



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



Collect and prepare training data

Built-in, high performance algorithms



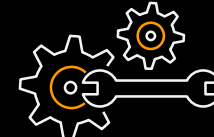
Choose and optimize your ML algorithm

One-click training



Set up and manage environments for training

Optimization



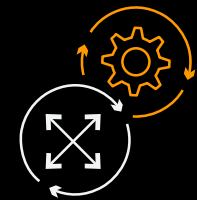
Train and tune model (trial and error)

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



Amazon SageMaker

Bringing machine learning to all developers

Pre-built notebooks for common problems



Collect and prepare training data

Built-in, high performance algorithms



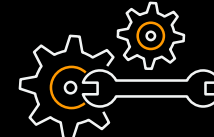
Choose and optimize your ML algorithm

One-click training



Set up and manage environments for training

Optimization



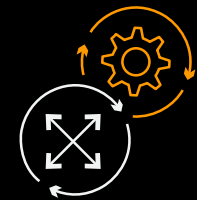
Train and tune model (trial and error)

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.)

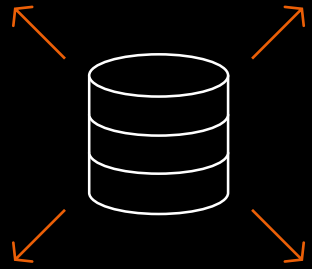
- **Compute**

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

- **Security**

Security-related parameters and configurations

Key capabilities - Amazon SageMaker Model Deployment



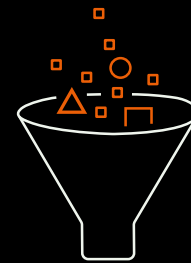
Auto-scaling

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



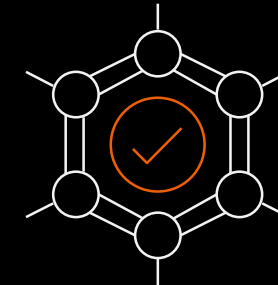
Production variants

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

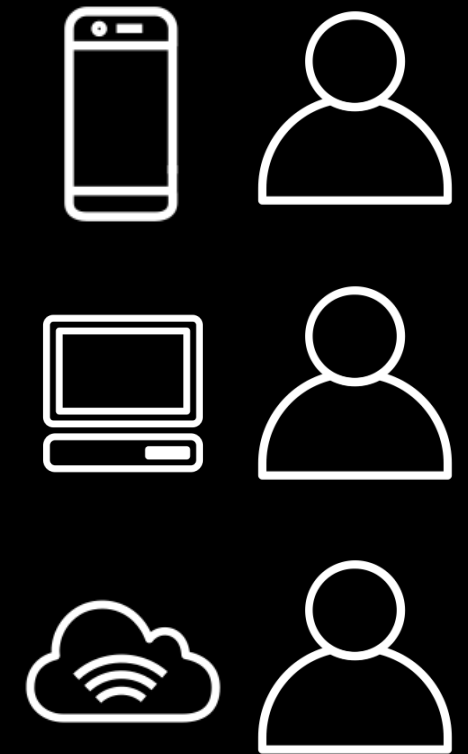
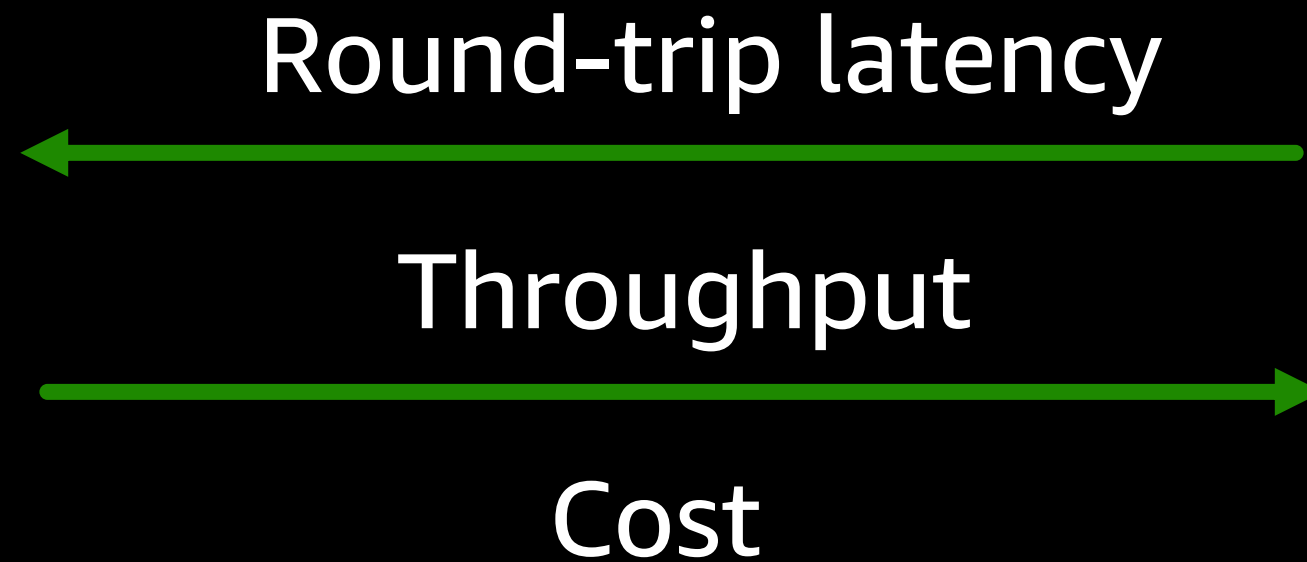
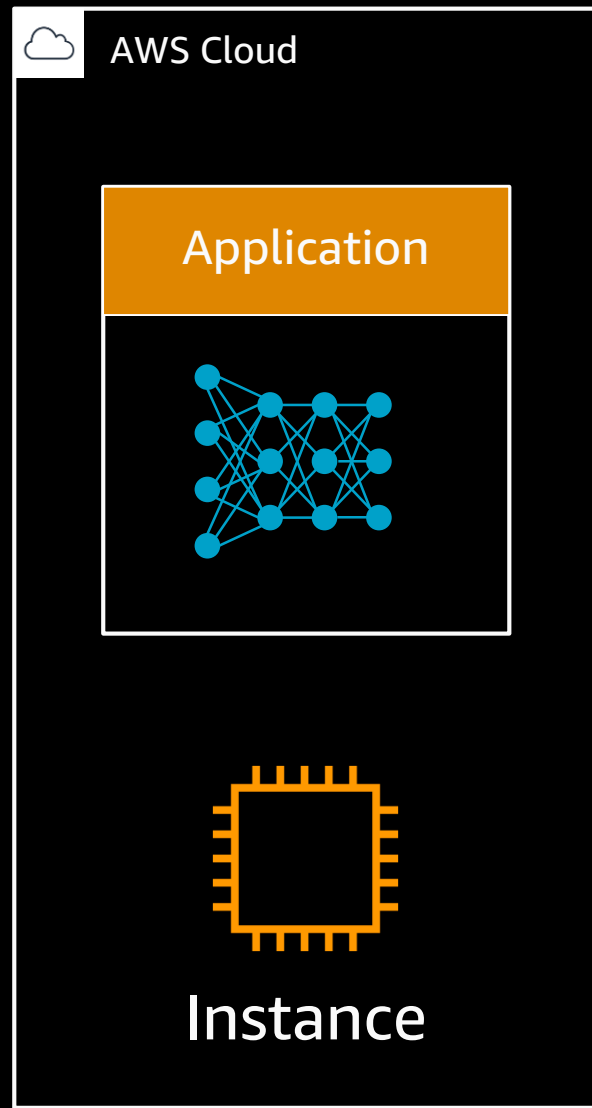


Model monitor

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



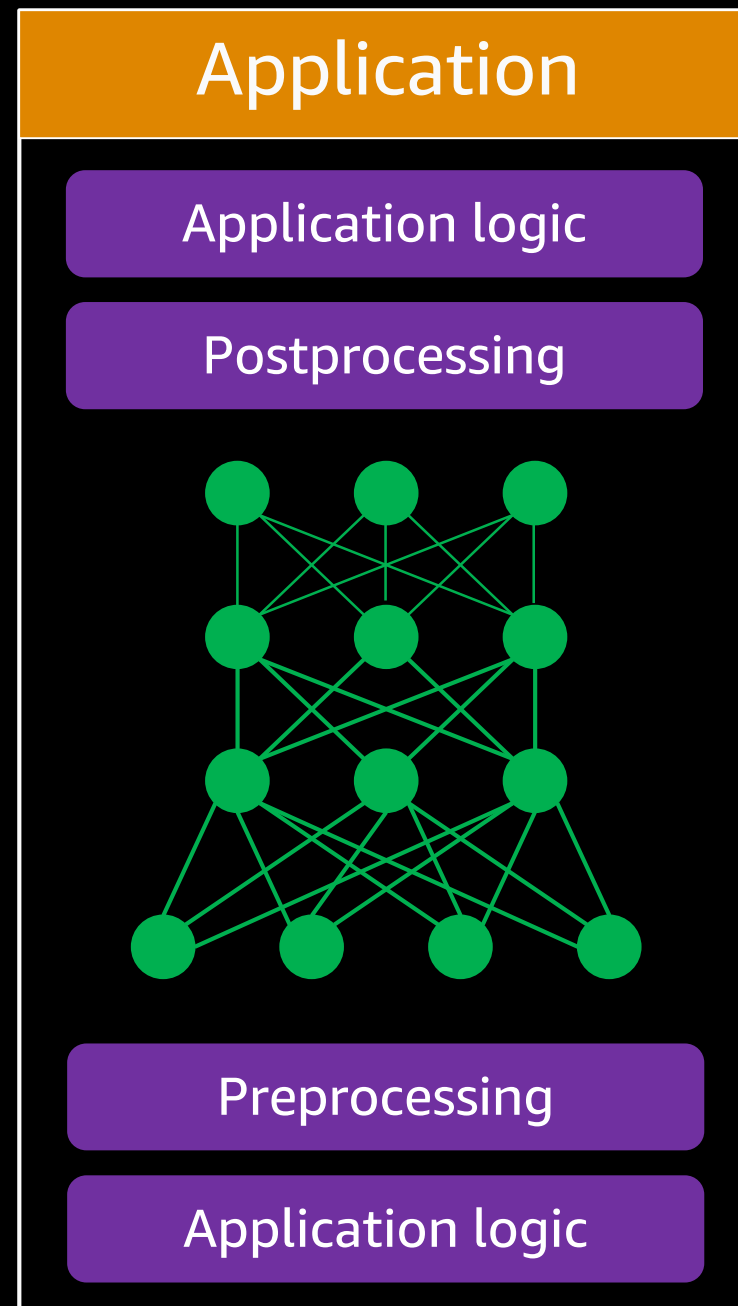
Inference Performance affects Customer Experience



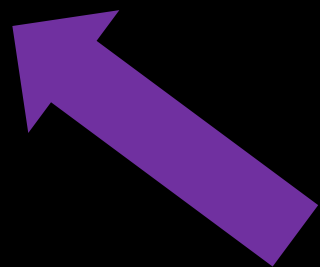
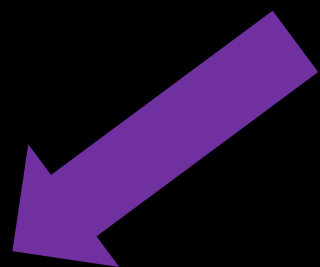
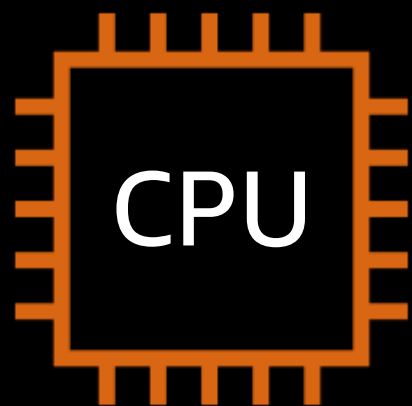
-
-
-



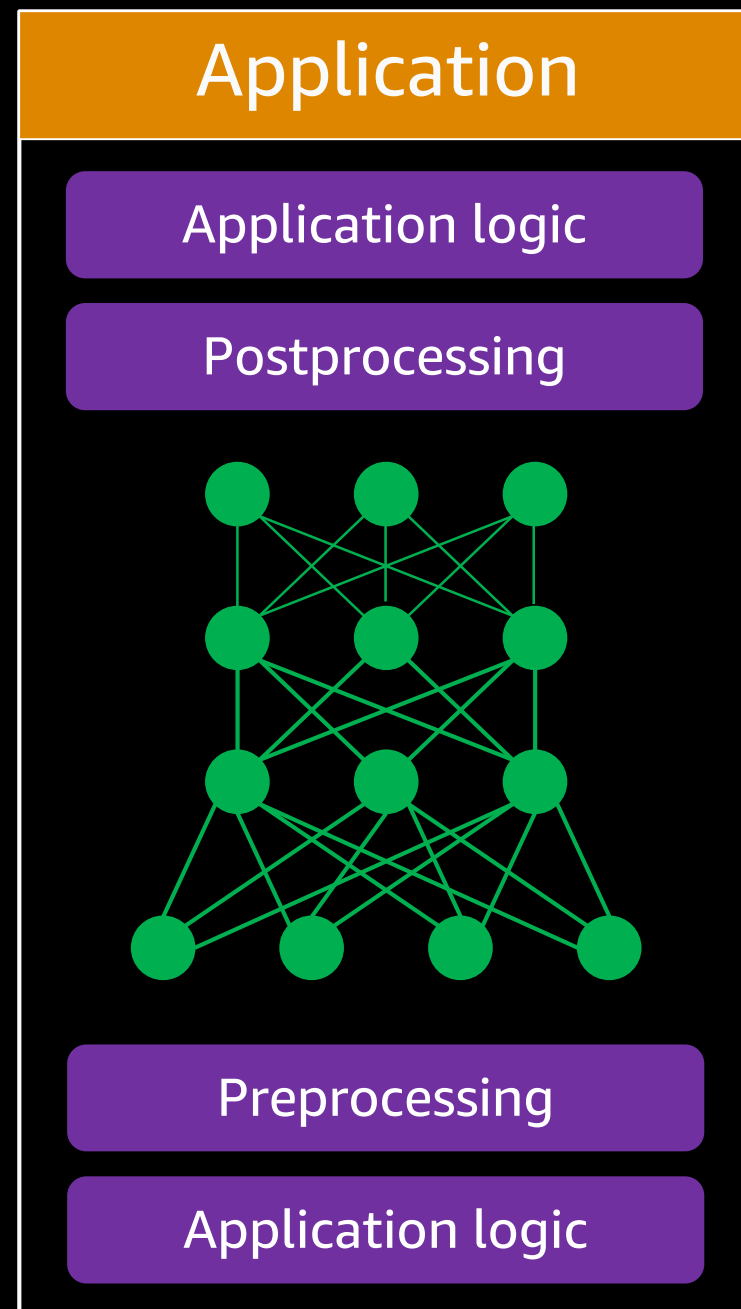
Speed up inference with an accelerator



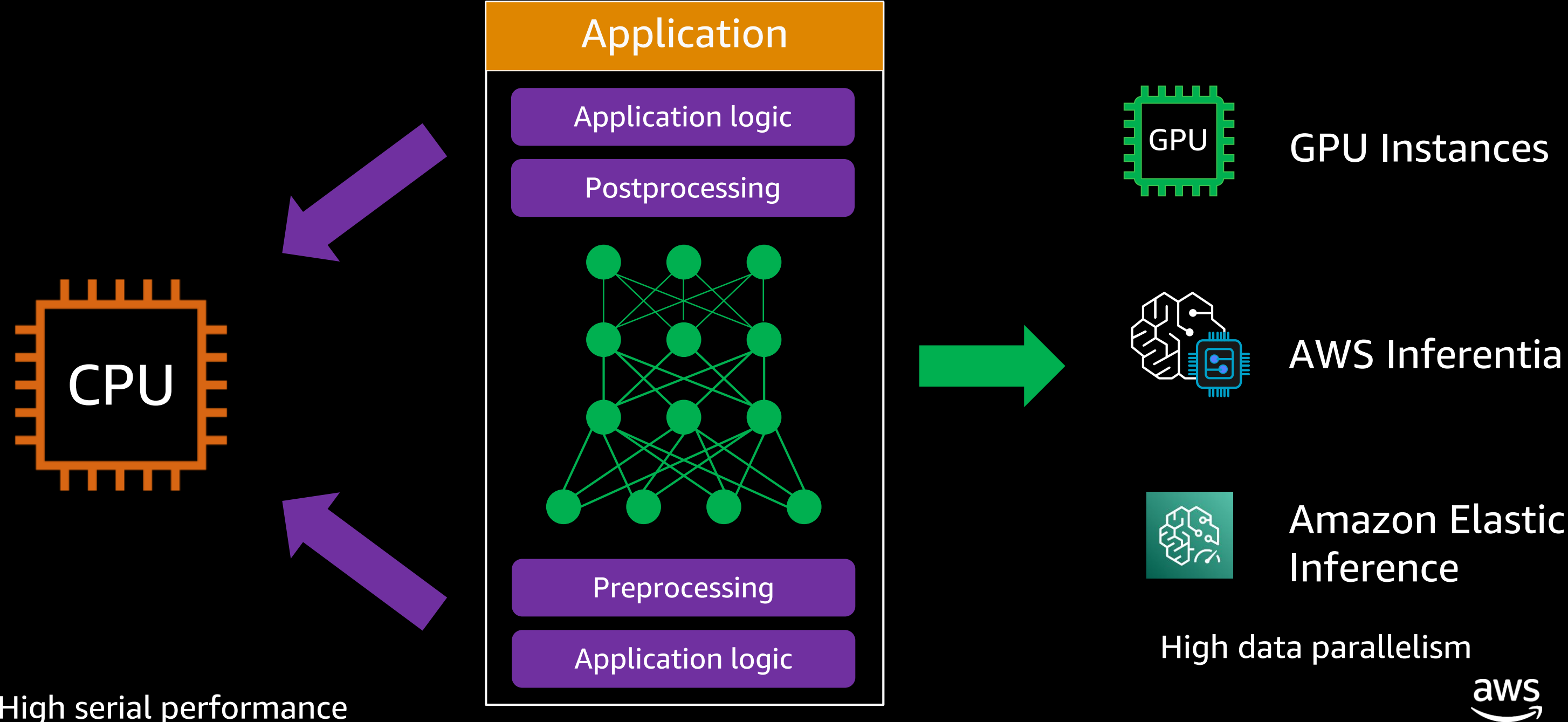
Speed up inference with an accelerator



High serial performance

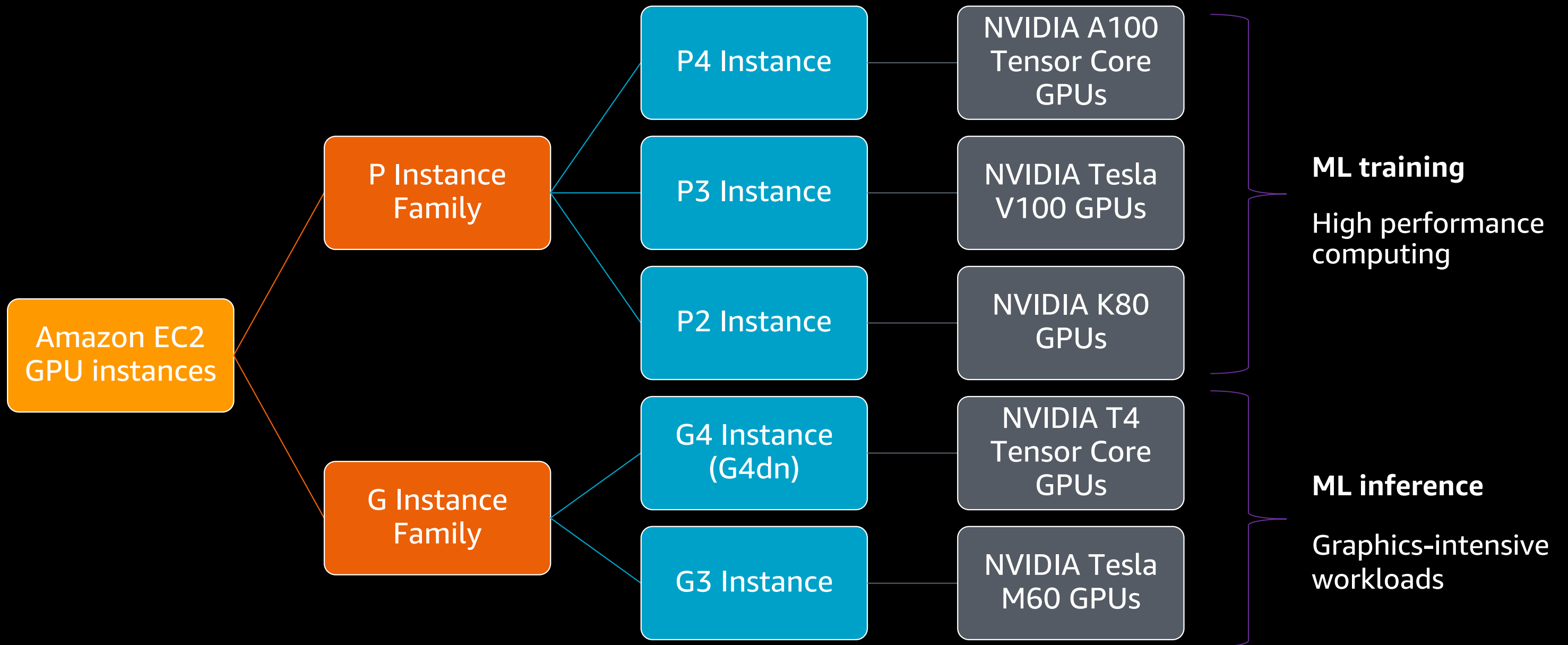


Speed up inference with an accelerator



GPU accelerated EC2 instances

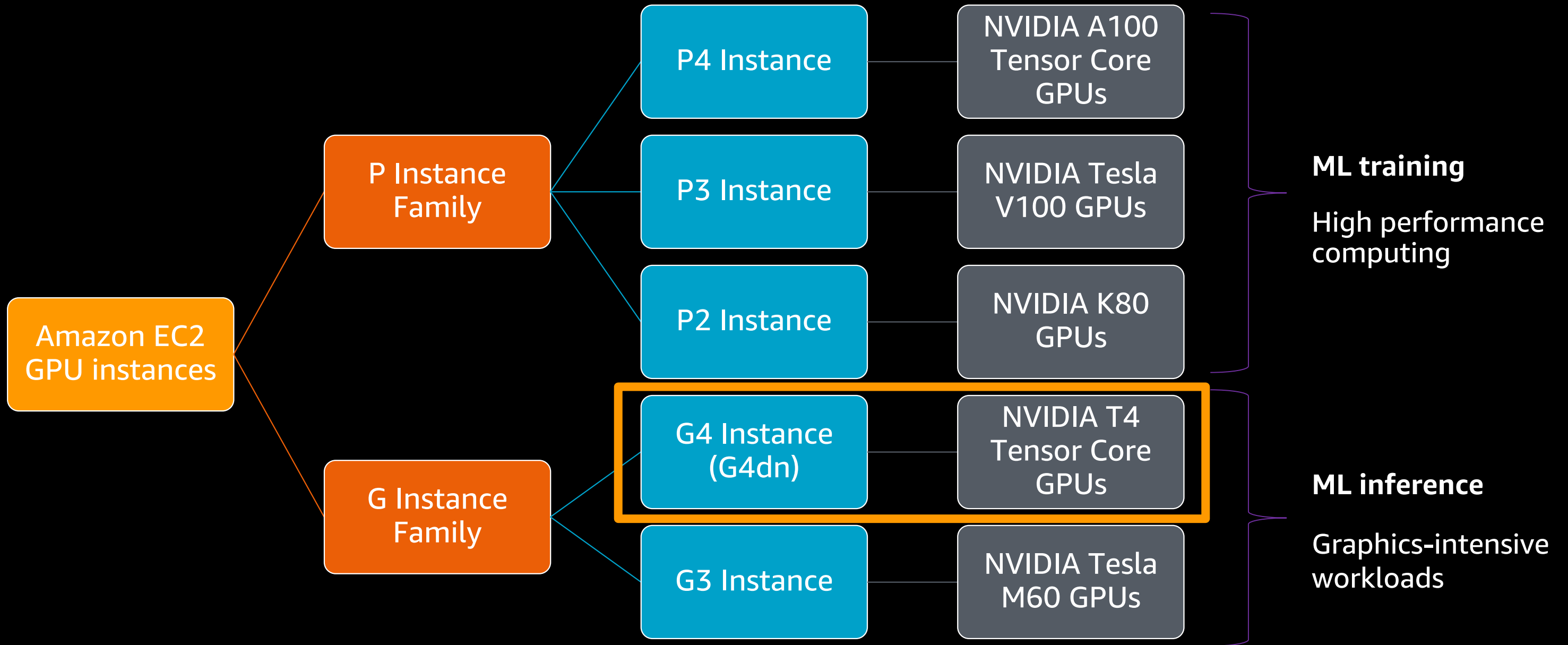
Amazon EC2 GPU instances for deep learning



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

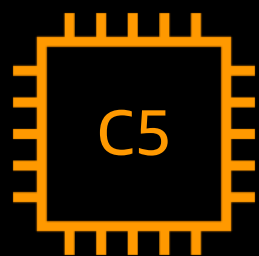
	g4dn.12xlarge	g4dn.metal
GPUs	4 x T4	8 x T4
vCPUs	4	8
System Mem (GiB)	16	32

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?



Low cost /
inference for

- Small DL models
- Traditional ML models

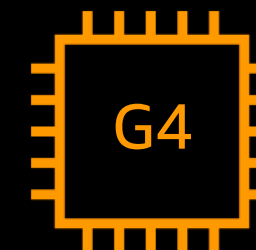
What about?

Mid-sized models

Need acceleration but not a
dedicated GPU

Lower throughput and higher
latency tolerance

Cost sensitive



Low cost /
inference for

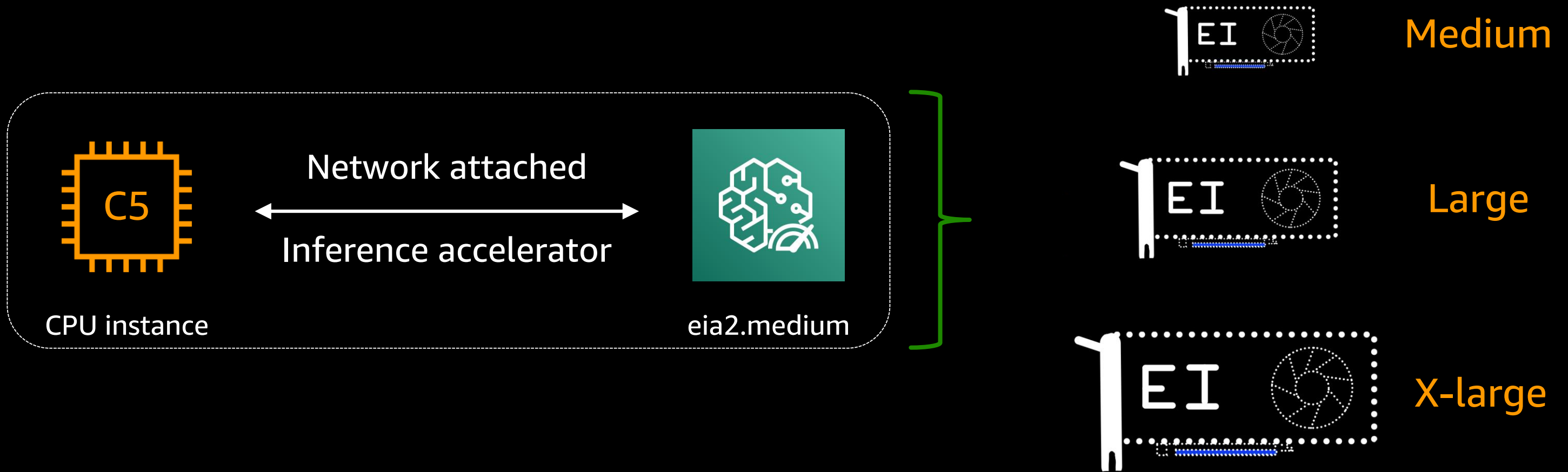
- Large DL models
- Large batch sizes
- High demand



Amazon Elastic Inference

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/>



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

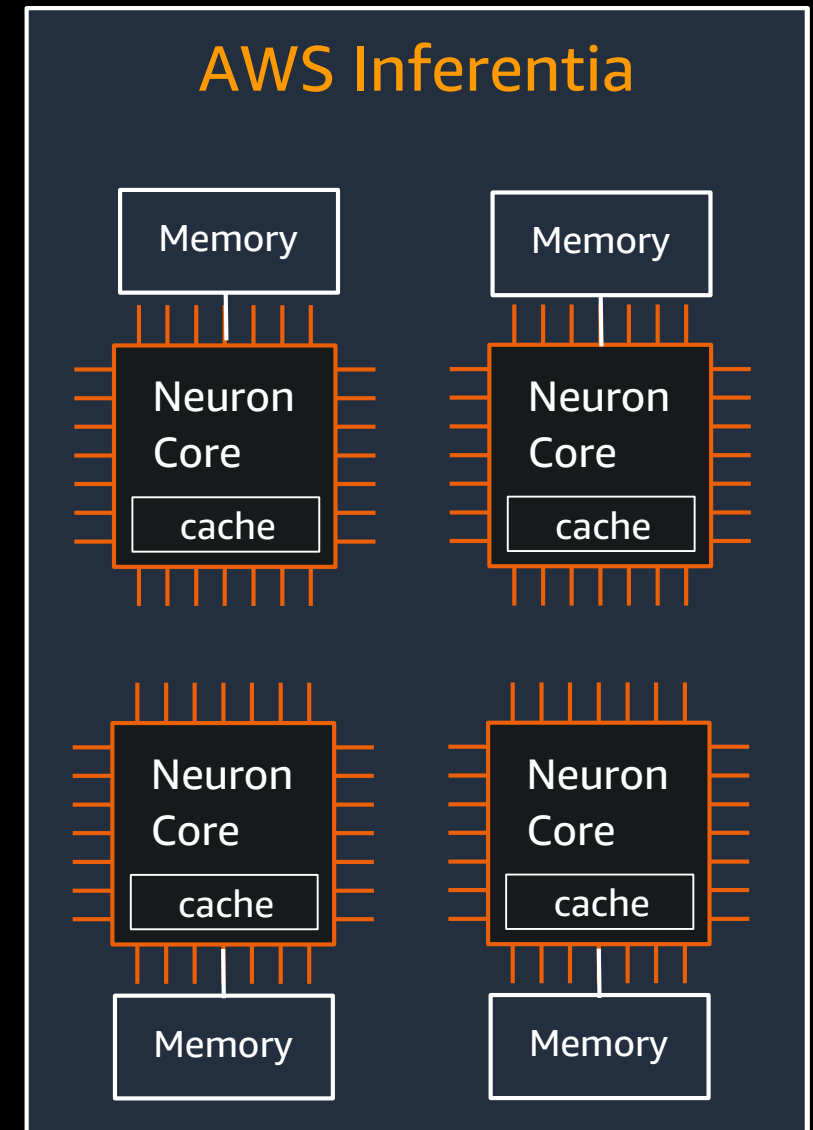
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

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



Deploying models at the edge



NEW

Amazon SageMaker Edge Manager

Model management
for edge devices

GENERALLY AVAILABLE

Improves performance by up to **25x**

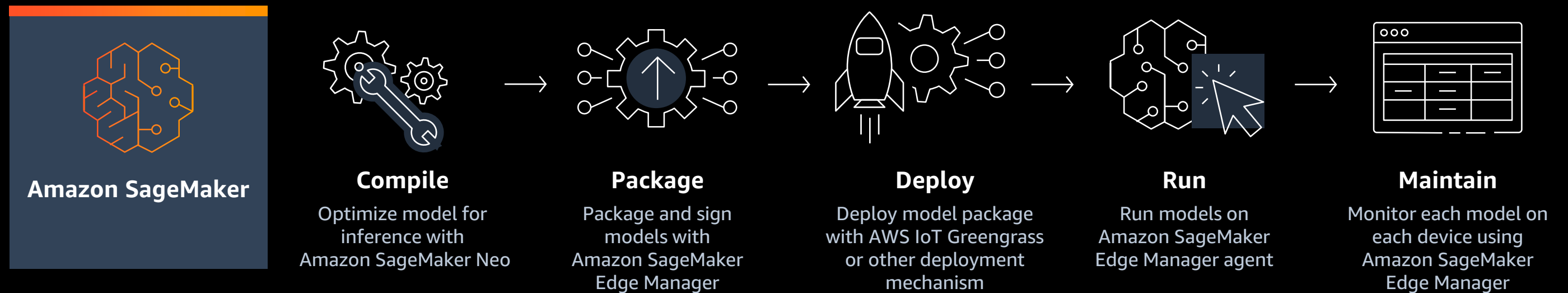
Easy **integration** with device applications

Continuous model **monitoring**

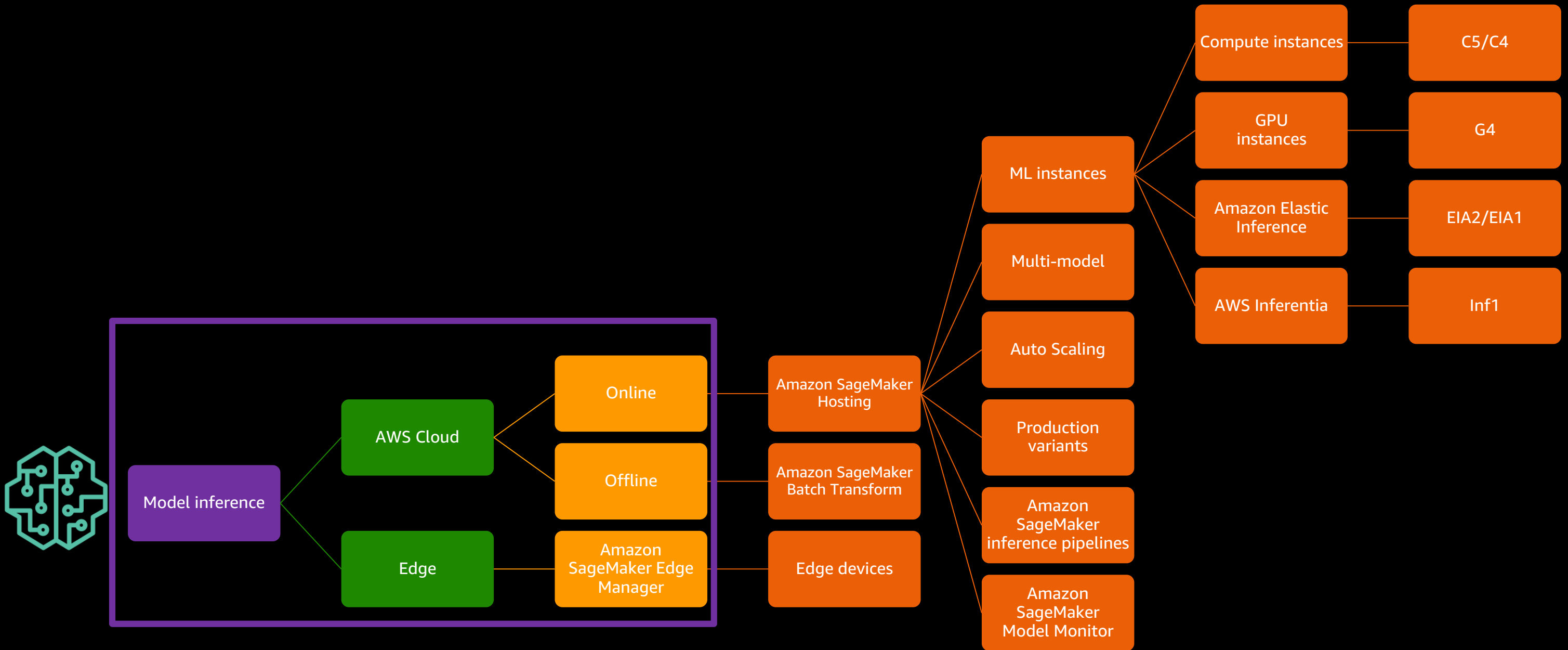
Run **multiple** models on each device

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

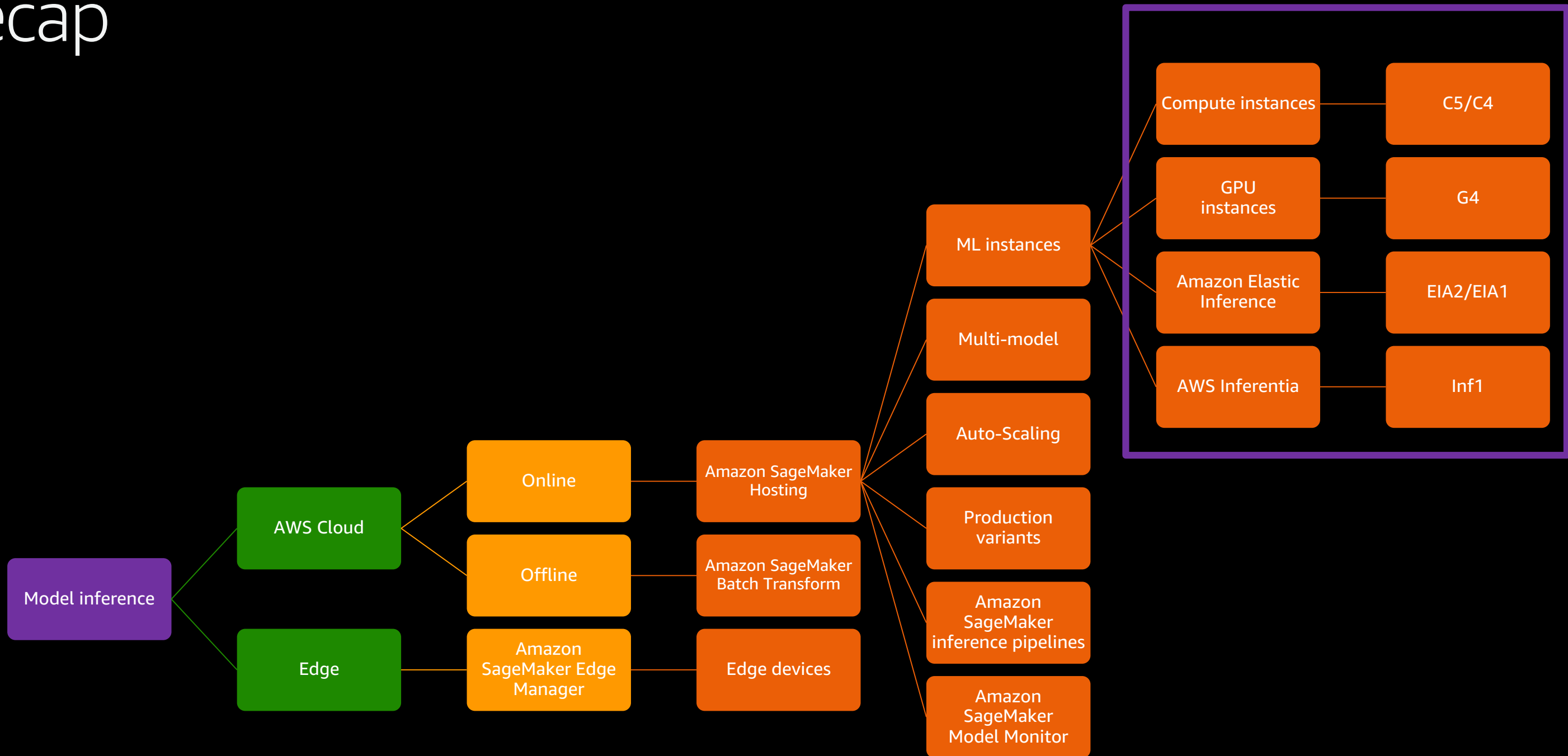
How Amazon SageMaker Edge Manager works



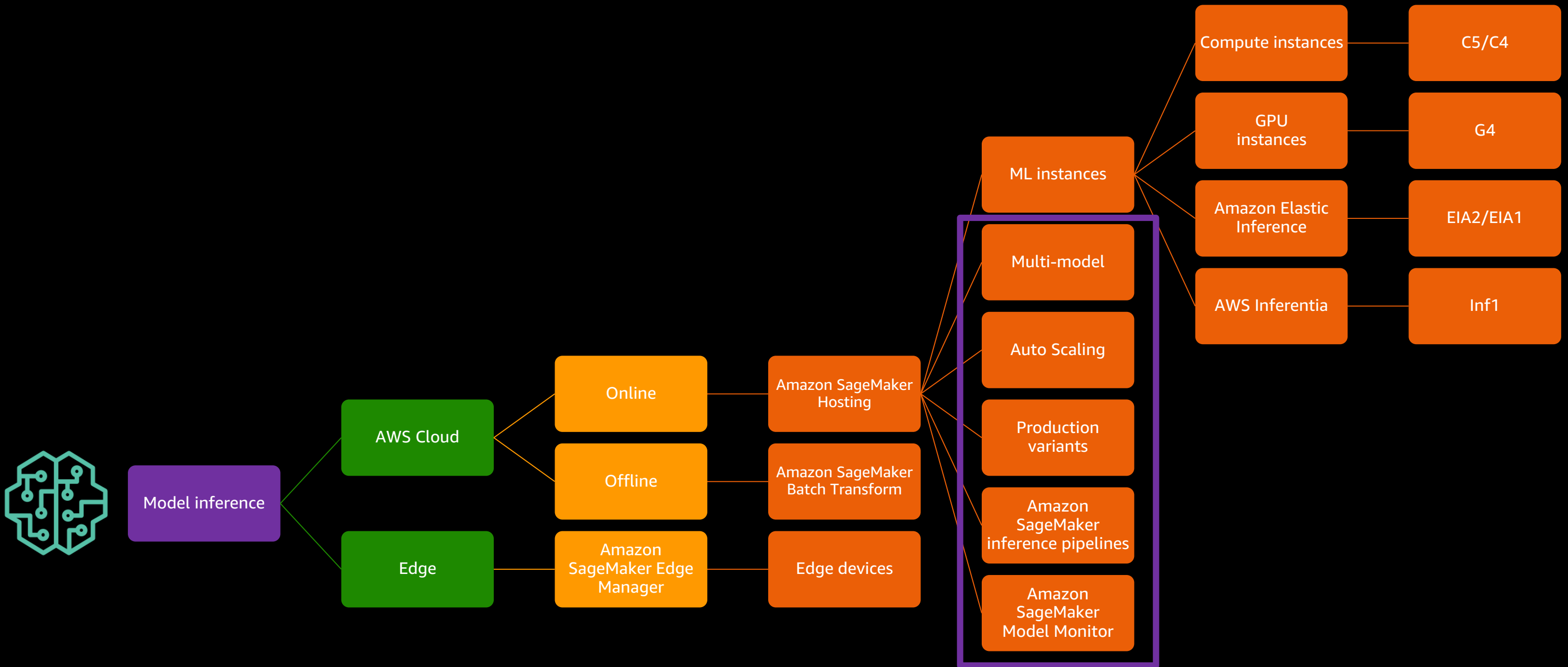
Recap



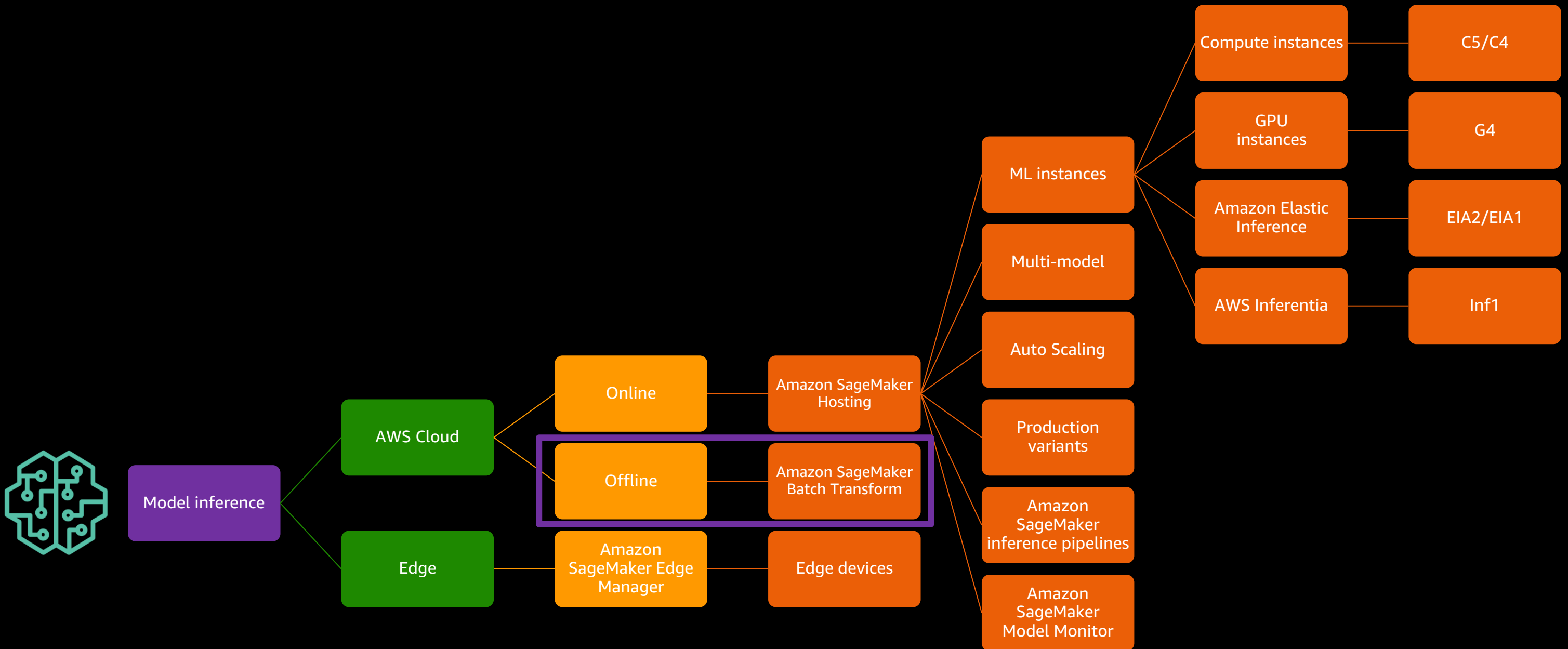
Recap



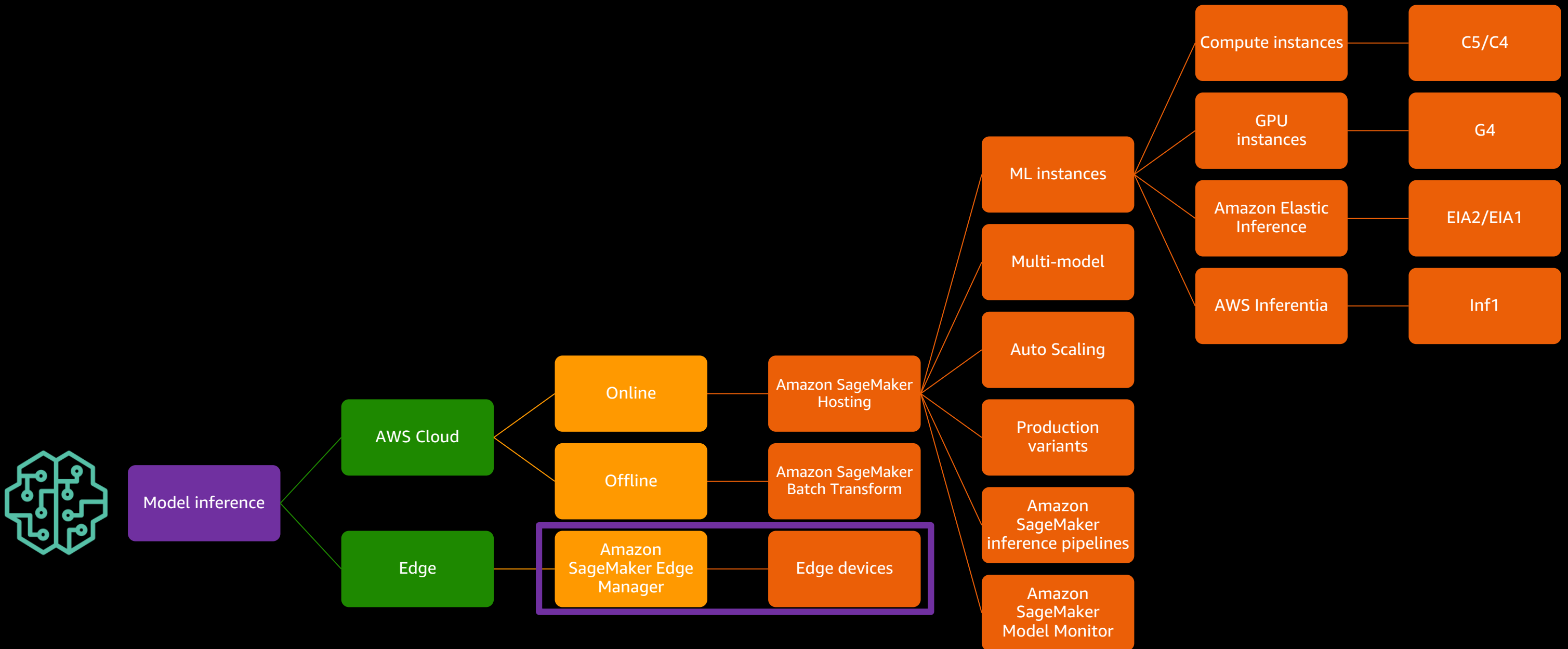
Recap



Recap



Recap



Resources

The screenshot shows the 'Amazon SageMaker Examples' README page. It includes a search bar, a list of examples, and an introduction to Ground Truth Labeling Jobs. The introduction states that these examples provide quick walkthroughs to get you up and running with the labeling job workflow for Amazon SageMaker Ground Truth. A list of examples follows, including 'Bring your own model for sagemaker labeling workflows with active learning', 'From Unlabeled Data to a Deployed Machine Learning Model: A SageMaker Ground Truth Demonstration for Image Classification', 'Ground Truth Object Detection Tutorial', 'Basic Data Analysis of an Image Classification Output Manifest', and 'Annotation Consolidation'.

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

The screenshot shows the 'What Is Amazon SageMaker?' page from the AWS Developer Guide. It defines Amazon SageMaker as a fully managed machine learning service. The text explains that data scientists and developers can quickly and easily build and train machine learning models, and then directly deploy them into a production-ready hosted environment. It also mentions that SageMaker provides an integrated Jupyter authoring notebook instance for easy access to your data sources for exploration and analysis. The page includes a table of contents, a list of topics (Amazon SageMaker Features, Amazon SageMaker Pricing), and a search bar.

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

The screenshot shows the 'Using the SageMaker Python SDK' page from the AWS documentation. It lists several high-level abstractions for working with Amazon SageMaker: Estimators, Models, Predictors, Session, and Transformers. It also provides a list of contents, including 'Using the SageMaker Python SDK', 'Train a Model with the SageMaker Python SDK', 'Using Models Trained Outside of Amazon SageMaker', 'SageMaker Automatic Model Tuning', 'SageMaker Batch Transform', 'Local Mode', 'Secure Training and Inference with VPC', 'Secure Training with Network Isolation (Internet-Free) Mode', 'Inference Pipelines', 'SageMaker Workflow', 'SageMaker Model Monitoring', 'SageMaker Debugger', 'SageMaker Processing', and 'FAQ'.

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

The screenshot shows the 'Build, train, and deploy a machine learning model with Amazon SageMaker' tutorial page. It includes a 'Getting Started Resource Center / 10-Minute Tutorial / ...' header, a 'Build, train, and deploy a machine learning model with Amazon SageMaker' title, and a table of contents. The table of contents lists 'About this Tutorial', 'Time', 'Cost', 'Use Case', 'Products', 'Audience', 'Level', and 'Last Updated'. The 'About this Tutorial' section states that the tutorial teaches how to use Amazon SageMaker to build, train, and deploy a machine learning (ML) model using the XGBoost ML algorithm. It also mentions that SageMaker is a fully managed service that provides every developer and data scientist with the ability to build, train, and deploy machine learning (ML) models quickly.

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

The screenshot shows the 'Deploy Models for Inference' page from the SageMaker Developer Guide. It explains that after building and training your models, you can deploy them to get predictions in one of two ways: to set up a persistent endpoint to get predictions from your models, or to use Amazon SageMaker hosting services. It includes a video tutorial titled 'Deploy Your ML Models to Prod...' and a diagram showing the SageMaker endpoints architecture. The diagram illustrates the flow from training to deployment and inference.

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

The screenshot shows the GitHub repository for 'aws/aws-neuron-sdk'. It includes a search bar, a list of issues, and a table of repository files. The table lists files such as 'docs-rtf', 'docs', 'misc/images', 'release-notes', 'src', '.gitignore', '.readthedocs.yml', 'CONTRIBUTING.md', 'FAQ.md', 'LICENSE-DOCUMENTA...', 'LICENSE-SAMPLECODE', 'LICENSE-SUMMARY-D...', 'README-origi...', 'README.md', and 'readmap-readme.md'. It also shows the repository's 'About' section, which describes it as a repository for Amazon custom machine learning chips, and lists the repository's statistics, including 34 watchers, 101 stars, and 52 forks.

<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



Thank you!

Eshaan Anand
Senior Partner Solutions Architect
Amazon Web Services