon FSx for
Lustre

Lustre



Amazon FSx for
openZFS

NFS

データ管理

エッジサービス (データ連携)

オフラインのデータ移動



AWS Backup

AWSサービス全体の
バックアップと
コンプライアンス



AWS Storage
Gateway

オンプレミスからク
ラウドストレージへ
のゲートウェイ



Amazon File Cache

NFSやAmazon S3
の高速キャッシュ
を提供



AWS DataSync

AWSストレージ
サービスへのデー
タ転送



AWS Transfer
Family

ファイル転送プロ
トコル (SFTP /
FTPS / FTP) EDI
(AS2) の提供

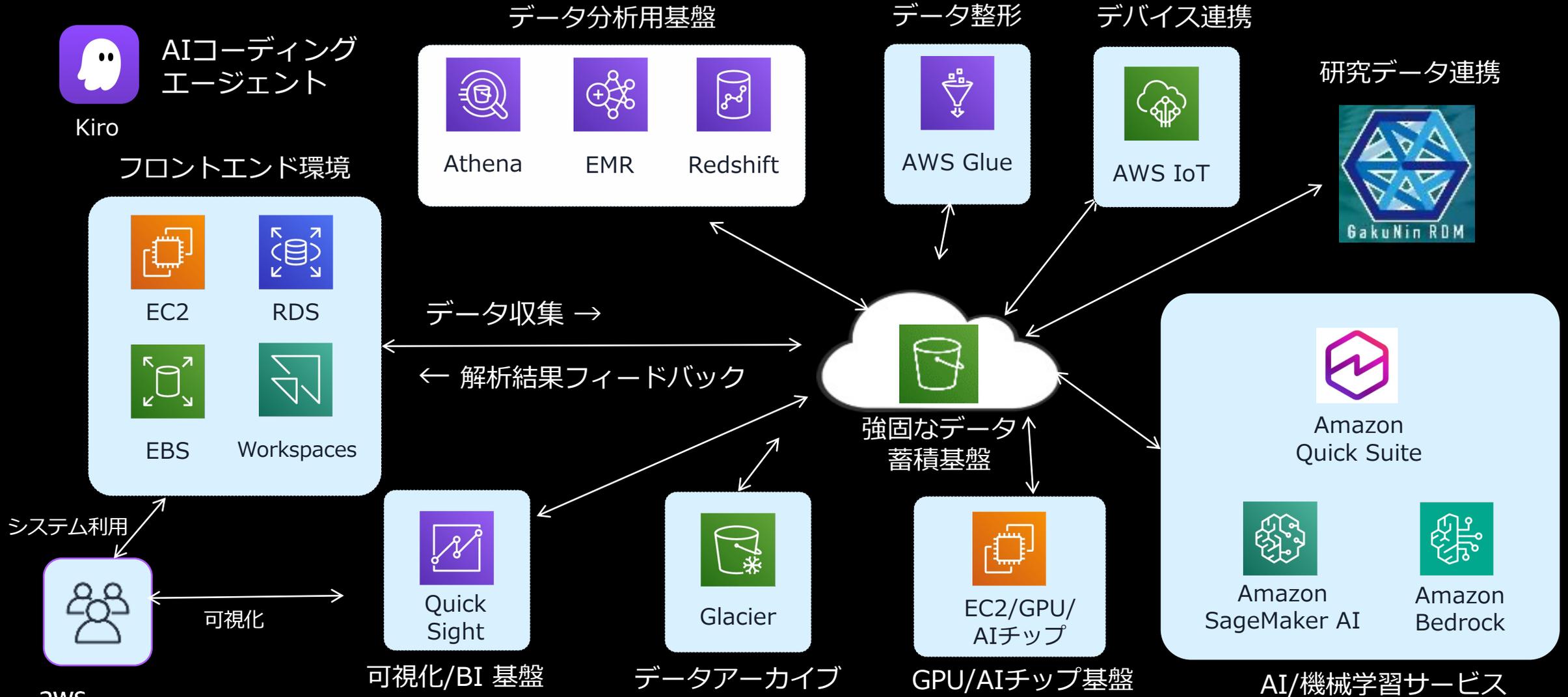


AWS Snowball

AWSストレージサービスとオンプレ
ミスの間でデータの移動と、エッジ
コンピューティングを提供するた
めの物理デバイス

研究データ活用のためのデータレイク連携基盤

Amazon S3 を中心に様々なデータ活用のサービスと連携が可能となり拡張性の高いシステム構成に



Amazon S3: 容量に関わらずデータを安全に長期保管

研究データを保管

- ✓ 長期保存が求められる (10年以上継続保存…、ストレージの入れ替え対応)
- ✓ 採来どれだけのデータになるか分からない (生データ、研究データ、ログ)
- ✓ 取得できるデータ量が膨大になり、オンプレミスストレージでは足りない etc

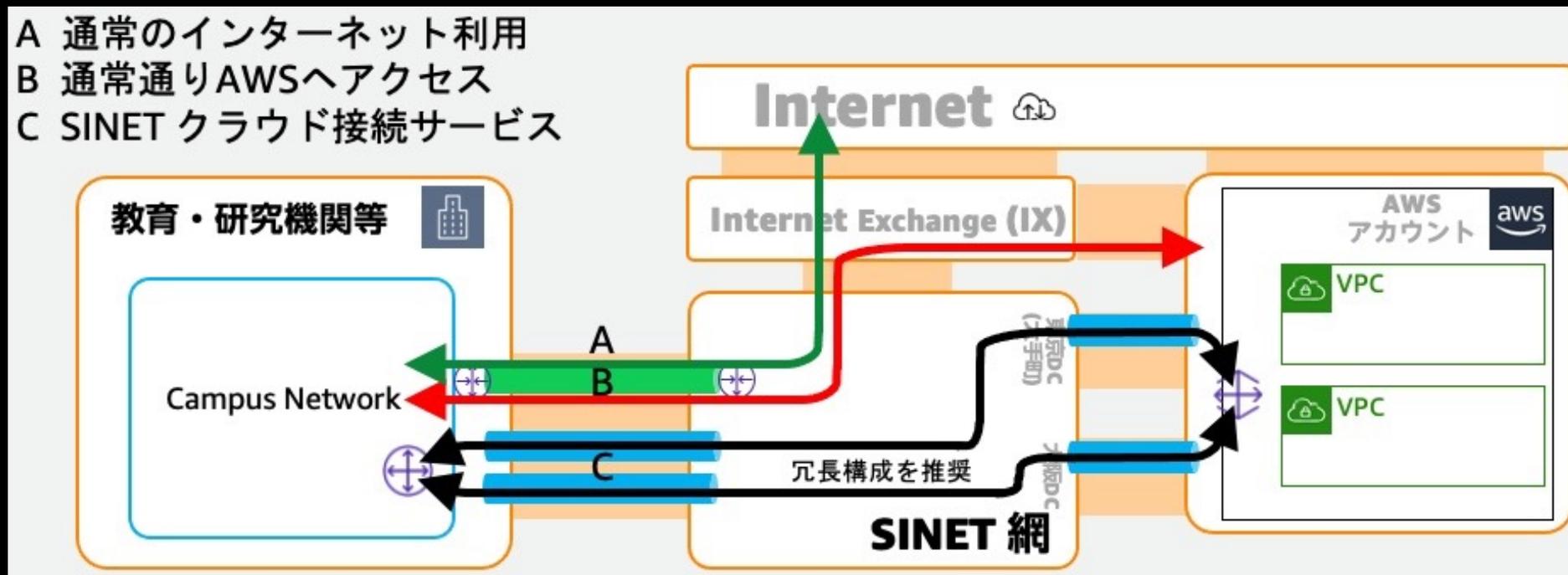
Amazon S3の活用

- ✓ **高信頼性**、**3ヶ所以上**に自動で**隔地**保存される
- ✓ **2006年**からサービススタート、長期稼働の実績
 - 国立情報学研究所 Gakunin RDMでの利用など
- ✓ 利用した分だけの課金、無制限の容量 → **スモールスタート**可能
- ✓ **データレイク**として様々なデータを集約、後から様々な方法で活用、S3へ直接クエリ等
- ✓ 利用頻度の低いデータは自動でアーカイブ階層としてコスト削減可能
- ✓ オンプレミスストレージのバックアップ先としてからはじめる、S3プラグインなどで対応しているストレージも多数
- ✓ 機関内と閉じた環境で接続、共同研究機関とのデータ共有での利用



SINET と AWS との接続の関係

- SINET とAWSはIX(インターネットエクスチェンジ)で直接ピアリングしており、(経路A)
- SINET 経由で Internet を利用している機関は申請不要で高速な帯域を活用可能 (経路B)
- SINET クラウド接続サービスを利用することで「閉じた環境」での利用も可能 (経路C)
- 2022年12月以降、 SINET AWS 間の接続には、**複数の 100 Gbps 専用回線も利用開始**



<https://aws.amazon.com/jp/blogs/news/aws-sinet-osaka-dc/>
<https://aws.amazon.com/jp/blogs/news/sinet6-aws-ur/SIN>

AWS Open Data Registry 活用 オープンデータ公開

無料でアクセスできるオープンデータ群

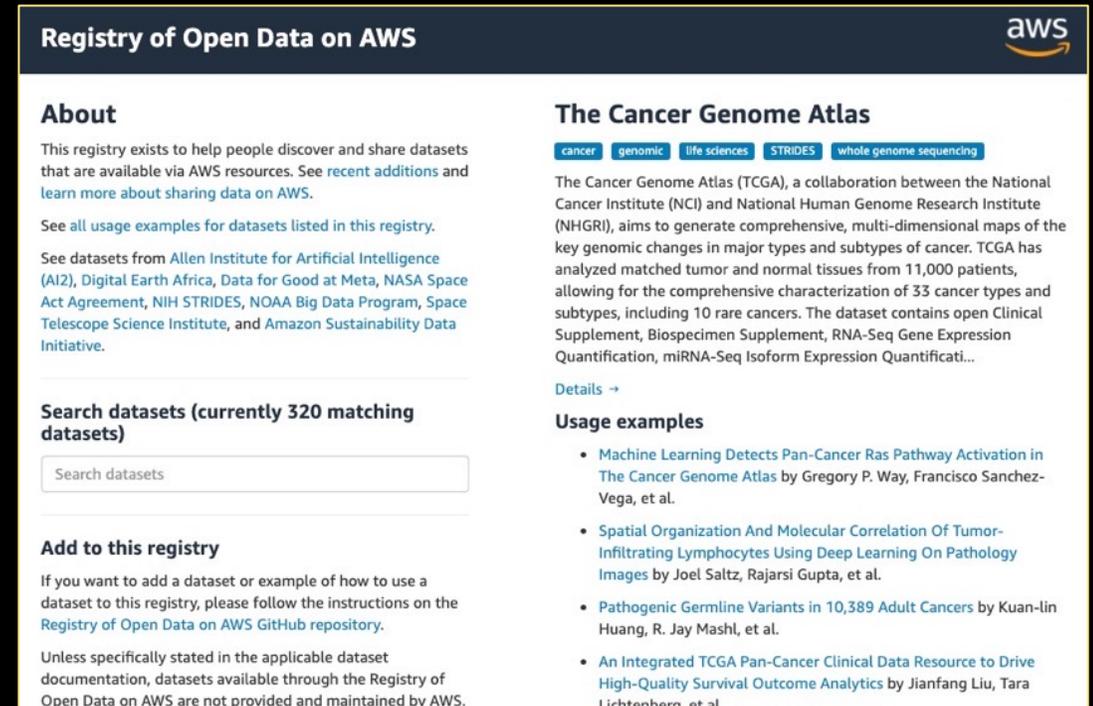
- 850のデータセット (2026年1月時点)
- データアクセス自体は無料
- <https://registry.opendata.aws/>

カテゴリ例：
 気候・気象データ (NOAA、NASA)
 ゲノミクスデータ (gnomAD)
 衛星画像 (Landsat、Sentinel-2)
 機械学習データセット (Common Crawl)

- **Open Data Sponsorship Program** により、オープンデータ向けのストレージコストをカバー可能
- <https://aws.amazon.com/jp/opendata/open-data-sponsorship-program/> (審査有)

国内研究機関の公開事例：

- 核融合科学研究所: [NIFS Large Helical Device \(LHD\) Experiment](#) (25年分の実験データ2PB+)
- 宇宙航空研究開発機構(JAXA) : [PALSAR-2 ScanSAR CARD4L \(L2.2\)](#)
- 理化学研究所 : [Brain/MINDS Marmoset Connectivity Resource on AWS](#)



The screenshot shows the AWS Open Data Registry interface. At the top, it says "Registry of Open Data on AWS" with the AWS logo. Below this, there are two main sections: "About" and "The Cancer Genome Atlas".

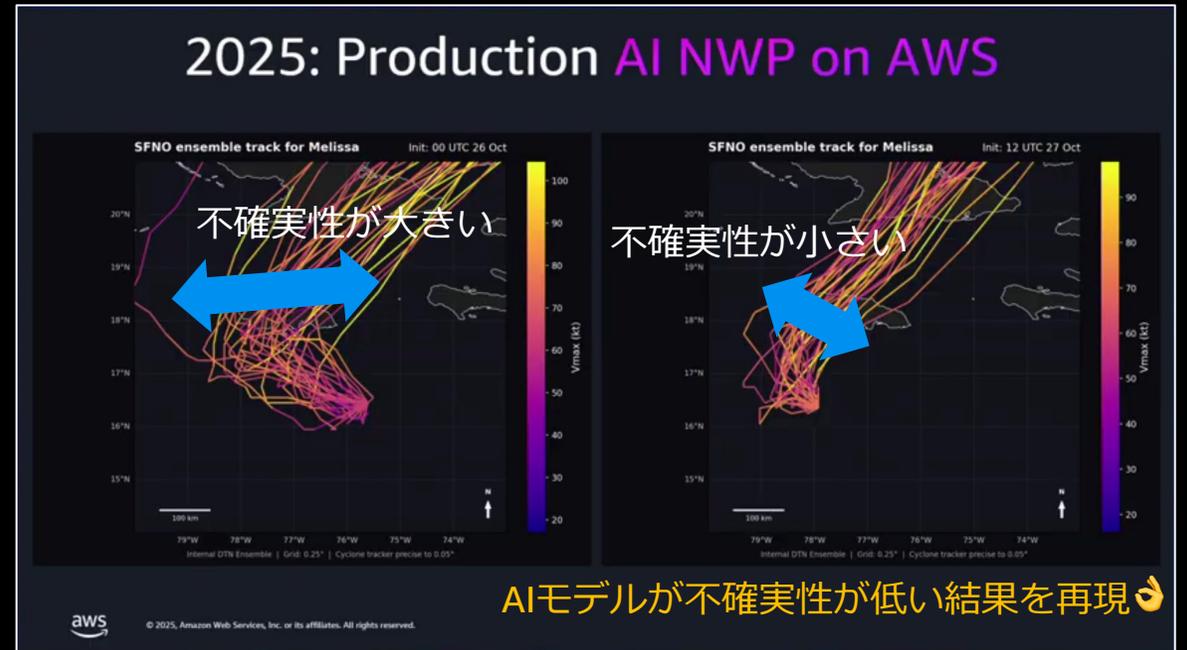
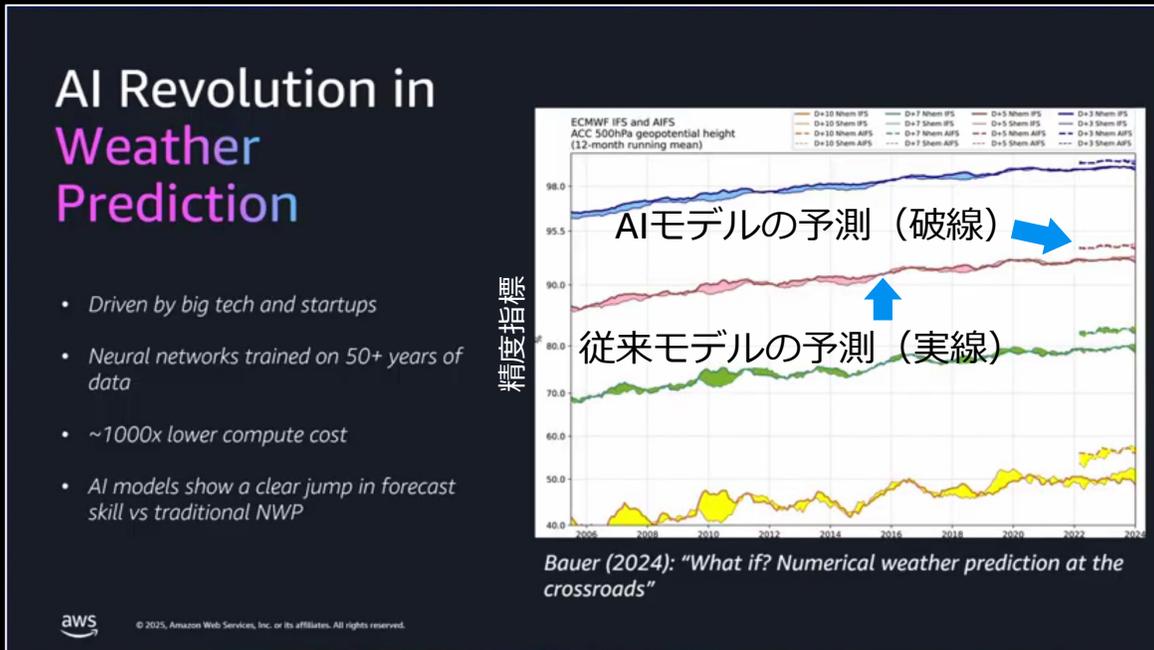
About
This registry exists to help people discover and share datasets that are available via AWS resources. See recent additions and learn more about sharing data on AWS.
See all usage examples for datasets listed in this registry.
See datasets from Allen Institute for Artificial Intelligence (AI2), Digital Earth Africa, Data for Good at Meta, NASA Space Act Agreement, NIH STRIDES, NOAA Big Data Program, Space Telescope Science Institute, and Amazon Sustainability Data Initiative.

The Cancer Genome Atlas
cancer genomic life sciences STRIDES whole genome sequencing
The Cancer Genome Atlas (TCGA), a collaboration between the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI), aims to generate comprehensive, multi-dimensional maps of the key genomic changes in major types and subtypes of cancer. TCGA has analyzed matched tumor and normal tissues from 11,000 patients, allowing for the comprehensive characterization of 33 cancer types and subtypes, including 10 rare cancers. The dataset contains open Clinical Supplement, Biospecimen Supplement, RNA-Seq Gene Expression Quantification, miRNA-Seq Isoform Expression Quantificati...
Details →
Usage examples
• Machine Learning Detects Pan-Cancer Ras Pathway Activation in The Cancer Genome Atlas by Gregory P. Way, Francisco Sanchez-Vega, et al.
• Spatial Organization And Molecular Correlation Of Tumor-Infiltrating Lymphocytes Using Deep Learning On Pathology Images by Joel Saltz, Rajarsi Gupta, et al.
• Pathogenic Germline Variants in 10,389 Adult Cancers by Kuan-lin Huang, R. Jay Mashl, et al.
• An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics by Jianfang Liu, Tara Lichtenberg, et al.

DTN社: 気象シミュレーションにおけるHPC on AWSとAI活用の拡大

- AWS Parallel Cluster、Amazon FSx for Lustreなどを利用中。今後PCSへ移行を計画。
- エネルギー・農業・天候など、天気や相場に強く依存する産業向けにデータと分析サービスを提供するグローバルテック企業であるDTNが、2020年にはAWSへフルクラウド化、HPC利用拡大と近年の気象シミュレーションにおけるAI活用について説明。
- 50年分の気象データでトレーニングしたAIモデルが、従来型シミュレーションと比較して、予測時間・コスト効率・精度の面で上回る。

ハリケーンの挙動予測シミュレーションの従来型/AIモデルの比較



AWS が提供する開発者のためのAI IDE

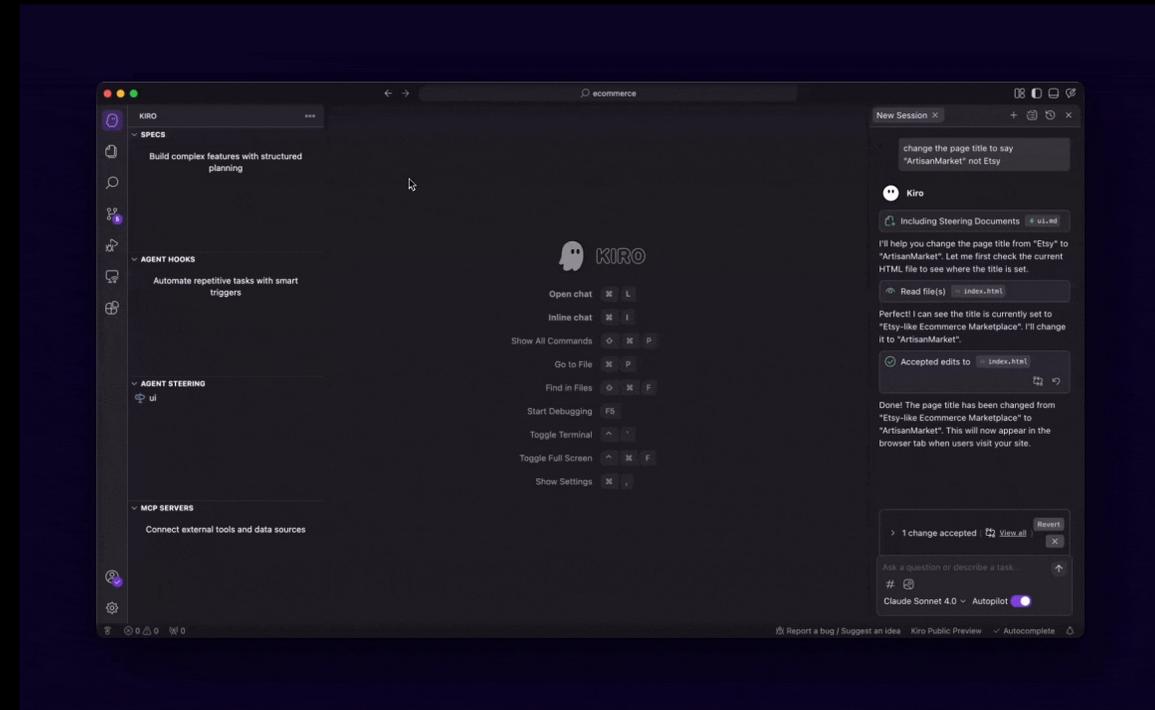
バイブコーディングでプロトタイプを作り、
従来の SDLC（ソフトウェア開発ライフサイクル）で製品を作る



Kiro は、開発者やエンジニアリングチームが
AI エージェントを使って
高品質なソフトウェアを提供することを支援します

利用シーン：仕様駆動開発

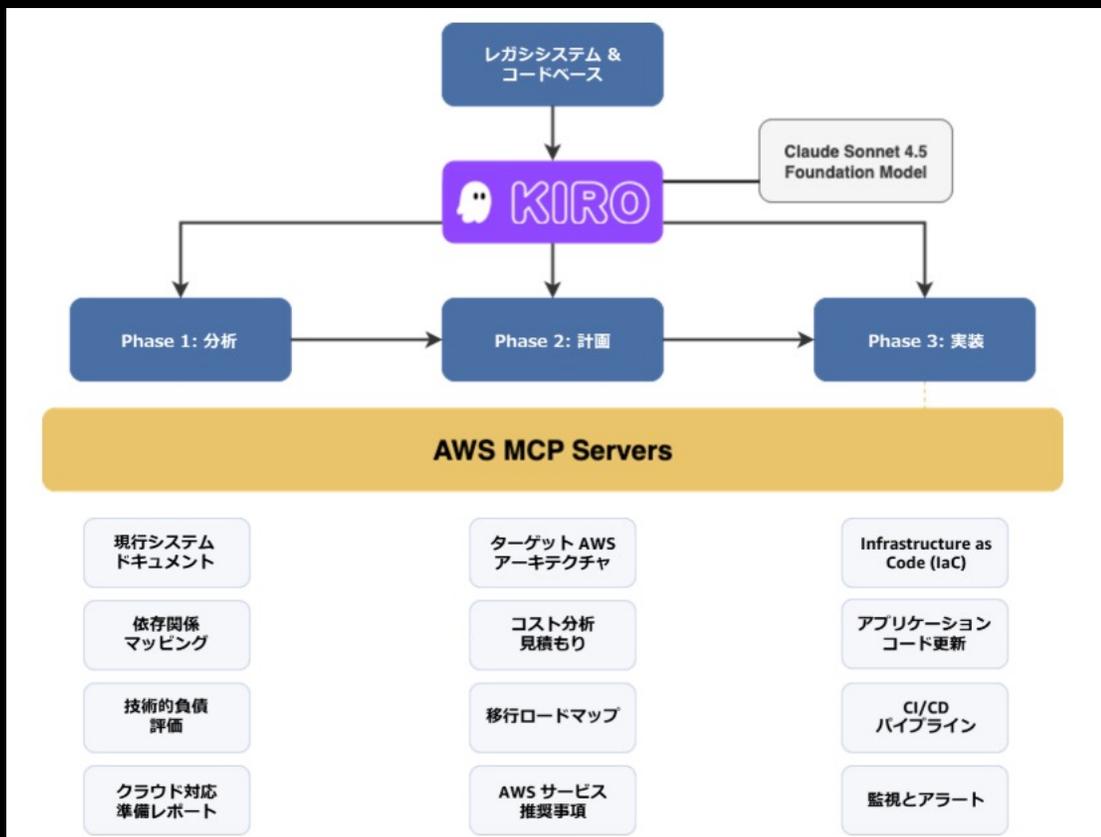
- 曖昧なアイデアを要件→設計→タスクに段階的に具体化
- 自然言語のプロンプトから構造化された仕様ドキュメントを自動生成する機能
- エージェントが自律的に機能実装、コードリファクタリング、ソフトウェアアップグレードを実現
- 高度なコンテキストを管理
 - ステアリングファイルで振るまいを定義
 - MCPサーバと統合し、ドキュメントやデータベースと接続
- IDE/CLIを利用可能
- AWS上で組織での生成 AI の安全な利用を実現



1. プロンプト入力「TODOアプリを作りたい」
2. requirements.md 生成（要件の明確化）
3. design.md 生成（技術設計）
4. tasks.md 生成（実装計画）
5. タスクに従って実装開始

<https://catalog.workshops.aws/kiro-intro/ja-JP>

利用シーン: クラウドインフラの移行・作成支援



移行前の環境がある場合、コードや仕様書から現状のアーキテクチャを把握

AWSのMCPサーバと連携させることで、KiroをAWSのスペシャリストとして活用

Kiroがサービス仕様やベストプラクティス知識を把握した上で環境構築を支援

運用フェーズではトラブルシューティングや報告用レポートも作成可能

開発中のコードを検証するためのクラウド環境構築をKiroと行う事で、効率よく検証環境を用意する事も可能

<https://aws.amazon.com/jp/blogs/news/agenic-cloud-modernization-accelerating-modernization-with-aws-mcps-and-kiro/>

<https://catalog.us-east-1.prod.workshops.aws/workshops/dd046665-c8d2-4eb0-b88c-d3519e5d3090/ja-JP>



例: Parallel Cluster with Kiro

バーチャル富岳：ポータブルHPC環境の構築

- ▶ セットアップ
- ▶ Lab 1: HPCクラスタの構築
- ▶ Lab 2: Apptainerコンテナのビルドと実行
- ▶ Lab 3: バーチャル富岳アプリの実行
- ▼ **Lab 4: Kiro CLI を使ったクラスター管理及びジョブプロファイリング**
 - AWS Builder ID にサインアップ
 - Kiro CLIを使ったクラスターのアップデート
 - GROMACSジョブのプロファイリング
 - ジョブプロファイルの評価
 - プロファイリングレポートの確認
- クリーンアップ
- まとめ

バーチャル富岳：ポータブルHPC環境の構築 > Lab 4: Kiro CLI を使ったクラスター管理及びジョブプロファイリング

Lab 4: Kiro CLI を使ったクラスター管理及びジョブプロファイリング

HPCアプリケーションのプロファイリング

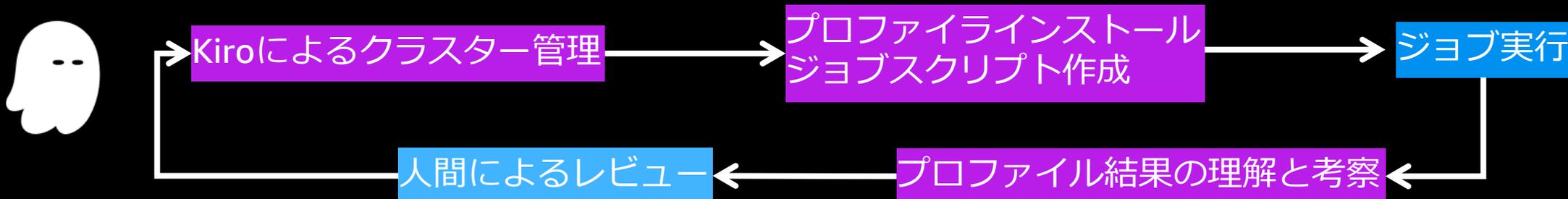
アプリケーションの性能は、HPCユーザーの生産性、および全体的なコスト (TCO) に直接影響します。HPCジョブは複数のノードと数万コア以上にまたがるため、最終的なパフォーマンスは個々のノードのパフォーマンス、ネットワーク、ストレージ、メモリなど複数の要因に依存します。プロファイリングデータは、HPC管理者、ユーザー、開発者がどの部分の最適化に取り組めば良いかの決定を下すのに役立ちます。これにより、改善が全体的な実行時間に影響を与えないセクションに時間を費やすのではなく、最も影響力のある領域をターゲットにすることができます。したがって、性能プロファイリングは、アプリケーションの動作を理解し、ボトルネックを特定し、インフラストラクチャを最適化するために重要です。

本セクションでは、エージェント型のコーディングアシスタントであるKiroを利用して、HPCクラスタに別のコンピュートリソースを追加し、その後Aperfを実行して、異なるコンピュートリソースで実行されたときのGROMACSジョブプロファイルの違いを評価します。

Kiro CLI

Kiroはユーザーのコードベース全体を理解し、プロンプトを構造化された仕様に変換し、反復的なタスクを自動化するAI搭載機能を持っています。Kiro CLI により、ユーザーはターミナルでAI支援開発を使用できます。Kiro CLIを使用してアプリケーションをプロファイリングし、Kiro CLIの基本機能を体験していただきます。

ParallelClusterのクラスタ管理とジョブの性能評価をKiroを使って実行できるシナリオ



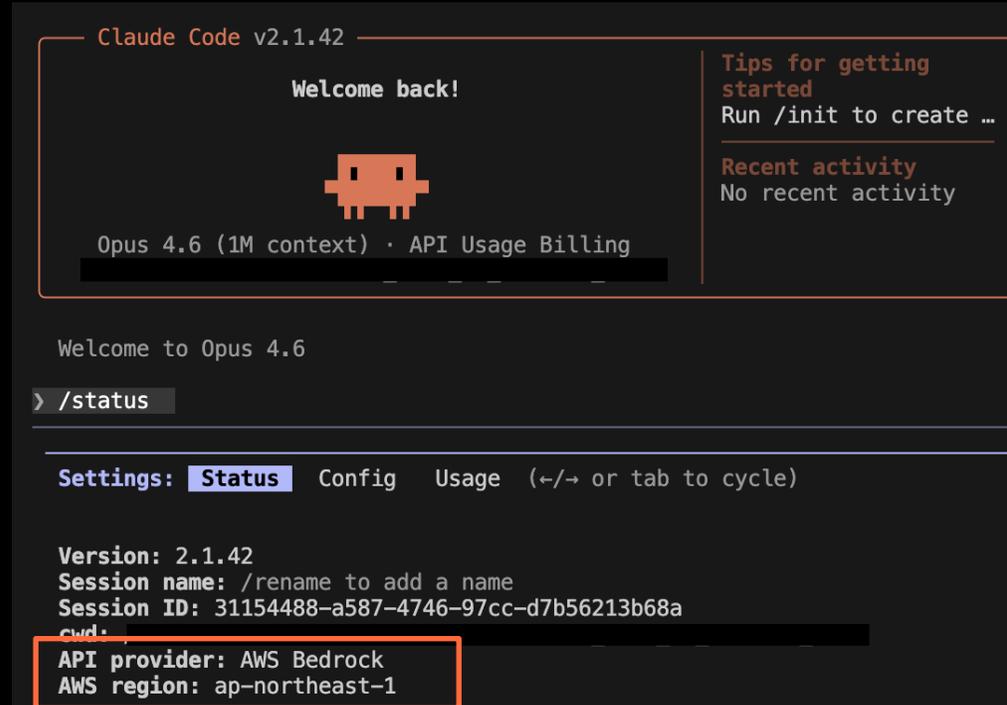
<https://catalog.us-east-1.prod.workshops.aws/workshops/04ffde5e-c2a6-4192-bd6d-2980c1f94e8b/ja-JP>

Claude Code on AWS (with Amazon Bedrock)

```
{ } settings.json
```

```
1 {
2   "env": {
3     "AWS_REGION": "ap-northeast-1",
4     "CLAUDE_CODE_USE_BEDROCK": "1",
5     "AWS_PROFILE": "ClaudeCode"
6   },
7 }
```

~/claude/settings.json



```
Claude Code v2.1.42
Welcome back!
Opus 4.6 (1M context) · API Usage Billing

Welcome to Opus 4.6
> /status

Settings: Status Config Usage (-/> or tab to cycle)

Version: 2.1.42
Session name: /rename to add a name
Session ID: 31154488-a587-4746-97cc-d7b56213b68a
cwd:
API provider: AWS Bedrock
AWS region: ap-northeast-1
```

- AWSへの認証設定と、環境変数を指定することでClaudeの接続先をAmazon Bedrockに指定することが可能
- サブスクリプションと比較して、安価に利用できる (最新のClaude Sonnet 4.6 / Opus 4.6も利用可能)
- ガバナンスの観点で、データをAWSや日本国内に留めることも可能
- 組織で導入するためのサンプルアーキテクチャ(CloudFormation)も公開



まとめ

1. AWSのサービスアップデート

1. 新インスタンス情報 (hpc8a、Graviton5、G7e、X8i)
2. AI Factory : AWSのAIインフラが顧客のデータセンターに来る時代
3. Parallel Computing Suiteによるマネージドなクラスタ環境の構築と事例

2. データ・AI活用方法

1. S3とオープンデータの活用 (DTN社 : Open Data + 気象AI予測運用)
2. Kiroを用いた研究活動への適用 — コーディング支援、計算リソースの移行や構築

最新のサービスに加えて、生成AIを駆使することで、より迅速に研究活動を進める動きが世界中で始まっています。



学術・研究機関向け問い合わせ先
aws-jpps-er@amazon.com

