

Reduce inference cost by up to 75% for TensorFlow models with Amazon Elastic Inference

Vikrant Kahlir, Solutions Architect, AMAZON WEB SERVICES

Rama Thamman, R&D Manager, AMAZON WEB SERVICES



Agenda

- ❖ Introduction
- ❖ Usage
- ❖ How EI works
- ❖ Performance
- ❖ Hands-on Lab for EC2 and SageMaker

The Amazon ML stack: Broadest & deepest set of capabilities

Vision

Speech

Language

Chatbots

Forecasting

Recommendations

AI SERVICES



REKOGNITION
IMAGE



REKOGNITION
VIDEO



TEXTRACT



POLLY



TRANSCRIBE



TRANSLATE



COMPREHEND
COMPREHEND
MEDICAL



LEX

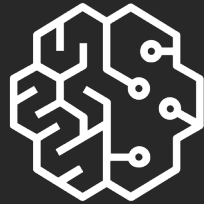


FORECAST



PERSONALIZE

ML SERVICES



AMAZON
SAGEMAKER

GROUND TRUTH

NOTEBOOKS

ALGORITHMS

AWS MARKETPLACE

REINFORCEMENT
LEARNING

TRAINING

OPTIMIZATION
(NEO)

DEPLOYMENT

HOSTING

Frameworks

Interfaces

Infrastructure

ML FRAMEWORKS & INFRASTRUCTURE



EC2 P3
& P3dn



EC2 C5



FPGAs



AWS IoT
Greengrass



ELASTIC
INFERENCE



© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.



The Amazon ML stack: Broadest & deepest set of capabilities

Vision

Speech

Language

Chatbots

Forecasting

Recommendations

AI SERVICES



REKOGNITION
IMAGE



REKOGNITION
VIDEO



TEXTRACT



POLLY



TRANSCRIBE



TRANSLATE



COMPREHEND
COMPREHEND
MEDICAL



LEX

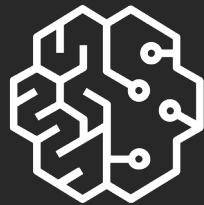


FORECAST



PERSONALIZE

ML SERVICES



AMAZON
SAGEMAKER

GROUND TRUTH

NOTEBOOKS

ALGORITHMS

AWS MARKETPLACE

REINFORCEMENT
LEARNING

TRAINING

OPTIMIZATION
(NEO)

DEPLOYMENT

HOSTING

ML FRAMEWORKS & INFRASTRUCTURE

Frameworks

TensorFlow

mxnet

PYTORCH

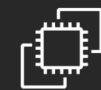
intel RL Coach

Interfaces

GLUON

K Keras

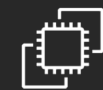
Infrastructure



EC2 P3
& P3dn



EC2 C5



FPGAs



AWS IoT
Greengrass



ELASTIC
INFERENCE

Deep learning model lifecycle

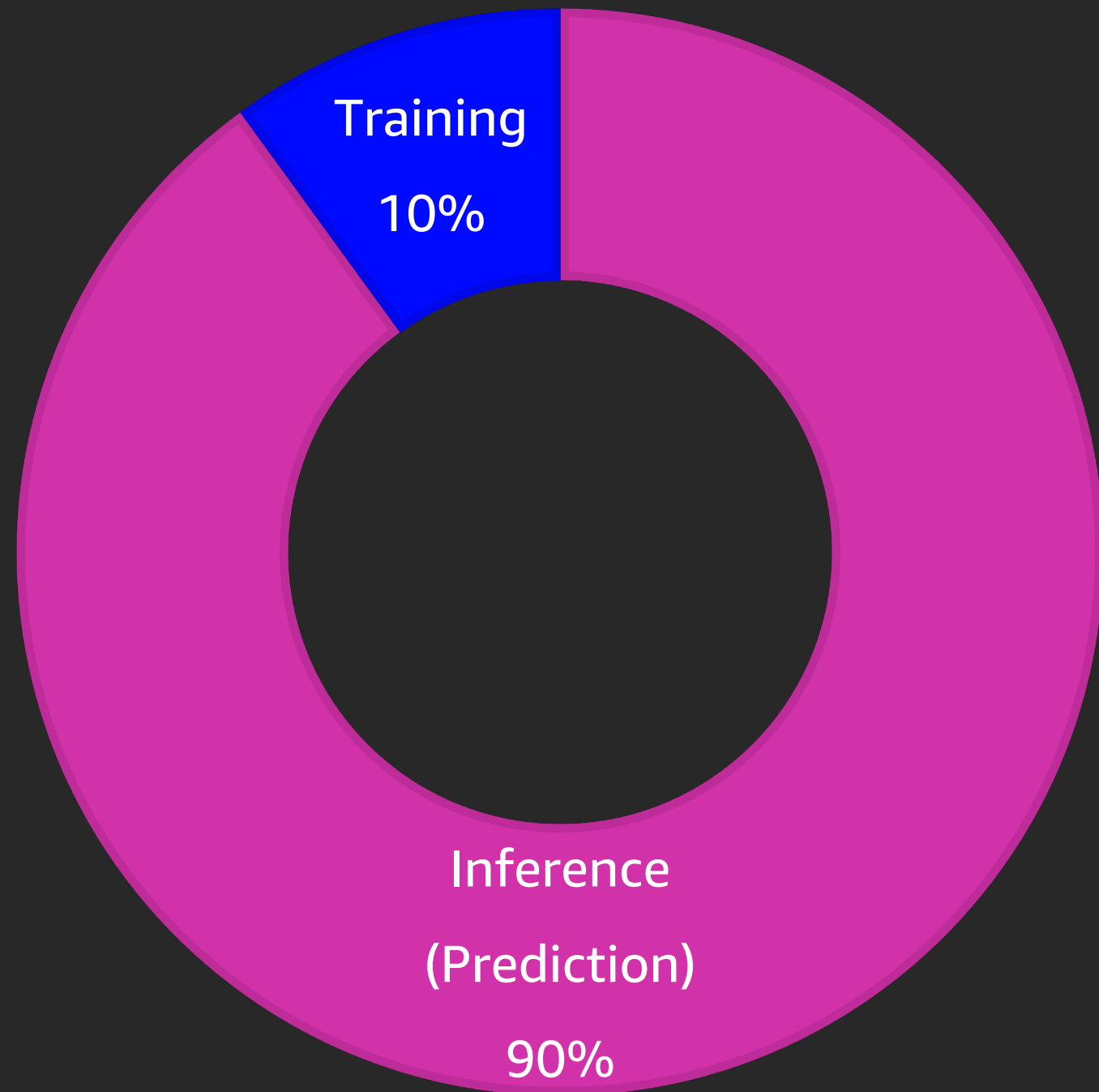
❖ Training

- ❖ Gather data for training and testing
- ❖ Architecture search
- ❖ Parameter tuning
- ❖ Distributed training using GPU's
- ❖ In the order of weeks

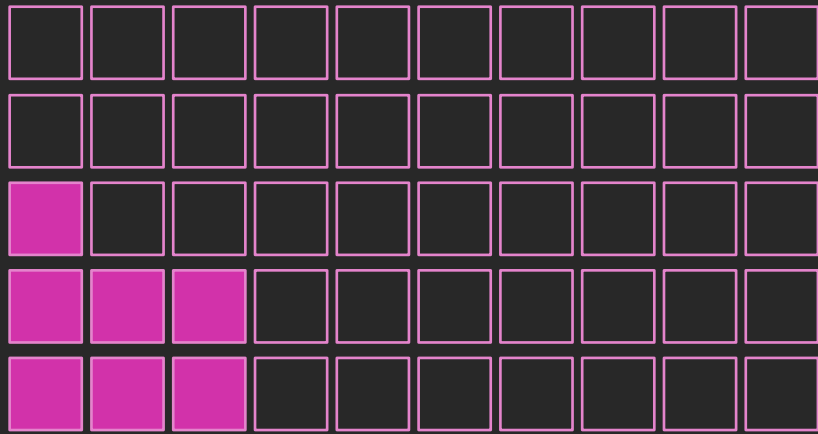
❖ Inference

- ❖ Hundreds of machines
- ❖ Different regions
- ❖ In the order of months

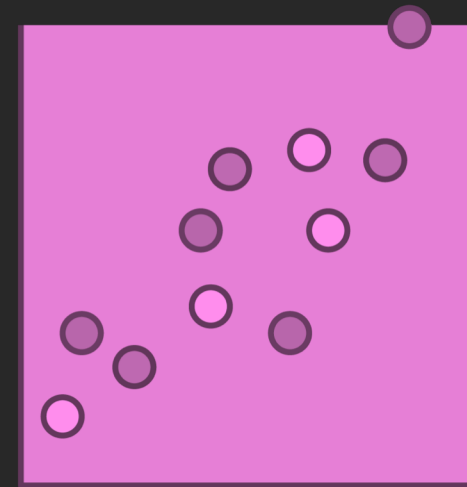
Predictions drive complexity
and cost in production



The challenges of inference in production

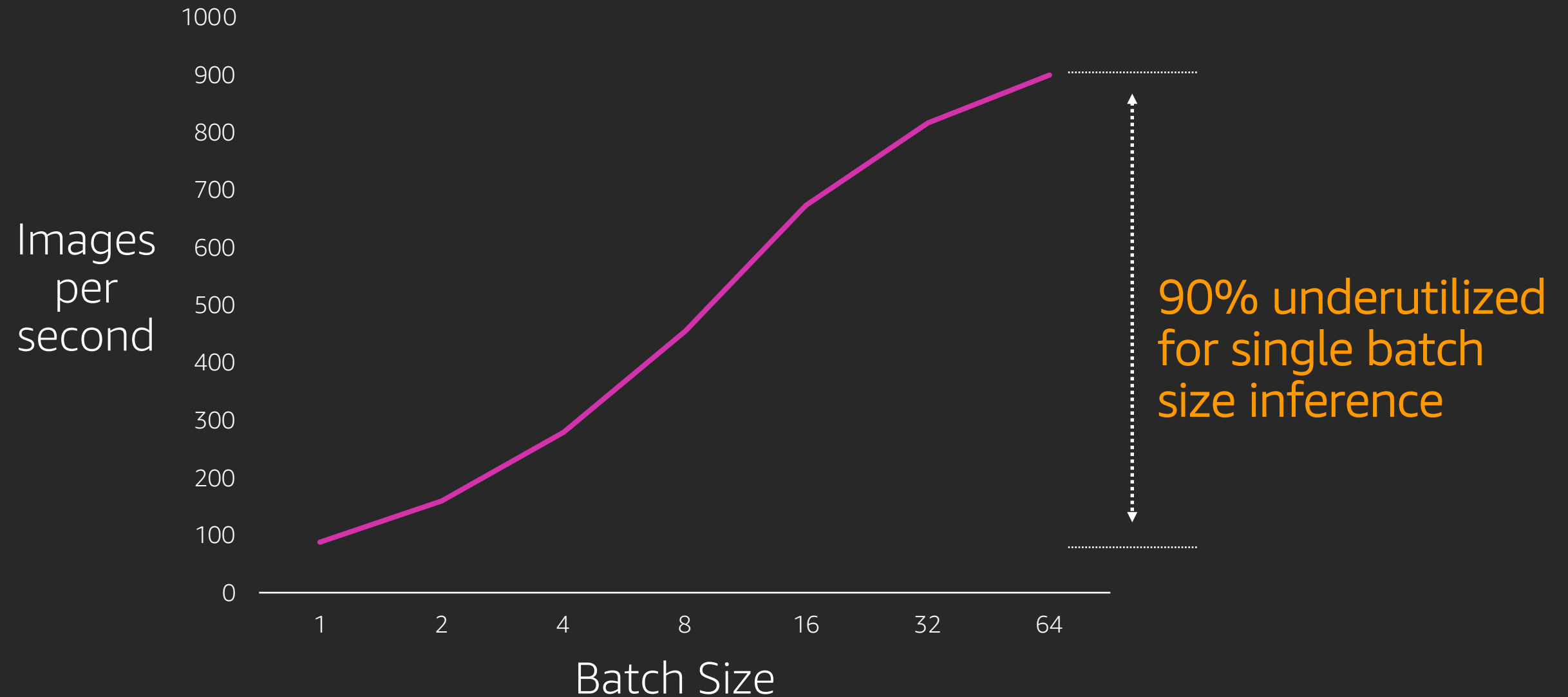


Low utilization and high costs



One size does not fit all

A closer look at GPU utilization for inference



Inception-v3 on a p3.2xlarge instance (using a V100 GPU)

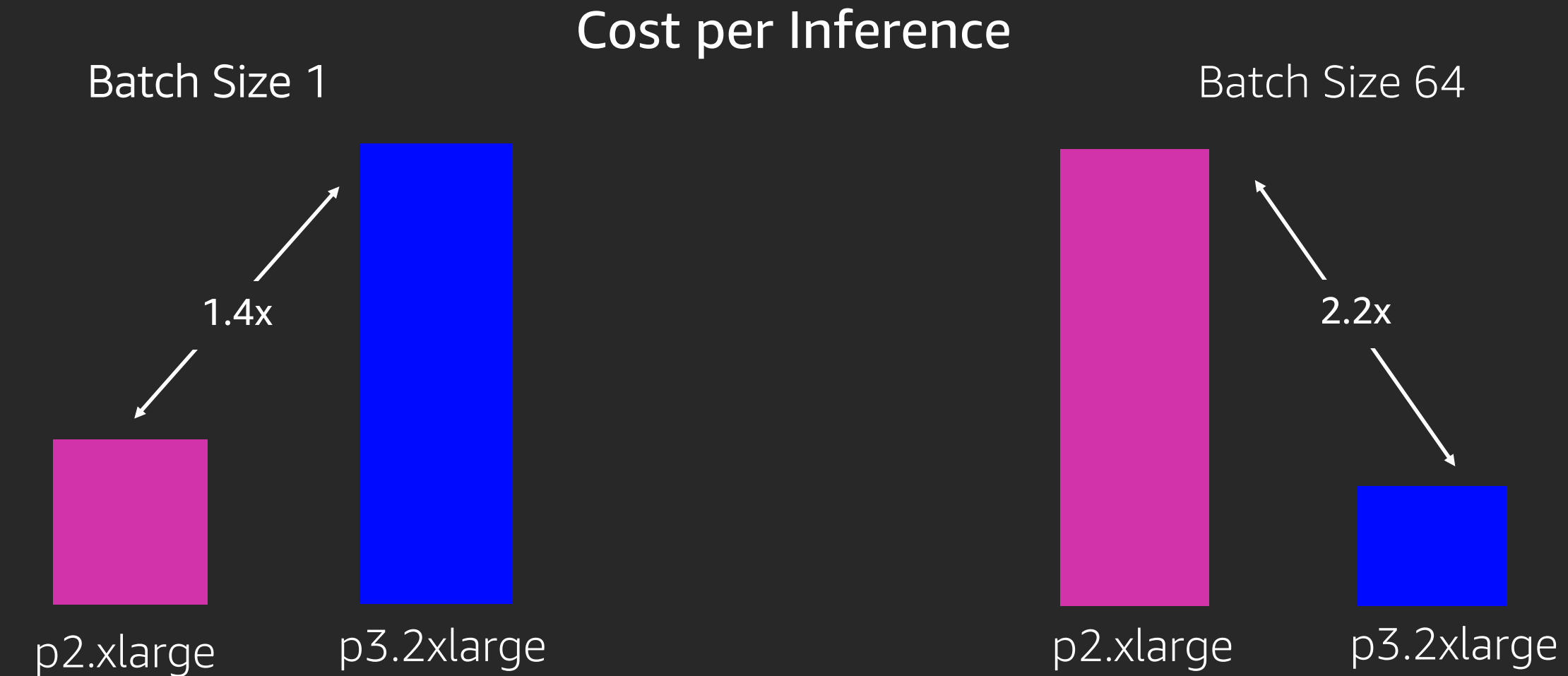
More sessions/processes doesn't solve the problem



Inception-v3 on a p3.2xlarge instance (using a V100 GPU) single batch inference

How cost effective are GPU instances for inference?

Smaller P2 instances are more effective for real time inference with small batch sizes



Inference deployment

- ❖ Run model inference on separate fleets of GPU instances and call out from main application
 - ❖ Requires heavy-lifting, can be expensive and inefficient
- ❖ Co-locate application stack along with model inference on GPU instance
 - ❖ Mismatch between host and accelerator resources can lead to over-provisioning of resources

What if you could keep your application on your familiar (CPU) instance and attach just the right amount of hardware acceleration for inference?

Amazon Elastic Inference

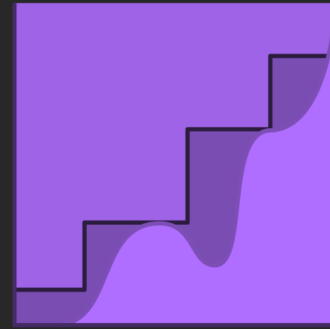


Amazon Elastic Inference

Reduce deep learning inference costs up to 75%



Lower inference costs



Match capacity to demand



Available between
1 to 32 TFLOPS per
accelerator

KEY FEATURES

Integrated with
Amazon EC2,
Amazon ECS and
Amazon SageMaker

Support for TensorFlow,
Apache MXNet, and ONNX
with PyTorch coming soon

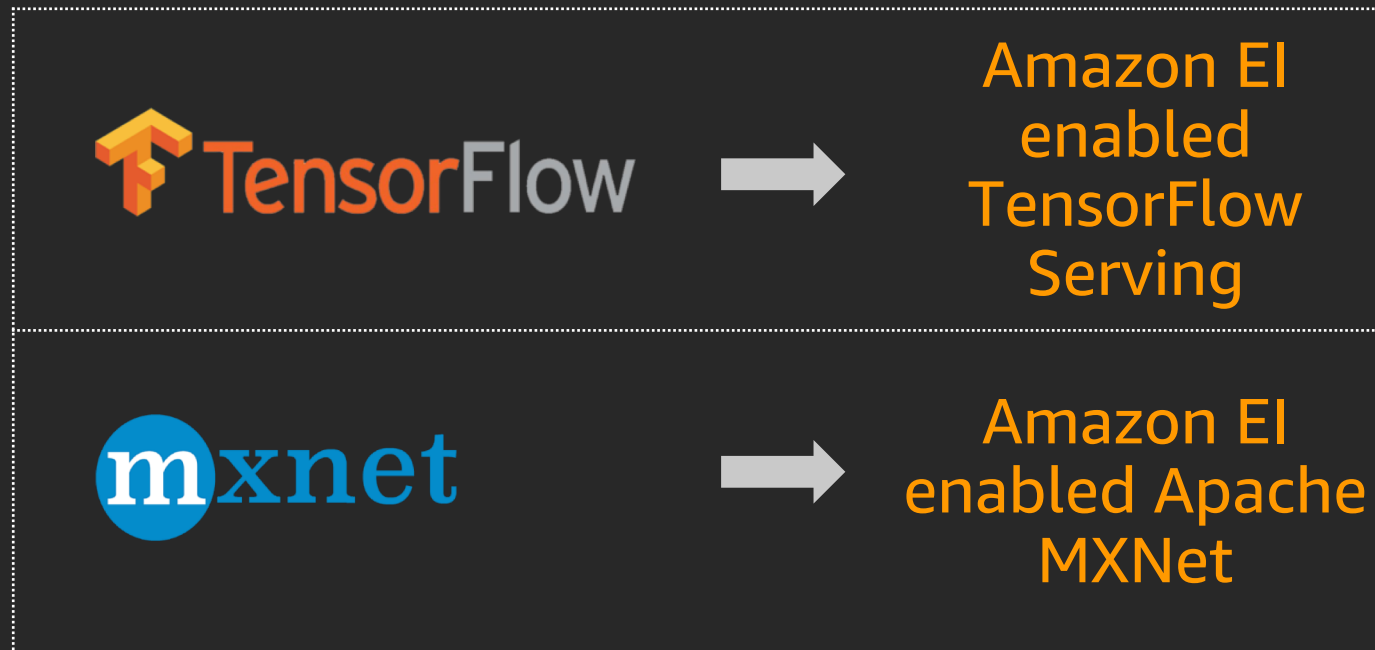
Single and
mixed-precision
operations

Acceleration sizes tailored for inference

Now available in N. Virginia, Ohio, Oregon, Dublin, Tokyo, and Seoul

Accelerator Type	FP32 Throughput (TOPS)	FP16 Throughput (TOPS)	Accelerator Memory (GB)	Price (\$/hr) (US)
eia1.medium	1	8	1	\$0.13
eia1.large	2	16	2	\$0.26
eia1.xlarge	4	32	4	\$0.52
eia2.medium	1	8	2	\$0.120
eia2.large	2	16	4	\$0.240
eia2.xlarge	4	32	8	\$0.340

Model Support

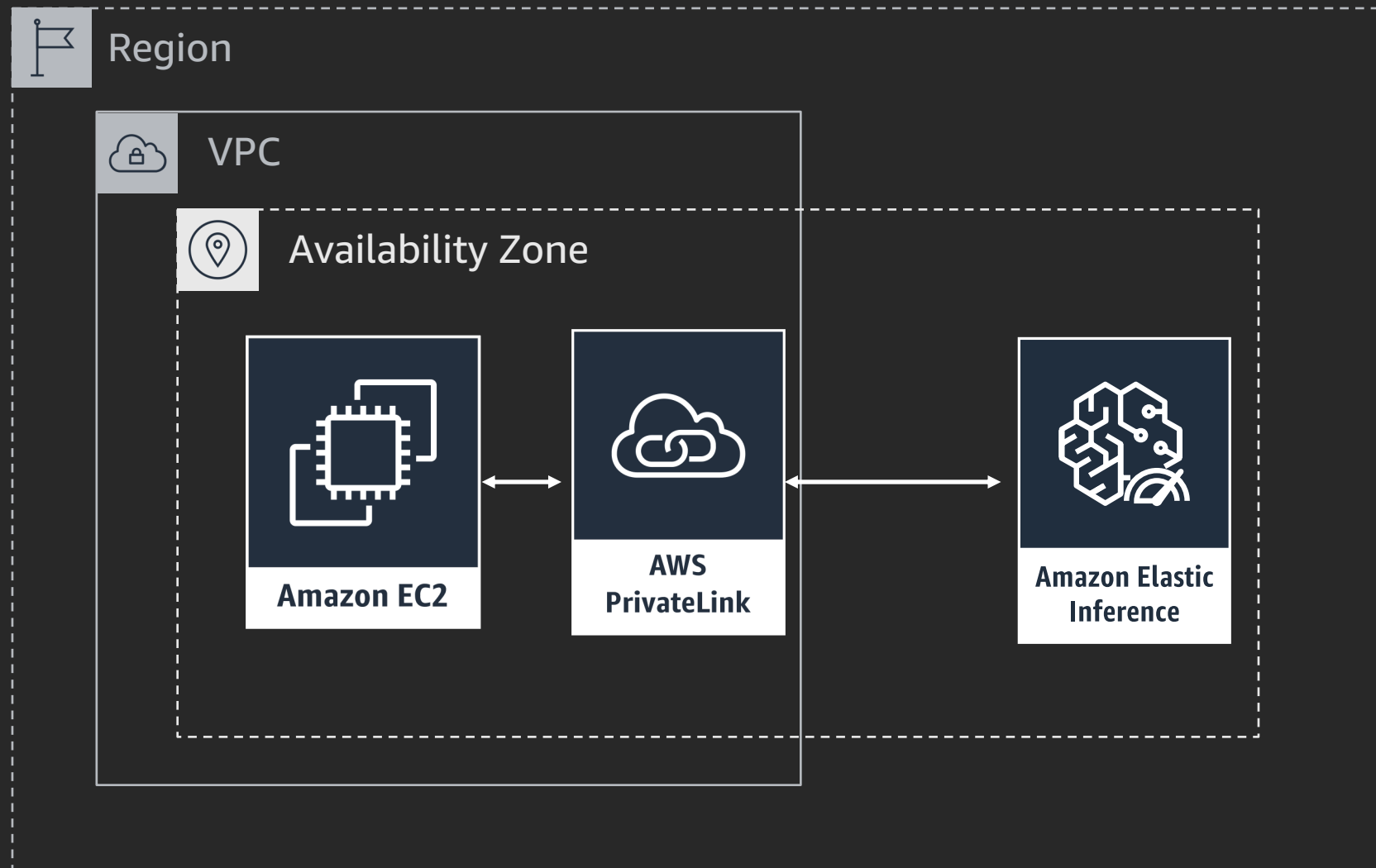


Amazon EI enabled TensorFlow Serving and Apache MXNet

- ❖ Auto discovery of accelerators
- ❖ IAM-based authentication
- ❖ Available via: the AWS Deep Learning AMIs, for download via S3 and automatically through SageMaker

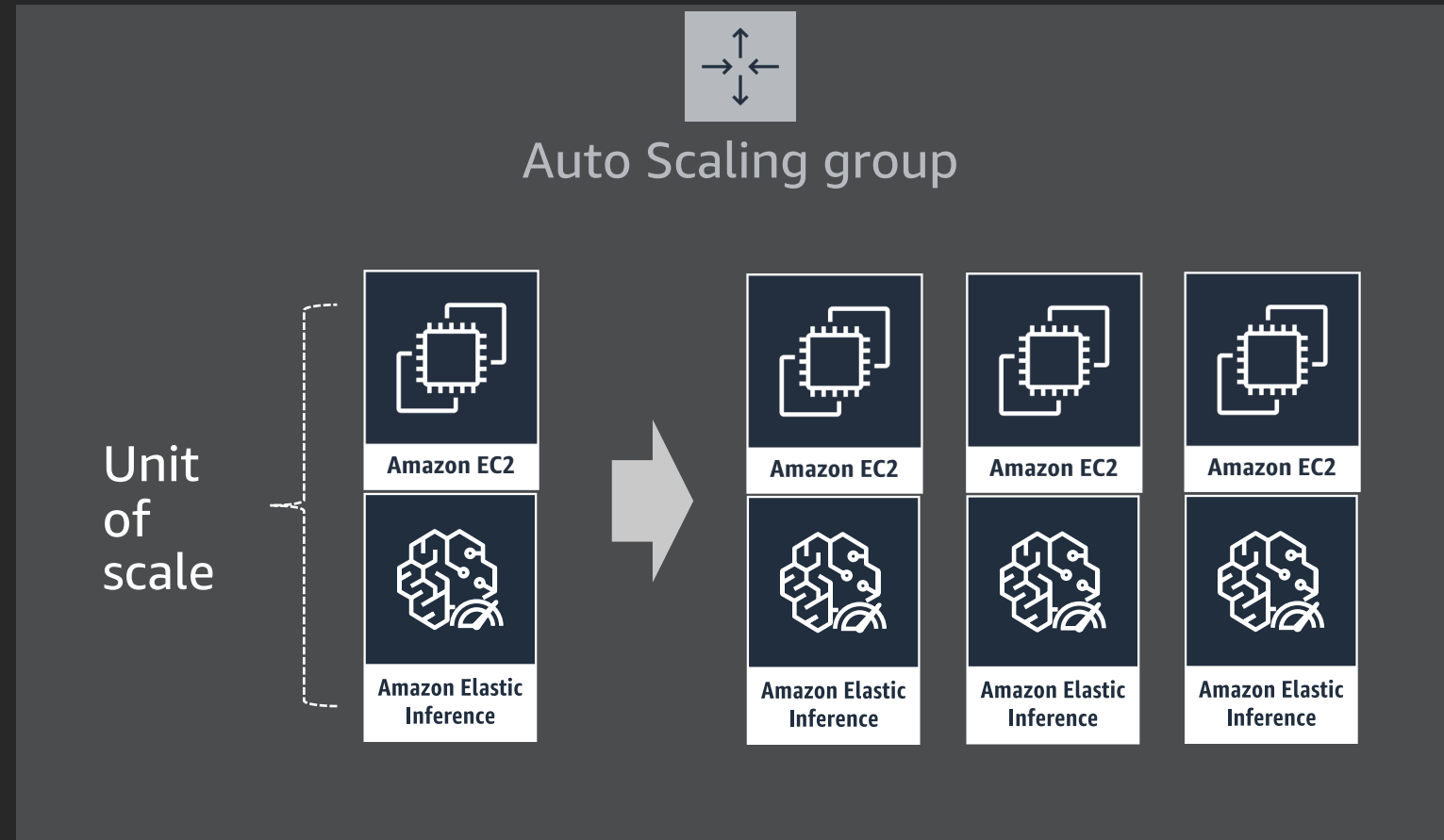
How Amazon Elastic Inference works

How does Elastic Inference work with Amazon EC2?



- Set up a AWS PrivateLink endpoint for your VPC to the EI service.
- Configure instances to launch with EI accelerator.
- Scale instances with accelerators with EC2 Auto Scaling – using Launch Templates

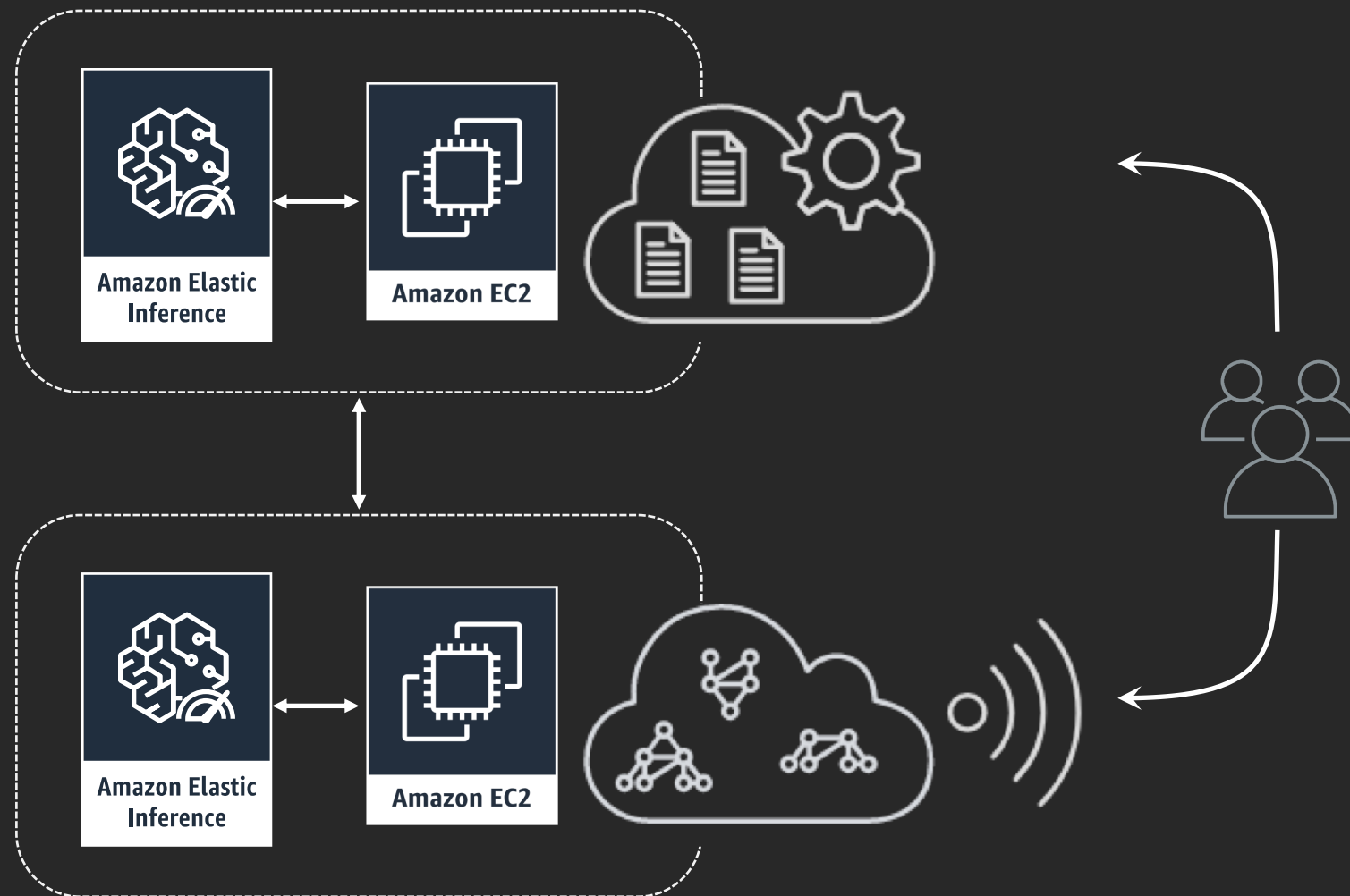
Scale capacity in EC2 Auto Scaling groups



Specify EI within launch templates

How does Elastic Inference work with SageMaker?

SageMaker Notebooks



❖ Prototype deployments with Notebooks in local mode

❖ Scale endpoints with low-cost Elastic Inference Acceleration

SageMaker Hosted Endpoints

Hands-on Lab

Lab 1: Attach Elastic Inference to Amazon SageMaker Inference Endpoint

Lab 2: Attach Elastic Inference to Amazon EC2

Lab Resources

<http://bit.ly/2p6aBzH>

Thank you!

Questions/Feedback:amazon-ei-feedback@amazon.com

