
aws
INNOVATE
AI/ML EDITION

aws INNOVATE
AI/ML EDITION

Fully-managed ML deployments on AWS

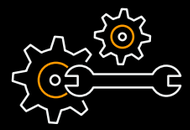
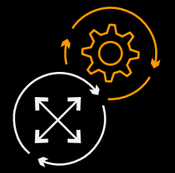
Eshaan Anand
Senior Partner Solutions Architect
Amazon Web Services

Agenda

- Amazon SageMaker
- Considerations & capabilities
- GPU instances for ML inference
- Accelerators for ML inference
- ML inference on the edge
- Resources

Amazon SageMaker

Bringing machine learning to all developers

Pre-built notebooks for common problems	Built-in, high performance algorithms	One-click training	Optimization	One-click deployment	Fully managed with auto-scaling, health checks, automatic handling of node failures, and security checks
					
Collect and prepare training data	Choose and optimize your ML algorithm	Set up and manage environments for training	Train and tune model (trial and error)	Deploy model in production	Scale and manage the production environment

Amazon SageMaker

Bringing machine learning to all developers

Pre-built notebooks for common problems

Collect and prepare training data

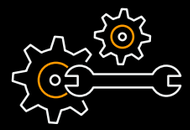
Built-in, high performance algorithms

Choose and optimize your ML algorithm

One-click training

Set up and manage environments for training

Optimization

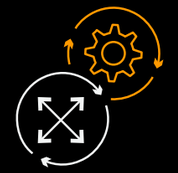


Train and tune model (trial and error)

One-click deployment

Deploy model in production

Fully managed with auto-scaling, health checks, automatic handling of node failures, and security checks



Scale and manage the production environment

Key considerations

- **Target performance**

  Target throughput under desired latency

- **Cost efficiency**

  Maximizing instance utilization to reduce cost / inference

- **Model and framework support**

  Support for custom and popular models (ResNet, BERT etc.)

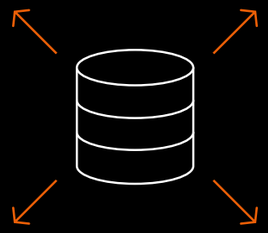
- **Compute**

  Use of CPU / GPU / accelerators / Edge for running inference

- **Security**

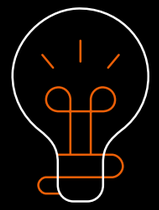
  Security-related parameters and configurations

Key capabilities - Amazon SageMaker Model Deployment



Auto-scaling

Scale inference endpoints based on traffic; set min and max instances and scaling criteria

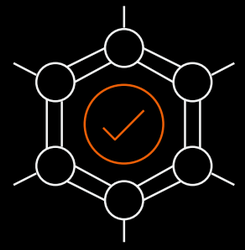


Production variants

Run different model versions on an endpoint and distribute traffic between model versions for AB testing

Inference pipelines

Run models sequentially in production with optional pre-processing and post-processing steps on each request



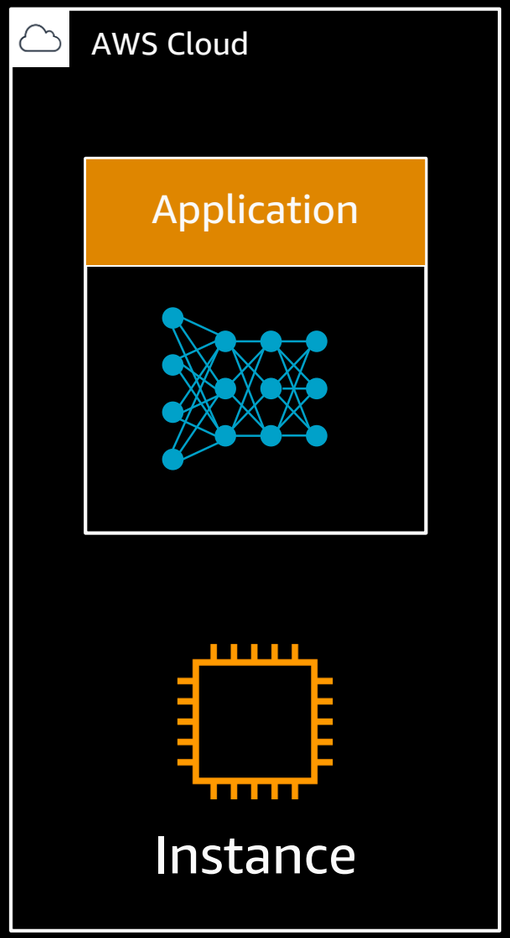
Multi-model endpoints

Deploy multiple (tens to thousands) models on an endpoint for significant cost savings

Model monitor

Collect requests/ responses from endpoints, get alerted on data drift through an automated, fully managed monitoring workflow

Inference Performance affects Customer Experience

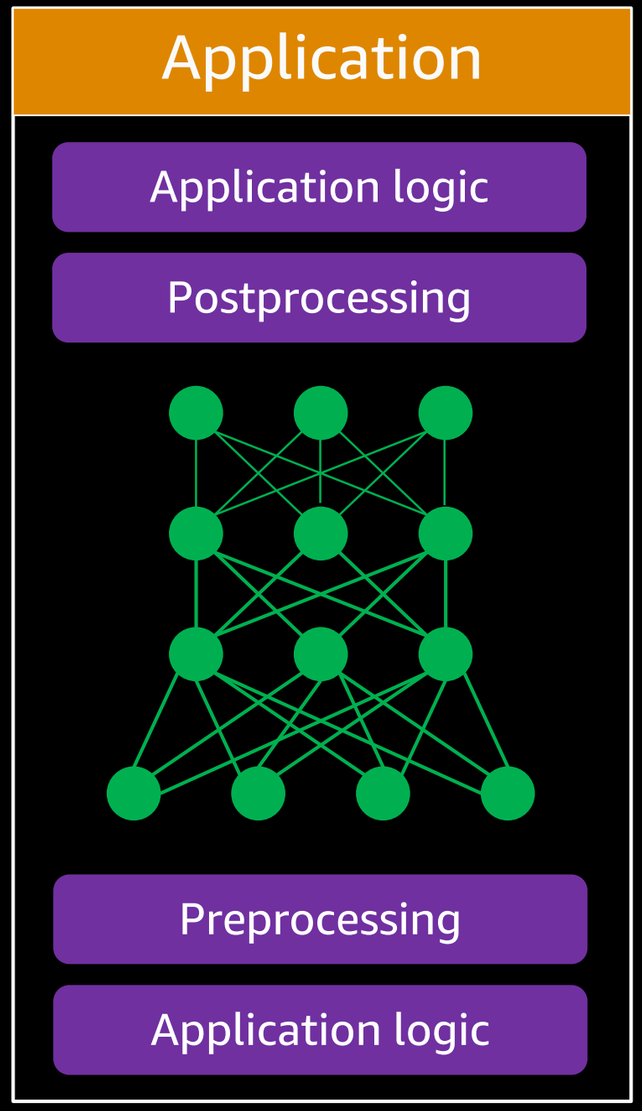


Round-trip latency

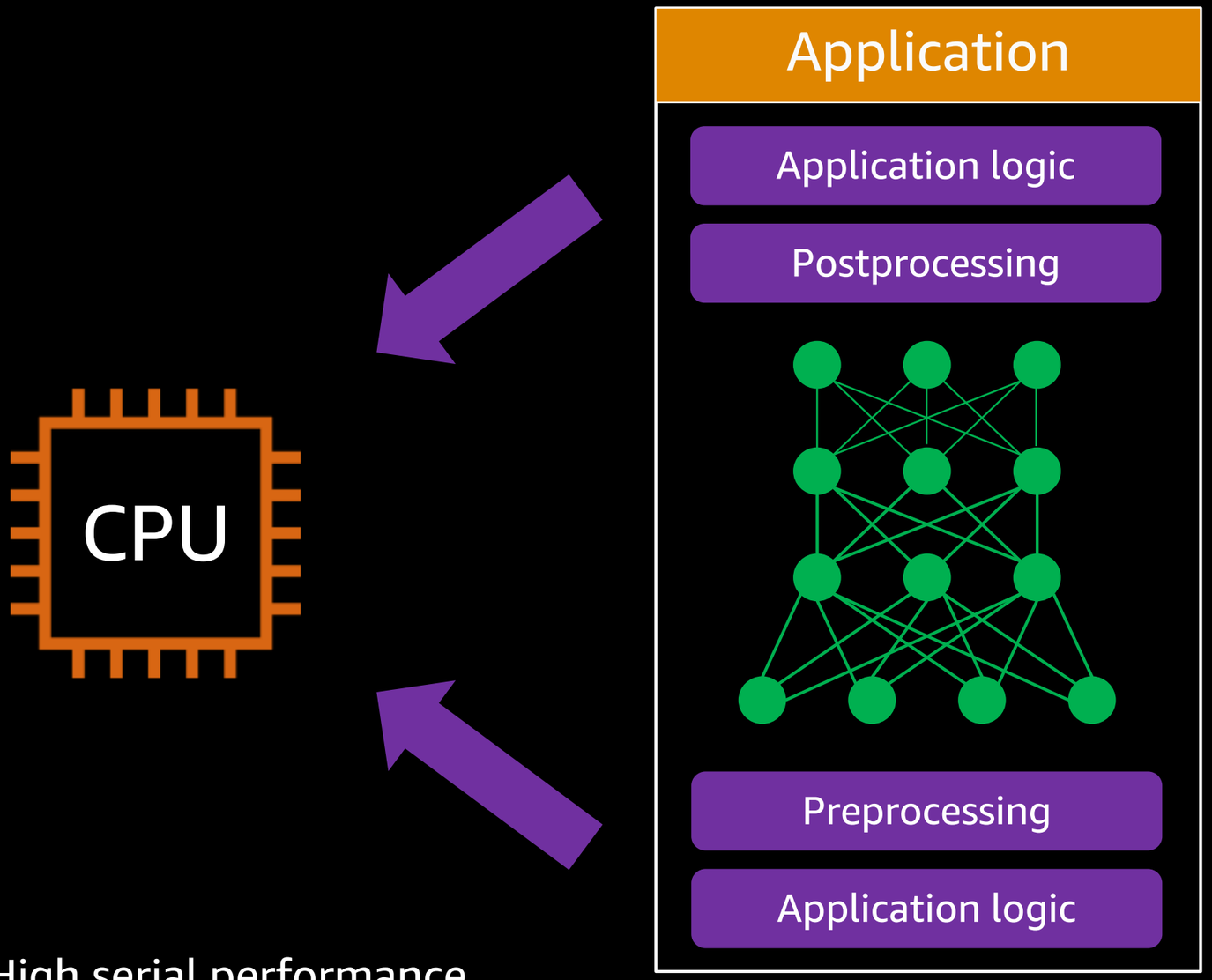
Throughput

Cost

Speed up inference with an accelerator

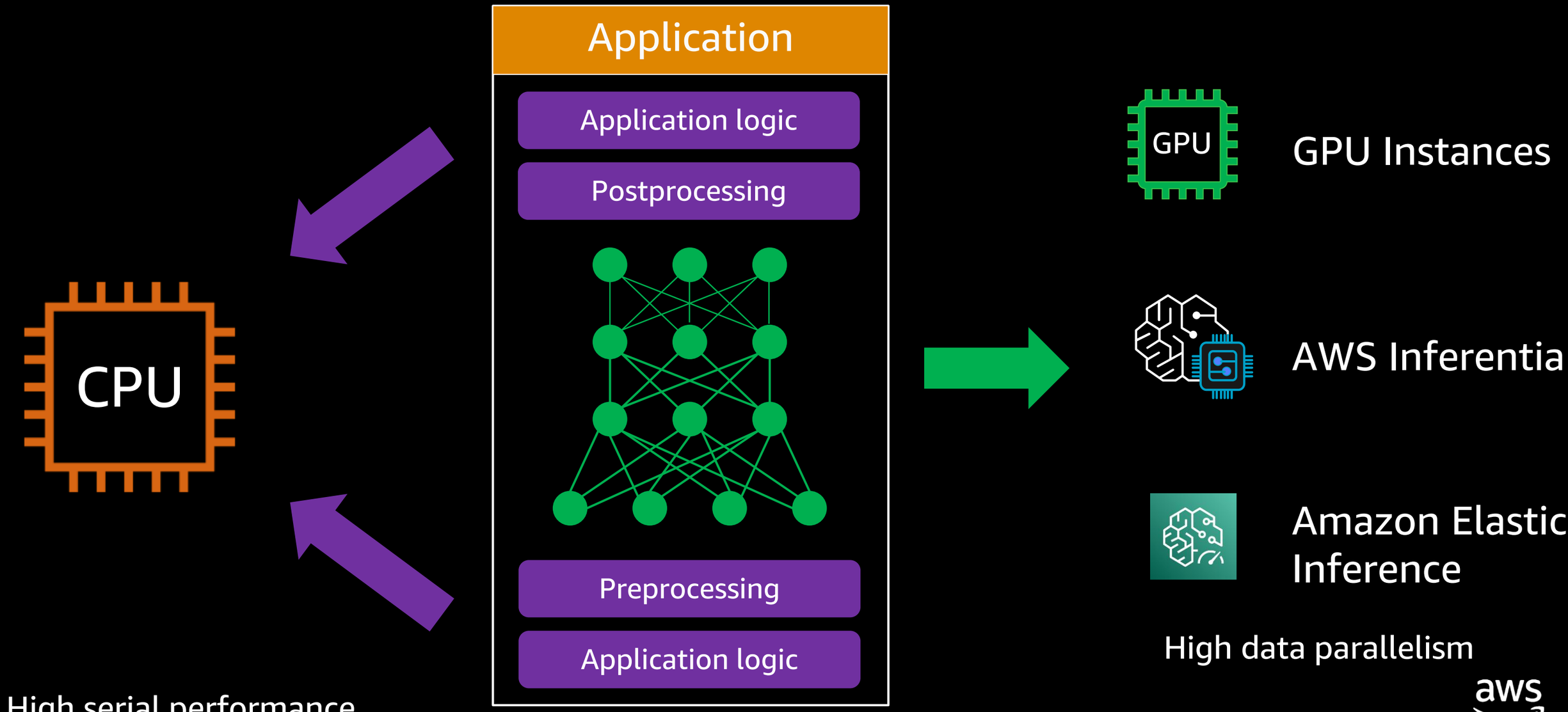


Speed up inference with an accelerator



High serial performance

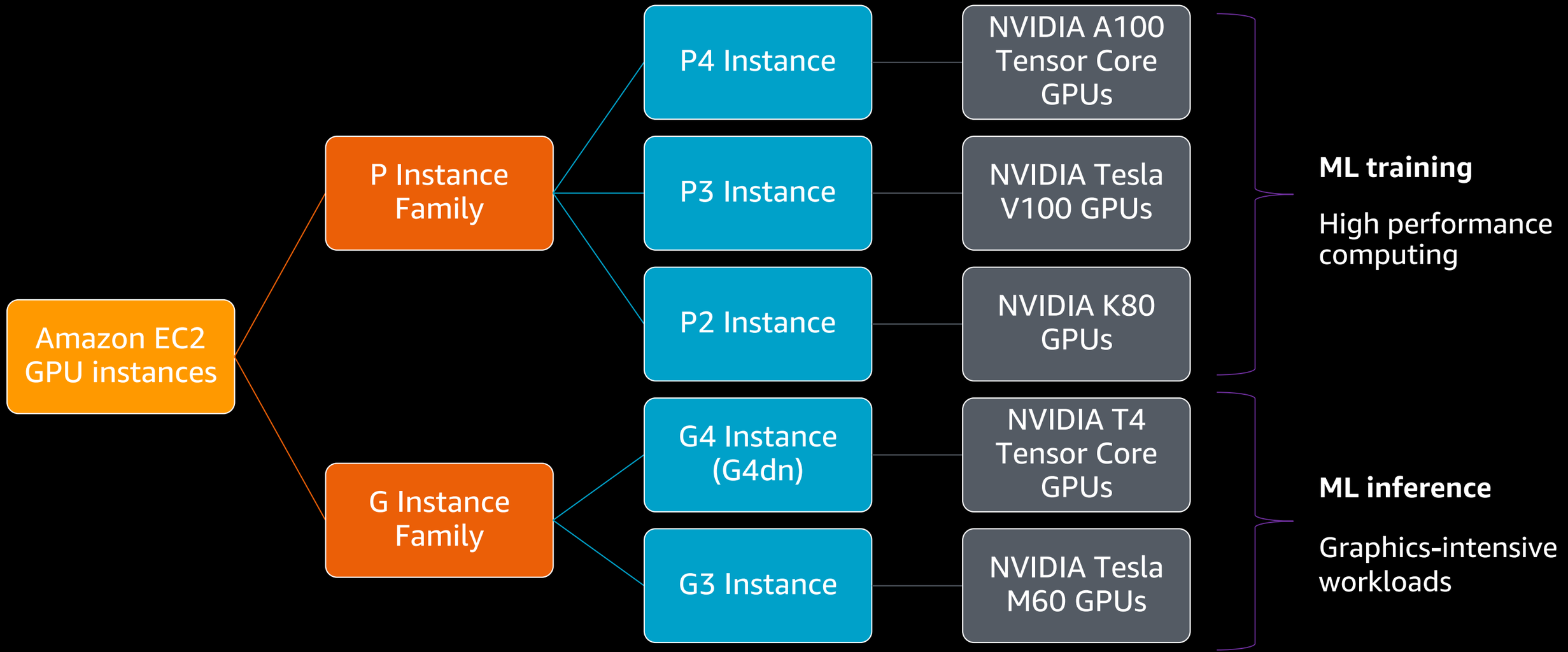
Speed up inference with an accelerator



GPU accelerated EC2 instances

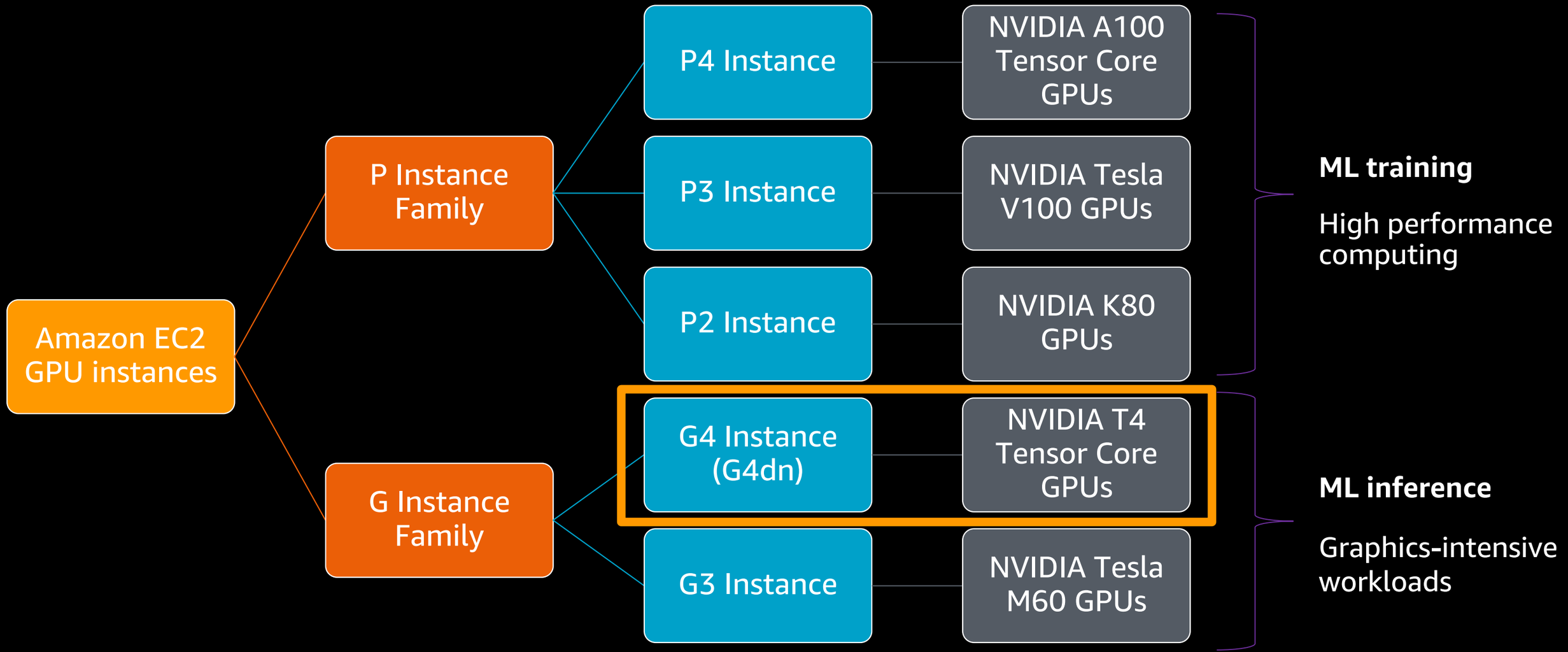
aws INNOVATE
ONLINE CONFERENCE

Amazon EC2 GPU instances for deep learning



https://aws.amazon.com/ec2/instance-types/#Accelerated_Computing
https://aws.amazon.com/ec2/instance-types/g4/

Amazon EC2 GPU instances for deep learning



https://aws.amazon.com/ec2/instance-types/#Accelerated_Computing
https://aws.amazon.com/ec2/instance-types/g4/

Amazon EC2 G4 instance family at a glance

BEST GPU INSTANCE FOR COST-EFFICIENT AND HIGH-PERFORMANCE INFERENCE DEPLOYMENTS

GPU memory: 16 GiB

Supported precision types

- FP32, FP16, INT8
- Tensor Cores (mixed-precision)

AWS optimizations for deep learning frameworks and GPU

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

Single GPU Instance

- g4dn.xlarge
- g4dn.2xlarge
- g4dn.4xlarge
- g4dn.8xlarge
- g4dn.16xlarge

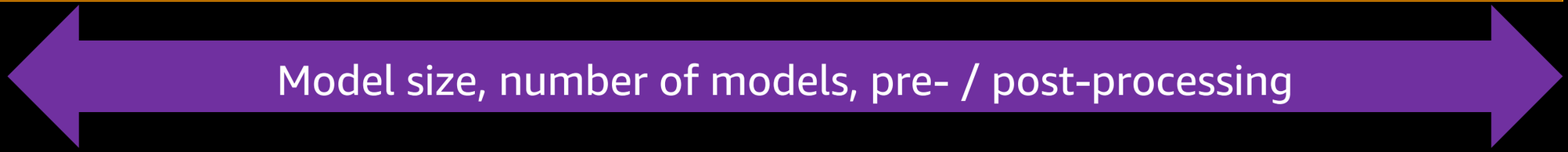
Multi-GPU Instances

- g4dn.12xlarge (4 GPUs)
- g4dn.metal (8 GPUs)

Choosing the right G4 instance size

Single-GPU instances: NVIDIA T4, 16 GB GPU memory

	g4dn.xlarge	g4dn.2xlarge	g4dn.4xlarge	g4dn.8xlarge	g4dn.16xlarge
vCPUs	4	8	16	32	64
System Mem(GiB)	16	32	64	128	256



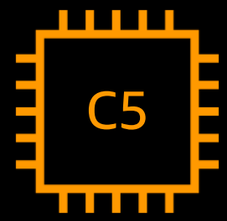
Multi-GPU instances

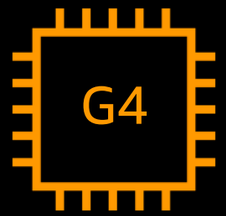
	g4dn.12xlarge	g4dn.metal
GPUs	4 x T4	8 x T4
vCPUs	4	8
System Mem (GiB)	16	32

https://aws.amazon.com/ec2/instance-types/#Accelerated_Computing

Start small, and scale up if you need more compute

What if you can't maximize GPU utilization?





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

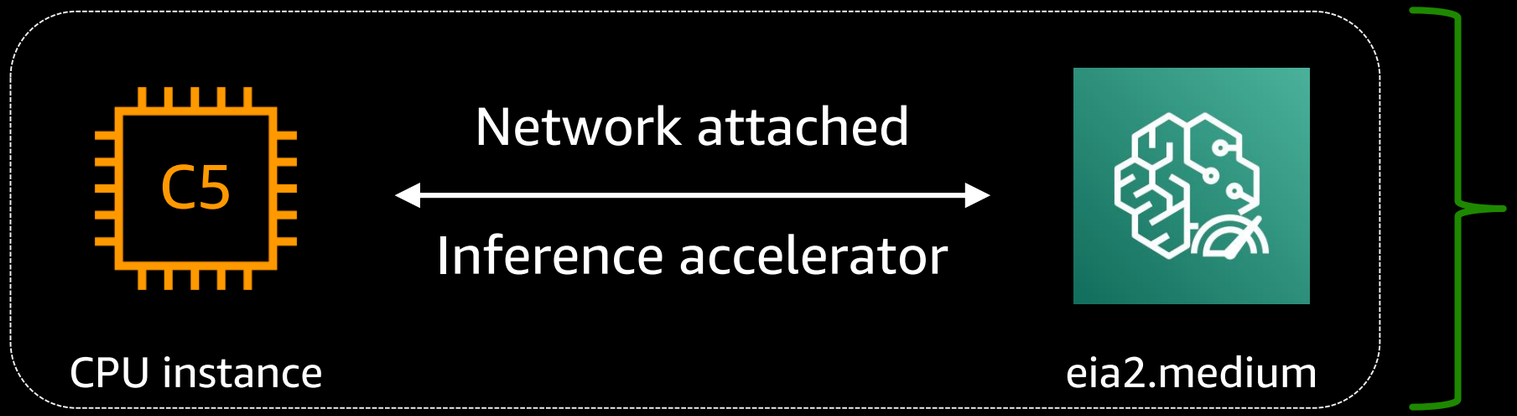
- Large DL models
- Large batch sizes
- High demand

Amazon Elastic Inference

aws INNOVATE
ONLINE CONFERENCE

Amazon Elastic Inference

LOWER MACHINE LEARNING INFERENCE COSTS BY UP TO 75%



Medium

Large

X-large

Reduce cost with access to
variable-size GPU acceleration

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

Choosing the right EI accelerator

Host CPU instances

Compute optimized C4 and C5 instance types

EI accelerator	FP32 – TFLOPS	FP16 – TFLOPS	Memory
EIA2 family:			
eia2.medium	1 TFLOPS	8 TFLOPS	2 GB
eia2.large	2 TFLOPS	16 TFLOPS	4 GB
eia2.xlarge	4 TFLOPS	32 TFLOPS	8 GB
EIA1 family:			
eia1.medium	1 TFLOPS	8 TFLOPS	1 GB
eia1.large	2 TFLOPS	16 TFLOPS	2 GB
eia1.xlarge	4 TFLOPS	32 TFLOPS	4 GB

Considerations for choosing CPU instance

- Number of custom layers and operators in your model
- Preprocessing steps
- Post processing steps

Considerations for choosing EIA

- Model size and memory
- Model complexity
- Target throughput

Start small, and scale up if you need more compute

AWS Inferentia

* Coming in 2021

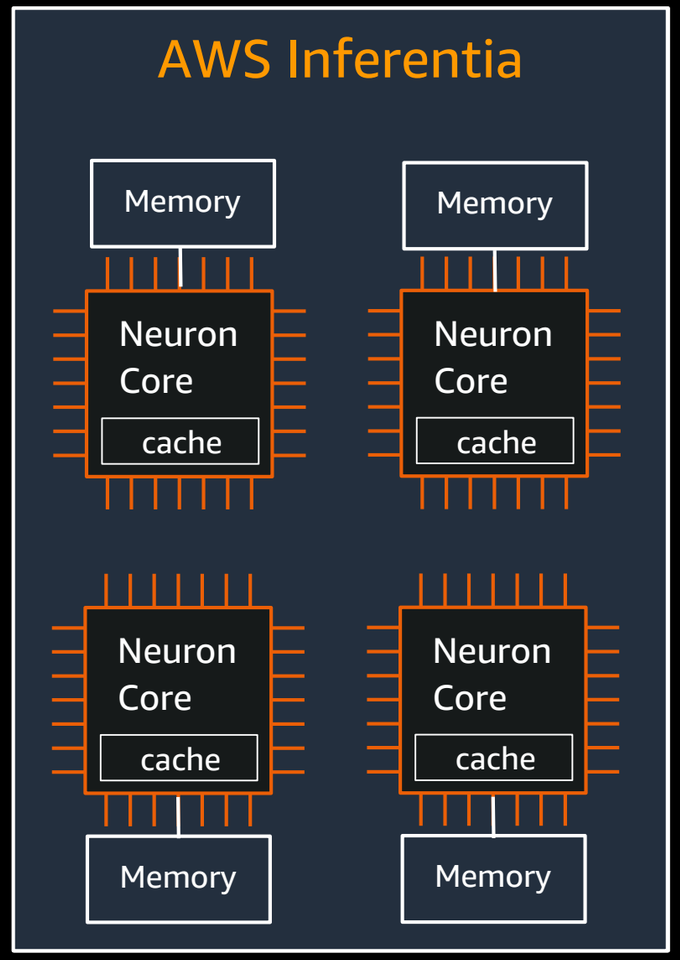
aws INNOVATE
ONLINE CONFERENCE

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/



Amazon EC2 Inf1 instance family at a glance

HIGH ML INFERENCE PERFORMANCE FOR THE LOW COST

- Accelerators – 1–16 AWS Inferentia chips
- Cores – 4–64 NeuronCores
- Up to 192 GiB of Memory
- Up to 100 Gbps networking bandwidth

AWS Neuron SDK enabled frameworks TensorFlow, MXNet, PyTorch available on

- AWS Deep Learning Containers
- AWS Deep Learning AMIs (DLAMI)
- Custom install with binary
- Amazon SageMaker hosting

Single Inferentia chip instance

- inf1.xlarge
- inf1.2xlarge

Multi-Inferentia chip Instances

- inf1.6xlarge (4 chips)
- inf1.24xlarge (16 chips)

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

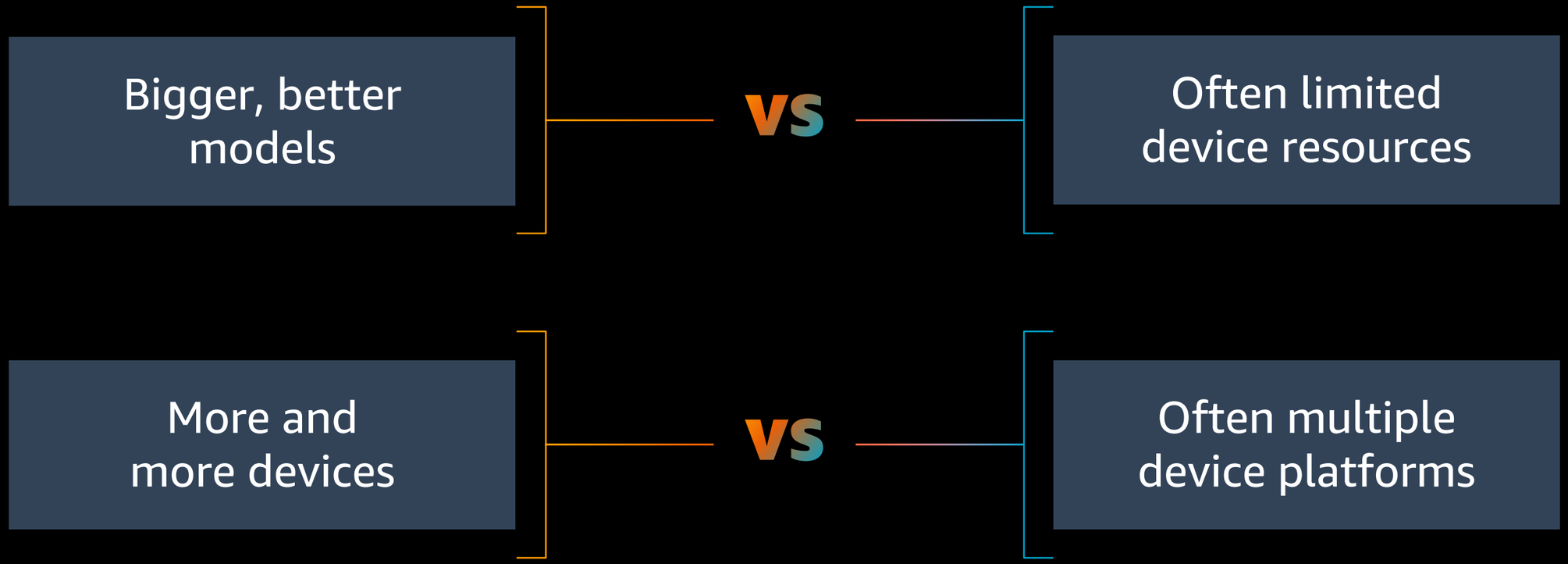
Instance size	vCPUs	Inferentia Chips	Number of NeuronCores
inf1.xlarge	4	1	4
inf1.2xlarge	8	1	4
inf1.6xlarge	24	4	16
inf1.24xlarge	96	16	64

Start small, and scale up if you need more compute

ML inference on the edge

aws INNOVATE
ONLINE CONFERENCE

Deploying models at the edge



NEW

Amazon SageMaker Edge Manager

Model management for edge devices

GENERALLY AVAILABLE

Improves performance by up to **25x**

Easy **integration** with device applications

Continuous model **monitoring**

Run **multiple** models on each device

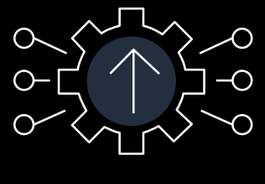
https://aws.amazon.com/sagemaker/edge-manager/

How Amazon SageMaker Edge Manager works

Amazon SageMaker

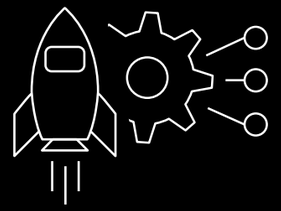
Compile

Optimize model for inference with Amazon SageMaker Neo



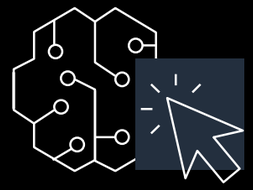
Package

Package and sign models with Amazon SageMaker Edge Manager



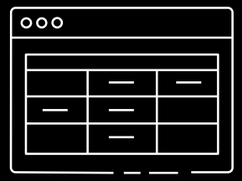
Deploy

Deploy model package with AWS IoT Greengrass or other deployment mechanism



Run

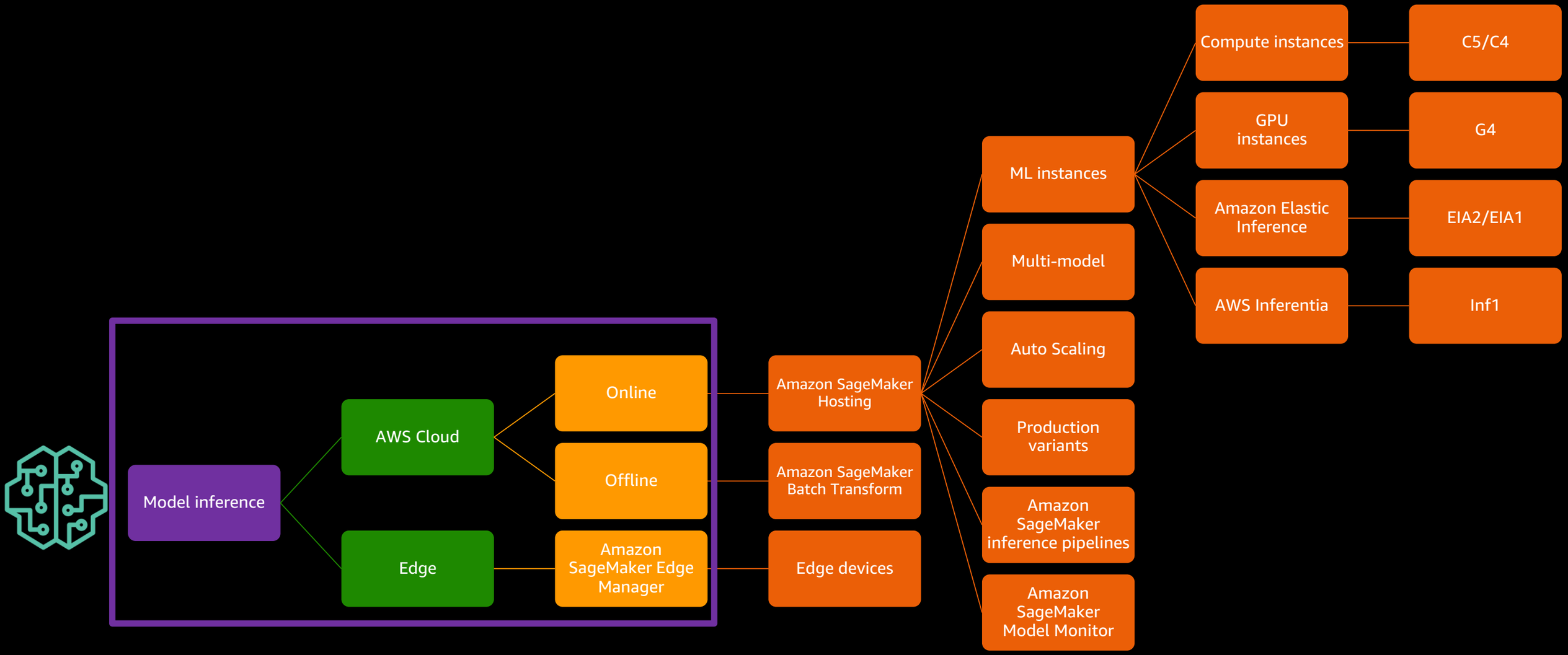
Run models on Amazon SageMaker Edge Manager agent



Maintain

Monitor each model on each device using Amazon SageMaker Edge Manager

Recap



Recap

Model inference

AWS Cloud

Edge

Online

Offline

Amazon SageMaker Edge Manager

Amazon SageMaker Hosting

Amazon SageMaker Batch Transform

Edge devices

ML instances

Multi-model

Auto-Scaling

Production variants

Amazon SageMaker inference pipelines

Amazon SageMaker Model Monitor

Compute instances

C5/C4

GPU instances

G4

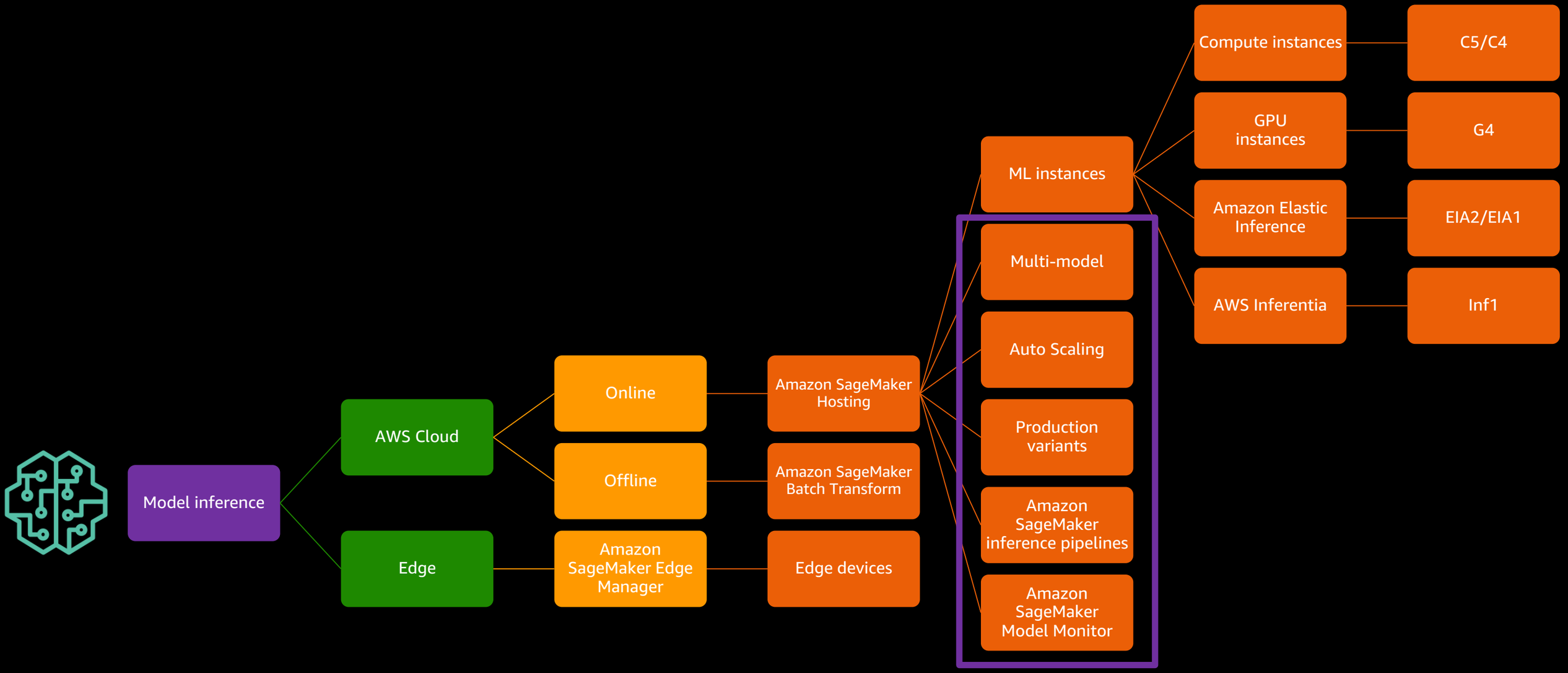
Amazon Elastic Inference

EIA2/EIA1

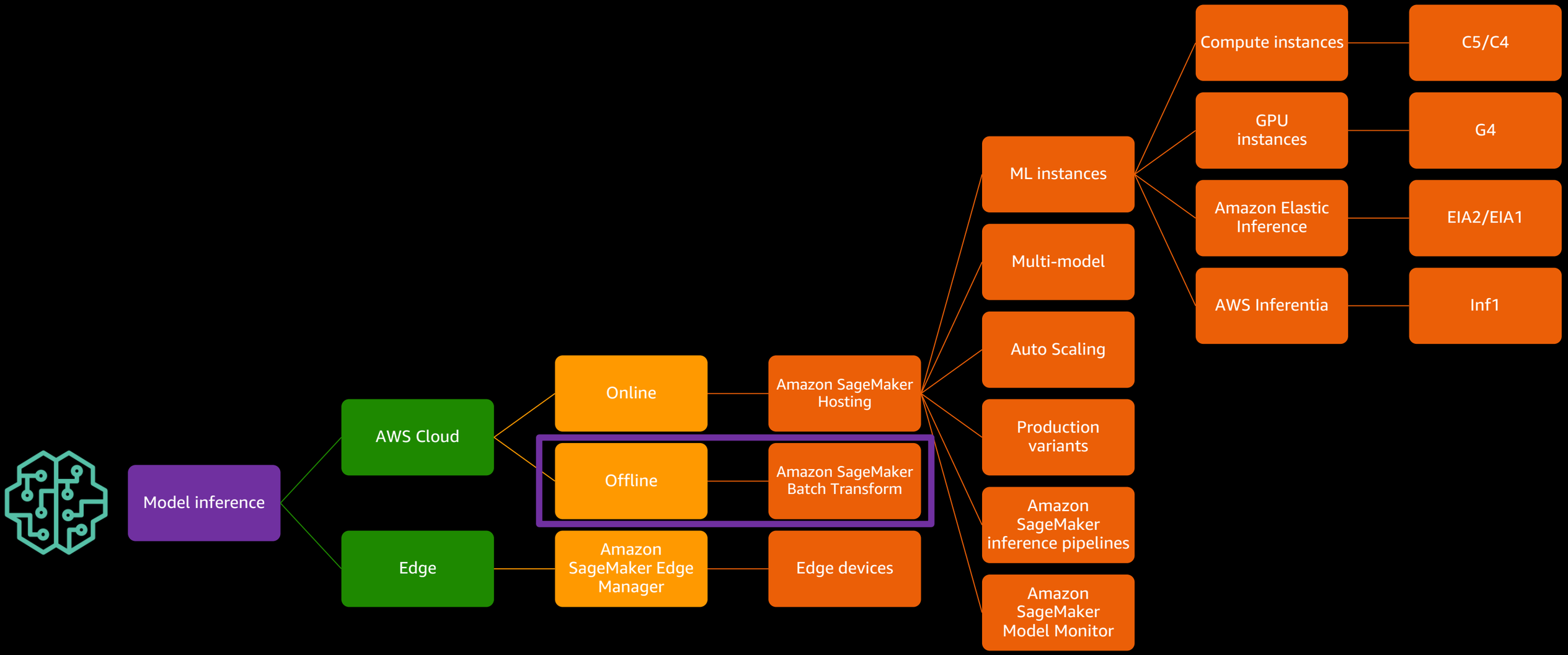
AWS Inferentia

Inf1

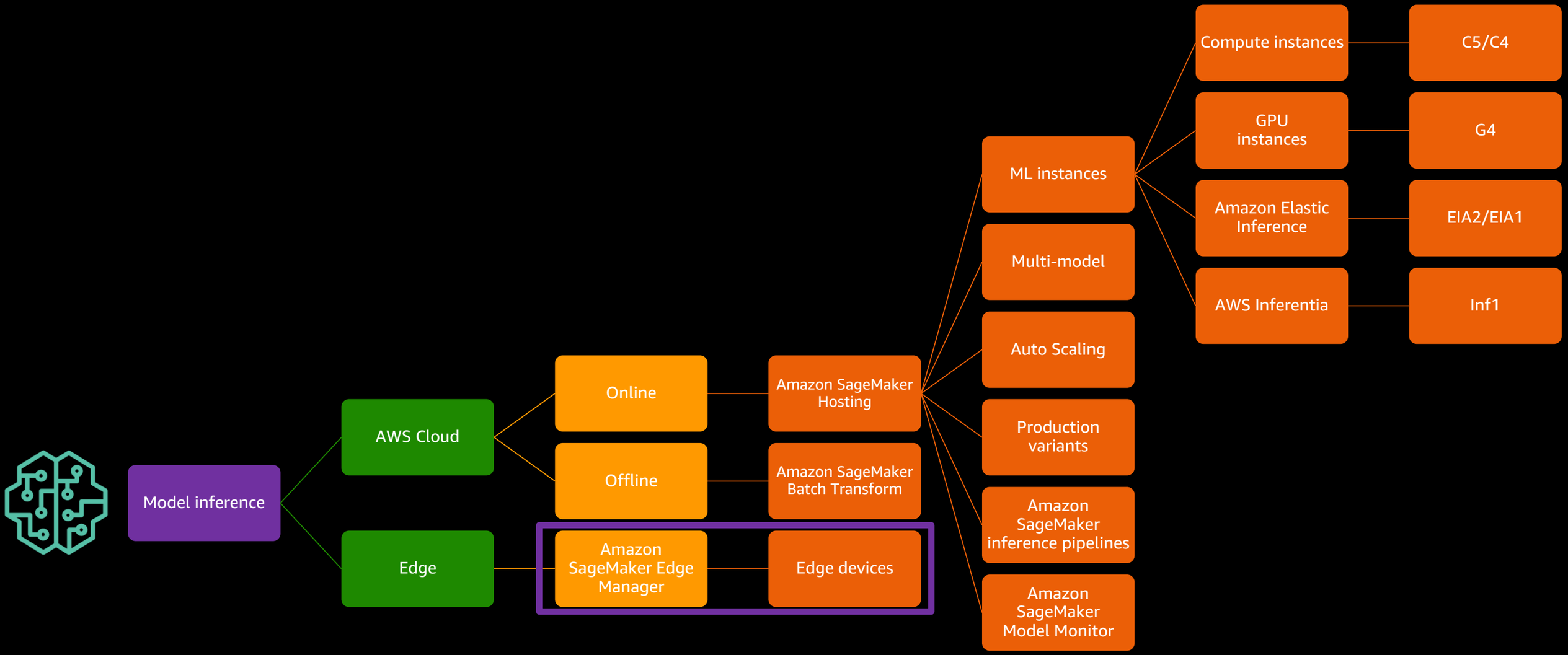
Recap



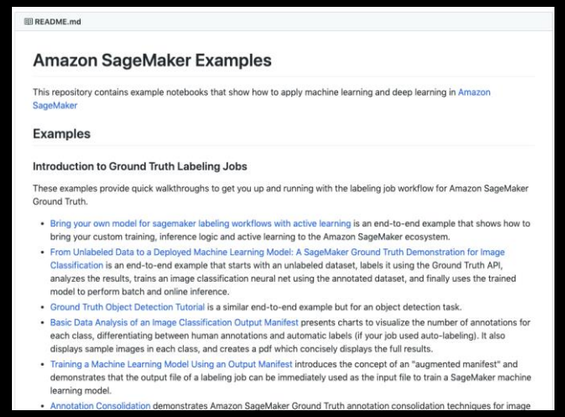
Recap



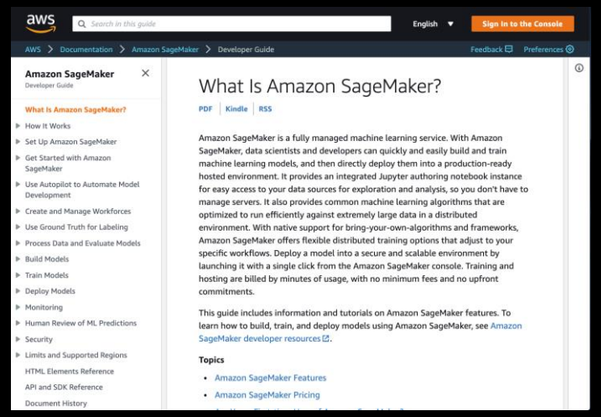
Recap



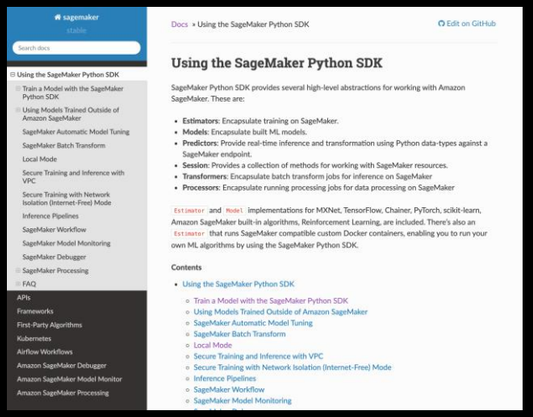
Resources



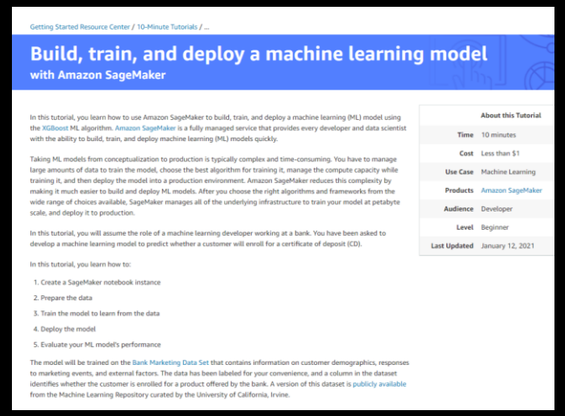
https://github.com/awslabs/amazon-sagemaker-examples



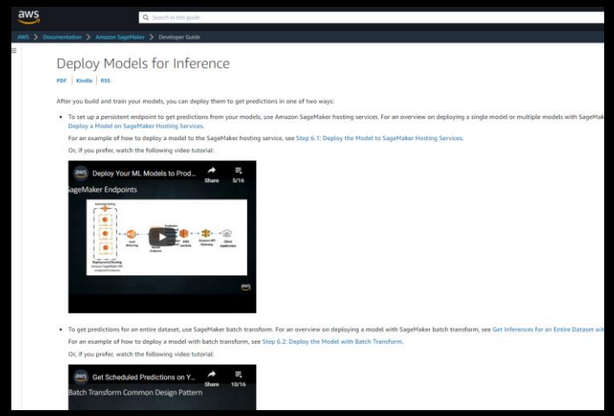
https://docs.aws.amazon.com/sagemaker/latest/dg/whatis.html



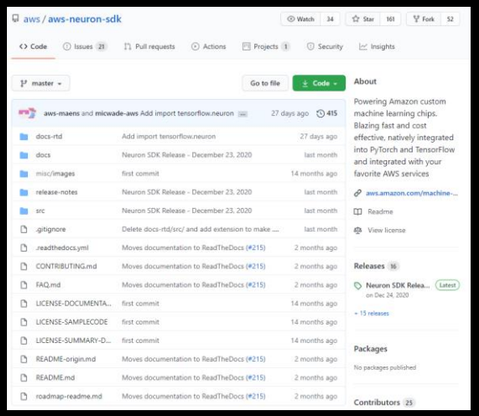
https://sagemaker.readthedocs.io/en/stable/overview.html



https://aws.amazon.com/getting-started/hands-on/build-train-deploy-machine-learning-model-sagemaker/



https://docs.aws.amazon.com/sagemaker/latest/dg/deploy-model.html



https://github.com/aws/aws-neuron-sdk

Visit the AI and Machine Learning Resource Hub for more resources

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

- The machine learning journey e-book
- Machine learning enterprise guide
- 7 leading machine learning use cases e-book
- A strategic playbook for data, analytics, and machine learning
- Accelerating ML innovation through security e-book
- … and more!

Visit resource hub »

https://tinyurl.com/aiml-aws

AWS Machine Learning (ML) Training and Certification

Learn like an Amazonian, based on the curriculum we've used to train our own developers and data scientists

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.

Flexibility to learn your way

Learn online with 65+ on-demand digital courses or live with virtual instructor-led training, plus hands-on labs and opportunities for practical application.

Validate your expertise

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

aws.training/machinelearning

Thank You for Attending AWS Innovate

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-apac-marketing@amazon.com


twitter.com/AWSCloud


facebook.com/AmazonWebServices


youtube.com/user/AmazonWebServices


slideshare.net/AmazonWebServices


twitch.tv/aws

aws INNOVATE
AI/ML EDITION

Thank you!

Eshaan Anand
Senior Partner Solutions Architect
Amazon Web Services