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

AI SERVICES

Vision
- REKOGNITION IMAGE
- REKOGNITION VIDEO
- TEXTRACT

Speech
- POLLY
- TRANSCRIBE

Language
- TRANSLATE
- COMPREHEND
- COMPREHEND MEDICAL

Chatbots
- LEX

Forecasting
- FORECAST

Recommendations
- 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
- Keras

Infrastructure
- EC2 P3 & P3dn
- EC2 C5 (intel XEON PLATINUM inside)
- FPGAs
- AWS IoT Greengrass (OpenVINO, intel inside)
- ELASTIC INFERENCE

aws

The Amazon ML stack: Broadest & deepest set of capabilities

AI SERVICES

- Vision: REKOGNITION IMAGE, REKOGNITION VIDEO, TEXTRACT
- Speech: POLLY, TRANSCRIBE
- Language: TRANSLATE, COMPREHEND, COMPREHEND MEDICAL
- Chatbots: LEX
- Forecasting: FORECAST
- Recommendations: 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, Keras
- Infrastructure: EC2 P3 & P3dn, EC2 C5, FPGAs, AWS IoT Greengrass, ELASTIC INFERENCE

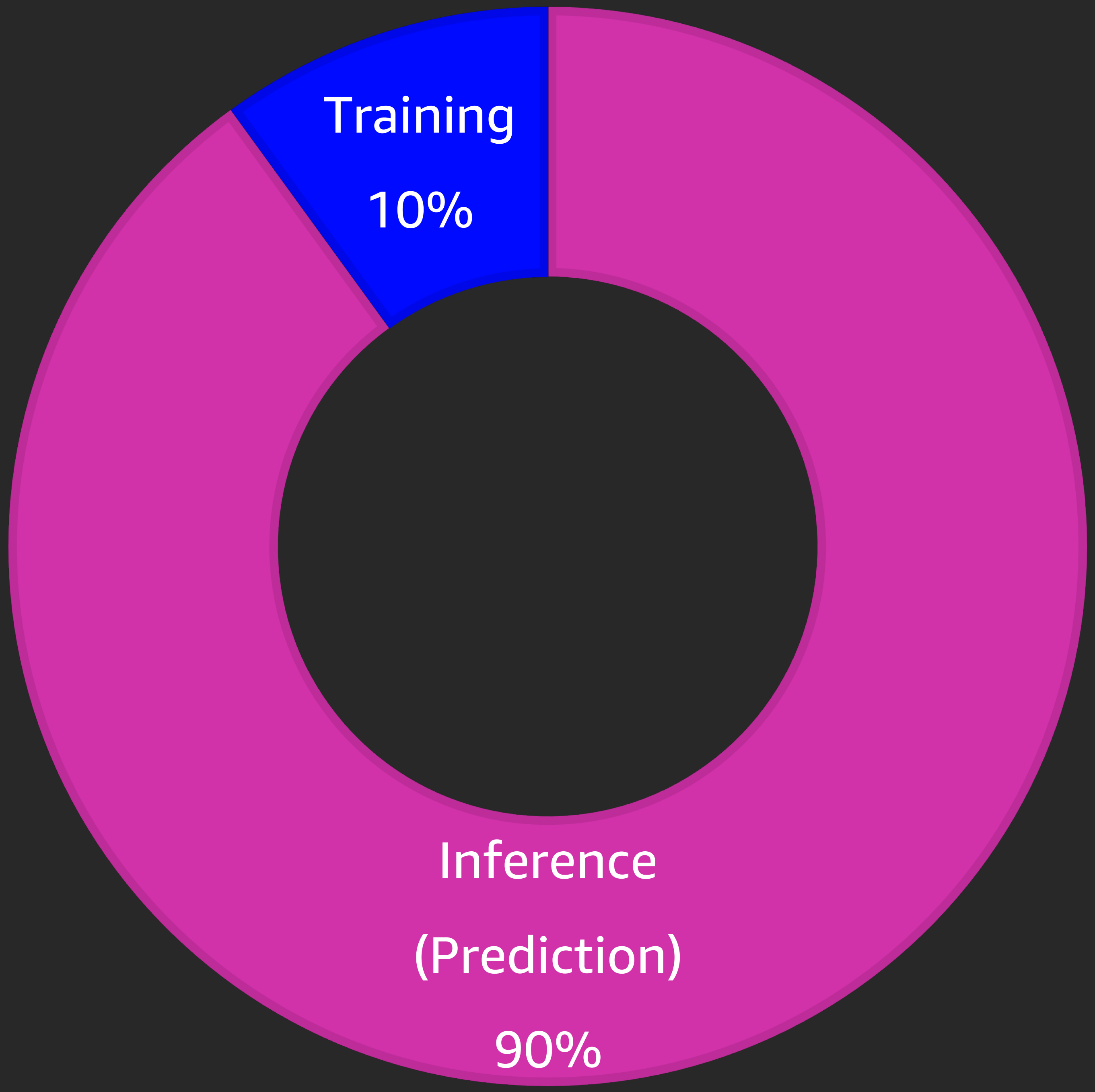
aws

Deep learning model lifecycle

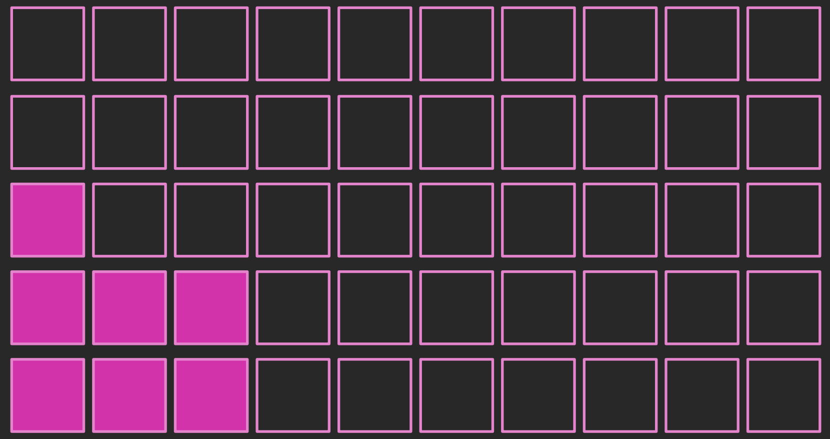
- Training
  - Gather data for training and testing
  - Architecture search
  - Parameter tuning
  - Distributed training using GPU’s
  - In the order of weeks

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

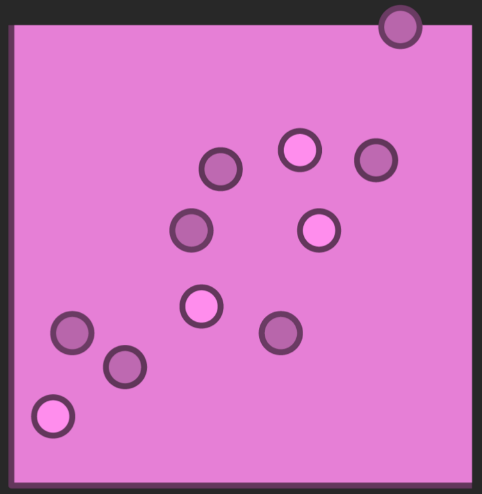
Predictions drive complexity and cost in production



The challenges of inference in production

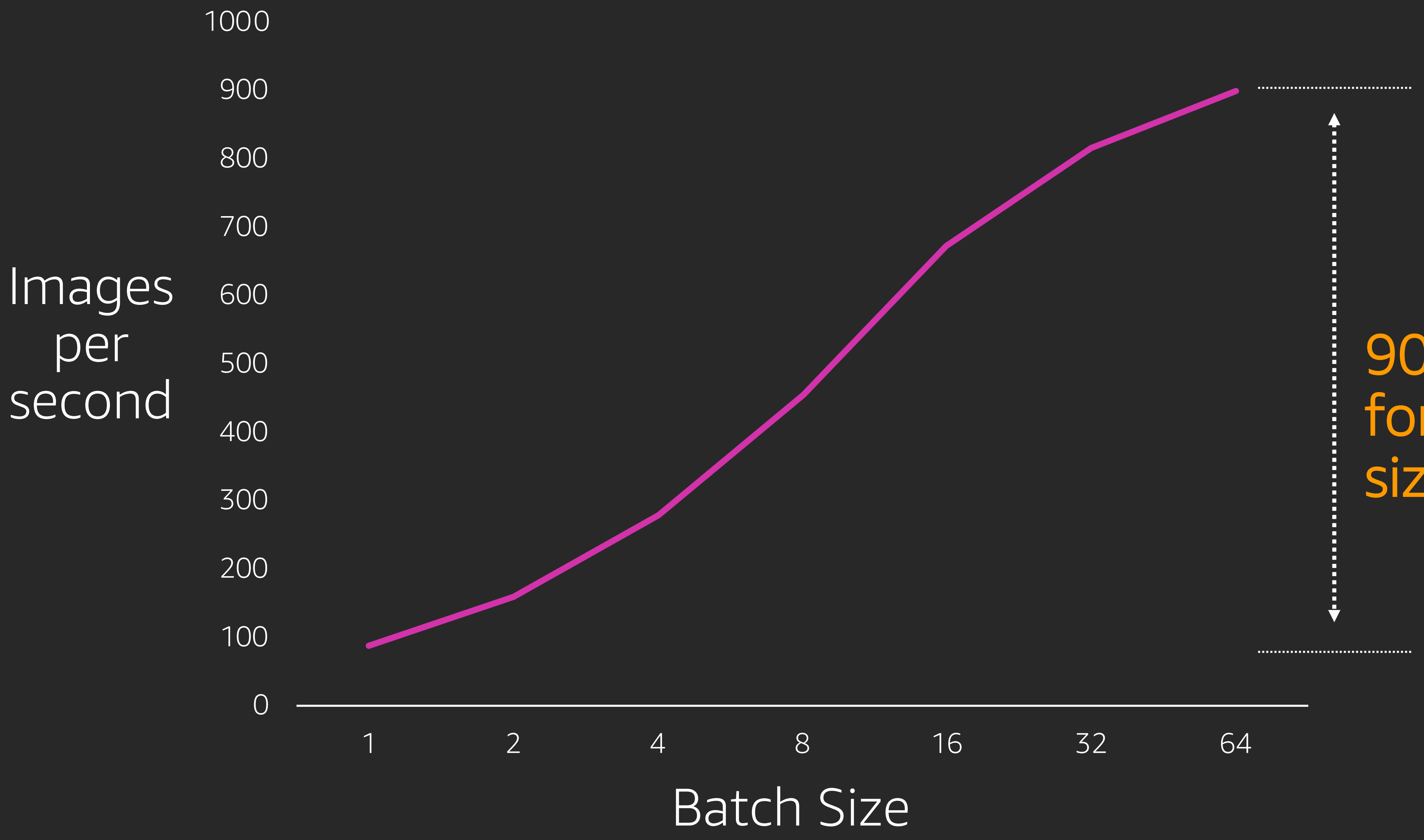


Low utilization and high costs



One size does not fit all

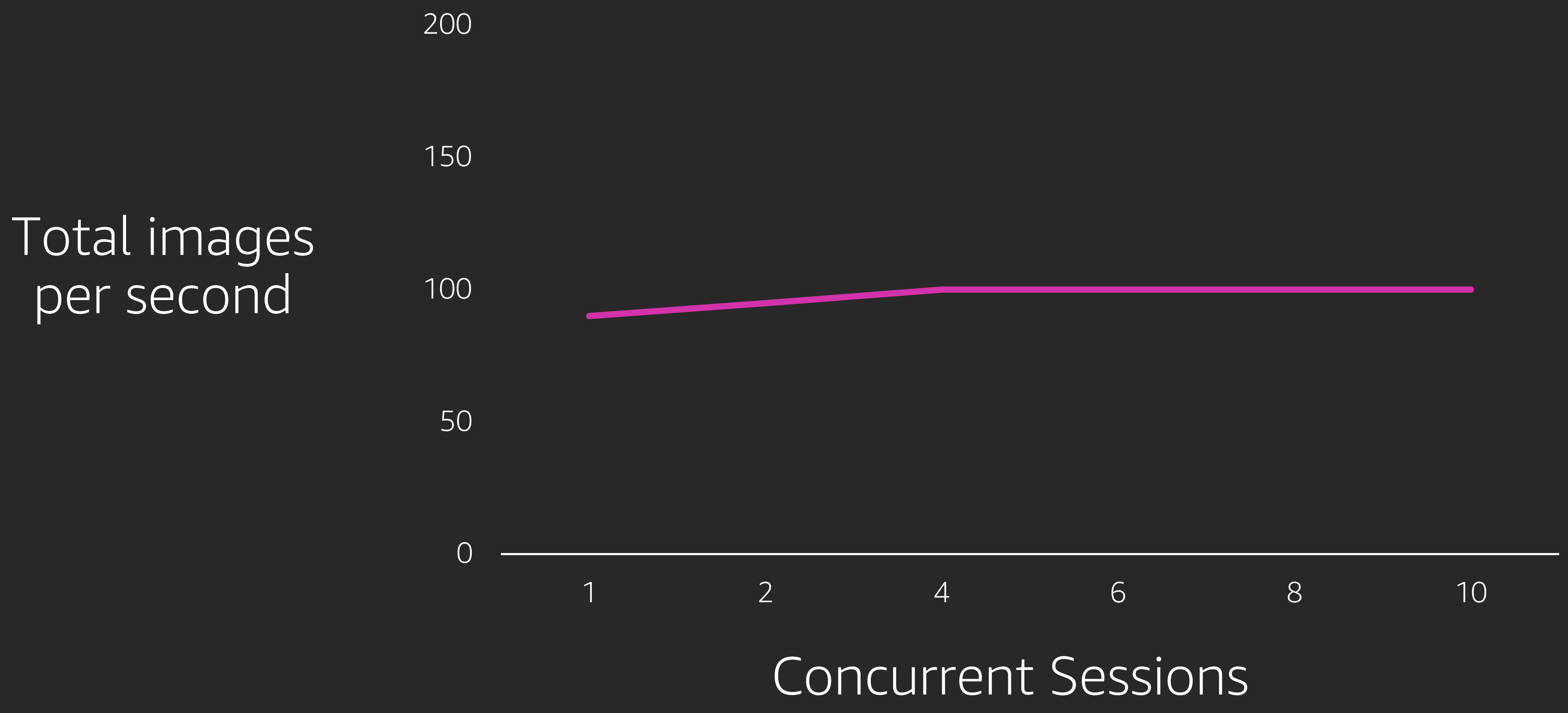
A closer look at GPU utilization for inference



90% underutilized for single batch size inference

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

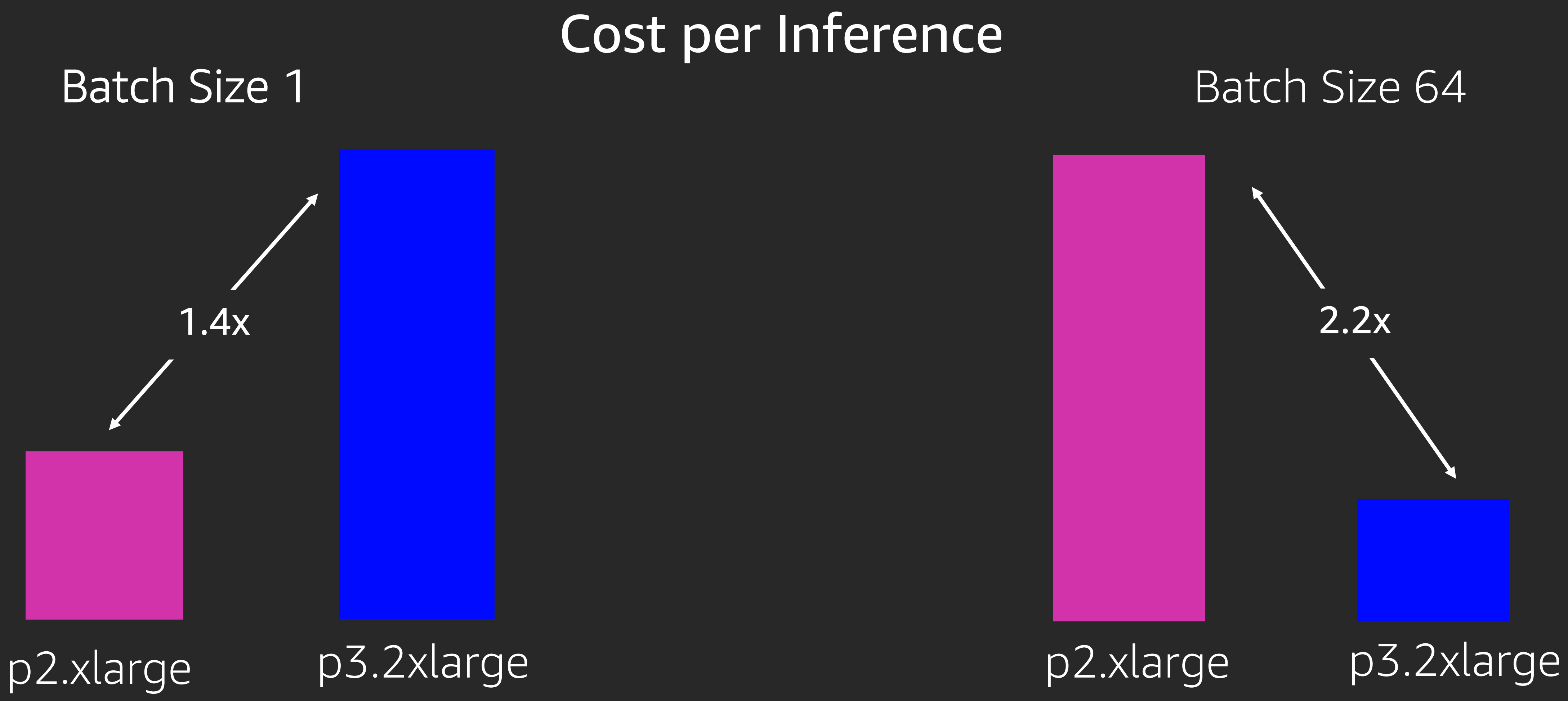
More sessions/processes doesn’t solve the problem



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


Cost per Inference

Lower numbers are better

Inception-v3 FP32 MXNet

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

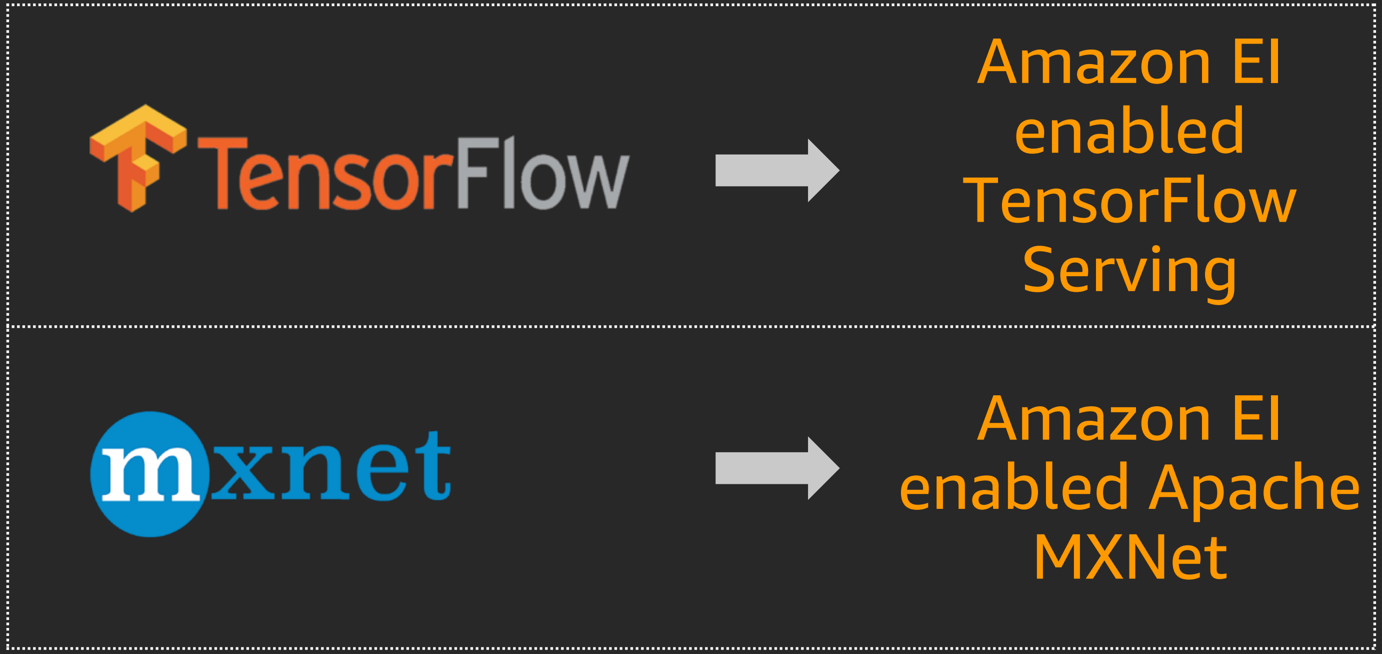
Single and mixed-precision operations

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

Amazon EI enabled Apache MXNet

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

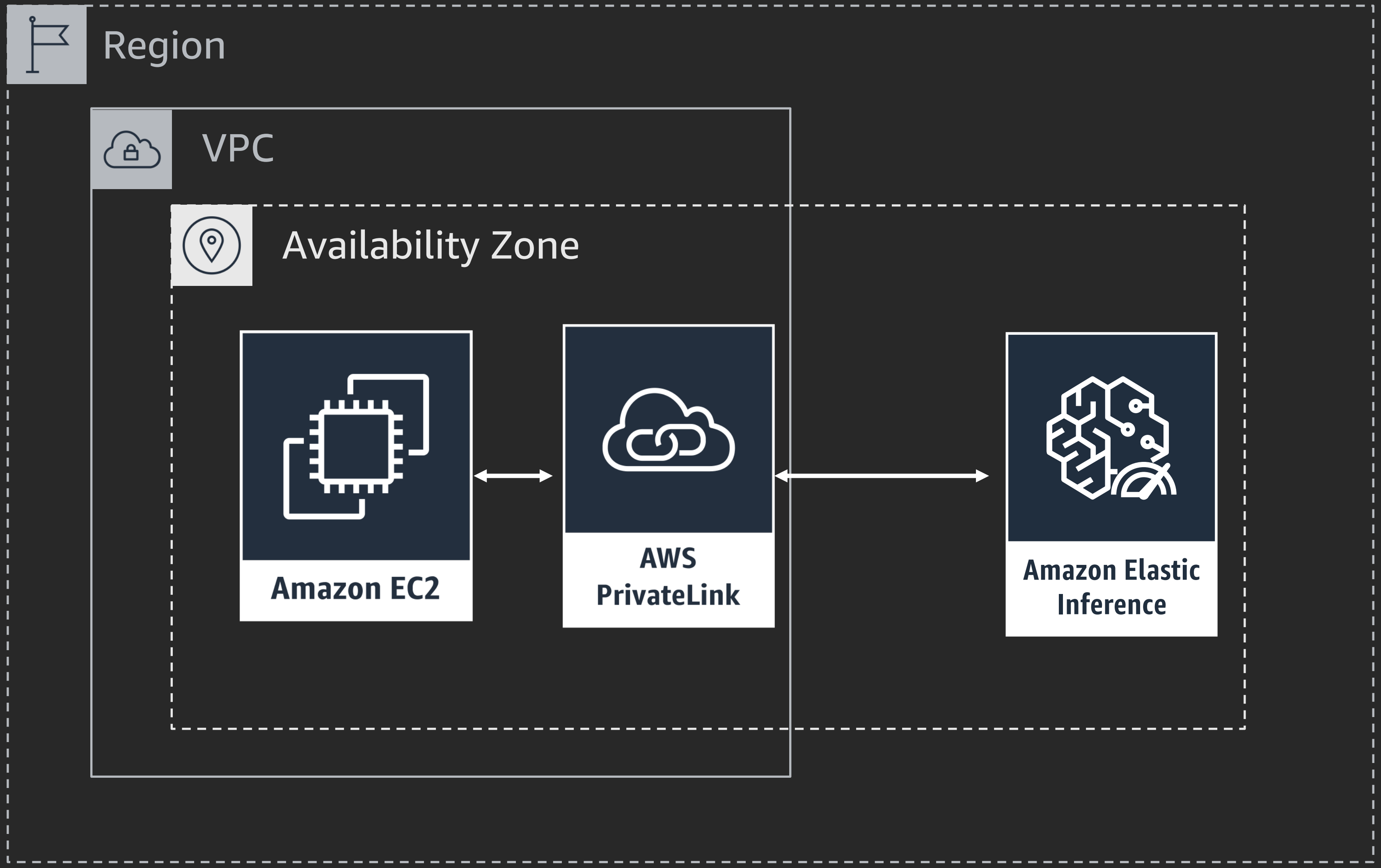
How to choose?

Considerations as you choose an instance and accelerator type combination for your model:

- What is your target latency SLA for your application, and what are you constraints?
- Start small and size up if you need more capacity.
- Input/output data payload has an impact on latency.
- Convert to Fp16 for lower latency and higher throughput.

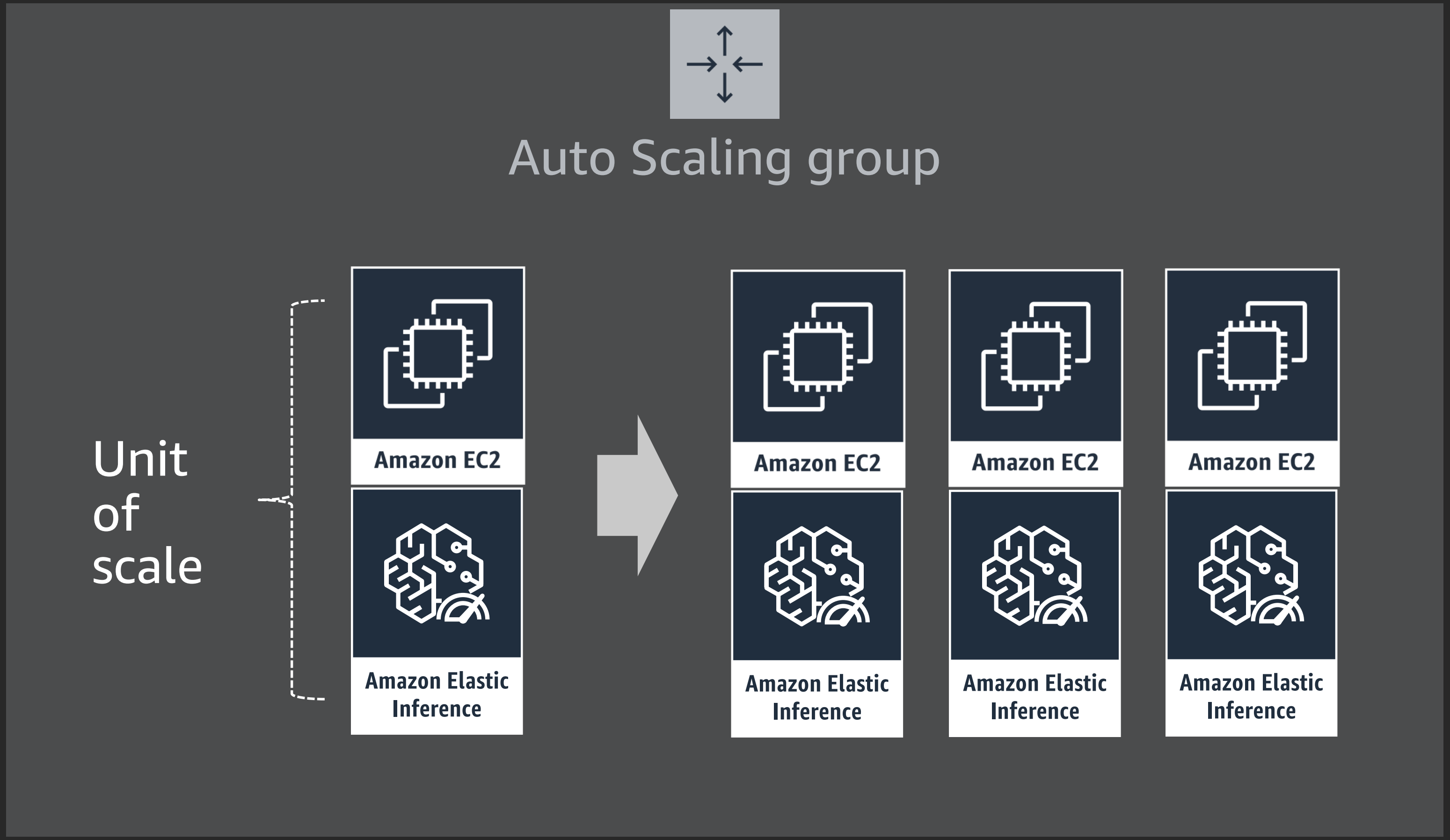
How Amazon Elastic Inference works

How does Elastic Inference work with Amazon EC2?



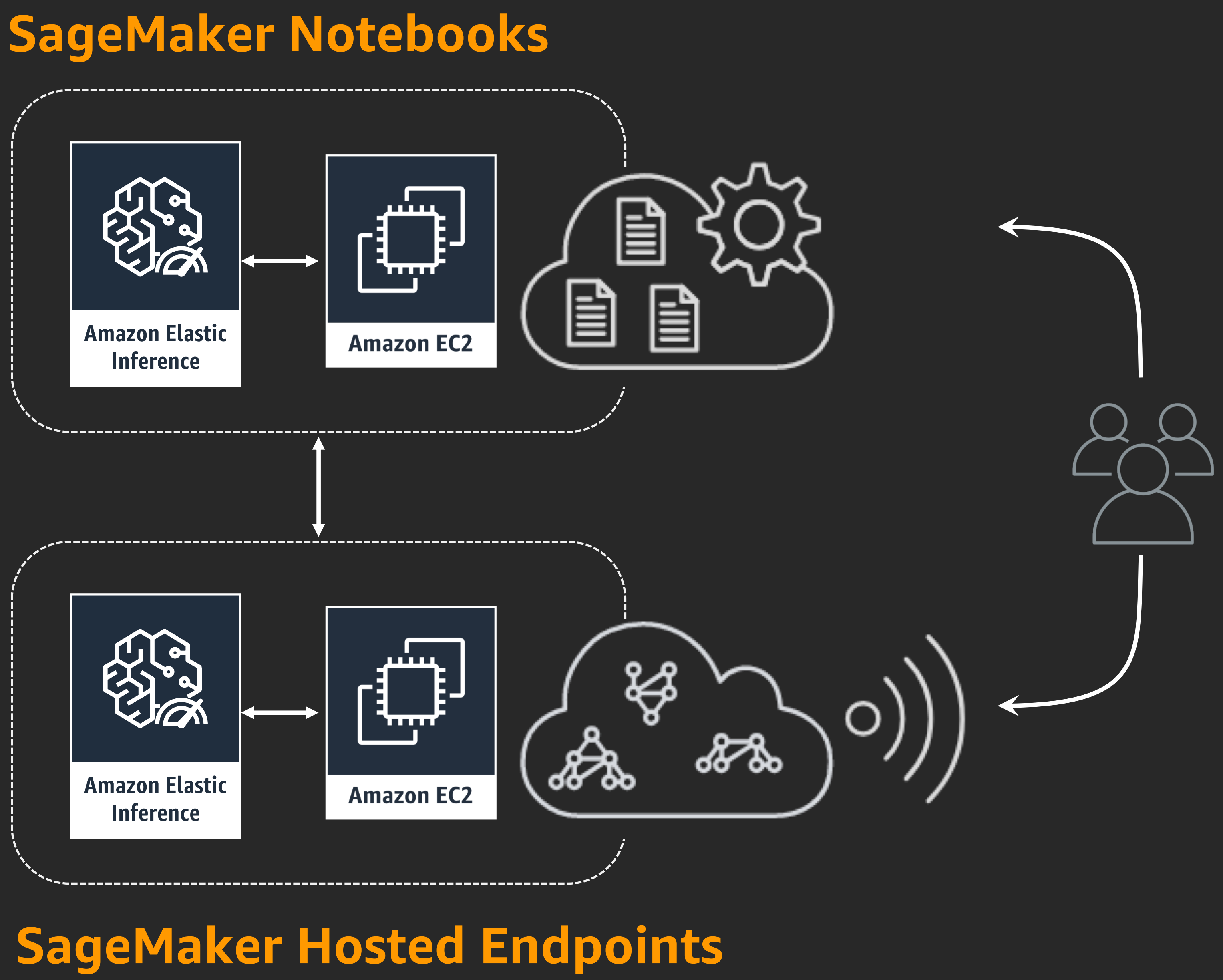
- Set up a AWS PrivateLink endpoint for your VPC to the EI service.
- Configure instances to launch with EI accelerator.
- Scale instances with accelerators with EC2 Auto Scaling – using Launch Templates

Scale capacity in EC2 Auto Scaling groups



Specify EI within launch templates

How does Elastic Inference work with SageMaker?



- ❖ Prototype deployments with Notebooks in local mode
- ❖ Scale endpoints with low-cost Elastic Inference Acceleration

EI vs. Local GPU

- When can EI latency be higher than local (whole) GPU?
  - Models with relatively less computation (single digit msec) (network roundtrip/transfer time becomes significant)
  - Models with large input/output tensor size (multiple MBs) (large network transfer time)
  - Models that exploit high GPU parallelism (EIA has reduced parallelism due to GPU slicing)
- When is local (whole) GPU not replaceable by EI?
  - CUDA based programming (custom CUDA kernels)
  - Acceleration for custom op in framework
  - Pre-/post- processing using custom GPU libraries (e.g., Nvidia DALI)

Summary

- EI accelerators available in a range of sizes suitable for inference workloads- Reduce inference costs by up to 75%
- Configure to launch with any EC2 instance type– scale capacity with autoscaling groups.
- EI configuration is also available though CloudFormation as you configure your instance resource.
- Deploy TensorFlow and MXNet models with no code changes.
- Integrated with SageMaker for a fully managed experience

aws.amazon.com/machine-learning/elastic-inference/

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