
aws
INNOVATE
DATA AND AI/ML EDITION
22 February 2023

aws INNOVATE
DATA AND AI/ML EDITION

Rapidly launch ML solutions at scale on AWS infrastructure

Santhosh Urukonda
Prototyping Architect
AWS India

AWS Prototyping

Looking for the right way to solve complex business challenges

Let's implement a prototype

But ...

How do we find the right team?
What is the process and investment required?
How do we know the success of measure?

Build and demonstrate a working solution in rapid scale

- Agile methodology
- Duration 4 -6 weeks
- No fee basis

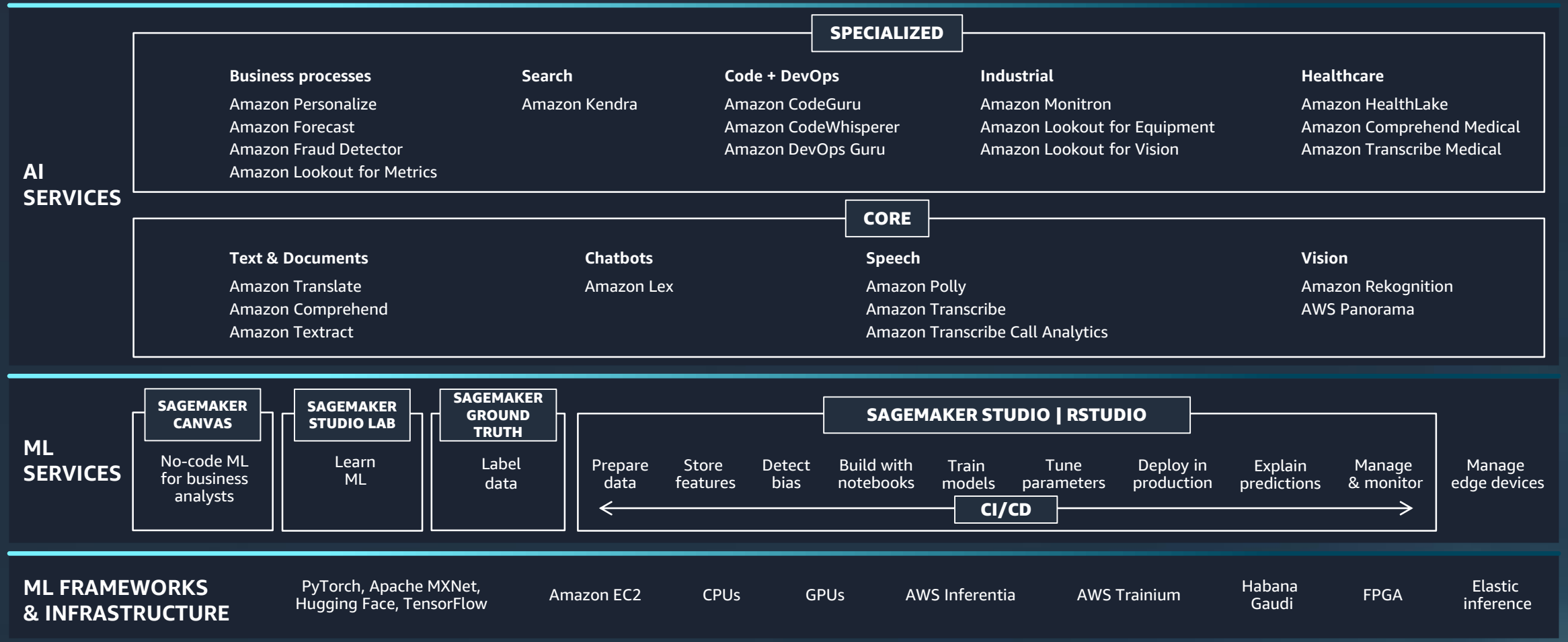
Prototyping

Agenda

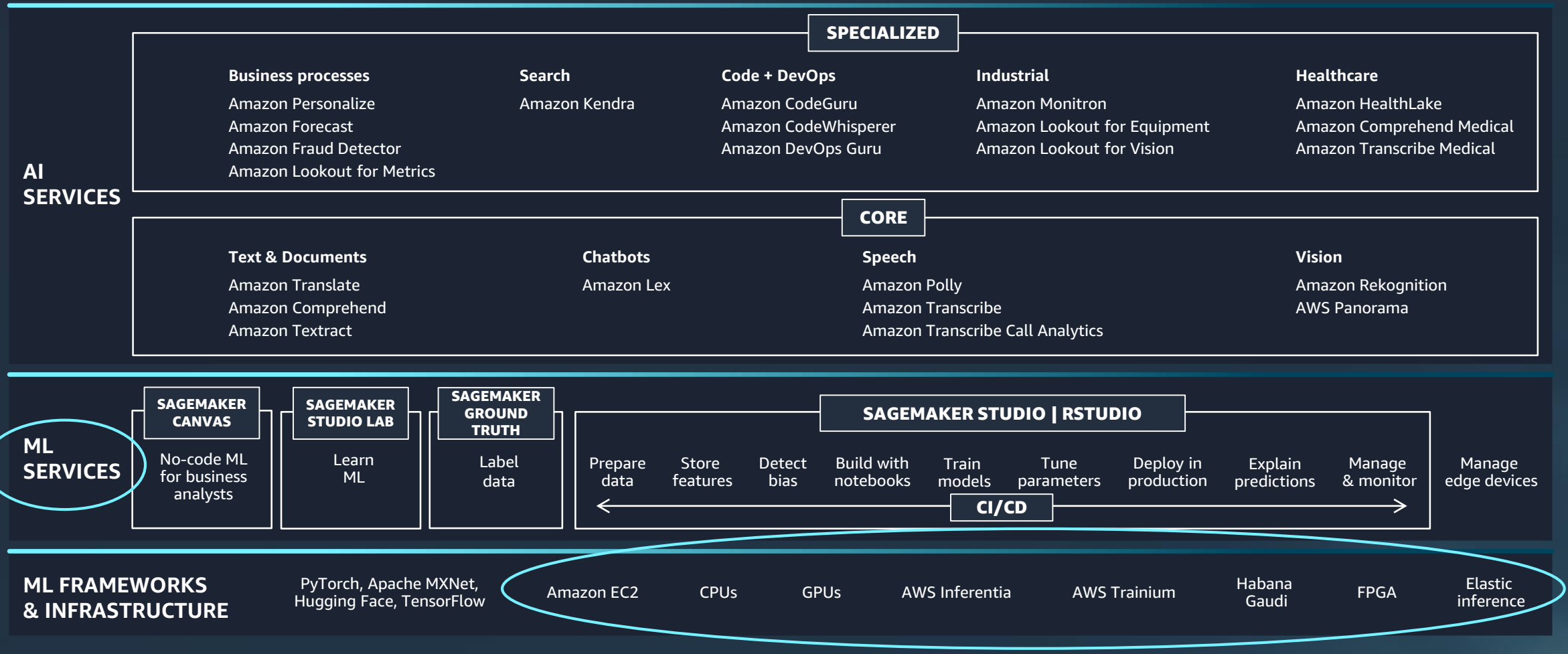
- Amazon SageMaker inference options
- Choosing the right Amazon EC2 instances for machine learning
- Demo - Low cost and high performance inference on AWS Inferentia

By 2026, one in every 3 applications will have machine learning infused in it

Meeting ML builders where they are



Meeting ML builders where they are



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

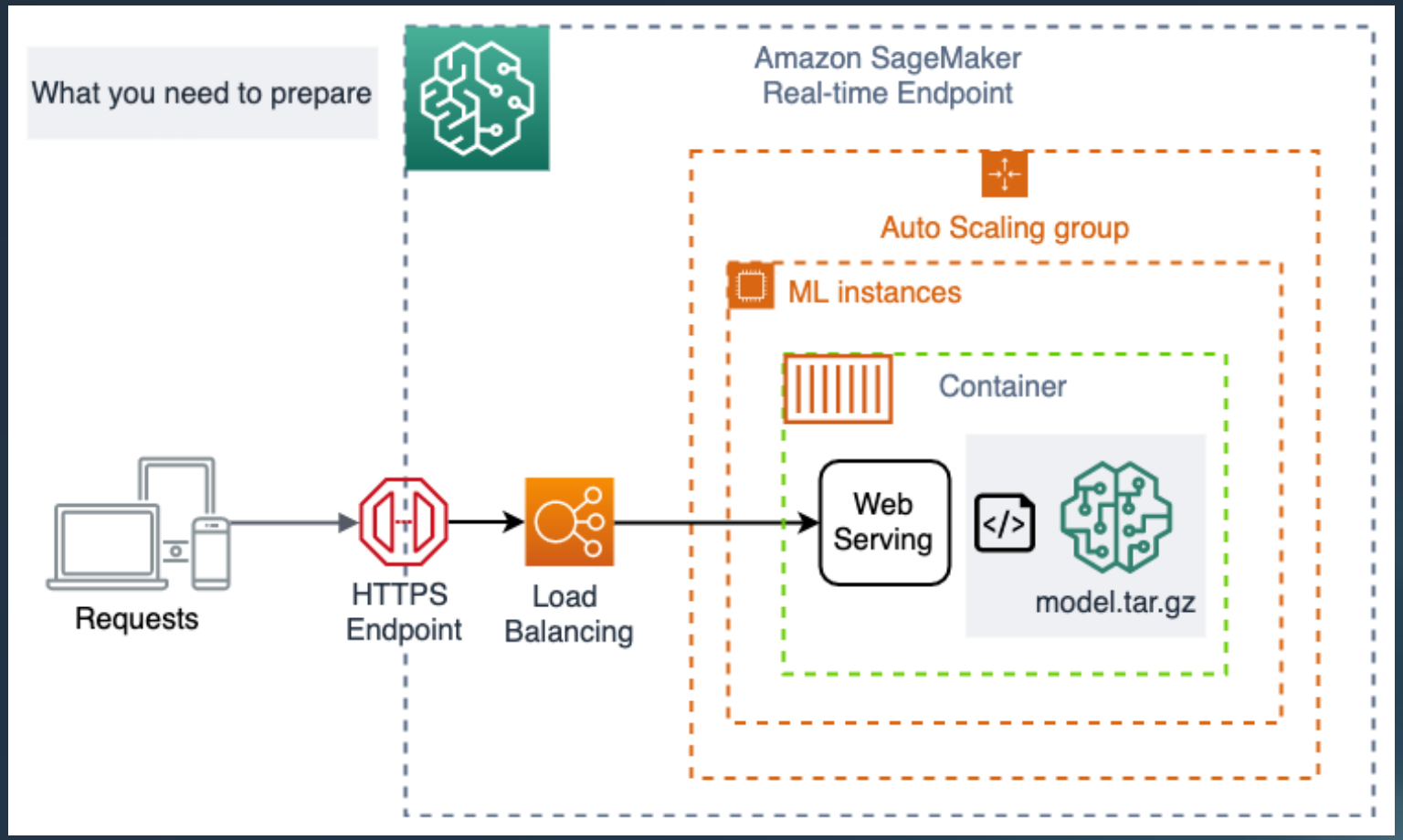
Amazon SageMaker Real-Time Inference

Create a long-running microservice

Instant response for payload up to 6MB

Accessible from an external application

Autoscaling



Amazon SageMaker Batch Transform

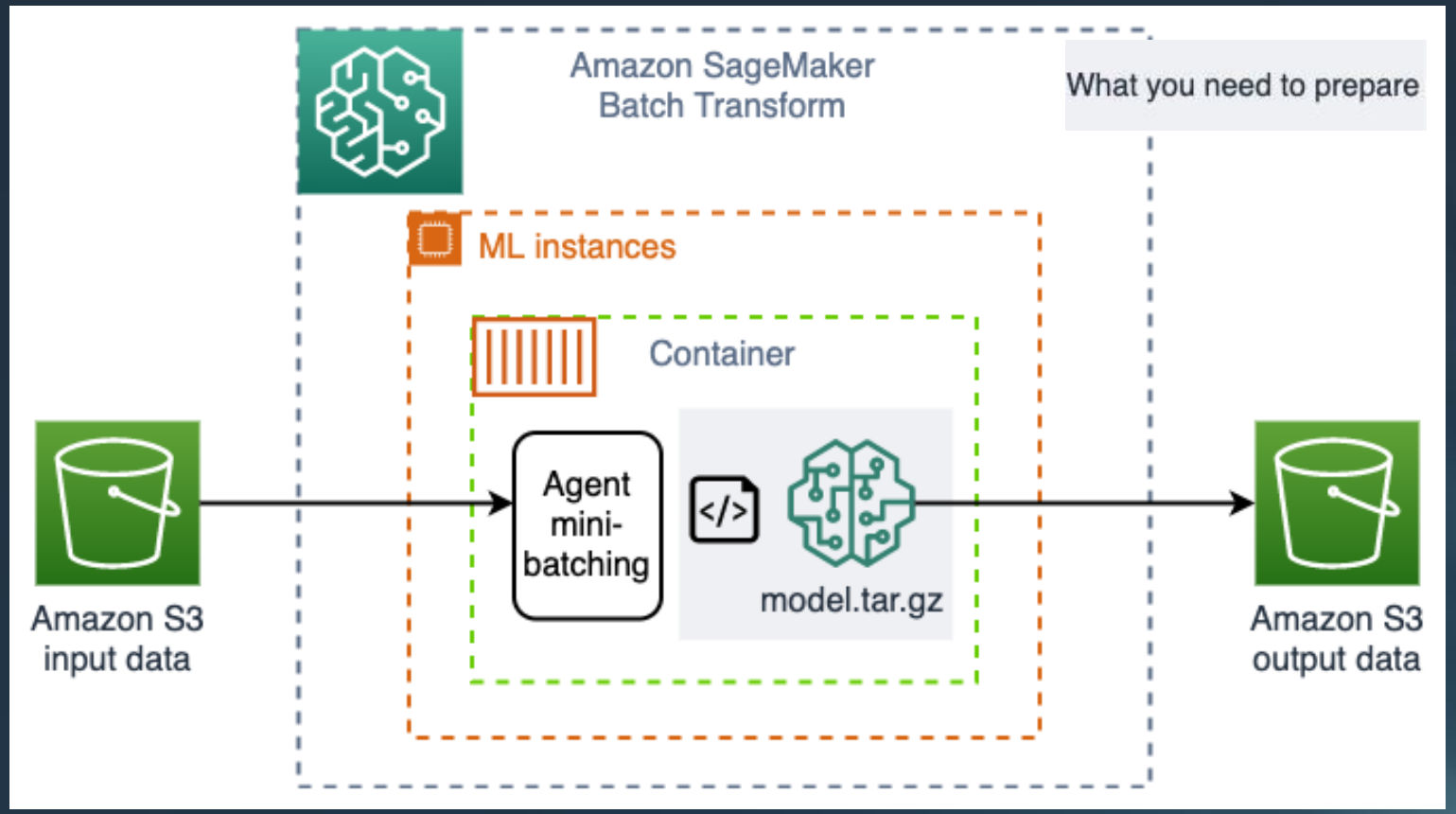
Amazon SageMaker Batch Transform

Provide Amazon S3 bucket path for input

Provide Amazon S3 bucket path for output

Provide compute resources

Name of the Amazon SageMaker model



Amazon SageMaker Asynchronous Inference

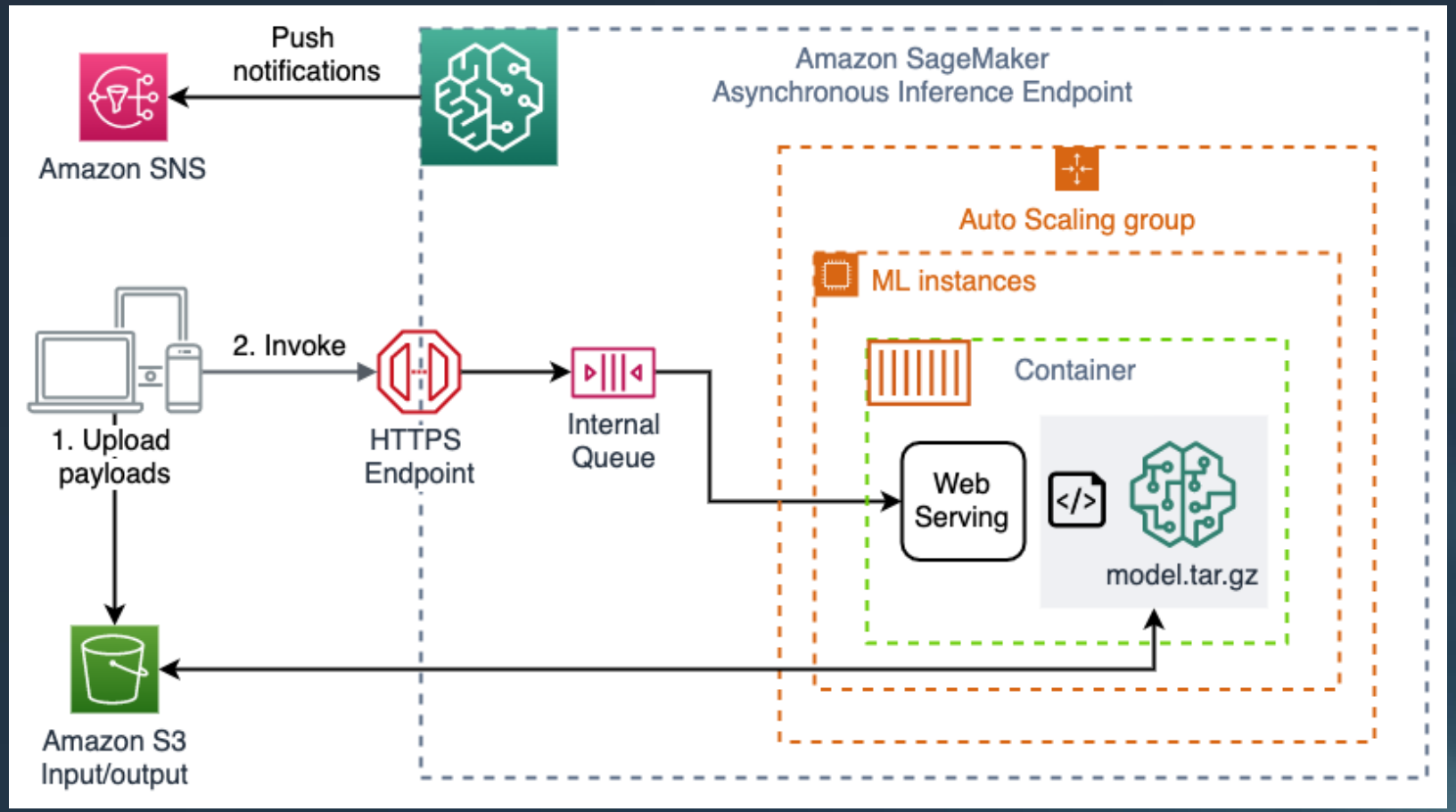
Amazon SageMaker Asynchronous Inference

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

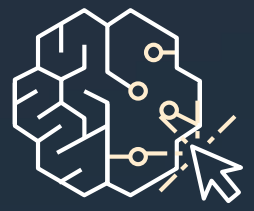
Amazon SageMaker Serverless Inference

First purpose built serverless ML inference in cloud

Fully managed

Pay only for what you use, billed in milliseconds

Amazon Inference Recommender

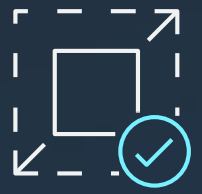


Instance recommendations

Instance type recommendation for initial deployments

Load tests

Run extensive load tests that include production requirements – throughput and latency



Endpoint recommendations

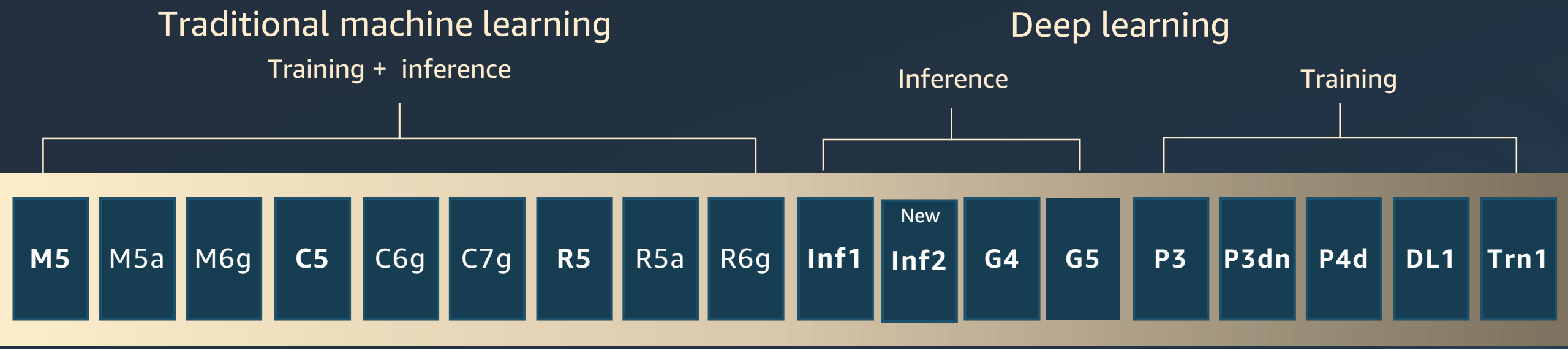
Get endpoint configuration settings that meet your production requirements

Designed for MLOps engineers and data scientists to reduce time to get models into production

ML solutions on Amazon EC2 instances

Broadest and deepest compute

CPU, GPU & CUSTOM EC2 INSTANCES FOR ML



NVIDIA

A100, V100, T4, A10 GPUs

AMD

EPYC CPU

annapurnalabs an amazon company

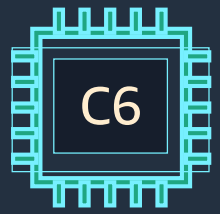
AWS Inferentia Chip
AWS Trainium Chip
AWS Graviton CPU

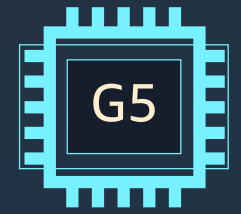
intel

Cascade Lake CPU
Skylake CPU
Habana accelerator

aws

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

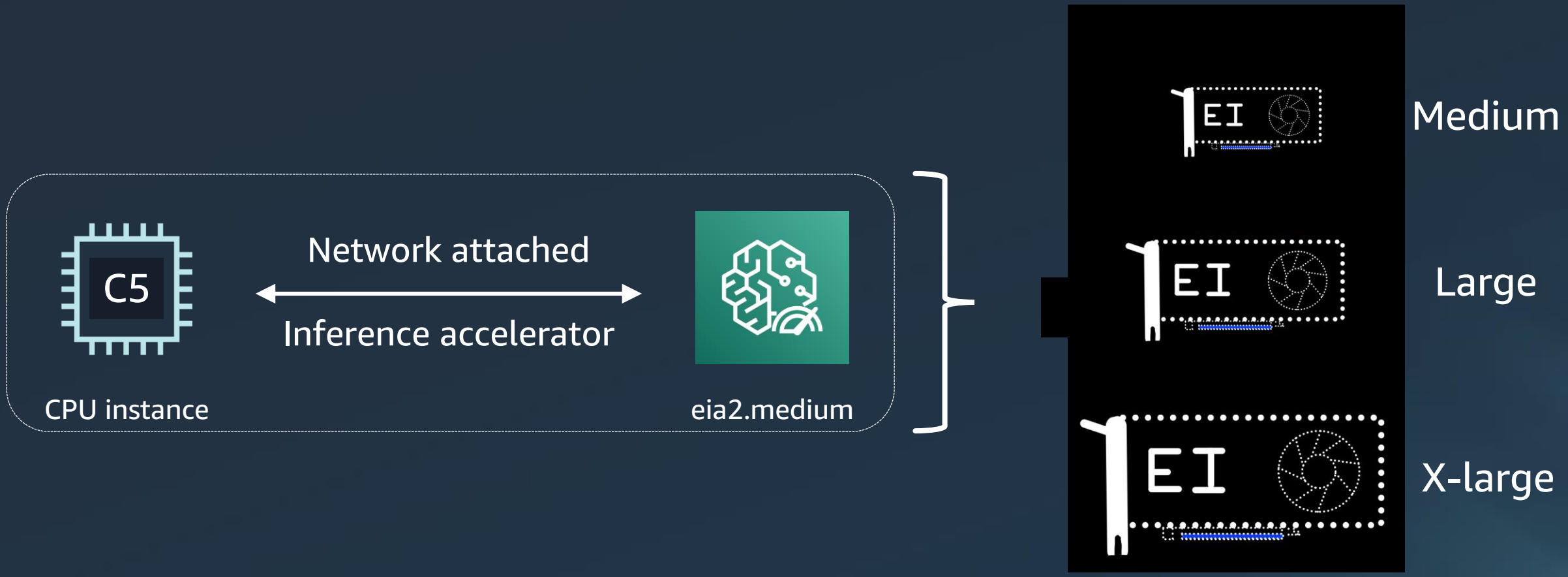
- Small DL models
- Traditional ML models

Low cost / inference for

- Large DL models
- Large batch sizes
- High demand

Amazon Elastic Inference

Lower machine learning inference costs by up to 75%



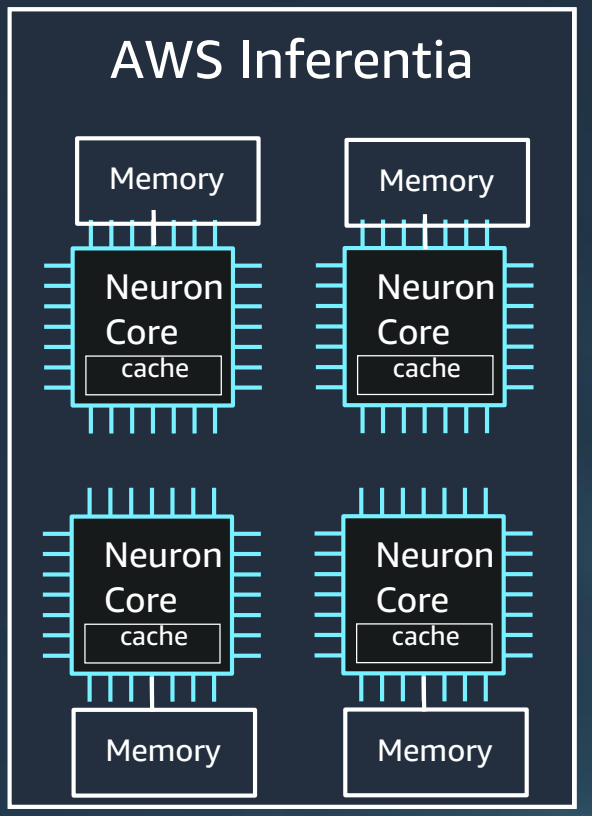
Reduce cost with access to
variable-size GPU acceleration

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

AWS Inferentia: Custom silicon for ML inference

First ML chip designed by AWS

- 4 Neuron Cores with up to 128 TOPS
- Two-stage memory hierarchy: Large on-chip cache + 8 GB DRAM
- Supports FP16, BF16, INT8 data types with mixed precision
- 1 to 16 Inferentia cores per instance with high-speed interconnect
- Optimized for high throughput and real-time low latency



Amazon EC2 Inf1 instances

HIGH PERFORMANCE, LOWEST-COST MACHINE LEARNING INFERENCE

- Featuring **AWS Inferentia**, the first ML chip designed by AWS
- **Lowest cost in the cloud** for running deep learning models – up to 80% lower cost than GPU instances
- Seamless software **integration with ML frameworks** like TensorFlow, PyTorch, and MXNet for quickly getting started and with minimal code changes
- Available through VMs, containers, Kubernetes, and Amazon SageMaker

AWS Inferentia

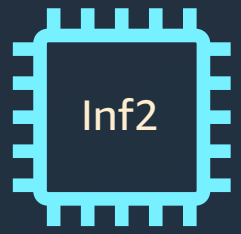
High performance ML inference chip, custom designed by AWS

Amazon EC2 Inf1 instances

Fastest and lowest-cost inference in the cloud

Amazon EC2 Inf2 instances

Preview



Inf2 instance

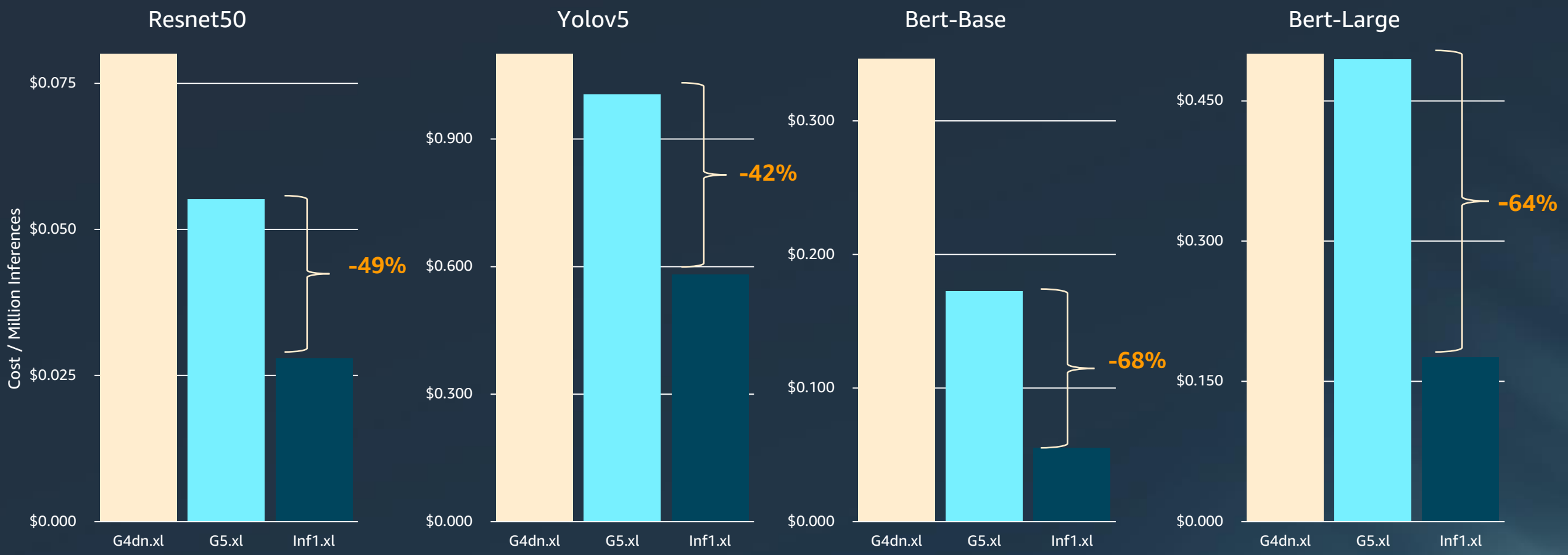
- Optimized to deploy 100B+ parameter models at scale
- Up to 4x higher throughput and up to 10x lower latency than Inf1 instances
- Up to 12 Inferentia2 accelerators and up to 384 GB of HBM2e high speed accelerator memory
- First inference platform with direct, 192 GB/s connectivity between multiple accelerators for distributed inference
- Native integration with PyTorch and TensorFlow
- 45% better performance/watt than GPU-based instances

Sign up for the Inf2 preview at https://aws.amazon.com/ec2/instance-types/inf2/

aws

Demo

Inference cost comparison



Lower is better

Recap

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

Additional 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 Data & AI/ML resource hub

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

- 6 steps to machine learning success e-book
- 7 leading machine learning use cases e-book
- Machine learning at scale e-book
- Achieving transformative business results with machine learning e-book
- Tackling our world's hardest problems with machine learning e-book
- Accelerating machine learning innovation through security e-book
- … and more!

https://bityl.co/FqdC

Visit resource hub

AWS Training and Certification

Access the AI & ML learning plan courses built by AWS experts on AWS Skill Builder

- Get started with digital self-paced, on-demand training and ramp-up guides to help you grow your technical skills
- Learn how to apply machine learning, artificial intelligence, and deep learning to unlock new insights and value in your role
- Take the steps today, towards validating your expertise with an AWS Certified Machine Learning – Specialty Certification

https://bit.ly/3FnxDH7

Learn your way explore.skillbuilder.aws »

Thank you for attending AWS Innovate – Data & 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!