AWS re:Invent

NOV. 29 – DEC. 3, 2021 | LAS VEGAS, NV

AIM412-R2

Deep learning applications with TensorFlow

Sam Palani
Sr. Machine Learning Specialist SA
AWS

Tim O'Brien
Principal Solutions Architect
AWS

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Topics

The AWS ML stack

Art of the possible

TensorFlow on AWS

Training

Inference

Q&A

Reminder!
This is a 400 level talk

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The AWS ML stack

BROADEST AND MOST COMPLETE SET OF MACHINE LEARNING CAPABILITIES

AI services

Health AI
- Amazon HealthLake
- Amazon Transcribe Medical
- Amazon Comprehend Medical

Industrial AI
- AWS Panorama + appliance
- Amazon Monitron
- Amazon Lookout for Equipment
- Amazon Lookout for Vision

Anomaly detection
- Amazon Lookout for Metrics

Code and DevOps
- Amazon DevOps Guru
- Amazon CodeGuru

Vision
- Amazon Rekognition

Speech
- Amazon Polly

Text
- Amazon Transcribe Medical
- Amazon Comprehend Medical
- Amazon Translate
- Amazon Textract

Search
- Amazon Kendra

Chatbots
- Amazon Lex

Personalization
- Amazon Personalize

Forecasting
- Amazon Forecast

Fraud
- Amazon Fraud Detector

Contact centers
- Contact Lens for Amazon Connect
- Voice ID

ML services

- Amazon SageMaker
- Label data

SageMaker Studio IDE

- Aggregate & prepare data
- Store & share features
- Auto ML
- Spark/R
- Detect bias
- Visualize in notebooks
- Pick algorithm
- Train models
- Tune parameters
- Debug & profile
- Deploy in production
- Manage & monitor
- CI/CD
- Human review

SageMaker JumpStart

Model management for edge devices

Frameworks and infrastructure

- TensorFlow
- mxnet
- PyTorch
- GLUON
- Keras
- scikit-learn
- Horovod
- DeepGraphLibrary
- AWS Deep Learning AMIs (DLAMI) and AWS Deep Learning Containers
- GPUs and CPUs
- Amazon Elastic Inference
- AWS Trainium
- AWS Inferentia
- FPGA

Art of the possible

- Text-based applications
- Image recognition
- Forecasting and time series
- Video detection
- Voice and sound recognition

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

TensorFlow on AWS

HTTPS://AWS.AMAZON.COM/TENSORFLOW/

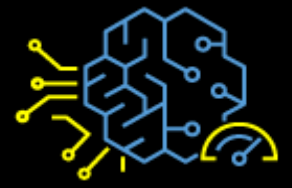
AWS optimizations for TensorFlow available on Amazon Elastic Compute Cloud (Amazon EC2) and Amazon SageMaker

- AWS Deep Learning Containers for training and inference
- AWS Deep Learning AMIs (DLAMI)

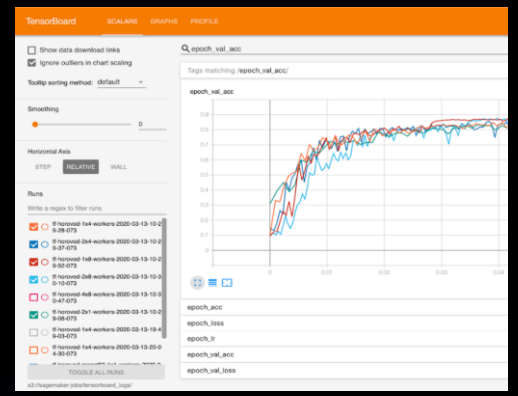
SageMaker benefits for TensorFlow

- Built-in support for TensorBoard, Debugger, local mode, hyperparameter tuning, managed spot training, Pipe mode, and Amazon Elastic Inference
- Distributed training – parameter server and Horovod
- Performance optimizations – GPUs, CPUs, and storage





Elastic Inference



TensorBoard



aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker training options

Script mode

- AWS deep learning containers for training

Custom container

- Extend pre-built containers
- Bring your own containers

Distributed training

- Data parallel
- Model parallel

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

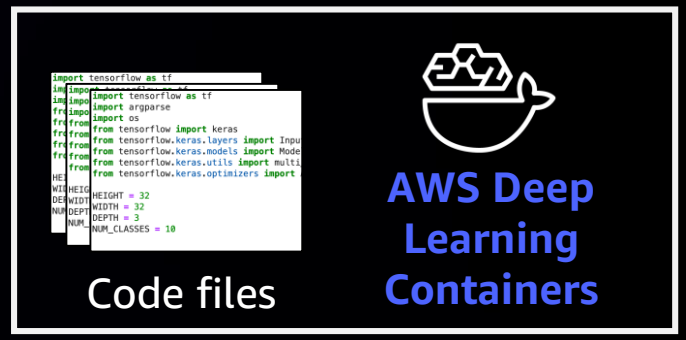
Training flow



SageMaker SDK

```
from sagemaker.tensorflow import TensorFlow

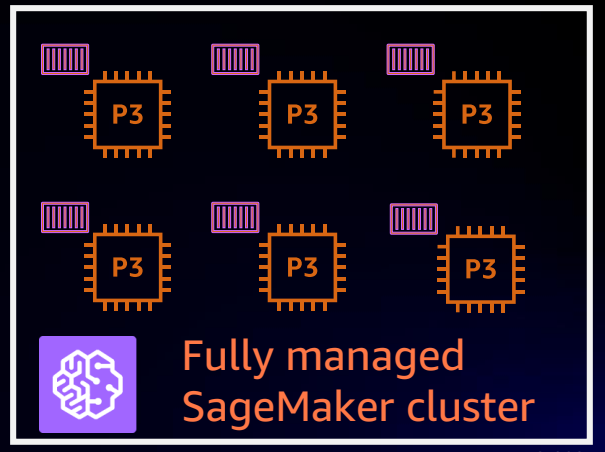
cifar10_estimator = TensorFlow(entry_point='source/cifar10.py',
                               role=role,
                               framework_version='1.14',
                               train_instance_count=1,
                               train_instance_type='ml.p3.xlarge')
```





Amazon ECR

Container registry





Amazon S3

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Distributed training solutions

Data parallelism

- Training dataset is split across multiple nodes.
- Each node has a replica of the model.
- Each node performs a forward and backward pass and synchronizes gradients.
- Still requires to fit the model and at least one record in memory (GPU/CPU).
- Additional complexity of computing gradients.

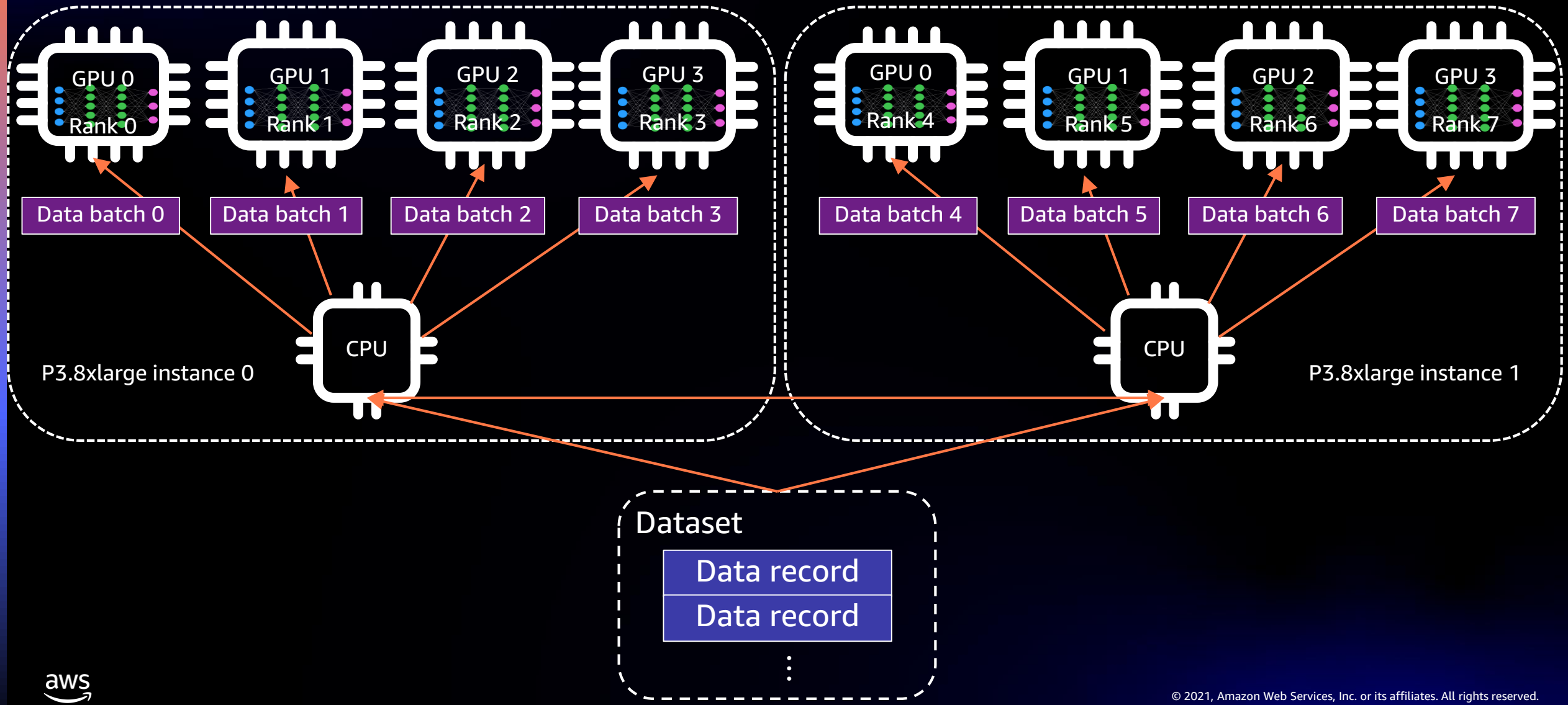
Model parallelism

- The model is partitioned across multiple nodes.
- Each node contains a subset of the model through which the data flows and the transformations are shared and complied.
- Efficiency driven by a pipeline execution schedule.

aws

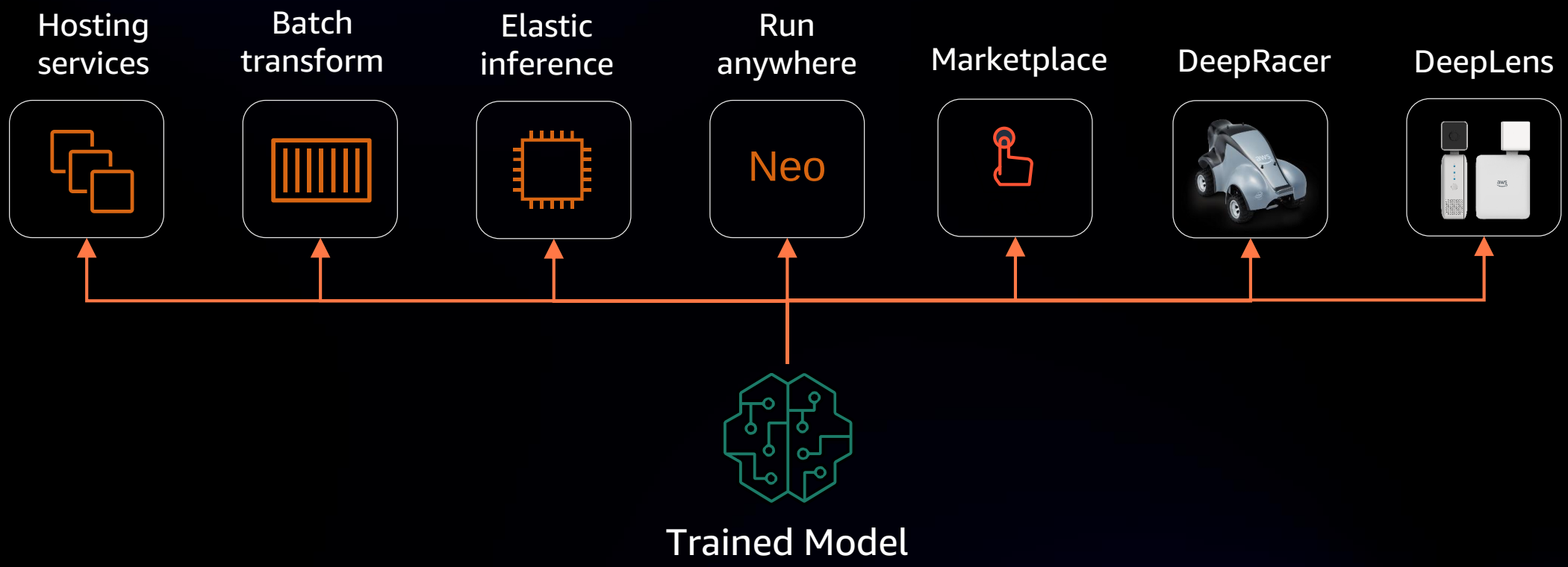
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Distributed training – Data parallel



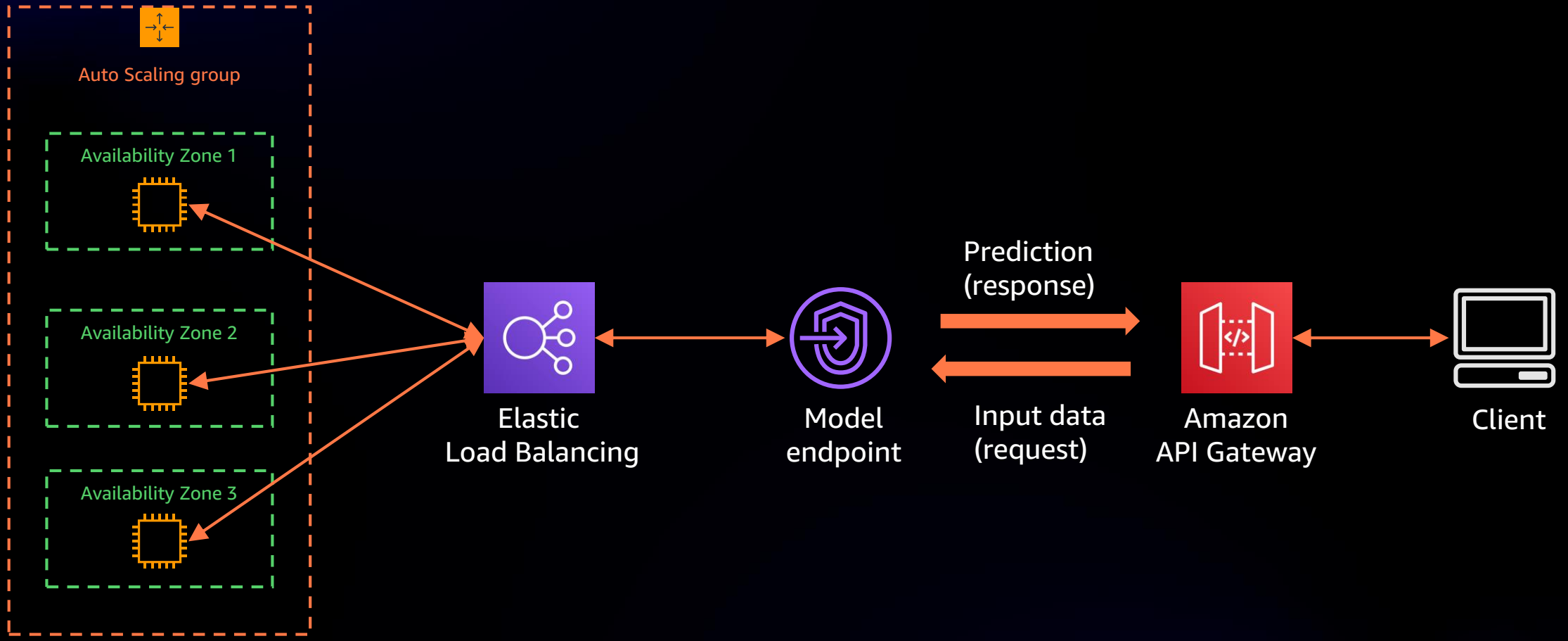
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon SageMaker inference options



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker endpoints

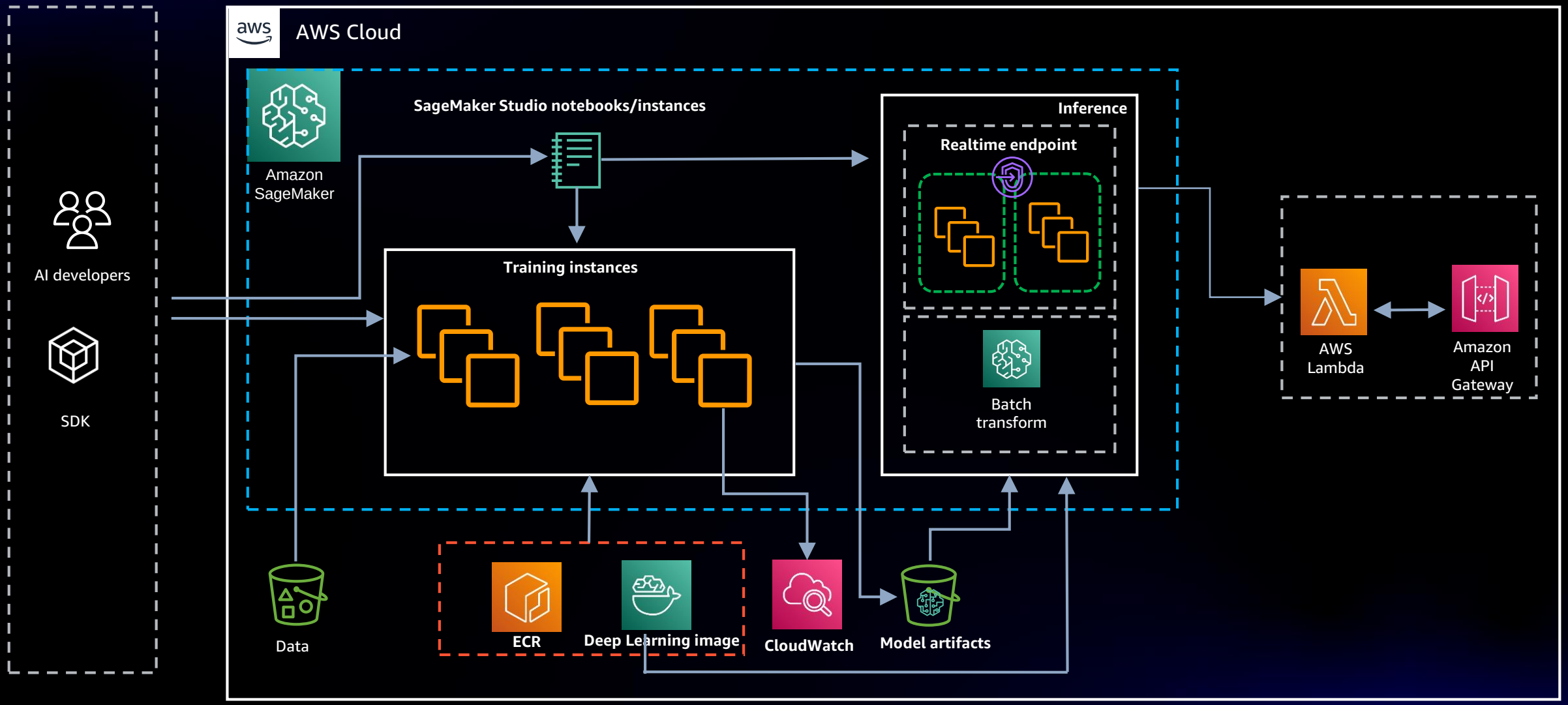


Deployment/hosting
Amazon SageMaker ML
compute instances

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Build, train, and deploy



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Sam Palani

@samx18

Tim O'Brien

tpobrien@amazon.com
linkedin.com/in/tobrien0

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.