
AWS
re:Invent
NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

CMP326

The convergence of HPC and ML on AWS

Ankur Srivastava

Sr. Solutions Architect, Frameworks, WWSO
AWS

Shubha Kumbadakone

Sr. Specialist, Frameworks, WWSO
AWS

Sachin Kadyan

Research Associate, OpenFold Research
Columbia University

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

- Machine learning on AWS for life sciences
- OpenFold: Trainable Alphafold2
- Inference architectures for running OpenFold
- Training architecture

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Machine learning on AWS for life sciences

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Trends in BioPharma R&D

Computational Chemistry & Structural Biology

Modeling and Simulation

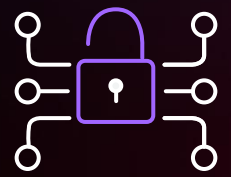
Imaging

Genomics

AI/ML

High Throughput Computing

High Performance Computing



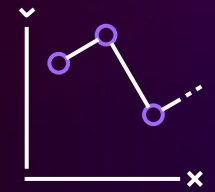
Compliant and Secure Environment



Global, Fault-Tolerant Infrastructure



Accelerated Time to Insight



Scalability and Dynamic Resourcing

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AI/ML in research and discovery



Research and discovery

Key machine learning use-case trends

- Assay hit optimization
- Predictive toxicology
- Predictive computational chemistry
- Target identification and validation
- Molecular structure predictions
- Process phenotypic, imaging, drug, and genomics data sets
- Predict synthesis pathways
- Antibody binding affinity predictions

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AI/ML across the HCLS value chain



Research and discovery

Molecular structure and activity prediction

Image labeling and screening

Target identification and validation

Clinical development

Optimization and forecasting trial timelines

Patient cohort, site local recruiting

Protocol optimization

Manufacturing and supply chain

Predictive modeling for input & output product demands

Predictive equipment maintenance

Shop floor IoT data for machine learning model building

Commercial

Next best action for sales and marketing

Prediction of disease progression or medication switching

Predict patient cohorts

Post-market surveillance and patient support

Prediction of patient health events

Automate adverse event detection

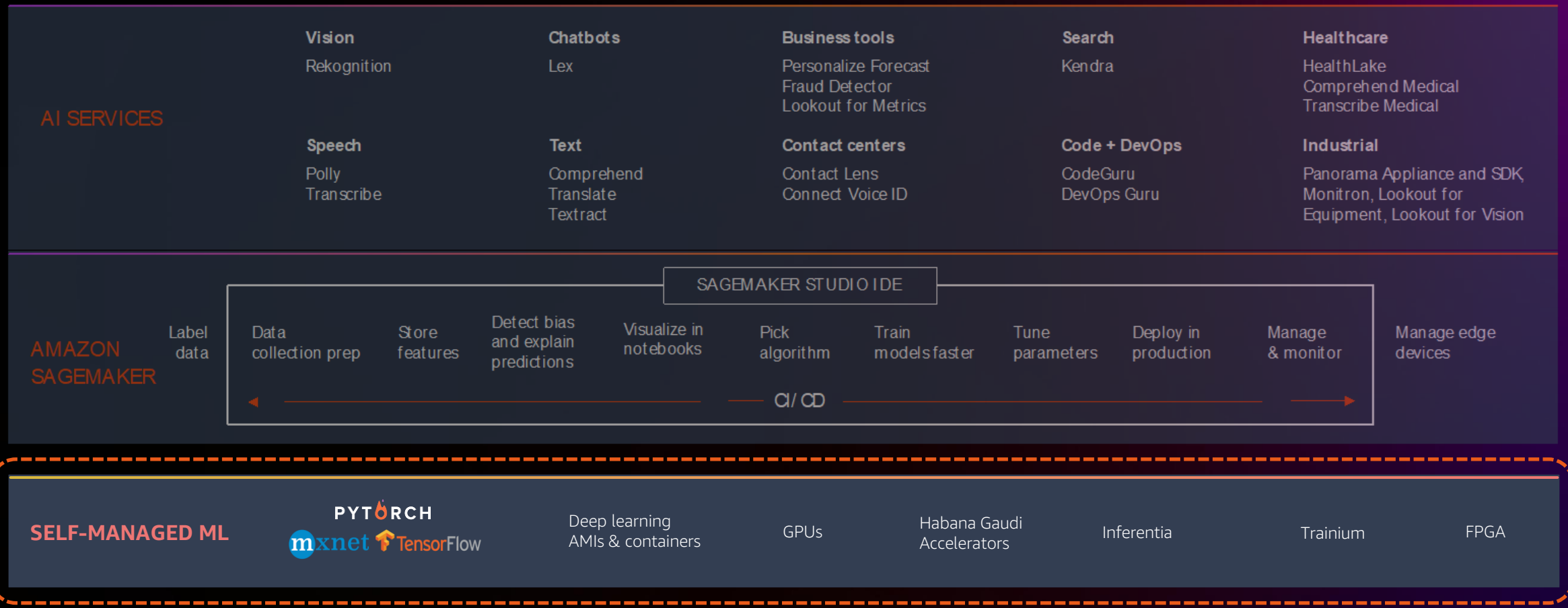
Pragmatic clinical trials

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The AWS ML stack

BROADEST AND MOST COMPLETE SET OF MACHINE LEARNING CAPABILITIES



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

ML Frameworks & Infrastructure Layer

TOOLS

DATA PROCESSING & LABELING


Kinesis


Glue


SageMaker Ground Truth

PREBUILT IMAGES


DLAMI


DL Containers

DEV, TRAINING, & TUNING



Notebooks



Horovod


Kubeflow Operators

INFERENCE SERVERS

Torch Serve

Nvidia Triton

TF Serving

MLOps


Kubeflow

ORCHESTRATION

AWS SERVICES


EKS


ECS


Fargate


Batch


Parallel Cluster

WORKFLOW

OPEN SOURCE SOFTWARE


Kubernetes


Kubeflow


Airflow


Ray

OTHER SERVICES

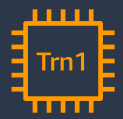

Lambda


Step Functions


Cloud Formation

CORE INFRASTRUCTURE

COMPUTE

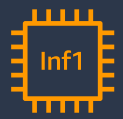












STORAGE


S3


EFS



FSxL

NETWORKING


EFA

DATA TRANSFER


Direct connect

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

*OSS – Open source software

Paige.ai accelerates training and deployment of Computer Vision Model for Digital Pathology

Model

Computer Vision model with <100M parameters and running MIL-based deep learning*

Images are tiled into slides, training dataset is 40K slides

Training

Compute: Amazon EC2 P4d instances and Docker images

Storage and File System: FSx for Lustre and Storage Gateway

Cluster Stack: Slurm, CUDA , PyTorch

Inference

Compute: Amazon EC2 G4dn instances

Orchestration: Kubernetes

Set their training and inference architectures in the cloud quickly while choosing the best compute options to meet their price performance needs

*https://www.nature.com/articles/s41591-019-0508-1

Paige

Company: Paige AI

Industry: Healthcare

Country: United States

Website: https://paige.ai/

About Paige AI

Paige.ai is a startup that uses the incredible power of tissue-based AI to develop a new generation of digital diagnostics and predictive tests, empowering pathologists and transforming oncology

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

"We used Amazon EC2 G5 instances to run inference using a large PyTorch model, ESM1-b, for protein sequence prediction. AWS offered agility and a strong partnership that helped us implement an architecture that quickly scaled up our workflow. Using this architecture, we were able to boost performance and operate on orders of magnitude more protein sequences."

Matt Biggs, PhD, Sr. Computational Biologist at Sestina Bio



Sestina Bio

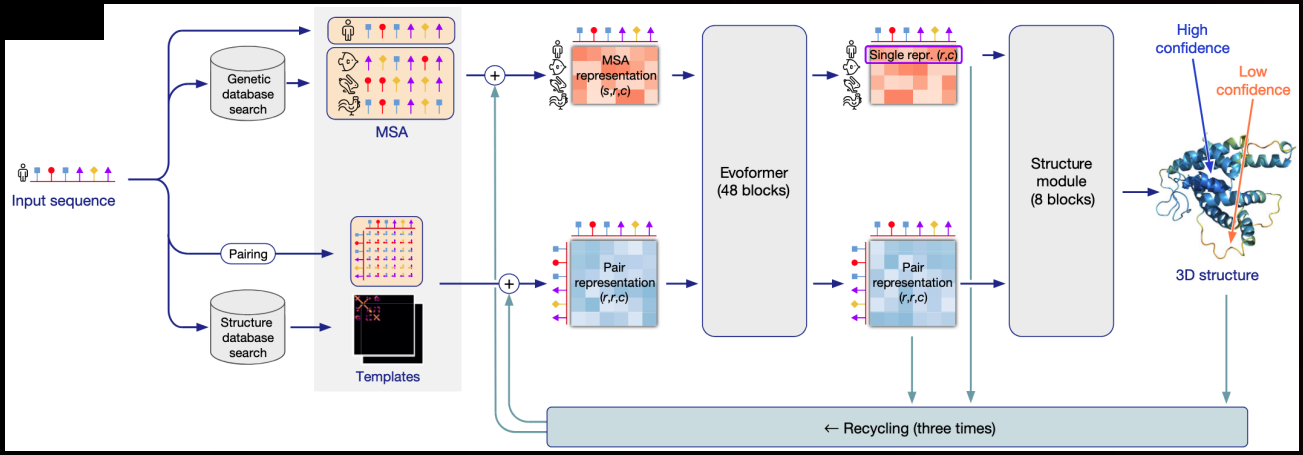
Sestina Bio is creating a new platform for data-driven synthetic biology applying machine learning to develop new tools for genome editing and single cell biology

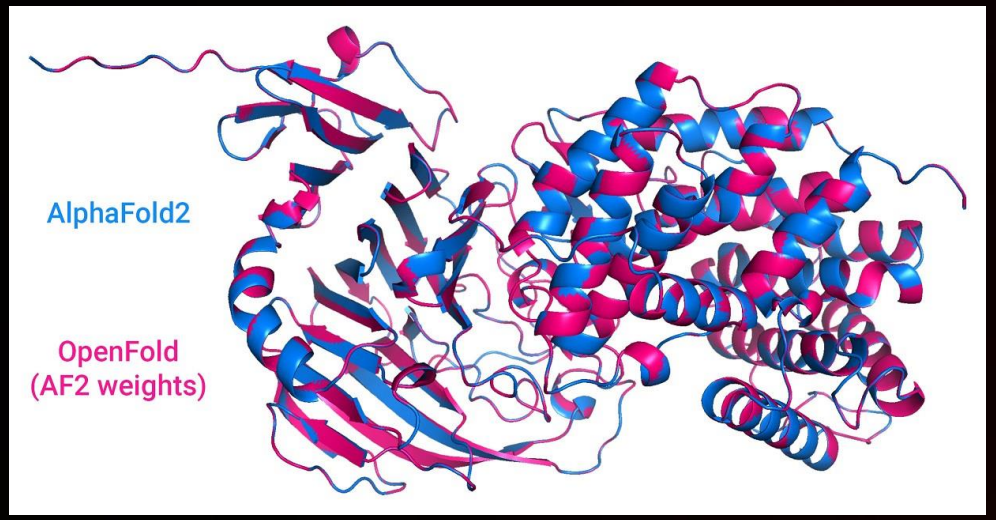
aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS and Columbia University collaboration

OpenFold: PyTorch-based open sourced trainable AlphaFold2





Registry of Open Data on AWS

OpenFold Training Data

Description

Multiple sequence alignments (MSAs) and template hits for 634,000 Protein Data Bank chains and approximately 350,000 UniClust30 clusters. MSAs were generated with HHBlits and JackHMMER against MGnify, BFD, UniRef90, and UniClust30 while templates were identified from PDB70 with HHSearch, all according to procedures outlined in the supplement to the AlphaFold 2 Nature paper, Jumper et al. 2021. We expect the database to be broadly useful to structural biologists training or validating deep learning models for protein structure prediction and related tasks.

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

OpenFold: Trainable Alphafold2

CASP 15 ENTRANT

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Importance of protein structures in biology

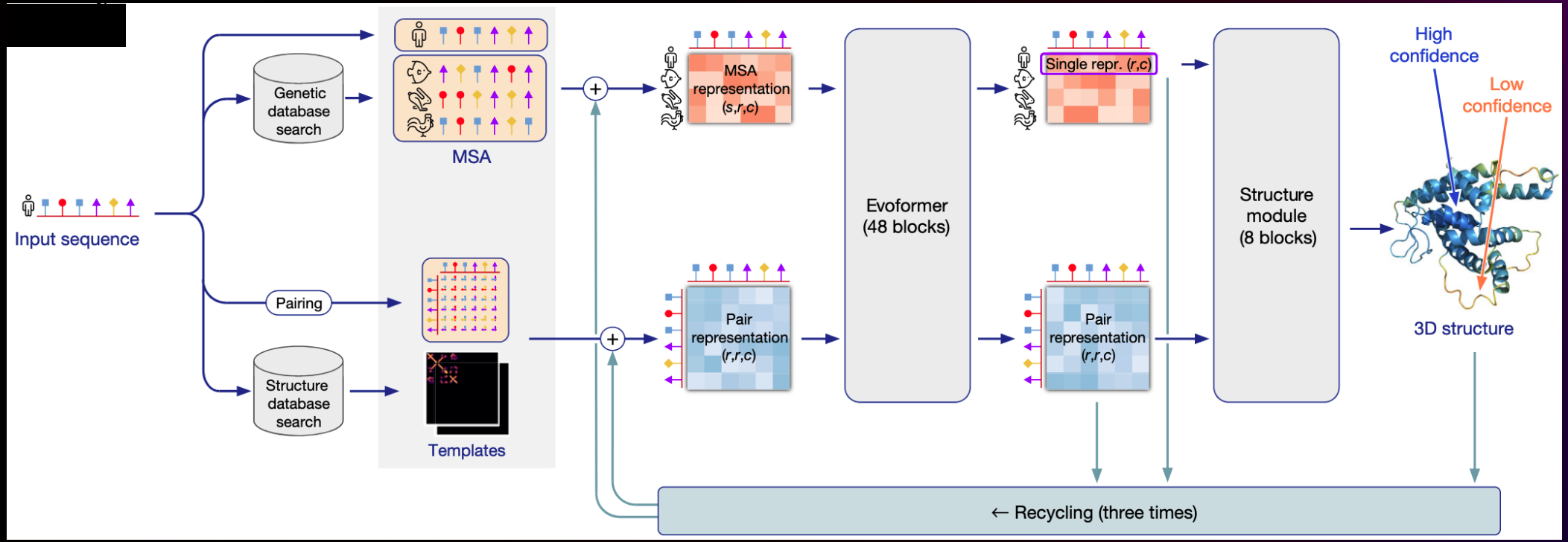
- Proteins are the foundational building blocks of all living organisms
- The 3-dimensional shape of a protein largely determines its function
- Proteins have deterministic, native, and replicable structures
- Upon synthesis, proteins “fold” into their native structures, within ms
- Predicting the way proteins fold is a central problem across molecular biology and other industries
- Traditional methods for determining protein structure (X-Ray Crystallography, Mass Spectrometry) are expensive and take years

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

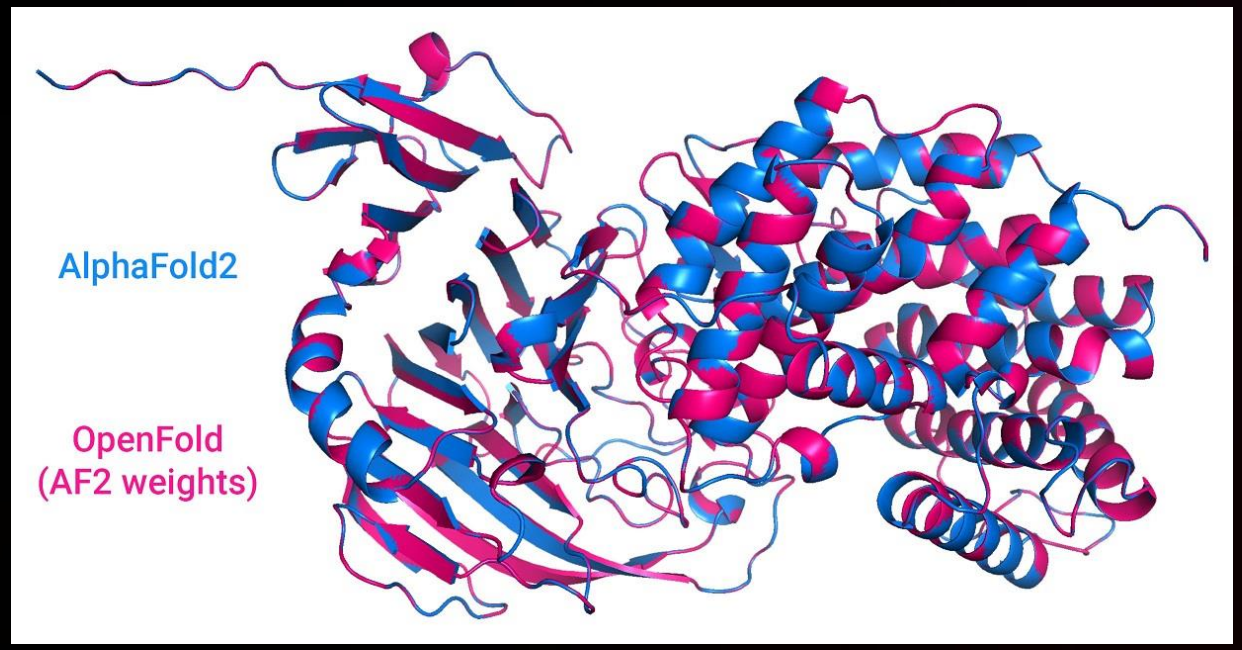
Machine learning for protein folding

- In recent years, computational methods like machine learning have been increasingly used to solve the protein structure prediction (PSP) problem
- AlphaFold, created by DeepMind, was a revolution in the field of PSP and recognized by Science magazine as the 2021 Breakthrough of the Year
-



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

OpenFold: PyTorch-based, open-source, trainable AlphaFold2



Performance at par with AlphaFold2

Supports FlashAttention memory optimizations that accelerate the MSA attention component

Numerous optimizations accelerate the inference pipeline, delivering twice the performance for shorter protein sequences than AlphaFold2

In-place attention and low-memory attention optimizations support predictions of protein structures up to 4,600 residue long, on Amazon EC2 p4d instances

MMSeqs2-based MSA generation pipeline is more efficient than AlphaFold2

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Extending scope: OpenFold single-sequence model design

Some prior enforcements on the model that influenced the design choices

Enforcement

- Must work on orphan proteins
- Not rely on evolutionary information
- Still perform well using the existing OpenFold architecture
- Work for protein designs for which similar structures are available
- Accelerated inference times

Implication

- No alignments can be used
- No attention between different protein sequences or alignments
- Uses embeddings from protein language models instead of generating embeddings from scratch
- Allows protein structure templates
- Eliminate computation heavy evolutionary information

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Results and comparison

Achieved 0.86 LDDT scores compared to the 0.91 LDDT scores for AlphaFold.

25% less parameters than AlphaFold, implying shorter training times.

Elimination of the majority of alignment process heavily speeds up inference jobs from hours to minutes.

Works for every kind of protein, including orphan proteins that don’t have evolutionary siblings.

Heavily optimized in time and memory using memory-efficient attention, chunking, checkpointing, and tensor offloading capabilities.

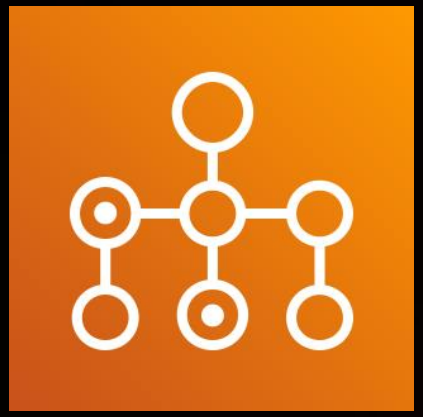
aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Running OpenFold inference

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

OpenFold inference



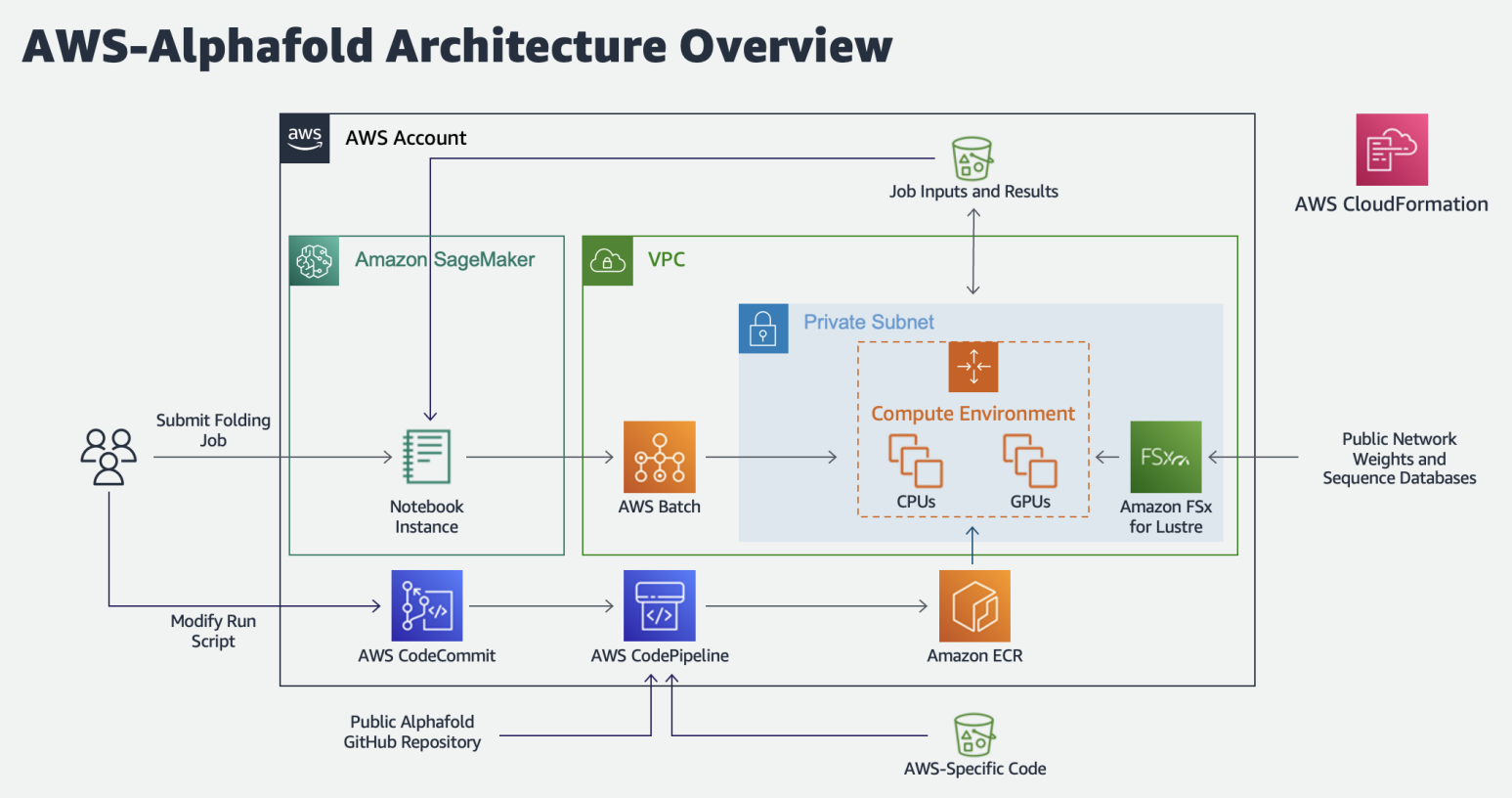
AWS Batch



Amazon EKS

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Protein folding inference reference architecture – AWS Batch



Runs each step as a containerized job using AWS Batch with the best fit of CPU, memory, and GPU resources

GitHub repo contains AWS CloudFormation template, Jupyter notebook, and supporting code to run protein analysis algorithms on AWS Batch

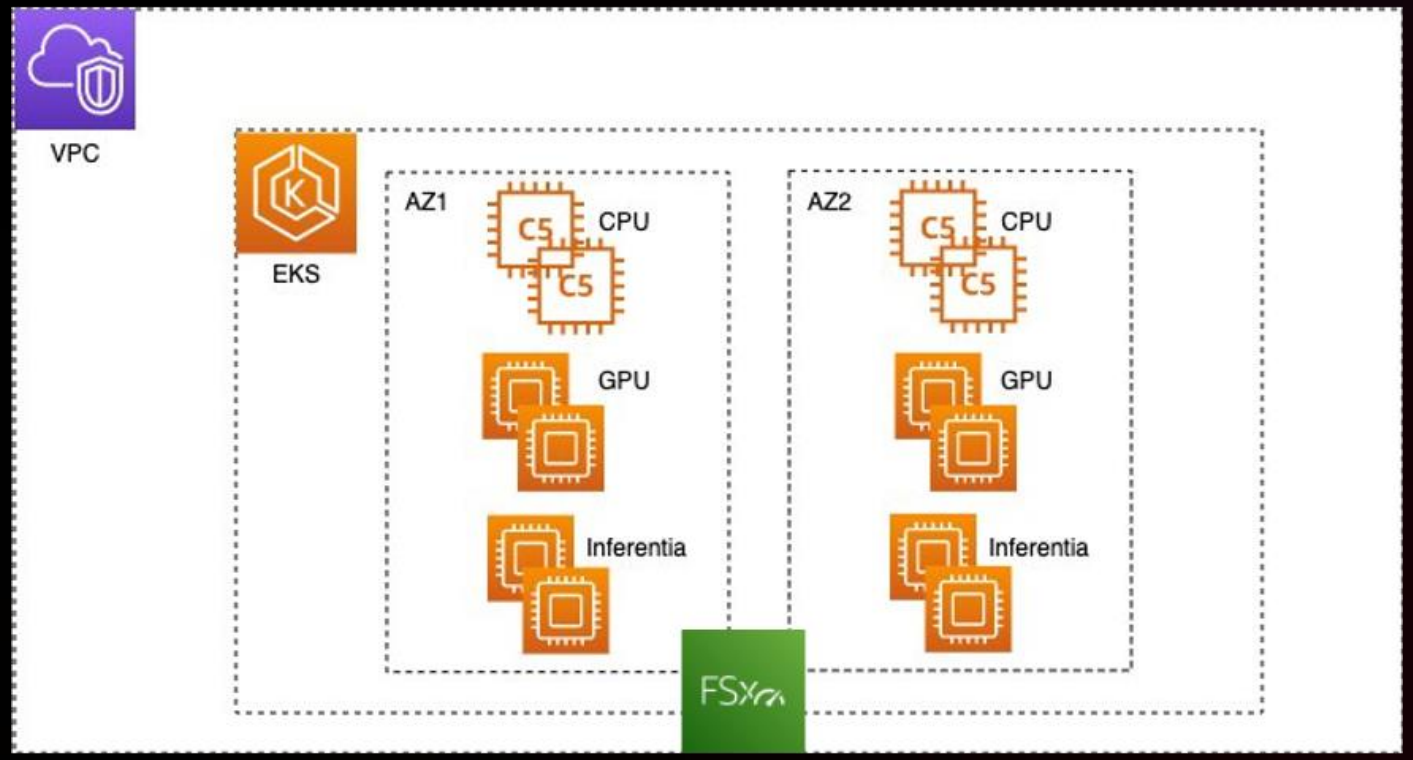
Modules include AlphaFold-2, AlphaFold-Multimer, OpenFold, and OmegaFold

https://github.com/aws-samples/aws-batch-arch-for-protein-folding

https://aws.amazon.com/blogs/hpc/optimize-protein-folding-costs-with-openfold-on-aws-batch/

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Protein folding inference reference architecture – Amazon EKS



- Deployed on CPU- and GPU-based instances
- MSA step runs on memory-optimized R6 instances
- Exposes MSA computation functionality as an API so it can be executed in a parallel manner
- Builds deep learning inferencing service on Amazon EKS using PyTorch FastAPI

Tested on OpenFold protein folding model

https://github.com/aws-samples/aws-do-openfold-inference

https://aws.amazon.com/blogs/machine-learning/run-inference-at-scale-for-openfold-a-pytorch-based-protein-folding-ml-model-using-amazon-eks/

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How to set up an Amazon EKS cluster

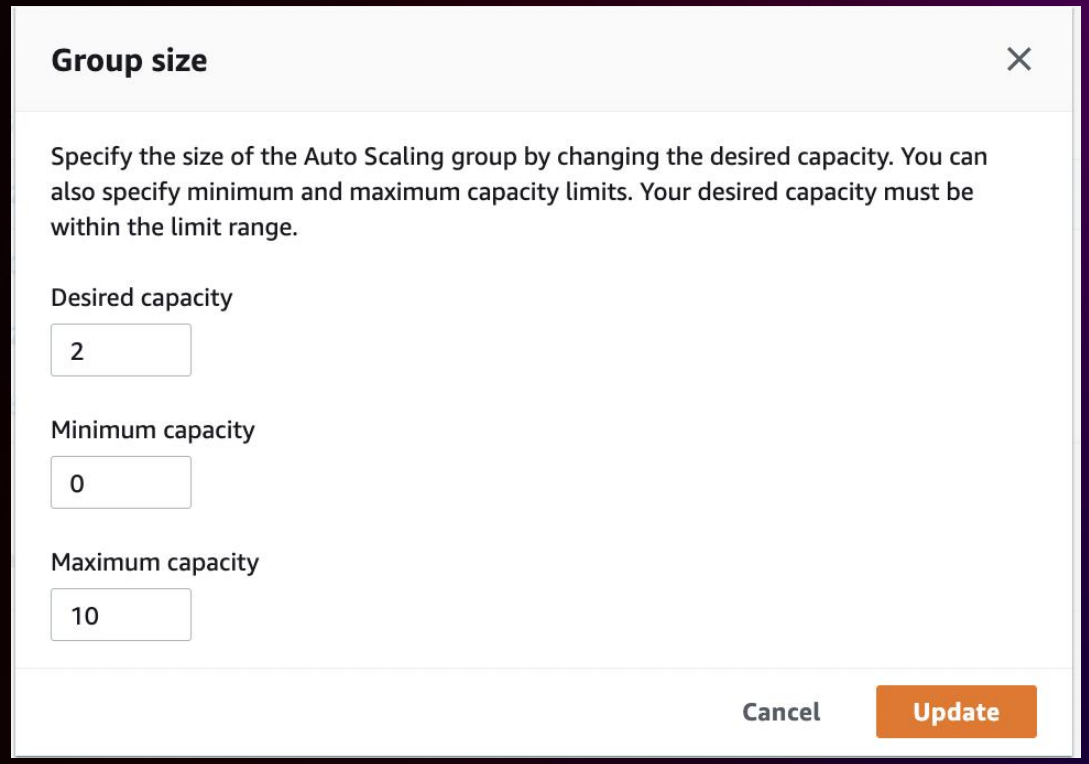
Use aws-do-eks framework (https://github.com/aws-samples/aws-do-eks)

Walkthrough config

- Steps:
  - Fill out config
  - Build, run, and execute in a container
  - eks-create.sh: Create Auto Scaling groups
  - eks-scale.sh: Scale nodes to desired capacity
- Spot Instances
- Custom AMIs

After a cluster is created:

- New nodegroups can be added after a cluster is created with eks-nodegroup-create.sh
- Nodegroups can be scaled via eks-scale.sh or via console



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Download data and model files

```
./download-openfold-data/build.sh
./download-openfold-data/push.sh
kubectl apply -f fsx-data-prep-pod.yaml
```

Download scripts:
https://github.com/aws-samples/aws-do-openfold-inference/tree/main/download-openfold-data/scripts

```
root@fsx-share-test:/fsx-shared/openfold/data# ls -l
total 325
drwxr-xr-x 2 root root 33280 Jul 12 20:29 bfd
drwxr-xr-x 2 root root 33280 Jul 12 20:56 mgnify
drwxr-xr-x 2 root root 33280 Jul 12 15:25 params
drwxr-xr-x 2 root root 33280 Jul 12 21:17 pdb70
drwxr-xr-x 3 root root 33280 Jul 12 23:04 pdb_mmcif
drwxr-xr-x 2 root root 33280 Jul 12 23:05 pdb_seqres
drwxr-xr-x 2 root root 33280 Jul 12 23:12 small_bfd
drwxr-xr-x 3 root root 33280 Jul 12 23:32 uniclust30
drwxr-xr-x 2 root root 33280 Jul 13 00:07 uniprot
drwxr-xr-x 2 root root 33280 Jul 13 01:07 uniref90
```

OpenFold data

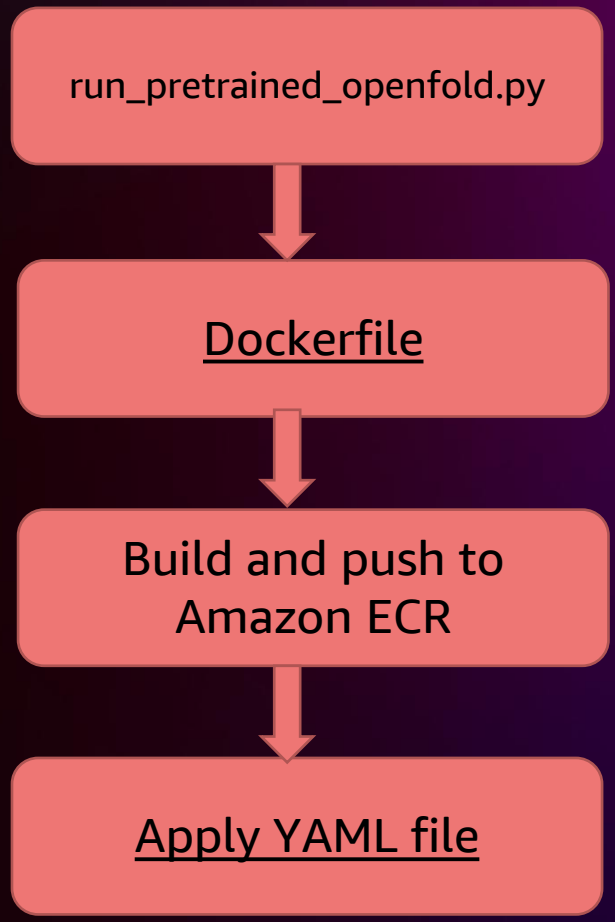
```
(base) awsankur@a07817abbc7e openfold_params % ls -l
total 5136008
-rw-r--r--  1 awsankur  staff      18657 Jul 12 09:36 LICENSE
-rw-r--r--  1 awsankur  staff       1613 Jul 12 09:36 README.md
-rw-r--r--  1 awsankur  staff  374586533 Jul 12 09:39 finetuning_1.pt
-rw-r--r--  1 awsankur  staff  374586533 Jul 12 09:39 finetuning_2.pt
-rw-r--r--  1 awsankur  staff  374586533 Jul 12 09:39 finetuning_3.pt
-rw-r--r--  1 awsankur  staff  374586533 Jul 12 09:39 finetuning_4.pt
-rw-r--r--  1 awsankur  staff  374586533 Jul 12 09:39 finetuning_5.pt
-rw-r--r--  1 awsankur  staff  374620131 Jul 12 09:39 finetuning_ptm_1.pt
-rw-r--r--  1 awsankur  staff  374620131 Jul 12 09:39 finetuning_ptm_2.pt
-rw-r--r--  1 awsankur  staff    7433216 Jul 12 09:36 initial_training.pt
```

OpenFold models

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Run OpenFold inference

- **run_pretrained_openfold.py**: End-to-end Python code to:
  - Load OpenFold models
  - Compute alignments
  - Run inference
  - Other functionality:
    - Model tracing
    - Relaxation
    - Multi-chain FASTAs
    - Jax model parameters
    - Deep speed checkpoint
    - Multi-model mode
    - subtract_plddt
    - Save output pickle files



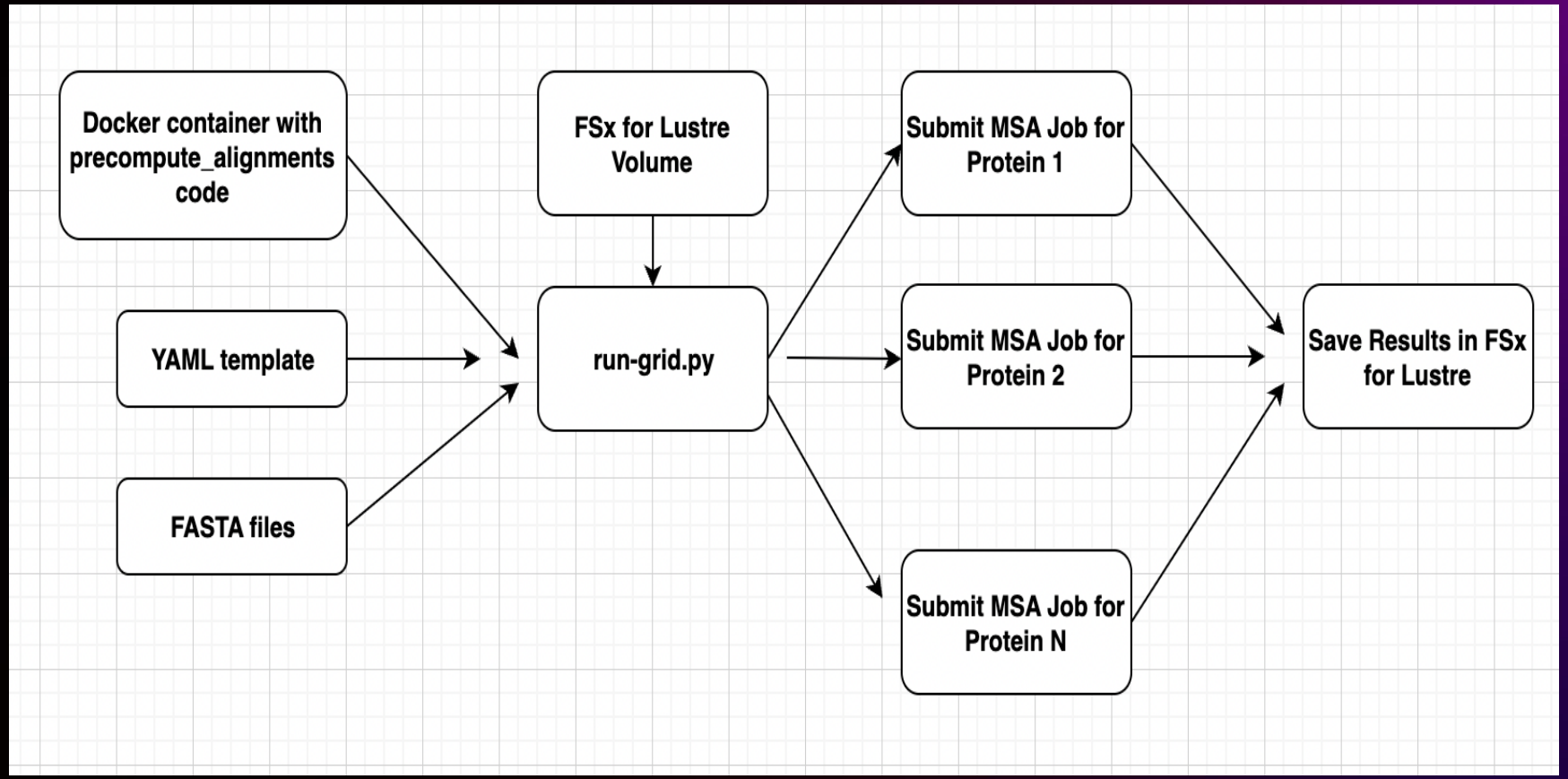
aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Multi sequence alignment computation

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Running alignment computation

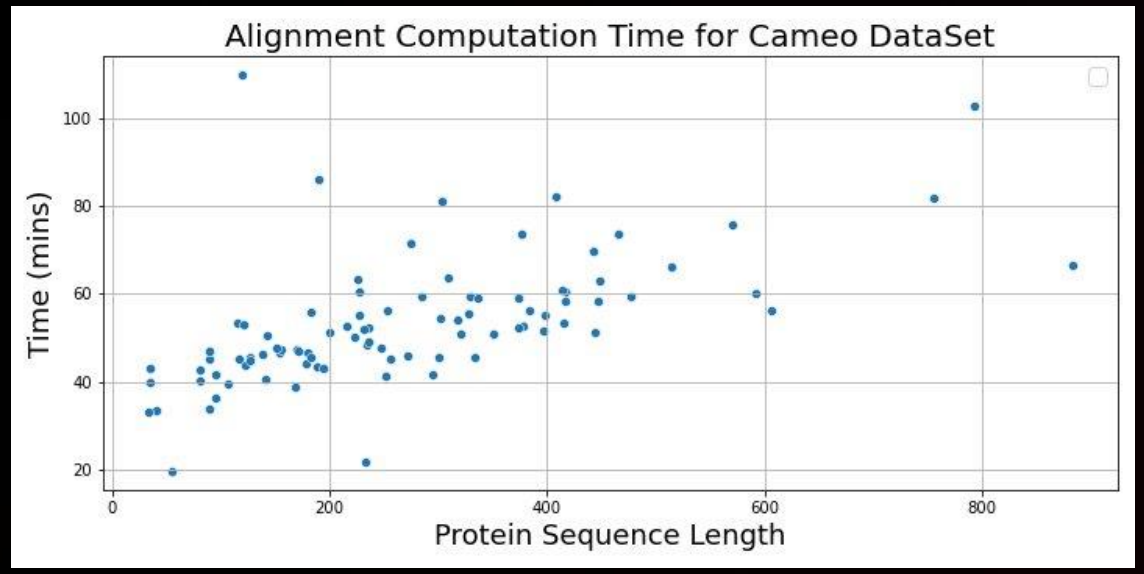
- Could take a couple of hours for some proteins
- Leverage the EKS cluster to run alignment computation for each protein's sequence in parallel in a separate Amazon EKS job
- Use memory-optimized instances such as r6i.xlarge

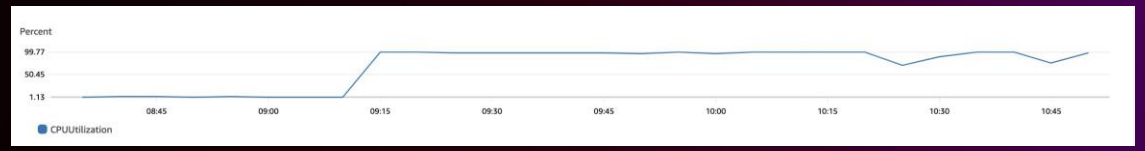


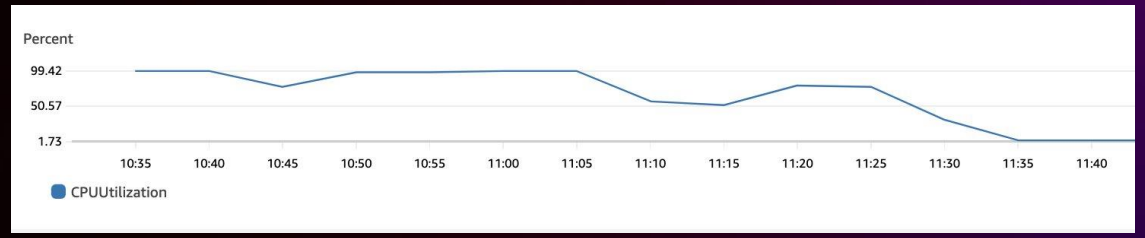
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Alignment computation using Cameo Dataset

Instance Type	Total Memory Available	Total vCPUs Available	Requested Pod Memory	Number of pods	Time Taken	On-Demand Hourly Cost	Total Cost
r6i.xlarge	32 GB	4	16 GB	92	2.5 hours	0.252 $/hr	$58
c6i.12xlarge	96 GB	48	Default	1	49 hours 43 mins	2.04 $/hr	$101







Max CPU Utilization

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

OpenFold PyTorch Model Fast API end points

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Exposing models with FastAPI

- Inference config
- FastAPI file
- ./deploy.sh
- Calling models

```
openfold_model_name=finetuning_ptm_2
processor=gpu
output_dir_base=/fsx-shared/openfold/cameo/inference_output/
alignment_dir=/fsx-shared/openfold/cameo/r6i_xlarge_run3_16GB/
config_preset=model_1_ptm
use_precomputed_alignments=True
skip_relaxation=True
no_cpus=4
# number of models per model server
num_models=1

# service_port=8080 - port on which model service will be exposed
service_port=8080

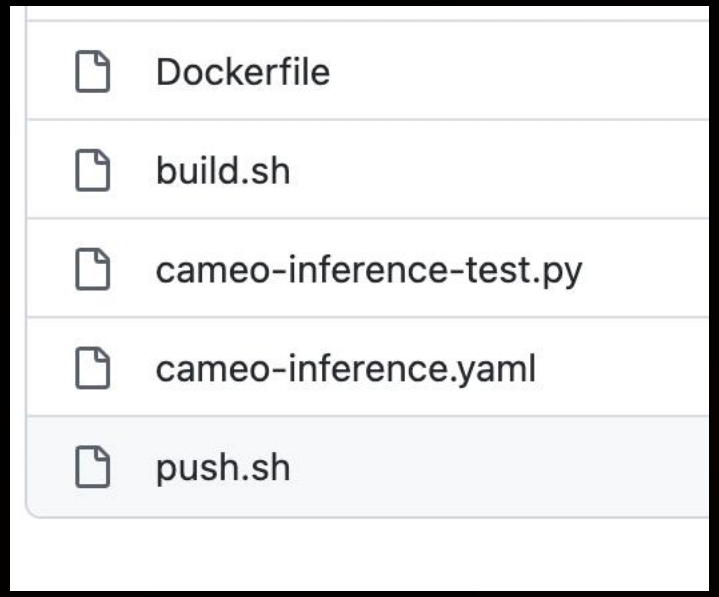
runtime=kubernetes
# Kubernetes-specific deployment settings
# instance_type = c5.xxx | g4dn.xlarge | g4dn.12xlarge | inf1.xlarge | inf1.6xlarge | ...
# A node group with the specified instance_type must exist in the cluster
# The instance type must have the processor type configured above
instance_type=g4dn.xlarge
# num_servers - number of model servers to deploy
# note that more than one model server can run on a node with multiple cpu/gpu/inferentia chips.
# example: 4 model servers fit on one inf1.6xlarge instance as it has 4 inferentia chips.
num_servers=1
```

Inference Config

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Exposing models with FastAPI

- GPU YAML template
- ./deploy.sh
- Scripts for PyTorch model inference




```
[ec2-user@ip-172-31-14-147 openfold]$ kubectl get svc -o wide
NAME             TYPE        CLUSTER-IP       EXTERNAL-IP   PORT(S)    AGE   SELECTOR
kubernetes       ClusterIP   10.100.0.1       <none>        443/TCP    55d   <none>
openfold-gpu-0   ClusterIP   10.100.126.119   <none>        8080/TCP   14d   app=openfold-gpu-0
openfold-gpu-1   ClusterIP   10.100.227.188   <none>        8080/TCP   14d   app=openfold-gpu-1
```

Deployed services

```
Alignment computation apis:
http://openfold-gpu-0.default.svc.cluster.local:8080/precompute_alignments/?tag="+tag+"&seq="+seq
http://openfold-gpu-1.default.svc.cluster.local:8080/precompute_alignments/?tag="+tag+"&seq="+seq

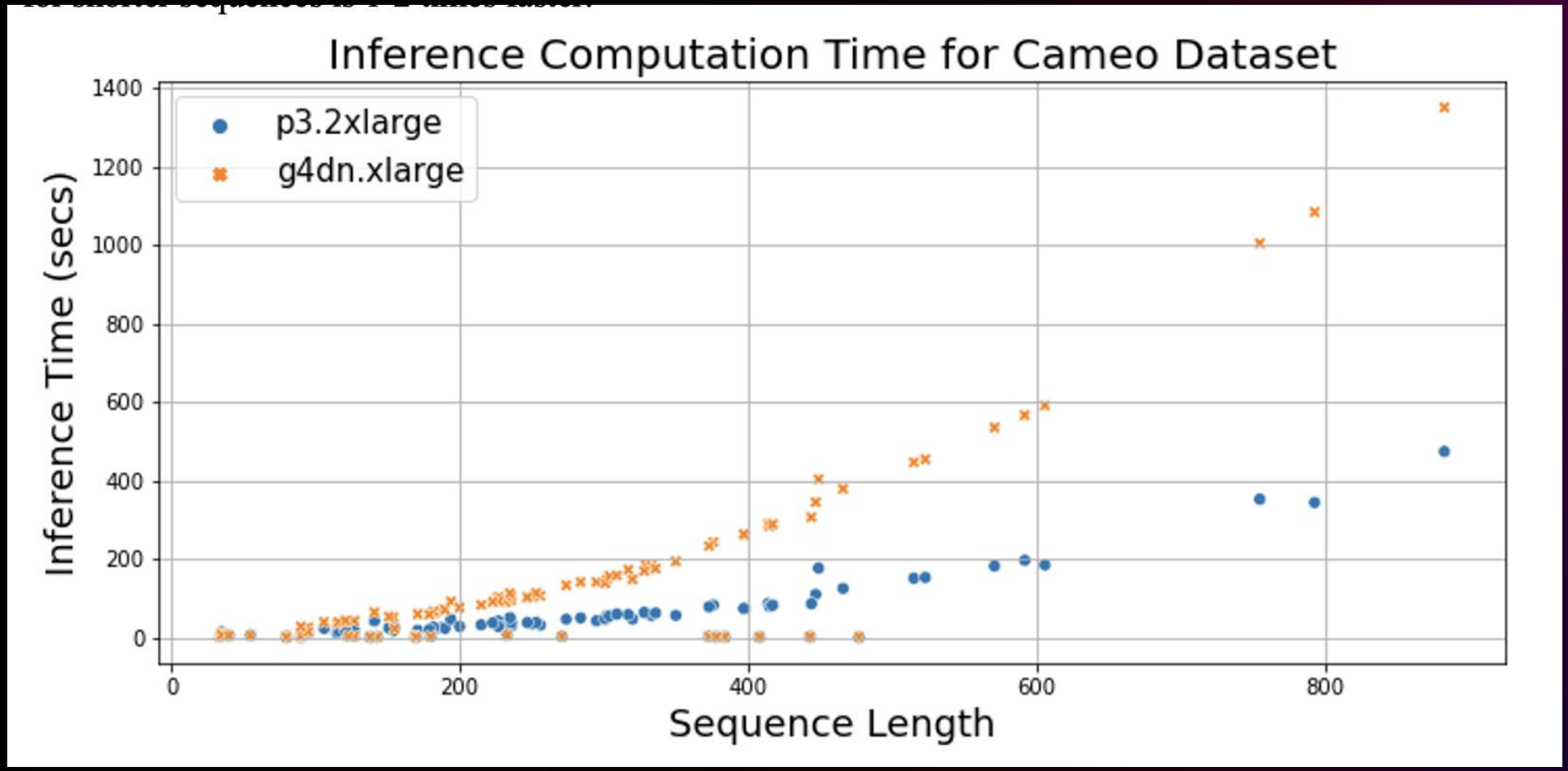
OpenFold inference apis:
http://openfold-gpu-0.default.svc.cluster.local:8080/openfold_predictions/model_0?tag="+tag+"&seq="+seq
http://openfold-gpu-0.default.svc.cluster.local:8080/openfold_predictions/model_1?tag="+tag+"&seq="+seq
http://openfold-gpu-1.default.svc.cluster.local:8080/openfold_predictions/model_0?tag="+tag+"&seq="+seq
http://openfold-gpu-1.default.svc.cluster.local:8080/openfold_predictions/model_1?tag="+tag+"&seq="+seq
```

Calling model apis

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Inference computation time

Instance Type	Number of GPUs	GPU Type	Total vCPUs Available	CPU Memory	GPU Memory	Total Inference Time on Cameo Dataset	On-Demand Hourly Cost	Total Cost
p3.2xlarge	1	Tesla V100	8	61 GiB	16	1.36 hours	3.06 $/hr	$4
g4dn.xlarge	1	Tesla T4	4	16 GiB	16 GiB	3.95 hours	0.526 $/hr	$2

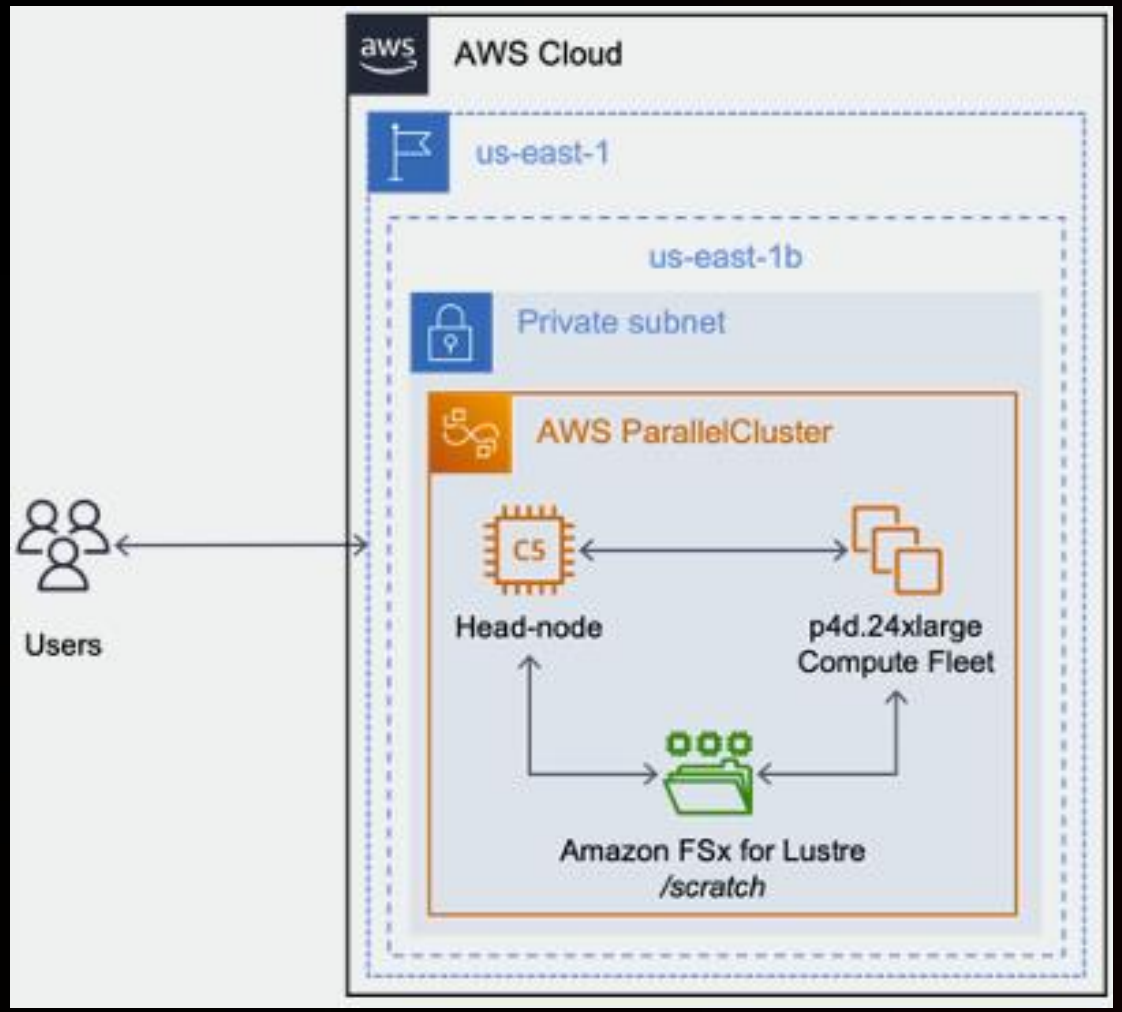


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

OpenFold distributed training architecture

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Distributed training for protein folding



- Compute nodes: V100 or A100 based EC2 instances (p3dn or p4d or p4de)

- Network: 400 Gbps ENA and EFA

- Storage: NVMe + Amazon EBS

- Shared file systems
  - NFS/apps
  - Amazon FSx for Lustre/scratch

- Cluster stack
  - Slurm
  - CUDA
  - PyTorch

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Next steps…

Team Mission: Meet customers where they are, help them adopt and scale ML workloads on AWS with EC2, key services, Open source tools, and frameworks (PyTorch, TensorFlow)

Architecting scalable solutions using EC2,EKS, ECS, Batch, Spot, KubeFlow, Ray, PyTorch, TensorFlow, etc.

EC2 Instance selection via benchmarking for specific models & workloads (e.g., P4d for training, G4dn, G5, Inf1 for inference).

Develop Proof-of-concepts to show the art-of-the-possible, unblock technical challenges, & advisory on best practices. POC credits available.

Guidance on PyTorch adoption, scaling, roadmap dive deeps in 3-way discussions with Meta (if needed).

Areas of interest: NLP, CV, Distributed training, ML inference, MLOPs

"Meta and AWS will jointly help enterprises use PyTorch on AWS to bring deep learning models from research into production faster and easier": Meta/AWS Press Release

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Ankur Srivastava
awsankur@amazon.com
@ankur_rice

Shubha Kumbadakone
shubkumb@amazon.com

Sachin Kadyan



Please complete the session survey in the **mobile app**

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.