



CBI学会 2024年大会 LS09-01

ドライ解析環境を AWS で構築し 創薬プロセスに組み入れるには

石尾 千晶

アマゾン ウェブ サービス ジャパン合同会社

自己紹介

石尾 千晶 (Chiaki Ishio)

アマゾン ウェブ サービス ジャパン合同会社
ソリューション アーキテクト

製薬企業のお客様のクラウド活用の技術支援をしております



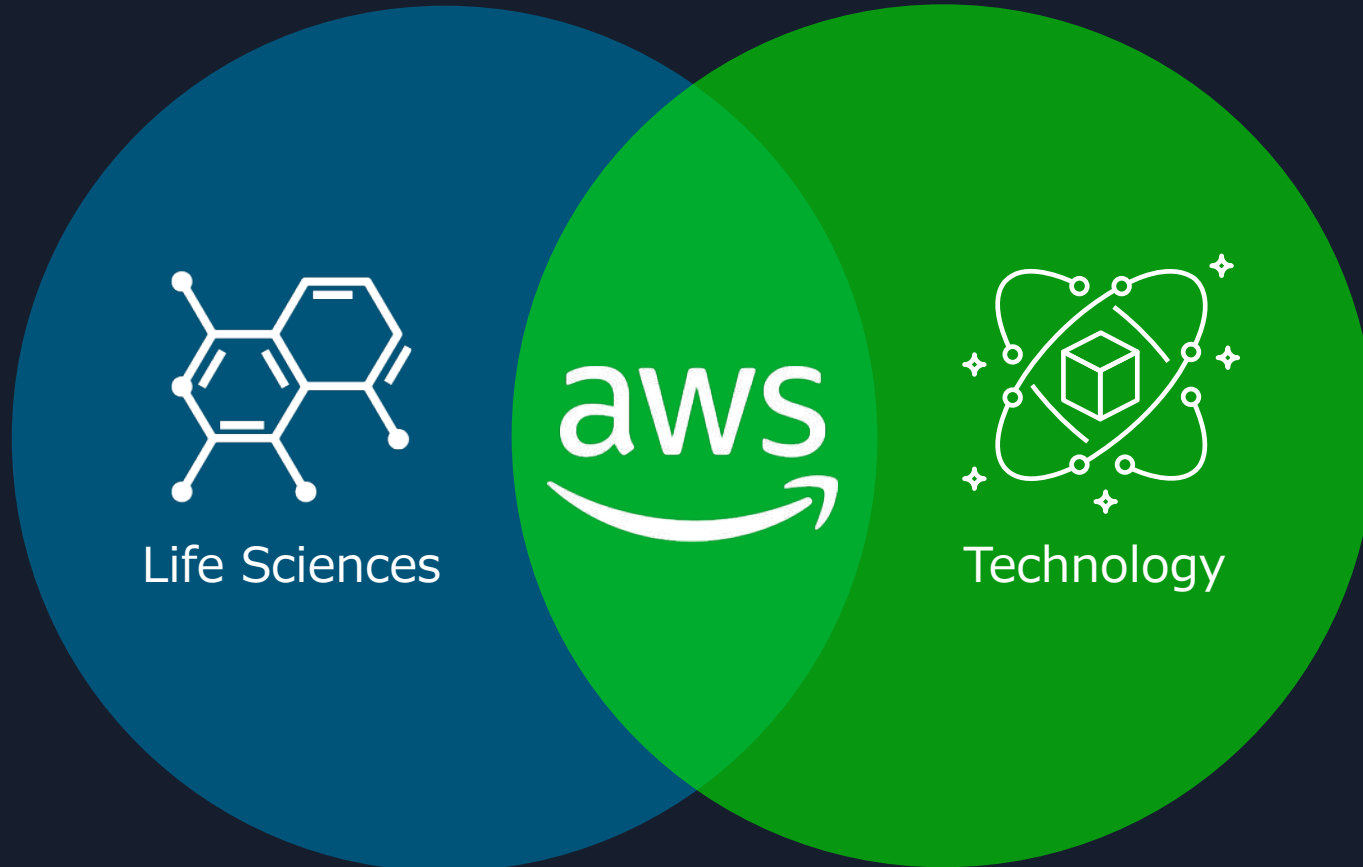
本日のアジェンダ

- ライフサイエンス業界における AWS の支援内容
- 国内外の製薬企業のお客様での AWS 活用事例
- 具体的な AWS 活用方法

AWS for Life Sciences



約20年にわたって、ライフサイエンス領域の お客様を支援しています



生成 AI の可能性を医薬品開発の現場に



創薬研究

タンパク質設計・
立体構造予測

論文検索



臨床開発

試験計画、患者集団、
施設の最適化

合成対象群のデザイン



製造 サプライチェーン

工場機器の予知保全

リソース最適化



マーケティング 営業 メディカル

患者の予後予測

コンテンツ生成

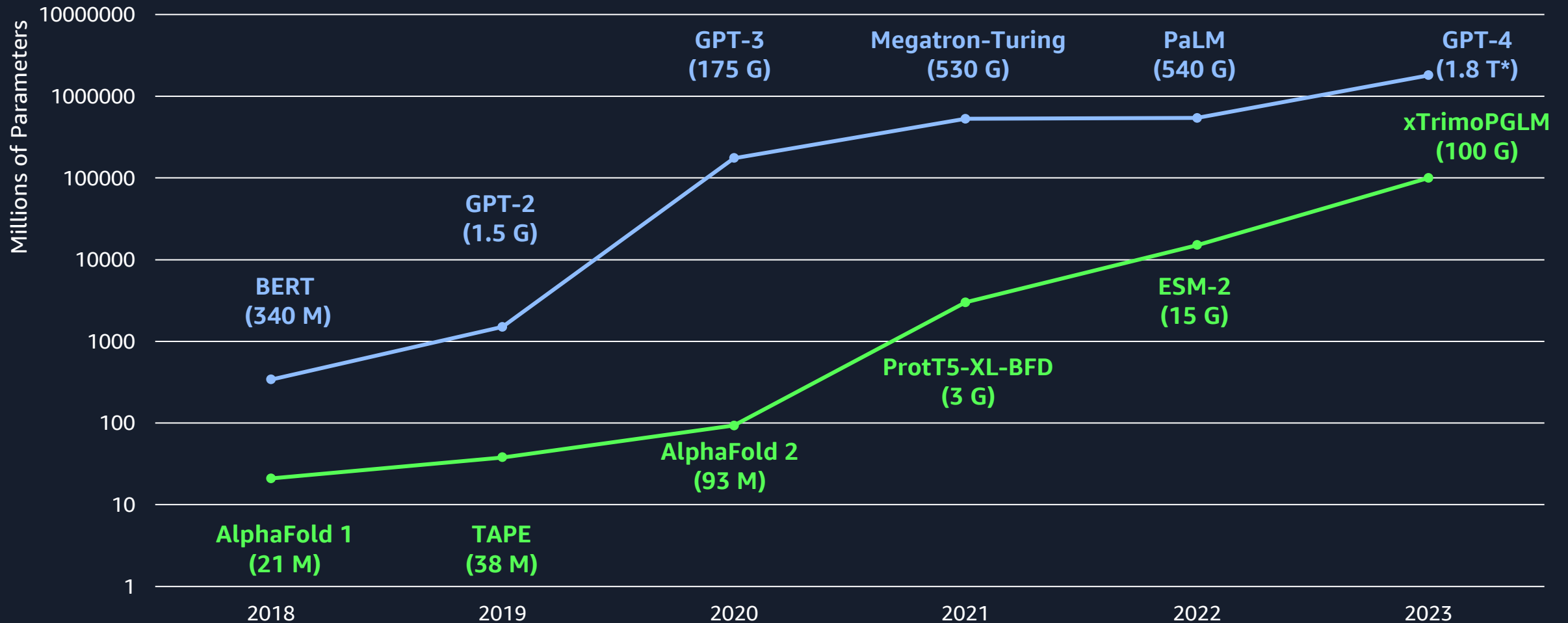


デジタルヘルス 患者サポート

患者ケア
コンシェルジュ

患者さんと治験の
マッチング

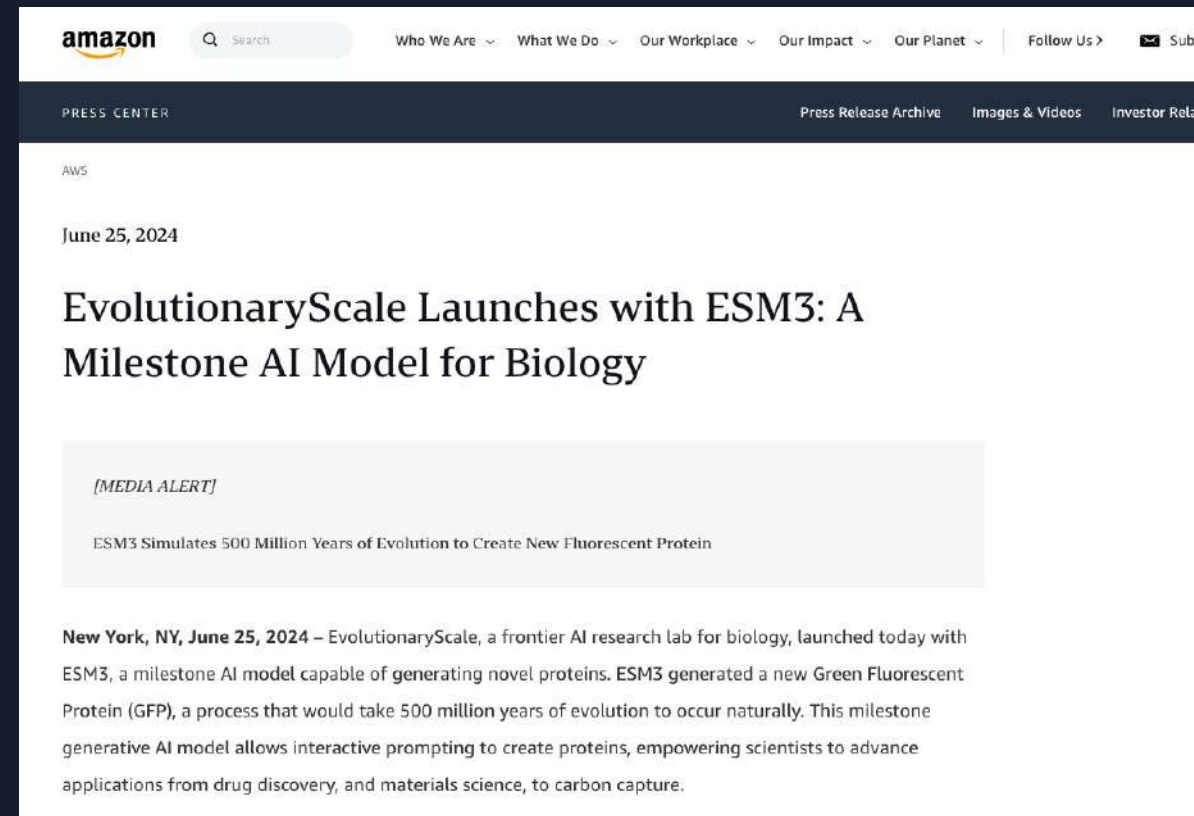
基盤モデルの進化



基盤モデルの進化：ESM3 の発表

EvolutionaryScale 社の 生物学向けの基盤モデルである ESM3 が AWS で利用可能に

- 新しい蛍光タンパク質（GFP）を生成
- オープン版はすでに AWS 上で利用可能 ([AWS Marketplace](#))
- 商用版も今後 AWS 上の複数のサービスで利用可能になる予定
 - Amazon SageMaker
 - Amazon Bedrock
 - AWS HealthOmics



The screenshot shows the Amazon Press Center page for EvolutionaryScale. The page is titled "EvolutionaryScale Launches with ESM3: A Milestone AI Model for Biology" and is dated June 25, 2024. The page includes a search bar, navigation links for "Who We Are", "What We Do", "Our Workplace", "Our Impact", "Our Planet", and "Follow Us". The main content area features a "PRESS CENTER" header, a "Press Release Archive" link, and an "Images & Videos" link. The main headline is "EvolutionaryScale Launches with ESM3: A Milestone AI Model for Biology". Below the headline is a "[MEDIA ALERT]" section with the text "ESM3 Simulates 500 Million Years of Evolution to Create New Fluorescent Protein". The main body of the press release starts with "New York, NY, June 25, 2024 – EvolutionaryScale, a frontier AI research lab for biology, launched today with ESM3, a milestone AI model capable of generating novel proteins. ESM3 generated a new Green Fluorescent Protein (GFP), a process that would take 500 million years of evolution to occur naturally. This milestone generative AI model allows interactive prompting to create proteins, empowering scientists to advance applications from drug discovery, and materials science, to carbon capture."

[AWS のプレスリリース 2024/06/25](#)

本日お伝えしたいこと



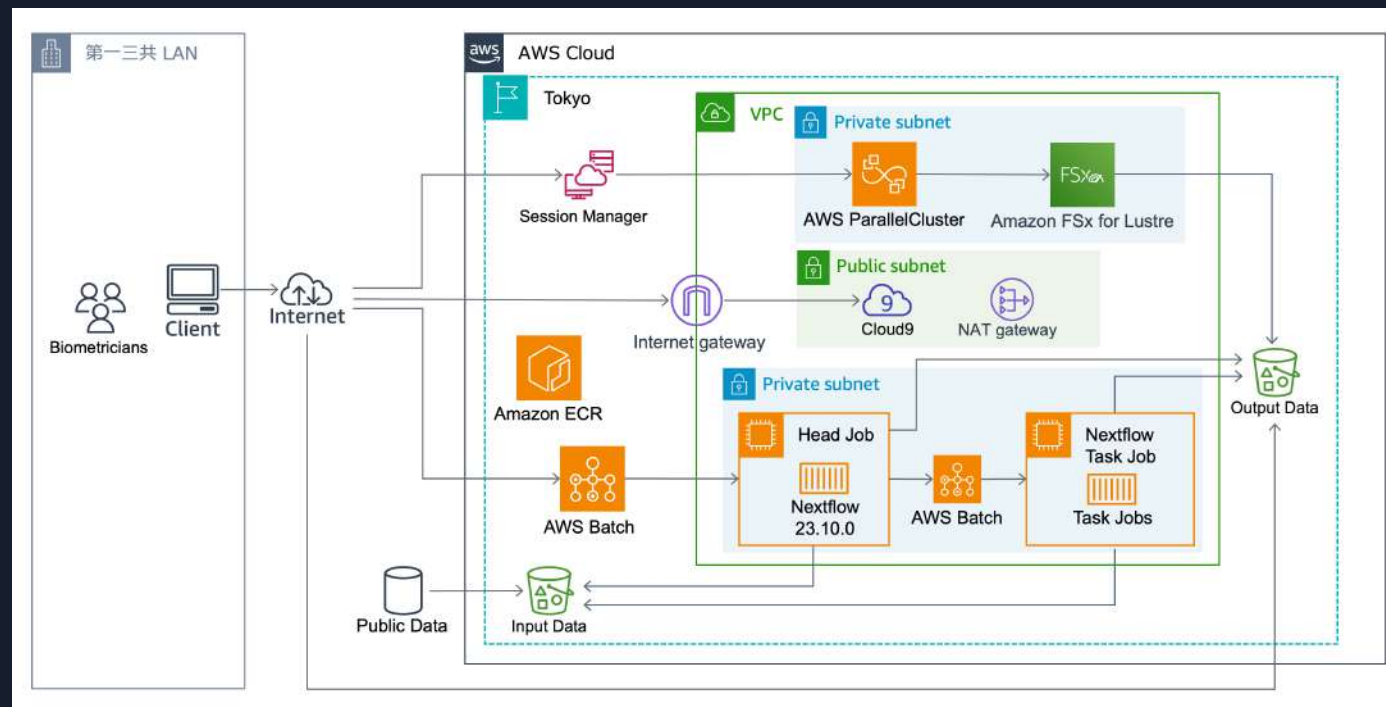
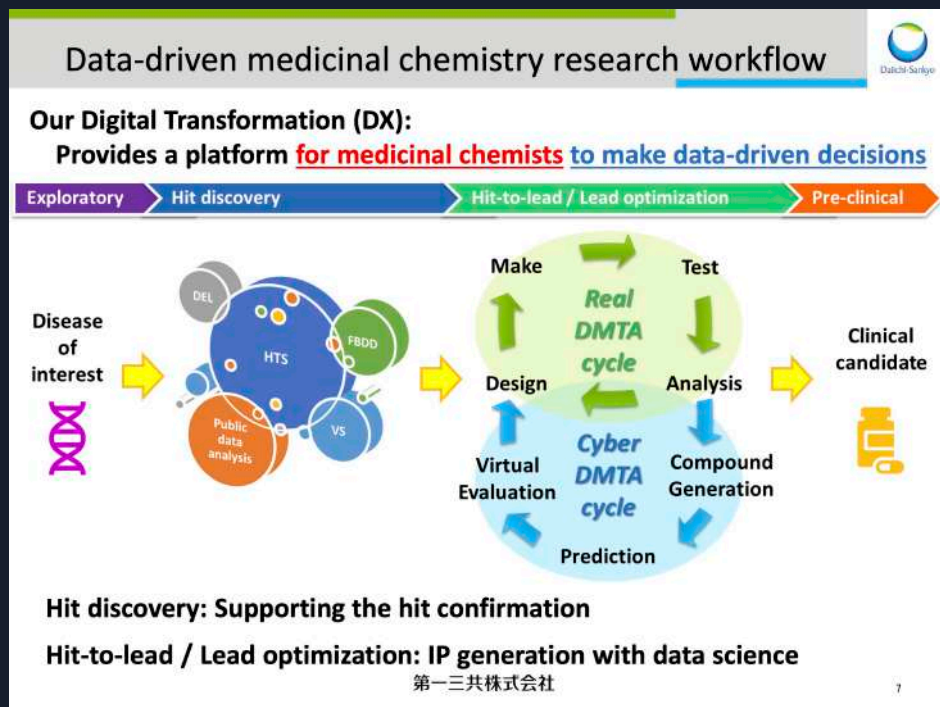
多様な基盤モデルが進化を続ける中で
どのようにドライ・ウェットの連携を進めるか？

- ① 抽象化・スケーラビリティ
- ② データ活用

AI 創薬での AWS 活用事例

第一三共様： データ駆動型創薬のプラットフォーム構築

実験データ転送・HPC 解析・データ分析のプラットフォームを AWS に構築、
研究者が自律的にデータ活用できる環境作り



創薬研究におけるクラウド活用の実例
～第一三共でのデータ駆動型創薬化学研究基盤～

AWS ブログ：「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」

中外製薬様： 機械学習パイプライン・人材育成の取り組み

ドライ・ウェットの橋渡しをするための機械学習パイプラインの自動化、
Python を用いた解析スキル獲得のための人材育成の取り組み

MALEXA®

パイプラインのまとめ

- 本システムのシステム概要について最後にまとめて説明します。

○ システムの構成図

○ システムのシーケンス図

本システムと同じ仕組みを横展開し、研究所のDX活用を推進していきます。

12

PyZAP

WET(実験) × DRY(アプリ開発) = 二刀流

- 当初PyZAPは「データ成型」および「可視化」スキルの獲得を目標としていた。
- 他の研究者に自身の開発したツールを使ってもらうためにアプリ化したいという声も多く、Jupyter × Streamlit環境 on Amazon EC2を提供したところ、Webアプリ開発もできる二刀流研究者が急増。

Jupyter × Streamlit on Amazon EC2
「Python & ブラウザだけ」で開発からサービス提供まで全て完結

ブラウザだけでWebアプリ開発

WET研究者が自らWebアプリを開発し、業務効率化 & 新たなInsight獲得

WET × DRYの二刀流

アプリ開発もできるWET研究者(二刀流研究者)が 20名以上！

9

AWSを活用した迅速なMALEXA機械学習パイプラインの内製開発、人材育成施策PyZAPを支えるインフラ

Evozyne 様: “Lab in the loop” の実現

- ウェットでの検証結果をもとに、pLM にフィードバックするサイクルを実装
- タンパク質設計のために、NVIDIA BioNeMo Framework を用いて pLM を開発
 - Amazon SageMaker 上の IDE を用いて、モデル開発・実験のオーケストレーション IDE から コンテナイメージ・公開モデルにアクセス
 - Amazon EKS (Elastic Kubernetes Service) 上で基盤モデルの学習を実行

Evozyne’s Machine Learning Platform: Mew

A Machine Learning Platform needs to tackle any or all of the following:

- Experiment tracking
- GPU compute orchestration
- Data and model versioning
- Interactive analysis (notebooks)
- Model deployment and inference

Fully managed service platform



- Plug and play with publicly available container images and pre-trained models
- A single IDE for running notebooks and pipelines
- Centralized management of data and models

Evozyne’s Mew Platform

We developed our own ML platform using bespoke open-source tools running on AWS infrastructure, named **Mew**



- A ML platform tailor-made for our unique approach to generative AI drug discovery
- A software platform that combines the power of NVIDIA BioNeMo and our own internal models
- A protein data store specific to our in-house lab capabilities

EVOZYNE 10

‘Lab-in-the-loop’ Paradigm for Leveraging Generative AI in Drug Design

Foundational Model

- Data Loading
- Optimized Training

Initial Candidates

Experimental Screen

Fine-Tune NIMs

- DNA
- Transcriptome
- Cell Images
- Small Molecule
- Structure



BioNeMo: Pretrained Model Stack | SageMaker: Compute & Orchestration | S3: Storage

EVOZYNE 16

A-Alpha Bio 様: BioNeMo の大規模実行環境

WETラボでの実験サイクルを削減し、タンパク質相互作用の測定・予測・設計の効率化



成果

- タンパク質の結合予測スループットが **12倍** 向上
- **2ヶ月で1億回以上** の推論処理（当初計画の**10倍**）
- Wet ラボでの実験サイクルを 1-2 回削減
→ コスト削減 & 開発効率向上

ソリューション

- AWS Batch 上で NVIDIA BioNeMo コンテナを展開
- Amazon EC2 P5 インスタンス（NVIDIA H100 GPU）の活用
- Amazon EC2 Capacity Blocks for ML を用いたキャパシティ確保

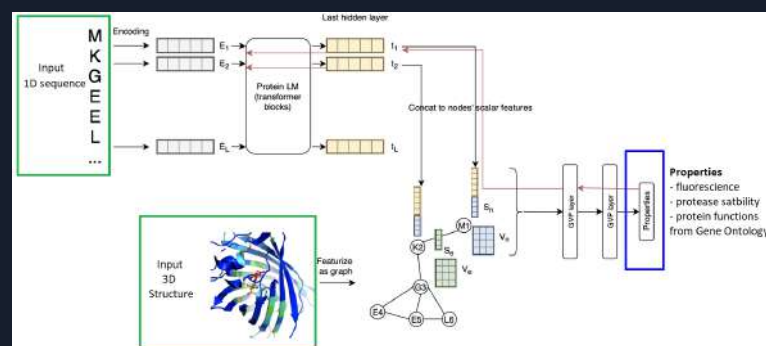


Janssen Biotherapeutics 様: AI 創薬研究へのご支援

AWS の Gen AI Innovation Center との共同研究

LM-GVP: an extensible sequence and structure informed deep learning framework for protein property prediction

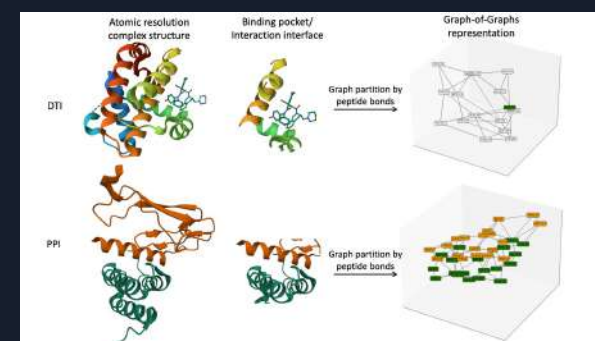
- LM-GVP* という深層学習フレームワークを構築し、タンパク質の性質を予測した
- フレームワークは pLM と GNN からなる



*Language Model – Geometric Vector Perceptrons,
<https://www.nature.com/articles/s41598-022-10775-y>

EGGNet, a generalizable geometric deep learning framework for protein complex pose scoring

- EGGNet* という深層学習フレームワークを構築し、タンパク質複合体のドッキングポーズのスコアリングを算出
- タンパク質の安定性予測を 12 %改善



*Equivariant Graph of Graphs neural Networks,
<https://www.biorxiv.org/content/10.1101/2023.03.22.533800v1>

AWS を活用して 創薬プロセスを加速するには？

本日お伝えしたいこと



多様な基盤モデルが進化を続ける中で
どのようにドライ・ウェットの連携を進めるか？

- ① 抽象化・スケーラビリティ
- ② データ活用

本日お伝えしたいこと



多様な基盤モデルが進化を続ける中で
どのようにドライ・ウェットの連携を進めるか？

① 抽象化・スケーラビリティ

AWS のサービス・ツール・実装サンプルを活用して
基盤モデルのアップデートに追従するための時間とコストを節約

ドライ解析を支える技術要素

ソフトウェア

オーケストレーション

コンピューート

ネットワーク

ストレージ

このレイヤだけ
見ていけば良い
状態にしたい

必要に応じて
自動的に
用意してほしい

より本質的な作業に集中するために

ソフトウェア

オーケストレーション

OS 導入・OS パッチ

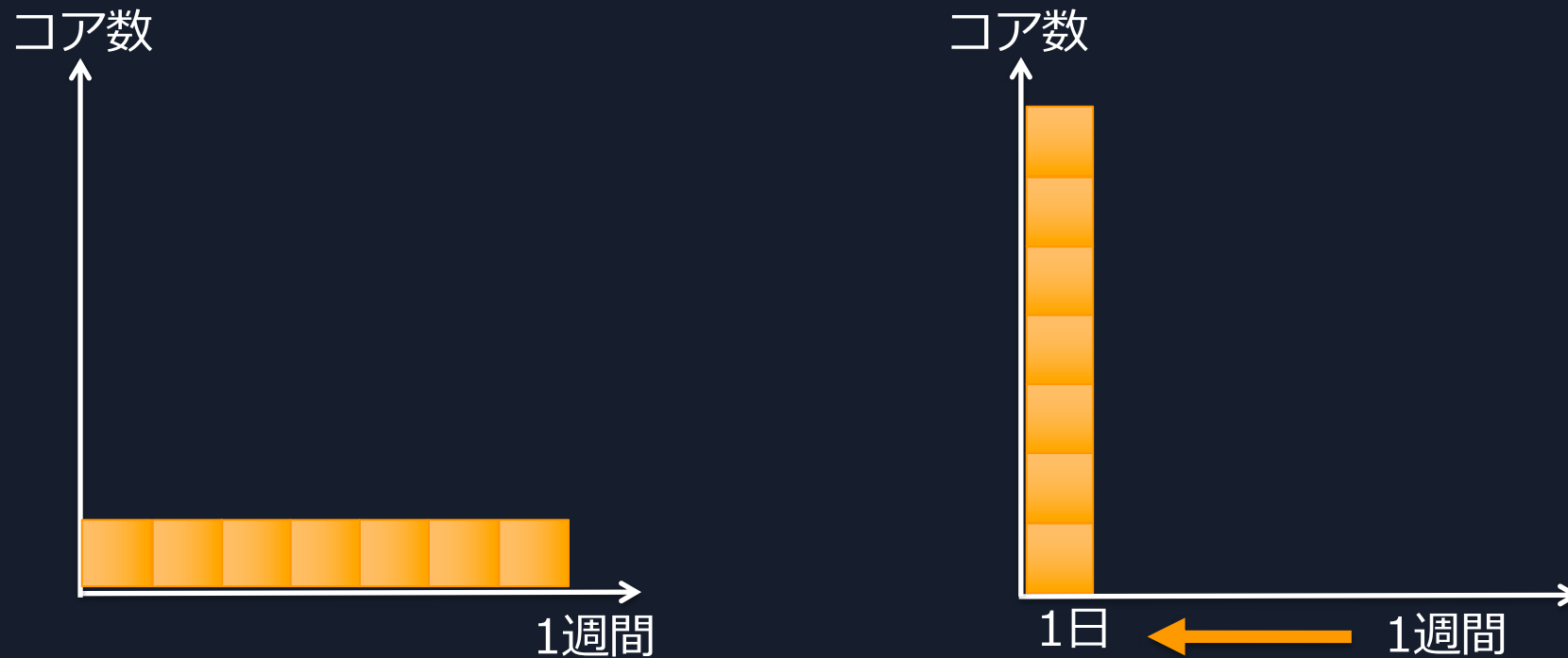
ハードウェアメンテナンス

電源・ラック・ネットワーク

このレイヤだけ
見ていけば良い
状態にしたい

必要に応じて
自動的に
用意してほしい

クラウド創薬の魅力： スケーラビリティをいかして計算時間を短縮



従来は手持ちの限られたリソースで、逐次処理していたジョブも
AWSなら必要な台数、インスタンスを起動して、一斉処理。
しかも費用は「時間×台数」なのでどちらも同じ。

ニーズに合わせて組み合わせる

使い慣れたインターフェース + スケーラブルな計算環境

ウェブアプリ

Jupyter Notebook

リモートデスクトップ

CLI

AWS ParallelCluster

AWS PCS

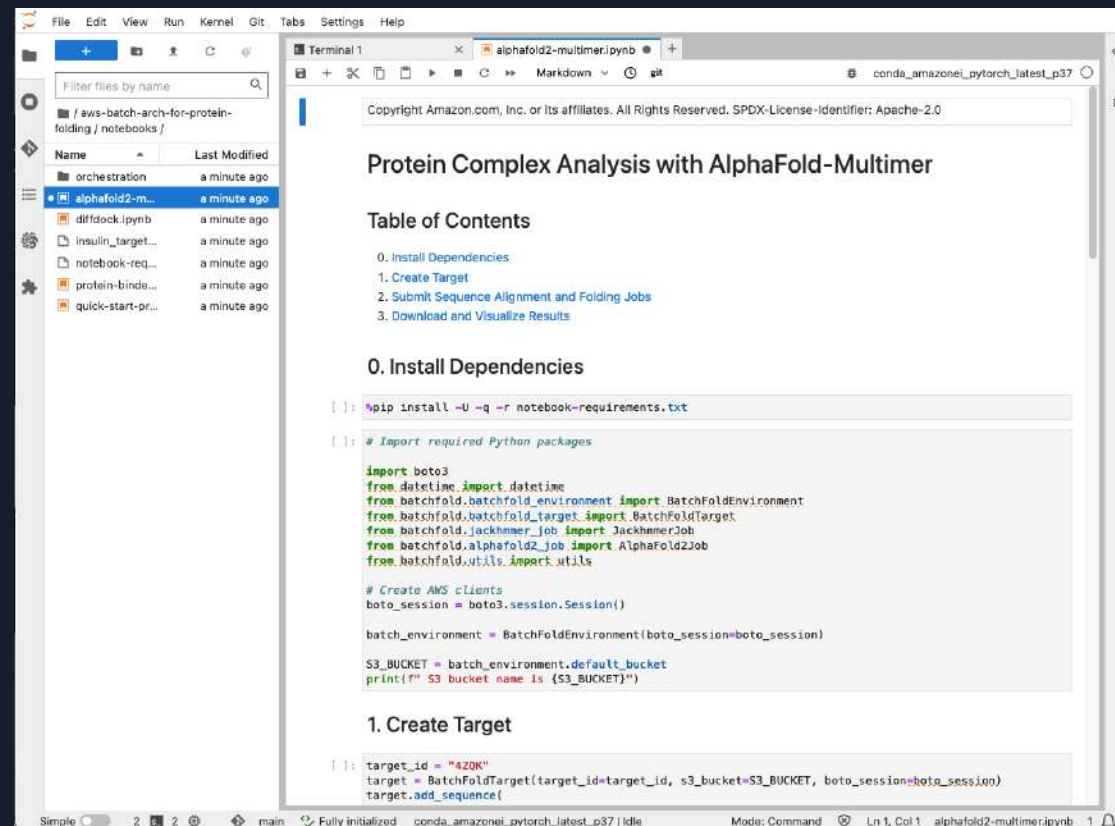
AWS Batch

AWS HealthOmics

実装サンプル：Jupyter でタンパク質構造解析

- タンパク質構造予測・設計用の機械学習モデルを SageMaker Notebook から実行できる
- オープンソースの実装例として提供
- 10個の多様なアルゴリズムに対応（後述）
- 裏側では AWS Batch 上でジョブ実行されて必要に応じて計算環境がスケールする
- 研究者は、SageMaker Notebook 経由でジョブの投入をおこなう
- IT 管理者は、起動テンプレートを使うことで簡単に構築可能

この後の塩野義製薬様のパートで詳しくご紹介いただきます



```
Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved. SPDX-License-Identifier: Apache-2.0

Protein Complex Analysis with AlphaFold-Multimer

Table of Contents

0. Install Dependencies
1. Create Target
2. Submit Sequence Alignment and Folding Jobs
3. Download and Visualize Results

0. Install Dependencies

[]: %pip install -U -q -r notebook-requirements.txt

[]: # Import required Python packages

import boto3
from datetime import datetime
from batchfold.batchfold_environment import BatchFoldEnvironment
from batchfold.batchfold_target import BatchFoldTarget
from batchfold.jackhammer_job import JackhammerJob
from batchfold.alphafold2_job import AlphaFold2Job
from batchfold.utils import utils

# Create AWS clients
boto_session = boto3.session.Session()

batch_environment = BatchFoldEnvironment(boto_session=boto_session)

S3_BUCKET = batch_environment.default_bucket
print(f" S3 bucket name is {S3_BUCKET}")

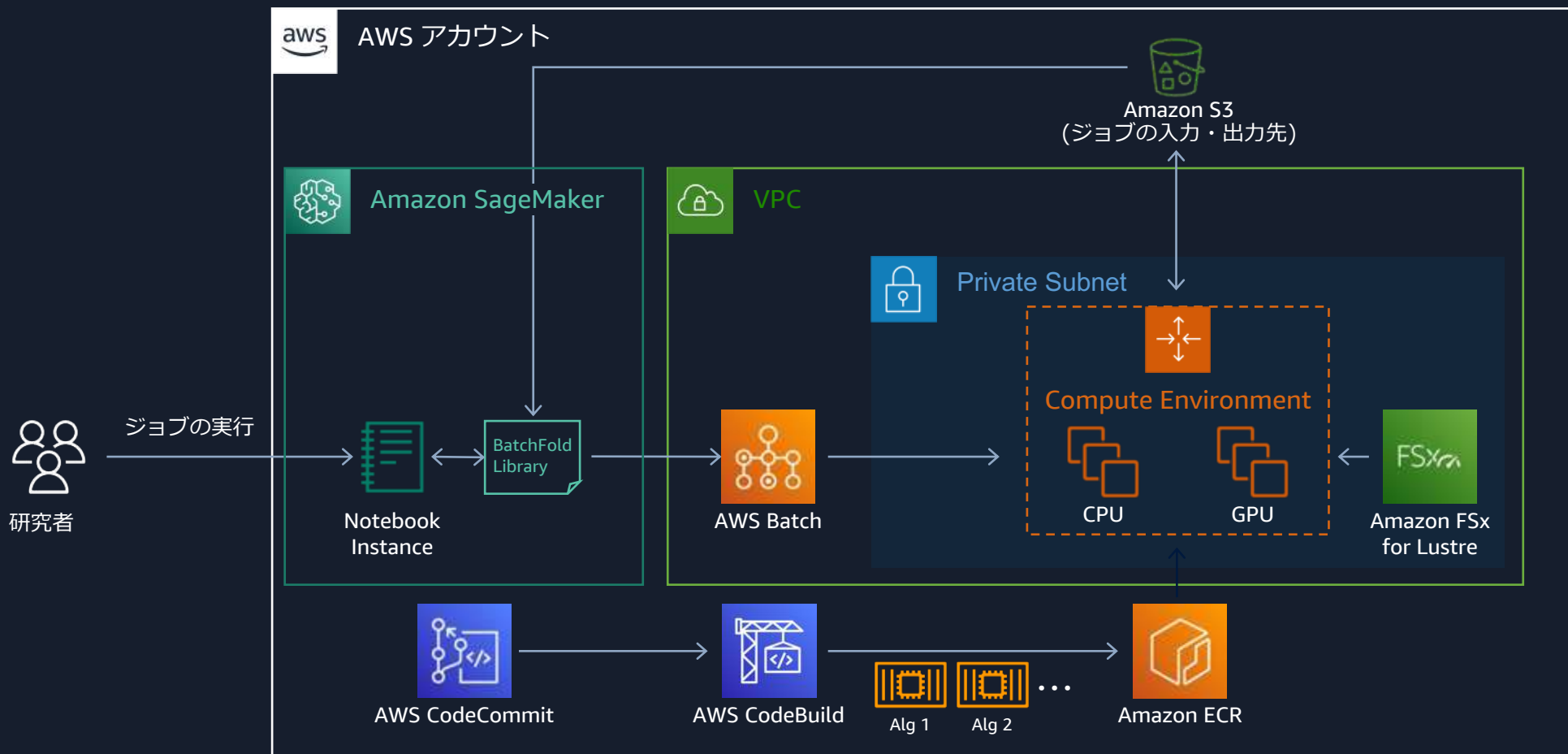
1. Create Target

[]: target_id = "420K"
target = BatchFoldTarget(target_id=target_id, s3_bucket=S3_BUCKET, boto_session=boto_session)
target.add_sequence()
```

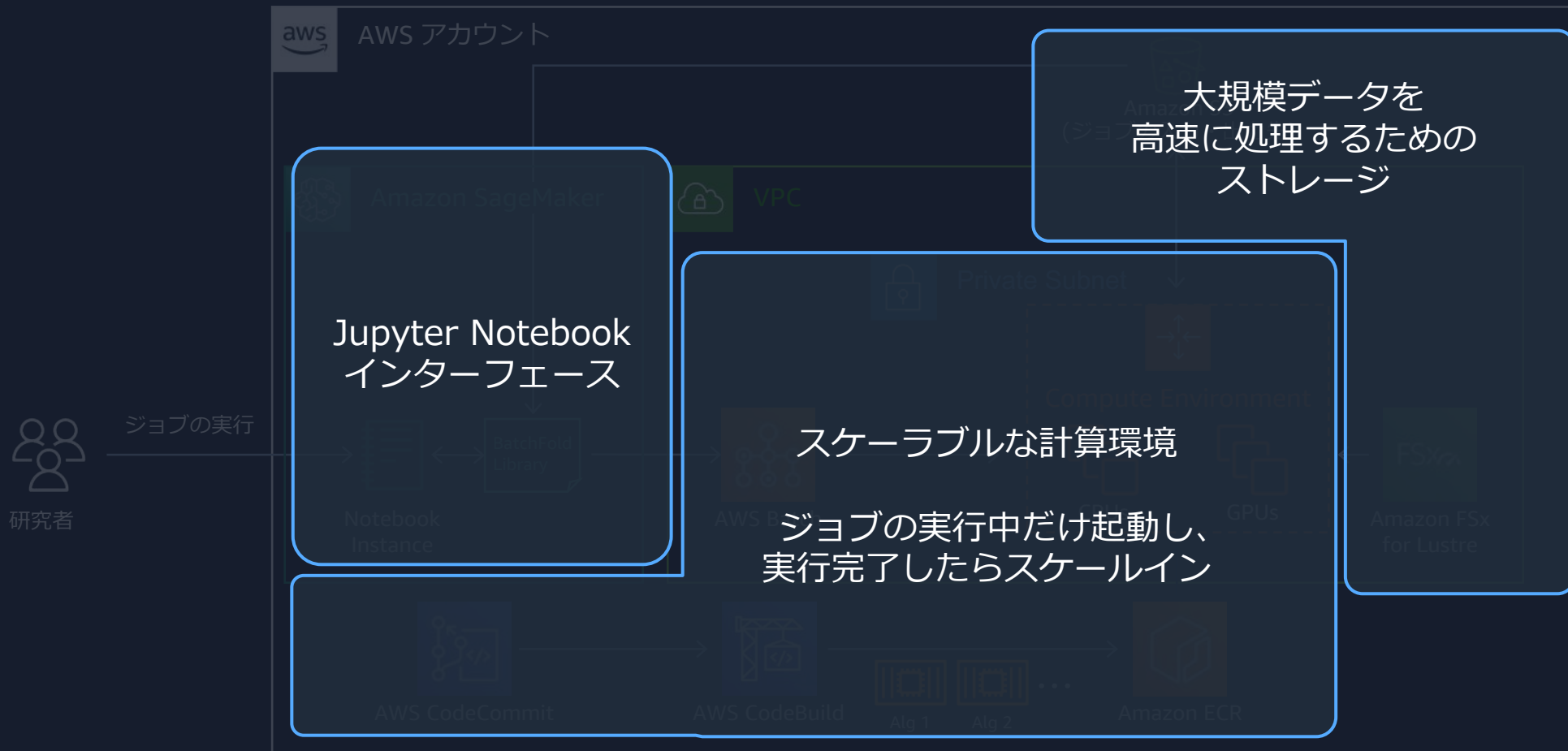
対応している主なアルゴリズム

- **MSA をベースとした構造予測**
 - **AlphaFold 2** (DeepMind)
 - **OpenFold** (Columbia University)
- **タンパク質言語モデル (pLM) をベースとした構造予測**
 - **OmegaFold** (Helixon US)
 - **ESMFold** (Meta Fundamental AI Research)
- **De novo タンパク質設計**
 - **RFDiffusion** (University of Washington)
 - **ProteinMPNN** (University of Washington)
- **ドッキングツール**
 - **DiffDock** (Massachusetts Institute of Technology)

アーキテクチャ全体像



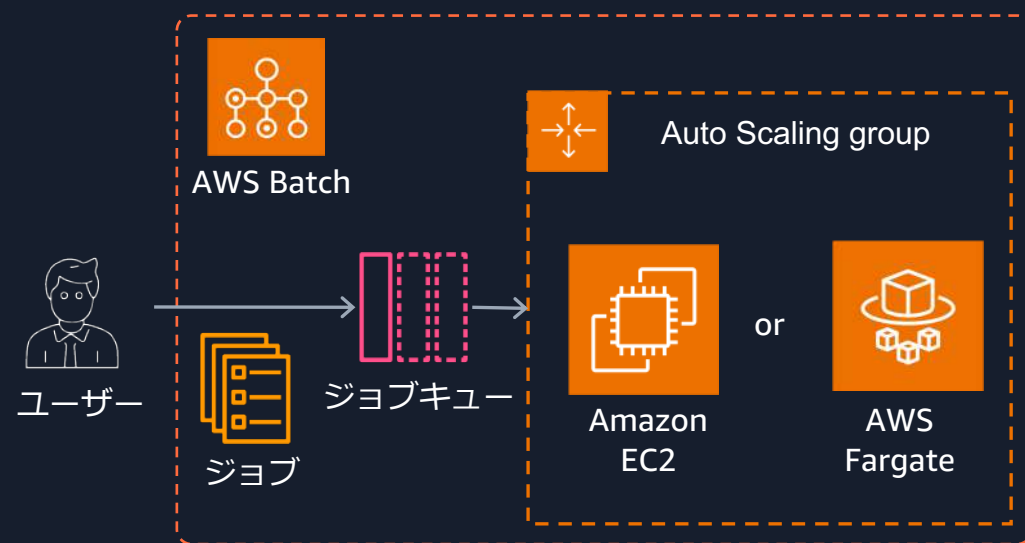
アーキテクチャ全体像



AWS Batch: スケーラブルな計算環境

大規模バッチ処理のため環境をフルマネージドで提供

- AWS Batch がインスタンスの起動や停止を行うため、スケジューラや計算ノードなどの管理が不要
- ジョブは Docker コンテナイメージ を元に作成し、自動でスケールするコンピューティング環境で実行する
- コンピューティング環境では
インスタンスタイプ・vCPU 数・
スポットインスタンス利用有無などを任意に指定可能



コンテナイメージを用意するだけで
スケーラブルな大規模バッチ処理環境が得られる

ニーズに合わせて組み合わせる

使い慣れたインターフェース + スケーラブルな計算環境

ウェブアプリ

Jupyter Notebook

リモートデスクトップ

CLI

AWS ParallelCluster

AWS PCS

AWS Batch

AWS HealthOmics

実装サンプル：ウェブアプリでタンパク質構造解析

タンパク質構造予測を手軽に実行できる
ウェブアプリ

オープンソースの実装例として提供

AlphaFold2 と ColabFold に対応

ジョブの投入・停止、結果の可視化

研究者は、AWS コンソールへのアクセス権限を
持っていないなくても簡単に使える

IT 管理者は、起動テンプレートを使うことで
ほぼ自動でウェブアプリを構築可能

The screenshot shows the AWS console interface for the 'AlphaFold2 Webapp on AWS' sample. The main content area is titled 'Create Protein Structure Prediction Job'. It features a text area for a FASTA file, a 'CREATE JOB' button, and a table of 'Protein Structure Prediction Jobs'. The table has columns for Job ID, Job Start Time, Job End Time, and Job Status. Below the table, there is a section for 'Protein Structure Prediction Job Result' showing a 3D protein structure model for Job ID 10.

Job ID	Job Start Time	Job End Time	Job Status
8	2023/08/26 15:24:40	2023/08/26 15:42:55	FAILED
9	2023/08/26 16:03:33	2023/08/26 16:03:36	FAILED
10	2023/08/26 16:07:22	2023/08/26 20:47:22	COMPLETED
11	2023/08/26 16:09:40	2023/08/26 21:34:48	COMPLETED
12	2023/08/26 22:26:16	2023/08/26 22:30:24	CANCELLED
13	2023/08/26 22:27:09	2023/08/26 22:30:25	CANCELLED

- [GitHub – AWS Samples “AlphaFold2 Webapp on AWS”](#)
- [AWS ブログ「大規模なタンパク質構造予測を行うための研究者向けウェブアプリを構築するには」](#)

本日お伝えしたいこと



多様な基盤モデルが進化を続ける中で
どのようにドライ・ウェットの連携を進めるか？

① 抽象化・スケーラビリティ

AWS のサービス・ツール・実装サンプルを活用して
基盤モデルのアップデートに追従するための時間とコストを節約

本日お伝えしたいこと



多様な基盤モデルが進化を続ける中で
どのようにドライ・ウェットの連携を進めるか？

① 抽象化・スケーラビリティ

② データ活用

本日お伝えしたいこと



多様な基盤モデルが進化を続ける中で
どのようにドライ・ウェットの連携を進めるか？

② データ活用

皆様の組織内のデータを活用し
最新のモデルをもとに素早くカスタマイズ

ニーズに合わせて組み合わせる

使い慣れたインターフェース + スケーラブルな計算環境

ウェブアプリ

Jupyter Notebook

リモートデスクトップ

CLI

AWS ParallelCluster

AWS PCS

AWS Batch

AWS HealthOmics

pLM のカスタマイズ：実装サンプル



AWS ParallelCluster

クラスタ管理の自動化ツール



[AWS Blog](#)



[AWSome Distributed Training \(BioNeMo\)](#)



Amazon SageMaker

機械学習プラットフォーム



[AWS Blog](#)



[Amazon SageMaker with NVIDIA BioNeMo](#)



Amazon EKS

Kubernetes マネージドサービス

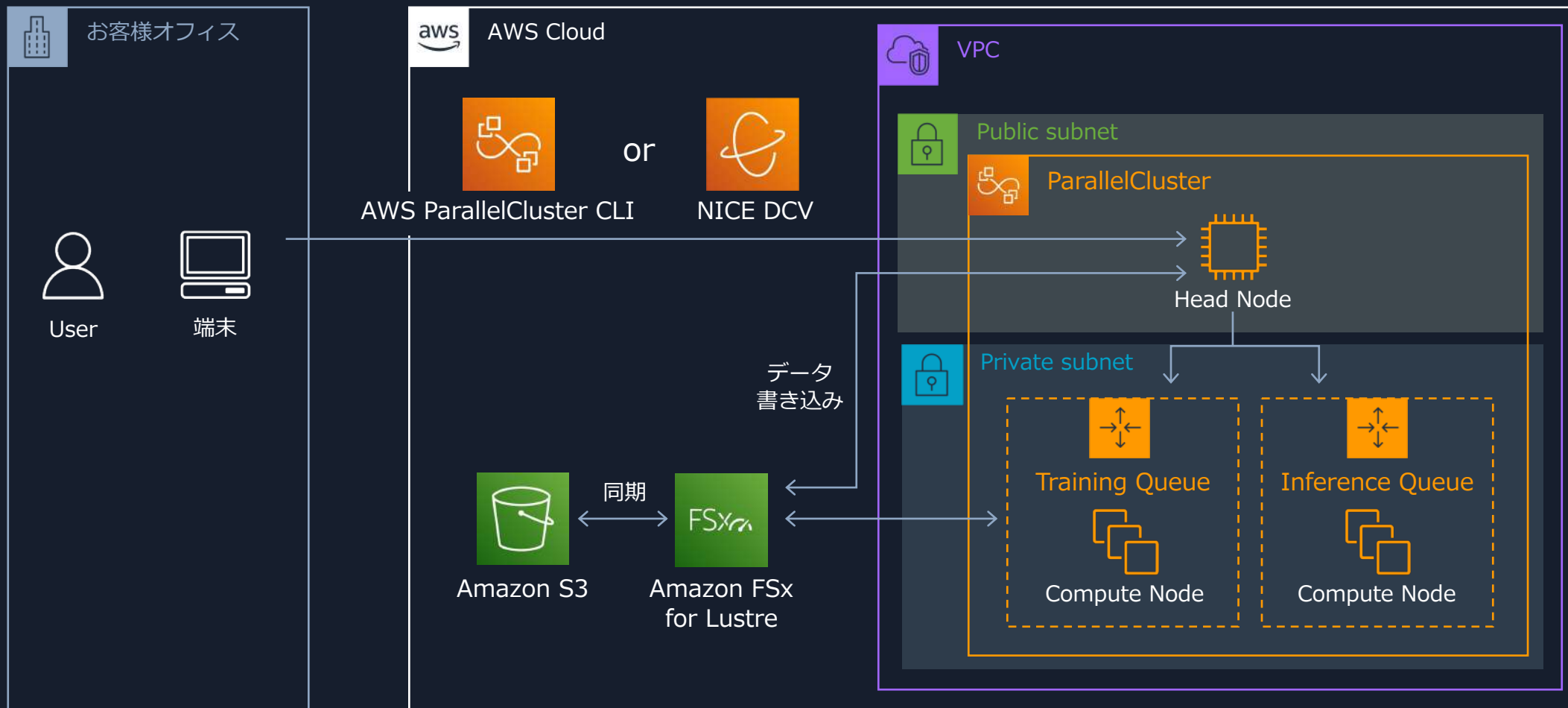


[AWS Blog](#)



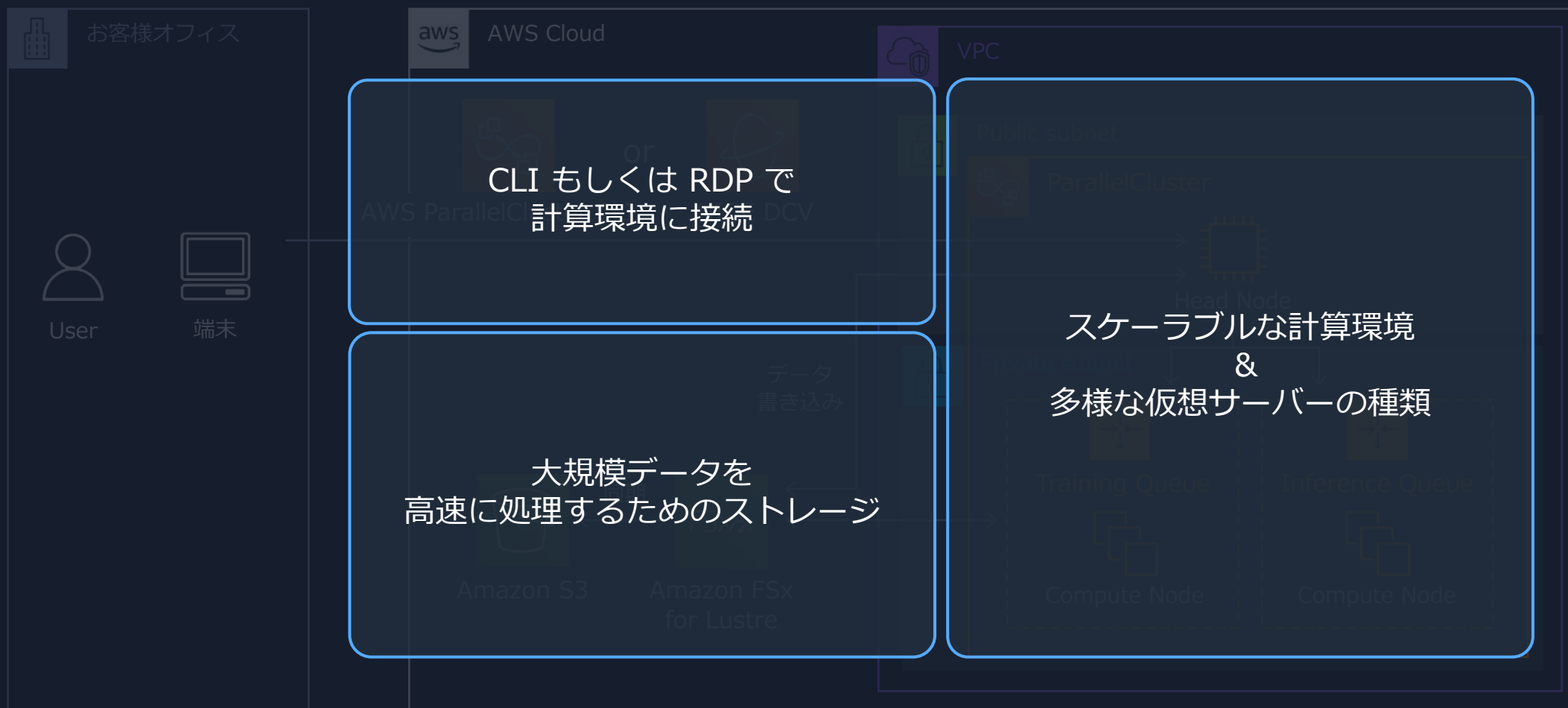
[Data on EKS \(BioNeMo\)](#)

分散学習のおすすめ構成例



[aws-samples/
awsome-distributed-training](https://github.com/aws-samples/awsome-distributed-training)

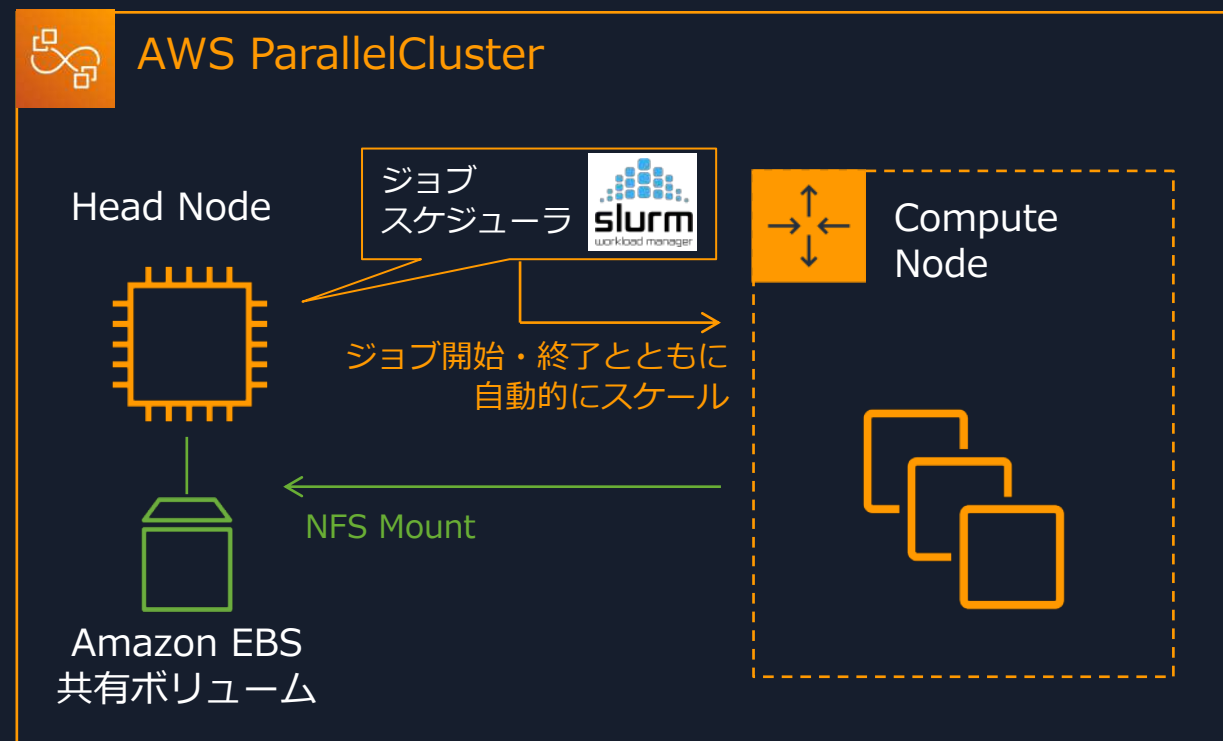
分散学習のおすすめ構成例



AWS ParallelCluster : クラスタ管理の自動化ツール

数コマンド操作でジョブ投入に応じて自動でスケールするクラスタを AWS 上に構築可能な AWS 公式のオープンソースソフトウェア

- 既存の HPC 向けスケジューラ (Slurm) と連動しスケラブルな計算環境を作成
- MPI/NCCL 環境がセットアップ済み
- x86/Arm 両方の環境に対応
- 使用するOSやネットワーク環境、ストレージ構成などを柔軟にカスタマイズ可能
- オープンソースのプロジェクトであり、誰でもソースコードを入手可能
<https://github.com/aws/aws-parallelcluster>



AWS では目的に応じて様々なアクセラレータの選択肢を提供



機械学習
(トレーニング)



NVIDIA H200
Tensor Core GPU

NEW



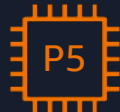
NVIDIA L40S GPU

GA!



AWS Trainium2

NEW



NVIDIA H100
Tensor Core GPU



NVIDIA A100
Tensor Core GPU



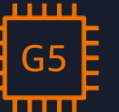
AWS Trainium

機械学習
(推論)



NVIDIA L40S GPU

GA!



NVIDIA A10G
Tensor Core GPU

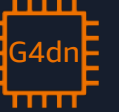


AWS Inferentia2



NVIDIA L4 GPU

GA!



NVIDIA T4 GPU



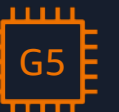
AWS Inferentia

グラフィックス

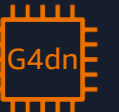


NVIDIA L4 GPU

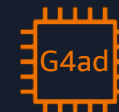
GA!



NVIDIA A10G
Tensor Core GPU



NVIDIA T4 GPU



AMD Radeon Pro
V520 GPU

内は2024年中にローンチ予定

内はアナウンス



本日お話ししたこと



多様な基盤モデルが進化を続ける中で
どのようにドライ・ウェットの連携を進めるか？

① 抽象化・スケーラビリティ

AWS のサービス・ツール・実装サンプルを活用して
基盤モデルのアップデートに追従するための時間とコストを節約

② データ活用

皆様の組織内のデータを活用し
最新のモデルをもとに素早くカスタマイズ


興味を持っていただいた方は…

1F 展示ホールのAWSブースに
ぜひお立ち寄りください





11/15(金) 第459回 CBI 学会講演会
「コンピュータが拓く創薬の新時代」
ぜひご参加ください

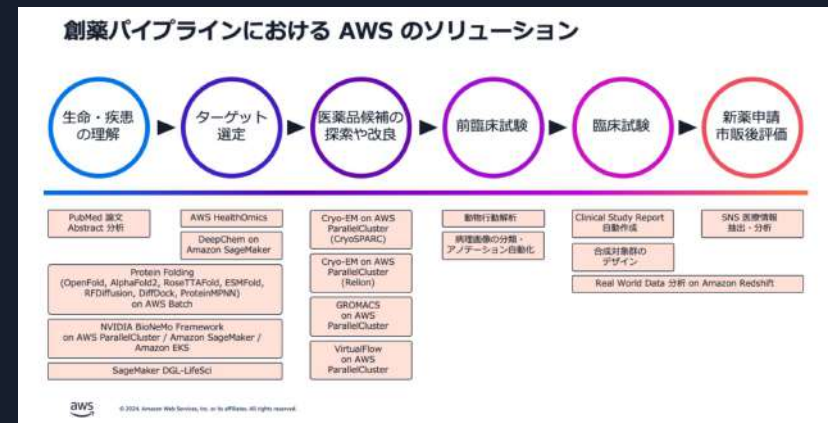
16:55-17:25

「創薬研究でAWSクラウドを活用する：
事例とデザインパターン」

 **創薬ワークロードに革新をもたらす
最新クラウドソリューション**

ご体感いただけるデモ

1. タンパク質立体構造予測・デザインにおける
AIモデルの大規模学習と推論 
2. マルチオミックス解析プラットフォーム
(データ解析、ワークフローの自動化・開発) 
3. AIEージェントによるバイオマーカー探索・
退院サマリー作成 
4. [新サービス] AWS Parallel Computing
ServiceでHPC環境の自動スケールが容易に 



<https://www.cbi-society.org/home/documents/seminar/2021to24/20241115.html>

Thank you!

Chiaki Ishio

ciishio@amazon.co.jp

