INNOVATE

AI/ML EDITION

24 February 2022

Rapidly launch ML solutions at scale on AWS infrastructure

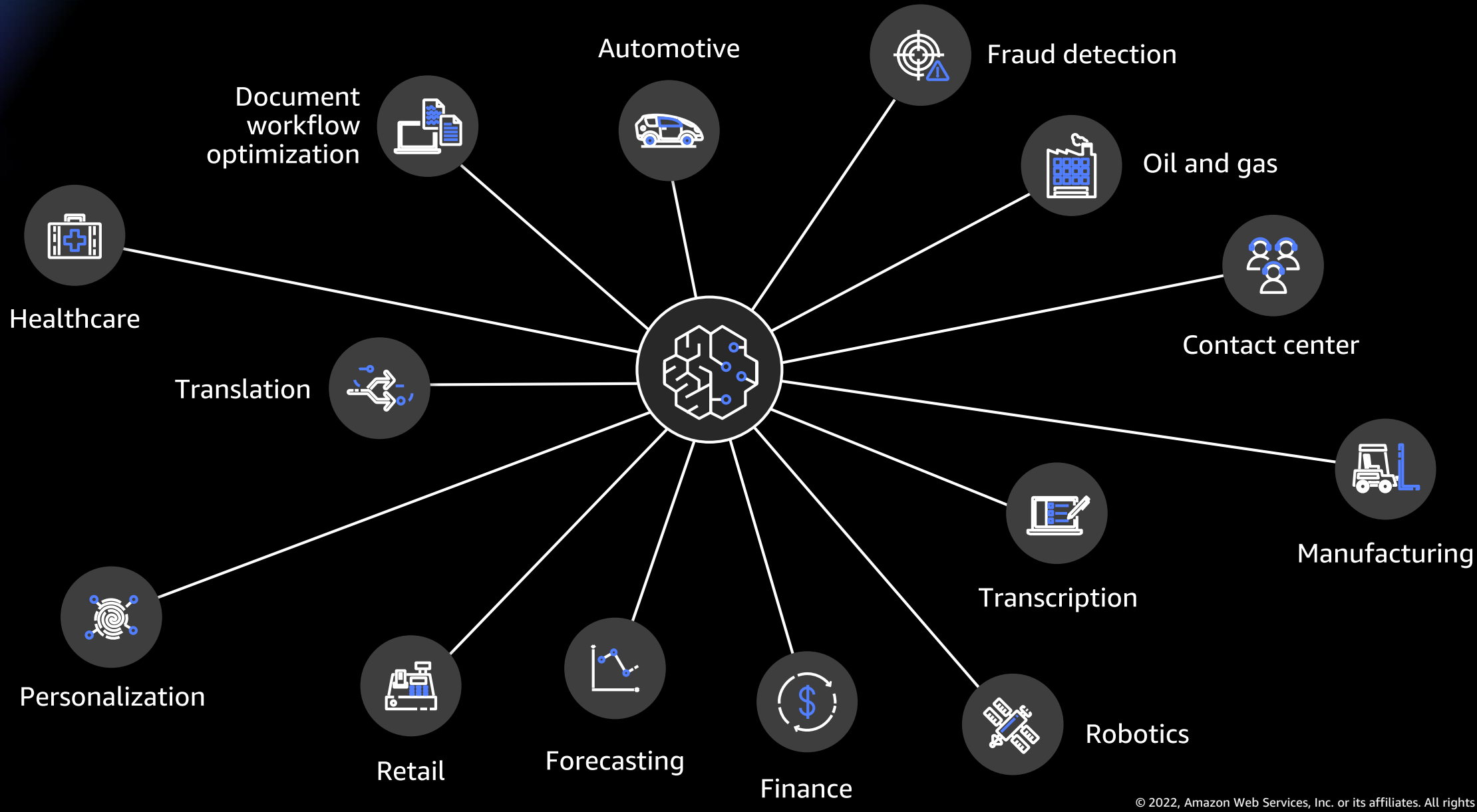
Santhosh Urukonda

Prototyping Architect,
AISPL

Takeaways

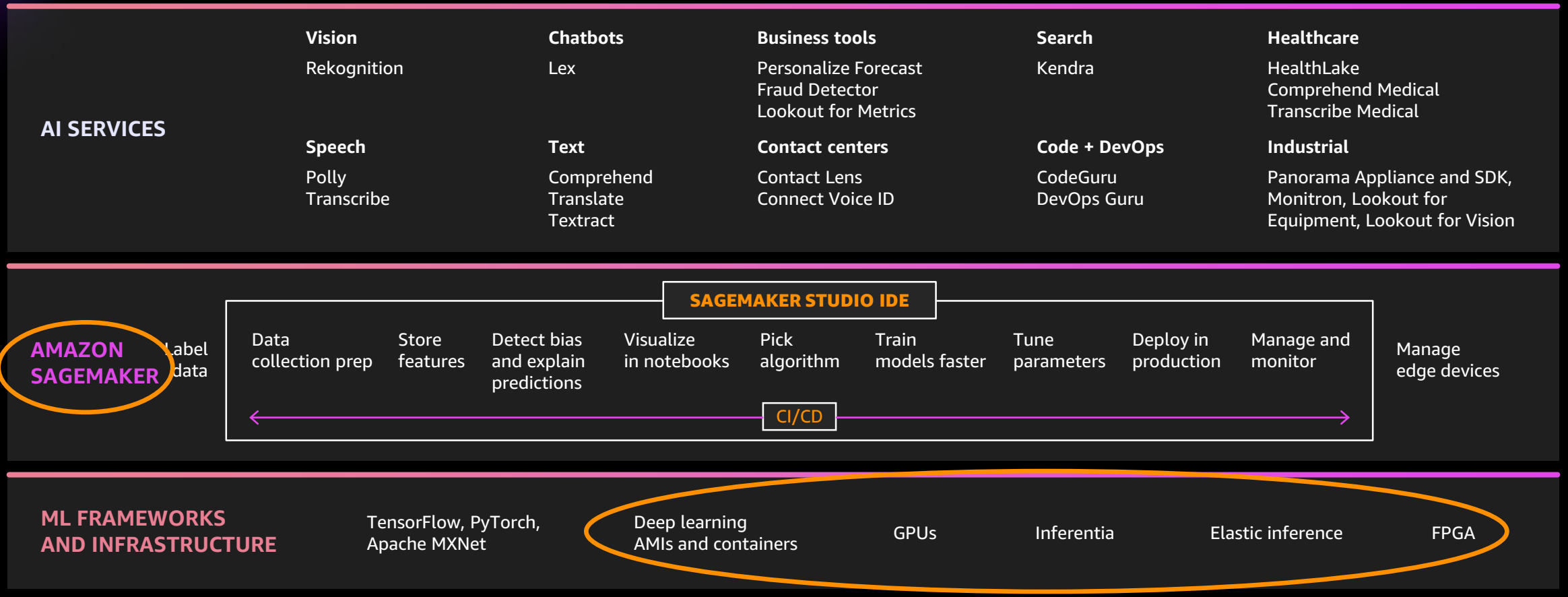
- Amazon SageMaker inference options
- Choosing the right Amazon Elastic Compute Cloud (Amazon EC2) instances for machine learning
- Demo - Low cost and high performance inference on AWS Inferentia

The reach of machine learning (ML) is growing



The AWS ML stack

BROADEST AND MOST COMPLETE SET OF MACHINE LEARNING CAPABILITIES



Amazon SageMaker inference options

Amazon SageMaker inference options

Real-time inference

- Low latency
- Ultra-high throughput
- Multi-model endpoints
- A/B testing

Example use cases:
ad serving, personalized recommendations, fraud detection

Batch transform

- Process large datasets
- Job-based system

Example use cases:
churn prediction, predictive maintenance, demand forecasting

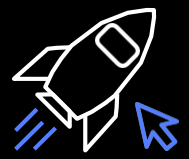
Asynchronous inference

- Near real-time
- Large payloads (up to 1 GB)
- Long timeouts (up to 15 min)

Example use cases:
computer vision, NLP

Amazon SageMaker Real-Time Inference

Amazon SageMaker Real-Time Inference

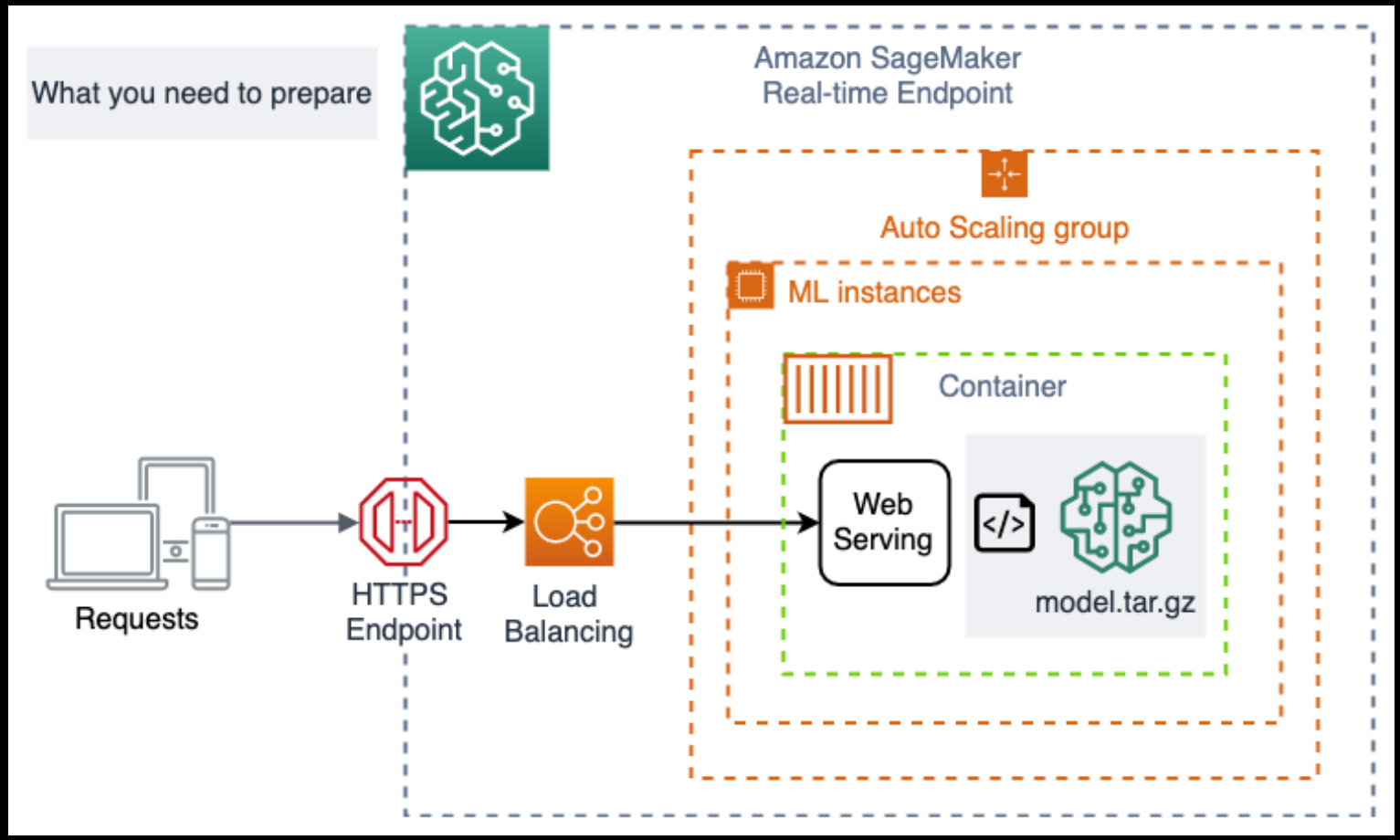


Create a long-running microservice

Instant response for payload up to 6MB

Accessible from an external application

Autoscaling



Amazon SageMaker Batch Transform

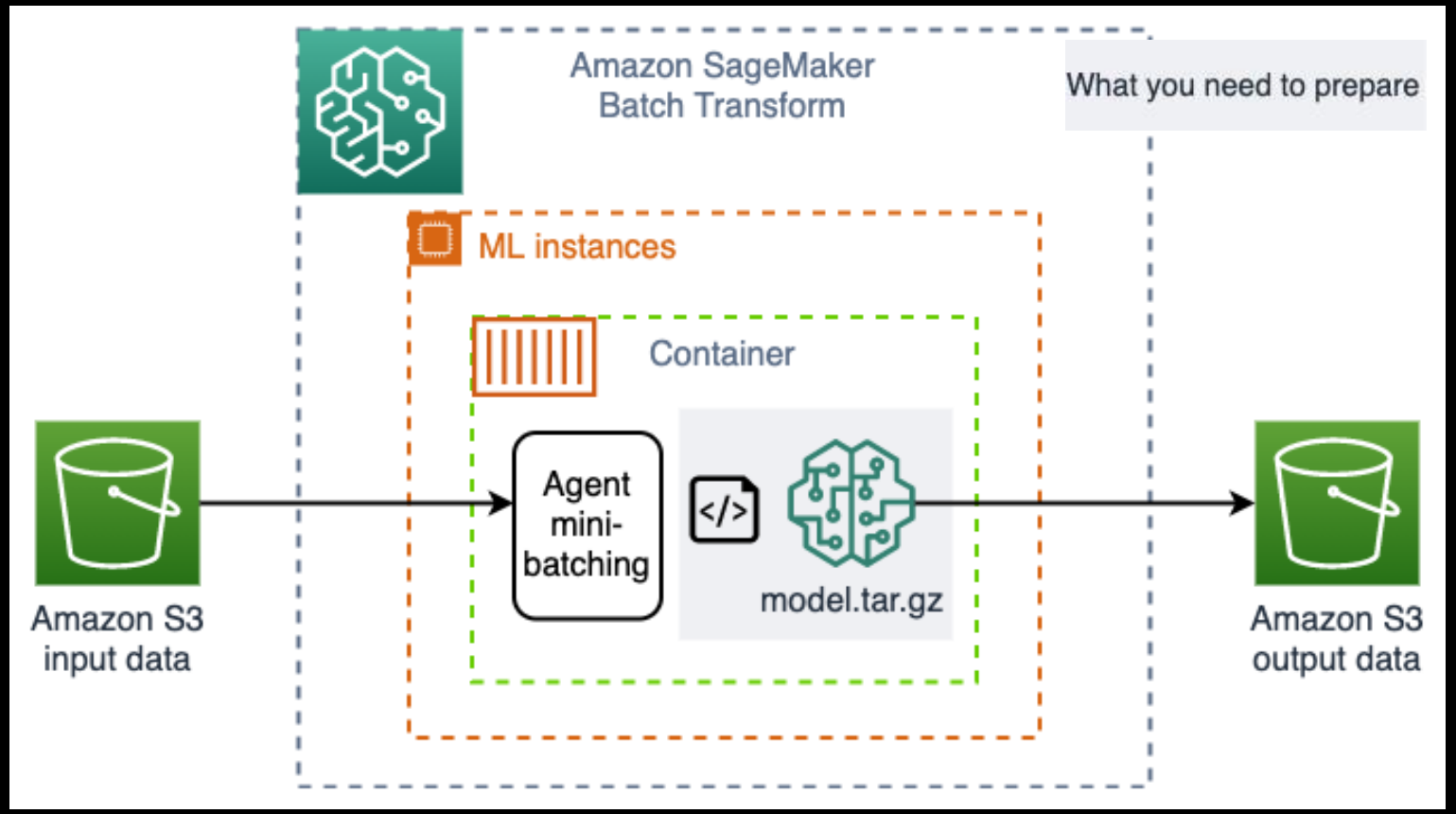
Amazon SageMaker Batch Transform

Provide S3 bucket path for input

Provide S3 bucket path for output

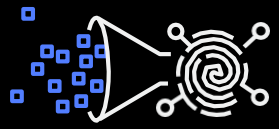
Provide compute resources

Name of the Amazon SageMaker model



Amazon SageMaker Asynchronous Inference

Amazon SageMaker Asynchronous Inference

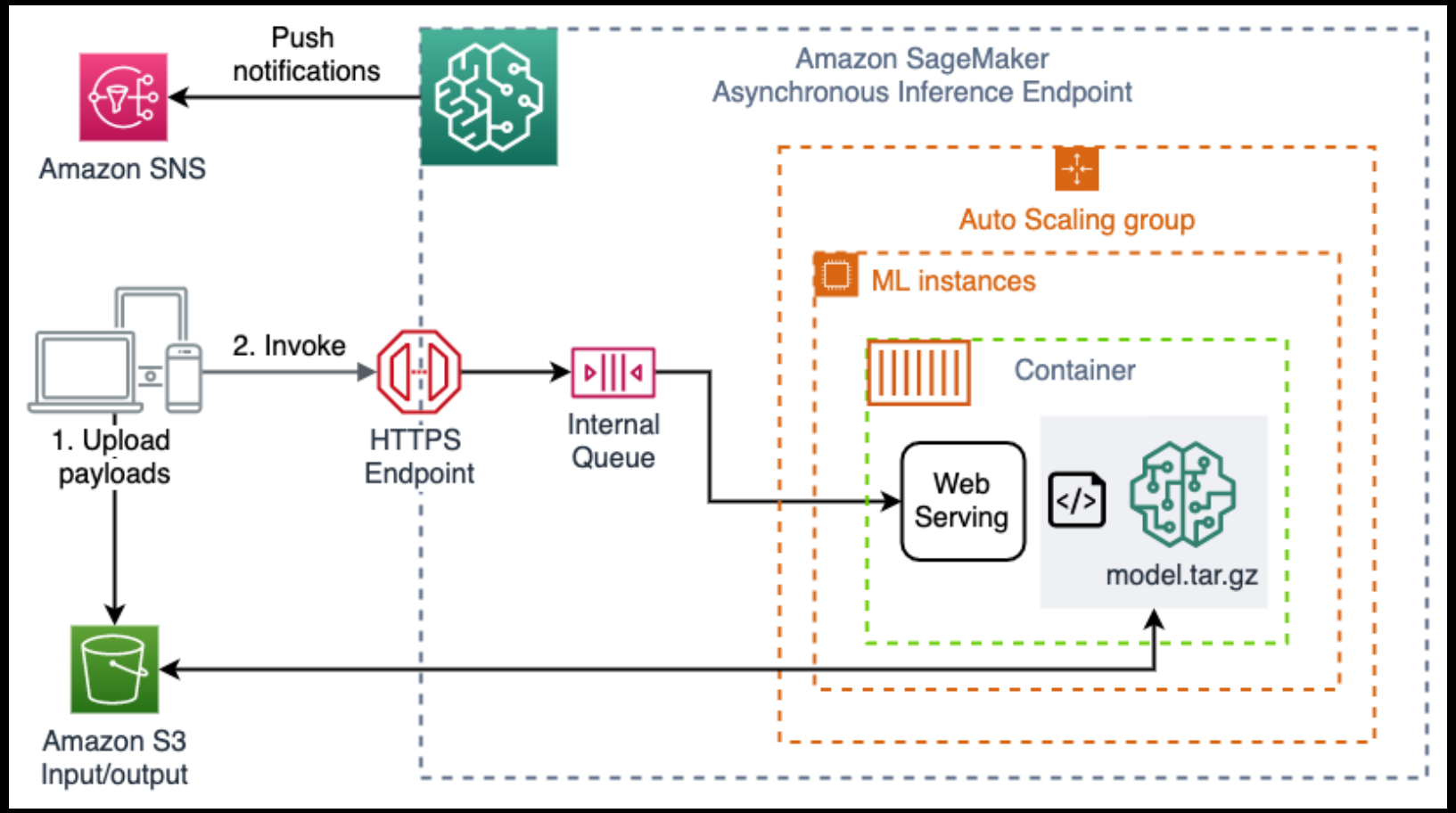


Ideal for large payload up to 1GB

Longer processing timeout up to 15 min

Autoscaling (down to 0 instance)

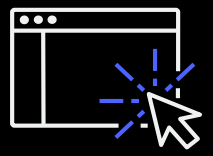
Suitable for CV/NLP use cases



Amazon SageMaker Serverless Inference

Amazon SageMaker Serverless Inference

First purpose built serverless ML inference in cloud



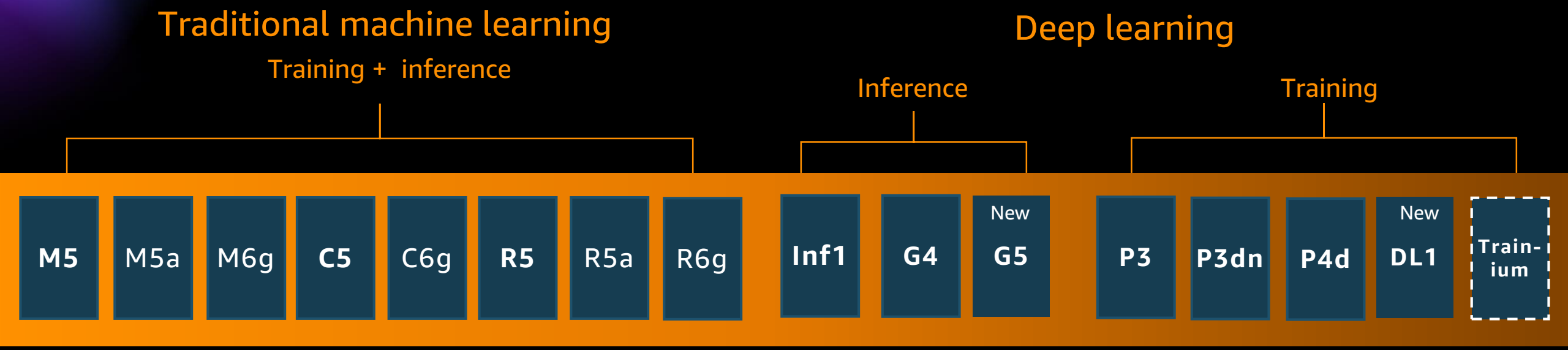
Fully managed

Pay only for what you use, billed in milliseconds

ML solutions on Amazon EC2 instances

Broadest and deepest compute

CPU, GPU & CUSTOM EC2 INSTANCES FOR ML



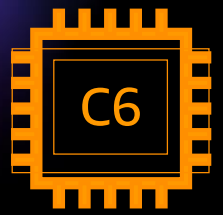
Cascade Lake CPU
Skylake CPU
Habana accelerator

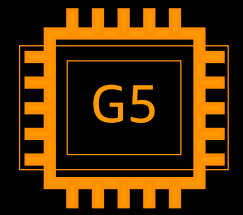
EPYC CPU

Inferentia Chip
Trainium Chip
Graviton CPU

A100, V100, T4, A10 GPUs

Choosing the right Amazon EC2 instance





Lower compute power

Higher compute power

What about?

Mid-sized models

Need acceleration but not a dedicated GPU

Lower throughput and higher latency tolerance

Cost sensitive

Low cost / inference for

- Small DL models
- Traditional ML models

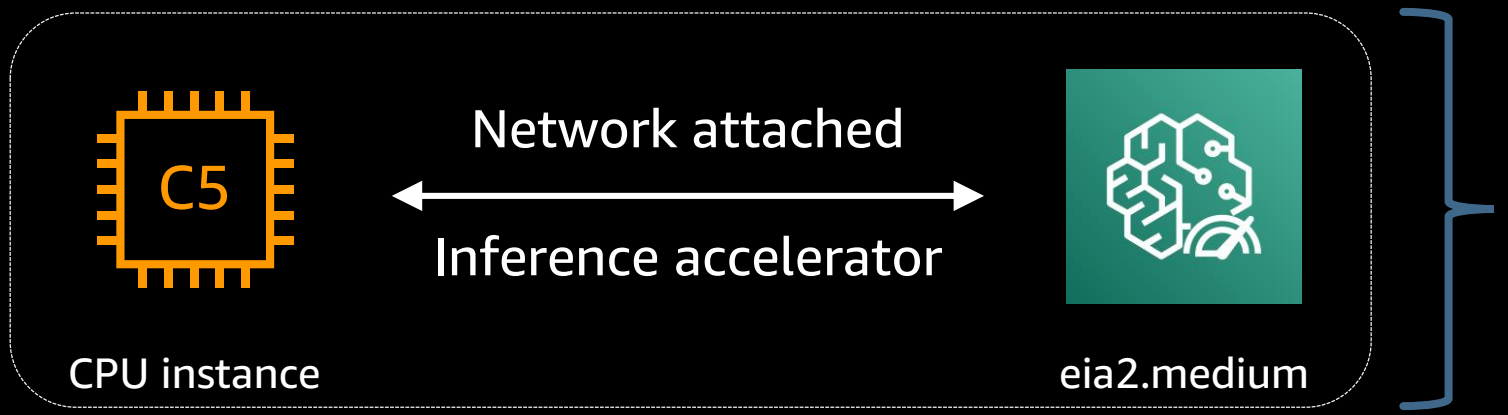
Low cost / inference for

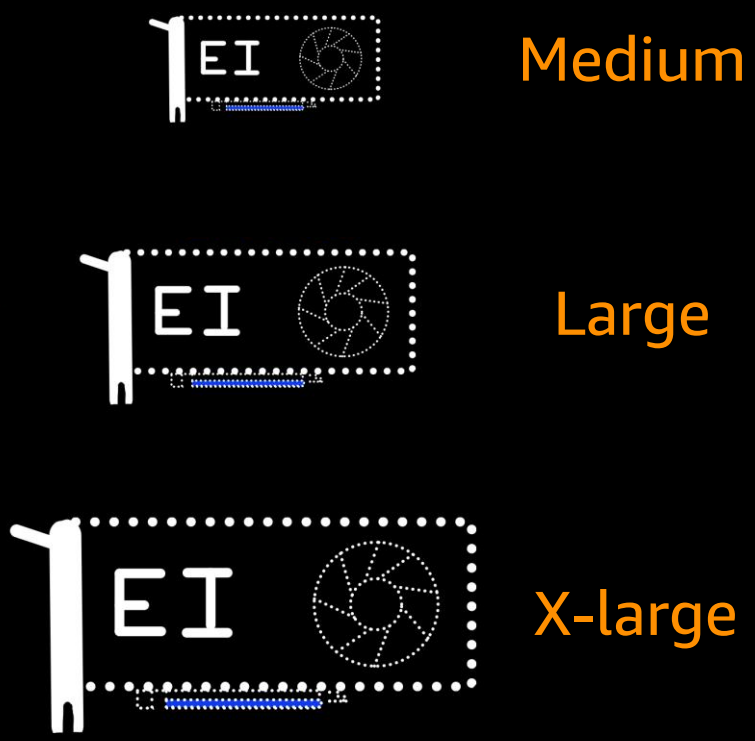
- Large DL models
- Large batch sizes
- High demand

aws

Amazon Elastic Inference

LOWER MACHINE LEARNING INFERENCE COSTS BY UP TO 75%





Reduce cost with access to
variable-size GPU acceleration

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

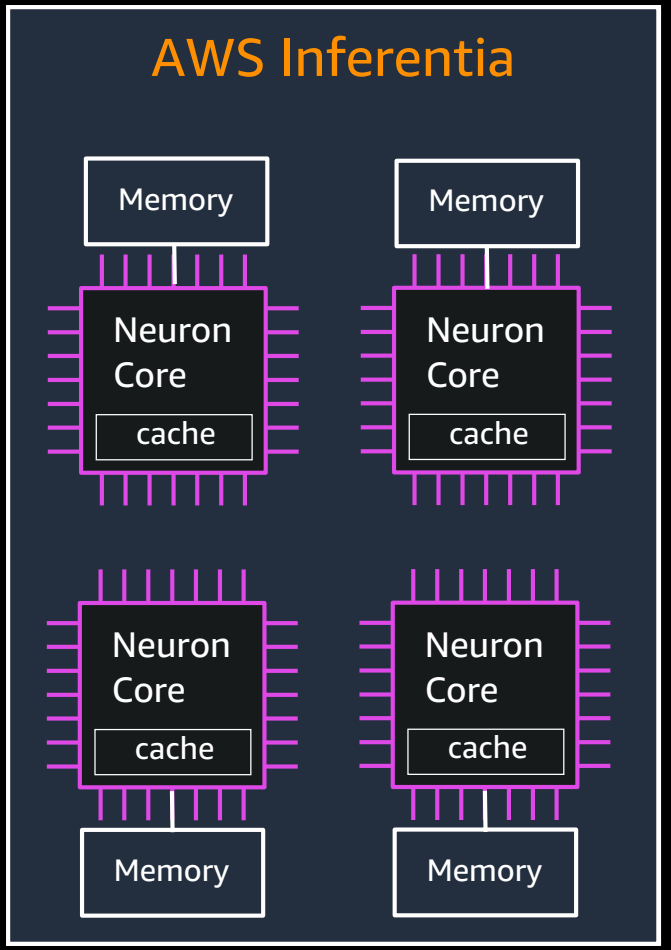
aws

AWS Inferentia: Custom silicon for ML inference

FIRST CUSTOM ML CHIP DESIGNED BY AMAZON

- 4 NeuronCores
- Up to 128 TOPS
- 2-stage memory hierarchy
  - Large on-chip cache and commodity DRAM
- Supports FP16, BF16, INT8 data types with mixed precision
- Fast chip-to-chip interconnect

https://aws.amazon.com/machine-learning/inferentia/



Choosing the right AWS Inf1 instance type

Considerations for Inf1 instances

- Optimizing for throughput or latency
  - Batching (batch inputs)
  - Pipelining (cache model)
- Number of models being deployed
- Number of custom layers and operators in your model
- Pre- and post-processing steps

Instance size	vCPUs	Inferentia Chips	Number of NeuronCores
inf1.xlarge	4	1	4
inf1.2xlarge	8	1	4
inf1.6xlarge	24	4	16
inf1.24xlarge	96	16	64

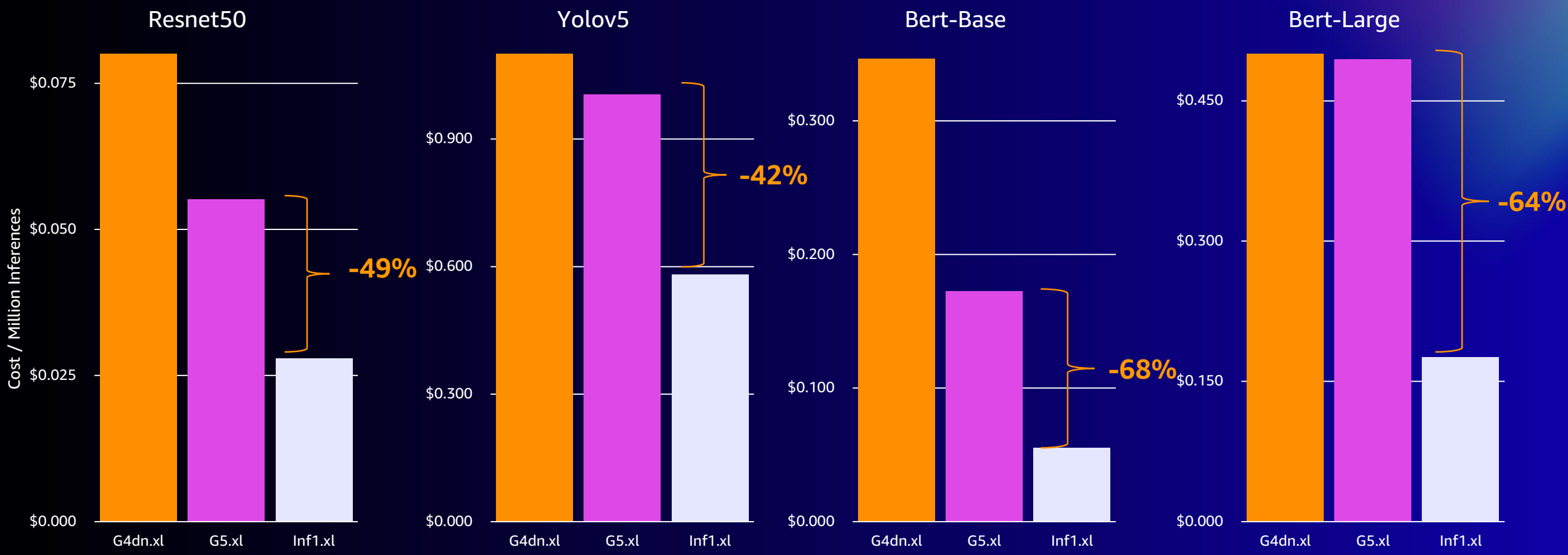
Start small, and scale up if you need more compute

https://aws.amazon.com/ec2/instance-types/inf1/

aws

Demo

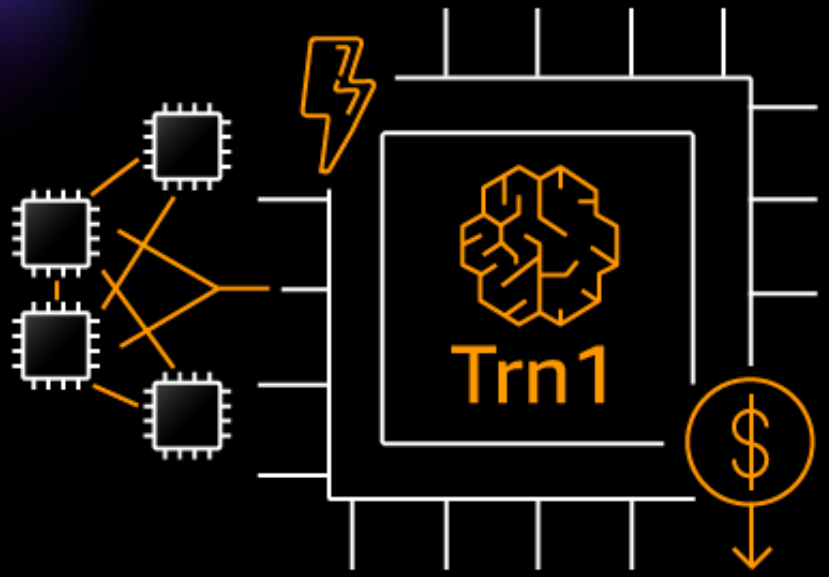
Inference cost comparison



Lower is better

Amazon EC2 Trn1 instances

THE MOST COST-EFFICIENT DL TRAINING IN THE CLOUD



- 60% higher accelerator memory (vs. P4d)
- 2x network BW (vs. P4d)
- Native support for PyTorch and TensorFlow
- Train on Trn1 and deploy anywhere

Instance size*	vCPUs	Trainium chips	Accelerator memory	NeuronLink	Instance memory	Instance networking	Local instance storage
Trn1.2xlarge	8	1	32 GB	N/A	32 GB	Up to 10Gbps	500 GB NVMe
Trn1.32xlarge	128	16	512 GB	**768 GB/sec**	512 GB	**800 Gbps**	8 TB NVMe

https://aws.amazon.com/ec2/instance-types/trn1/

aws

Recap

- Amazon SageMaker inference options
- Choosing the right EC2 instances for machine learning
- Demo - Low cost, high performance inference on AWS Inferentia

Resources

Amazon SageMaker developer guide

go.aws/32fyqck

Amazon SageMaker hands-on lab

go.aws/3rvXSTp

Amazon SageMaker examples

bit.ly/3Aeefbb

AWS Neuron SDK

bit.ly/3Ijv2g0

AWS Neuron/ Inf1/ quick start guide

bit.ly/3roMk4o

Visit the AI & Machine Learning resource hub for more resources

Dive deeper into these resources, get inspired and learn how you can use AI and machine learning to accelerate your business outcomes.

- The machine learning journey e-book
- 7 leading machine learning use cases e-book
- A strategic playbook for data, analytics, and machine learning e-book
  Accelerate machine learning innovation with the right cloud services & infrastructure e-book
- Choosing the right compute infrastructure for machine learning e-book
- Improving service and reducing costs in contact centers e-book
- Why ML is essential in your fight against online fraud e-book
- … and more!

https://bit.ly/3mwi59V

Visit resource hub

AWS Machine Learning (ML) Training and Certification

AWS is how you build machine learning skills

Courses built on the curriculum leveraged by Amazon's own teams. Learn from the experts at AWS.

aws.training/machinelearning

Flexibility to learn your way

Learn online with on-demand digital courses or live with virtual instructor-led training, plus hands-on labs and opportunities for practical application.

explore.skillbuilder.aws/learn

Validate your expertise

Demonstrate expertise in building, training, tuning, and deploying machine learning models with an industry-recognized credential.

aws.amazon.com/certification

aws

Thank you for attending AWS Innovate – AI/ML Edition

We hope you found it interesting! A kind reminder to **complete the survey.**
Let us know what you thought of today's event and how we can improve the event experience for you in the future.


aws-apj-marketing@amazon.com


twitter.com/AWSCloud


facebook.com/AmazonWebServices


youtube.com/user/AmazonWebServices


slideshare.net/AmazonWebServices


twitch.tv/aws

Thank you!

Santhosh Urukonda