AWS

Deploy ML models for inference at high performance and low cost

State of the Art: ML Models in Production

Hyper-personalization

Recommendation systems built with “one model per customer” patterns

Generative AI

Multi-modal applications

ML applications built with multiple types of Models

Agility

Update ML models in minutes, not weeks

Deploy models to serve inference

Amazon SageMaker

SAGEMAKER STUDIO IDE

Real-time inference | Async inference | Serverless inference | Batch inference | Multi-model endpoints | Multi-container endpoints | Inference pipelines | Manage and version models | MLOps | Model monitoring | Shadow Testing | Metrics and logging in CloudWatch

SageMaker JumpStart

CONTAINERS

TensorFlow | PyTorch | mxnet | Keras | scikit learn | GLUON | ONNX | DJL DeepJavaLibrary | LMI | NVIDIA TRITON INFERENCE SERVER | BYOC

ML COMPUTE INSTANCES & ACCELERATORS

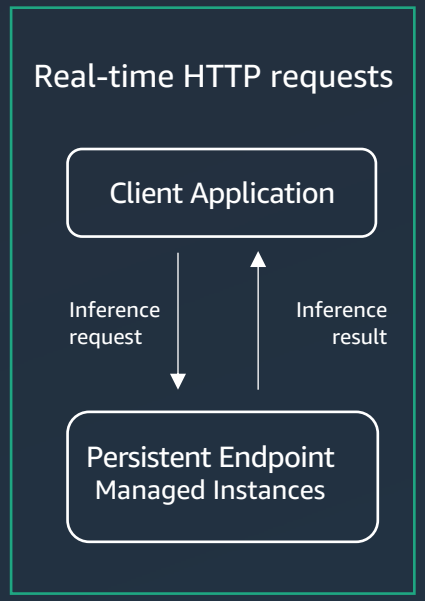
CPUs | GPUs | Inferentia | Graviton (ARM)

DEEP LEARNING COMPILERS AND RUNTIMES

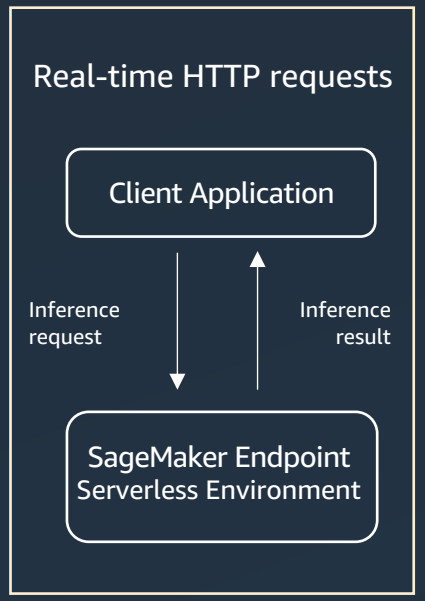
SageMaker Neo | NVIDIA TensorRT/cuDNN | Intel oneDNN | ARM Compute Library

Amazon SageMaker deployment modes

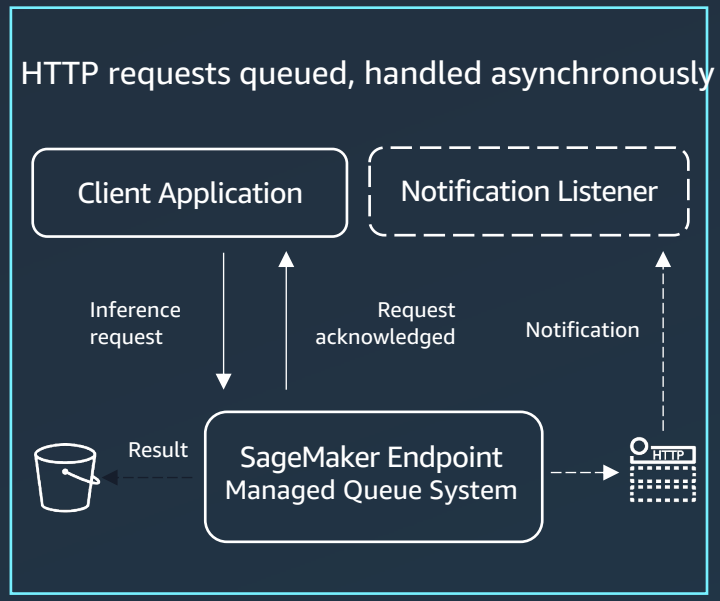
Real-time Inference



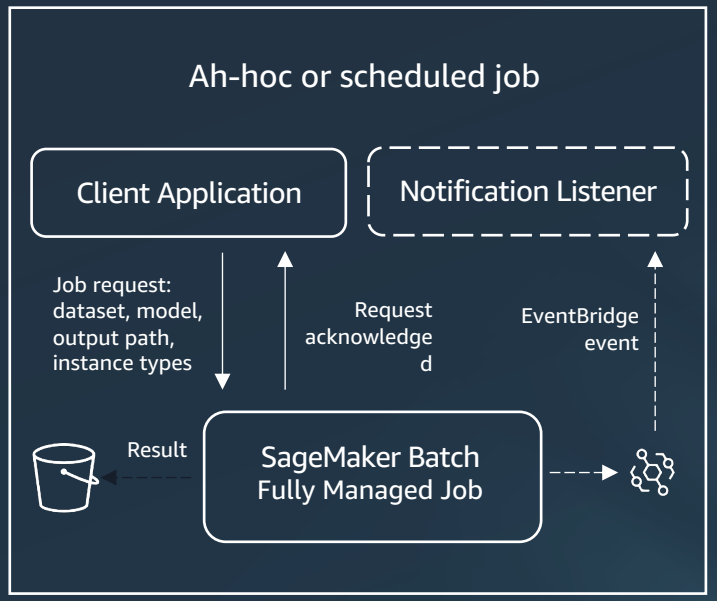
Serverless Inference



Asynchronous Inference



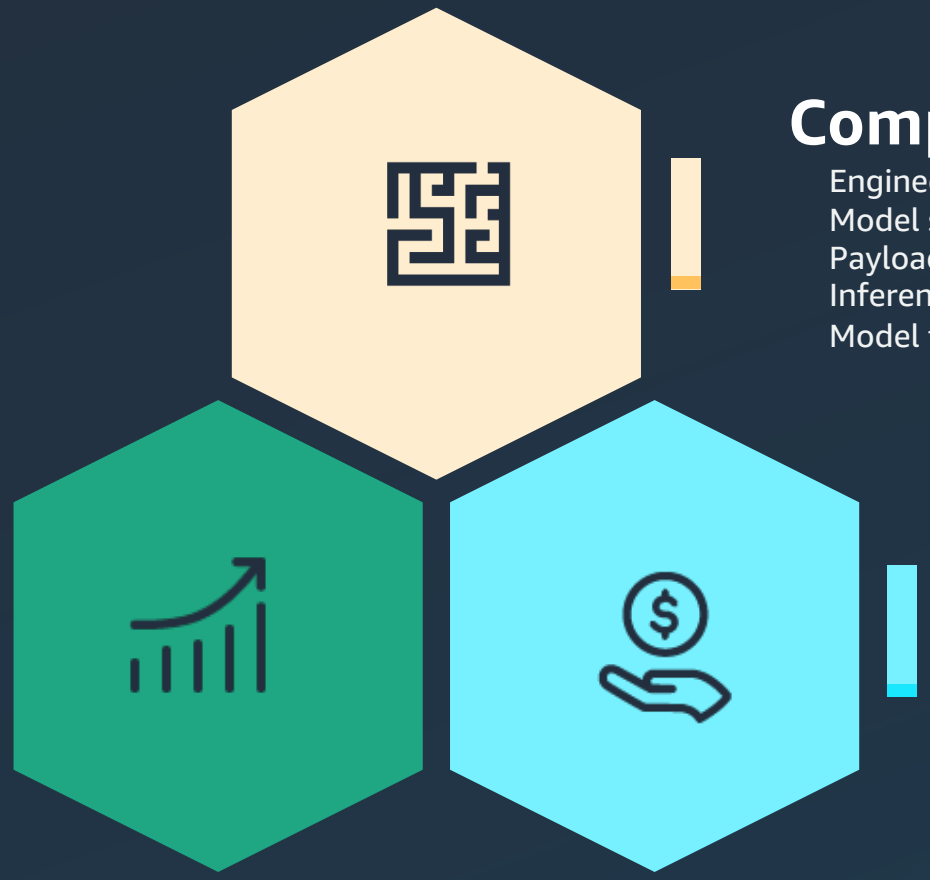
Batch Transform



Example use cases and technical considerations

Ad serving, search, personalized recommendations, Generative AI	Extract data from documents, form processing, chatbots, model dev/test	Video processing, large image processing, decoupled applications and systems	Business forecasting, propensity modeling, churn prediction, predictive maintenance

Model deployment – How do you strike the right balance?



Complexity

Engineering effort
Model size
Payload size
Inference workflows
Model testing and upgrade

Performance

Latency
Throughput
Availability

Cost

Model hosting cost
Operational overhead
Number of models to deploy and manage

Complexity

Deployment complexity

SageMaker JumpStart

- Easily access ML assets and quickly bring ML applications to market

Machine Learning Hub for SageMaker
Browse through ~400 contents including, built-in algorithms with pre-trained models, Gen AI Models, solution templates, and example notebooks

Pre-built inference scripts
Compatible with SageMaker

UI as well as API based machine learning
Use UI for single click model deployment or API for Python SDK based workflow

Notebook with examples
JumpStart lets you jump into notebook to use selected model with examples to guide customers through entire ML workflow

Share and collaborate within an organization
Share models and notebooks with others within your organization, and allow them to train with their own data or deploy as-is for inferencing

Deployment complexity

1) Choose Gen AI models offered by model providers

AI21 labs

LightOn

stability.ai

co:here

alexa

2) Try and deploy the model

a) Try out models directly on the AWS Console

b) Deploy to ML instances on SageMaker with one click

Data stays in your account – model, instances, logs, model inputs, model outputs

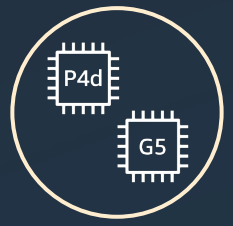
Fully integrated with the suite of SageMaker services and features

Deployment complexity

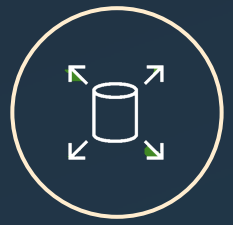
Large ML models

with 100 billion+ parameters

Easily parallelize models across multiple GPUs to fit models into the instance and achieve low latency



Deploy models on the most performant and cost-effective GPU-based instances or on AWS Inferentia



Leverage 500GB of Amazon EBS volume per endpoint

Deployment complexity

Large Model Inference container (LMI)

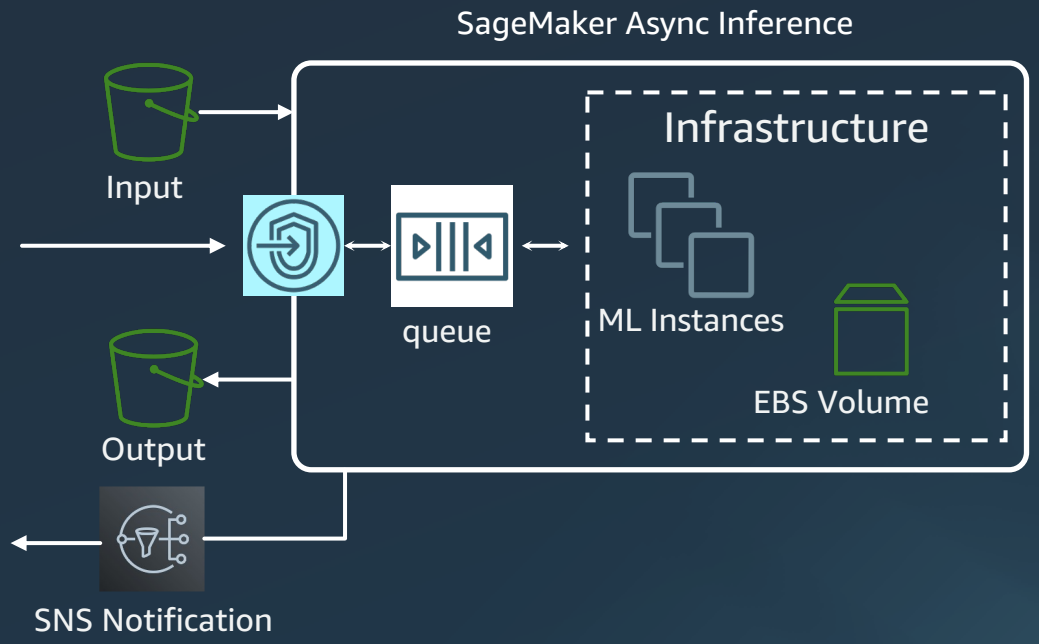
- Model sharding across accelerators
  (Model parallel inference)
- High performance model server
- BF16 support on DeepSpeed
- Faster model downloads from S3

  (4-mins to download 360-GB BLOOM-176B model)
- Supports DeepSpeed, Hugging Face Accelerate, and Stable Diffusion

  (5 lines model setup bundled with optimization strategy)

Deployment complexity

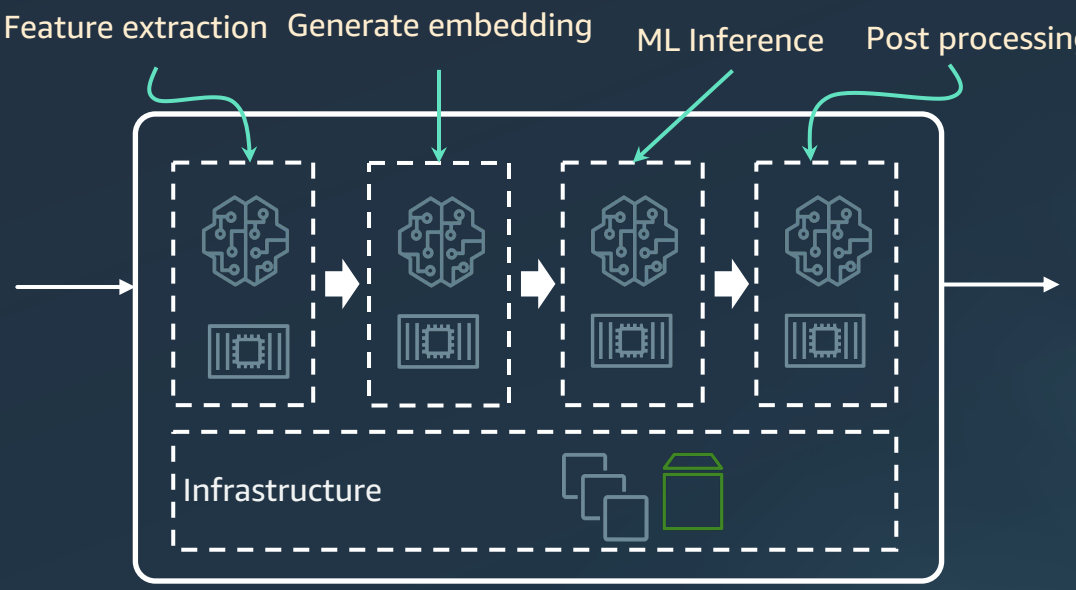
Large payloads or long running inference



Deployment complexity

Inference workflows with pre- and post processing steps

SageMaker serial inference pipelines

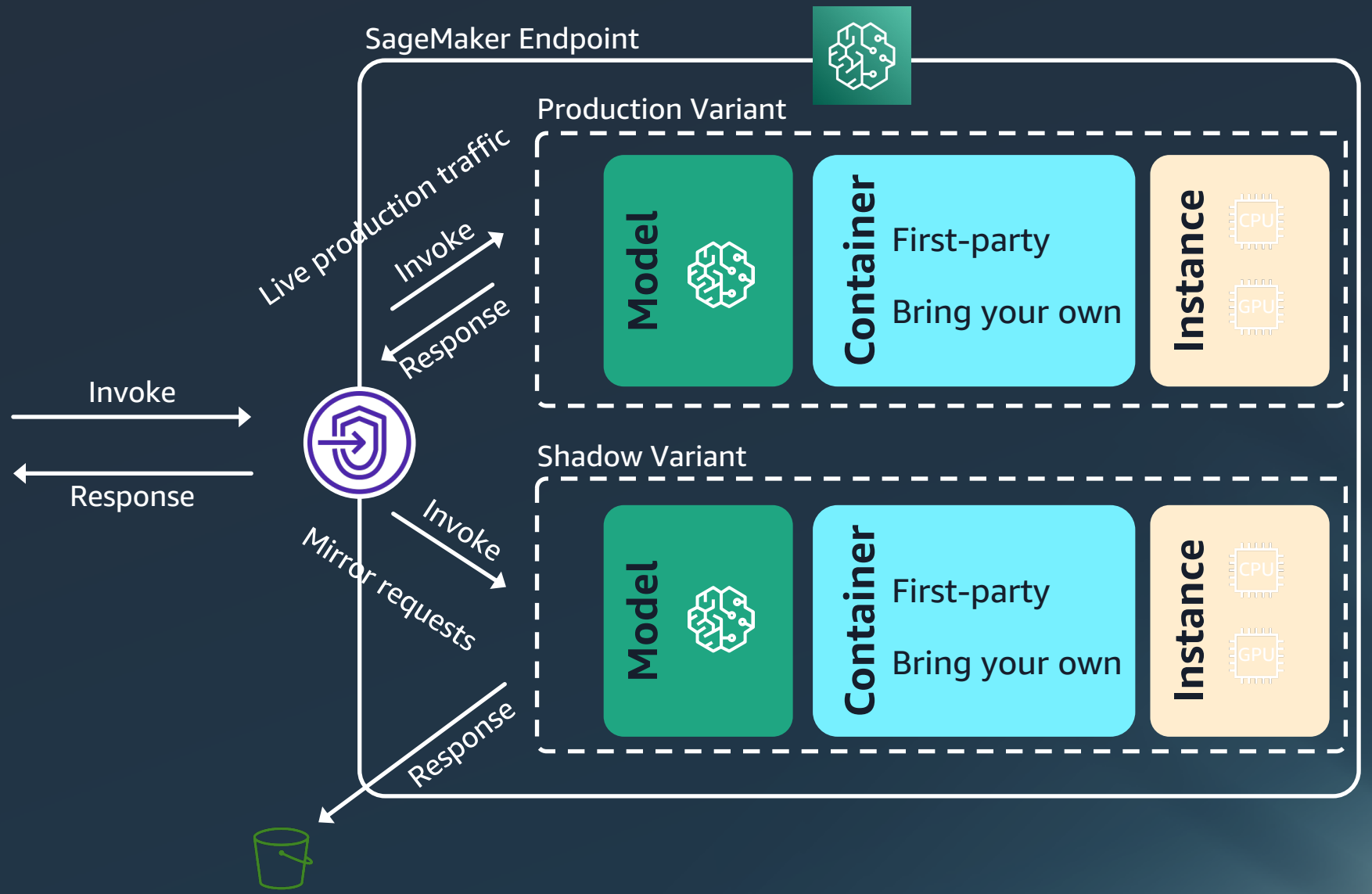


SageMaker now supports shadow testing

SageMaker takes care of mirroring requests

Start small and dial up to control costs

Accessible through AWS console, CLI, APIs

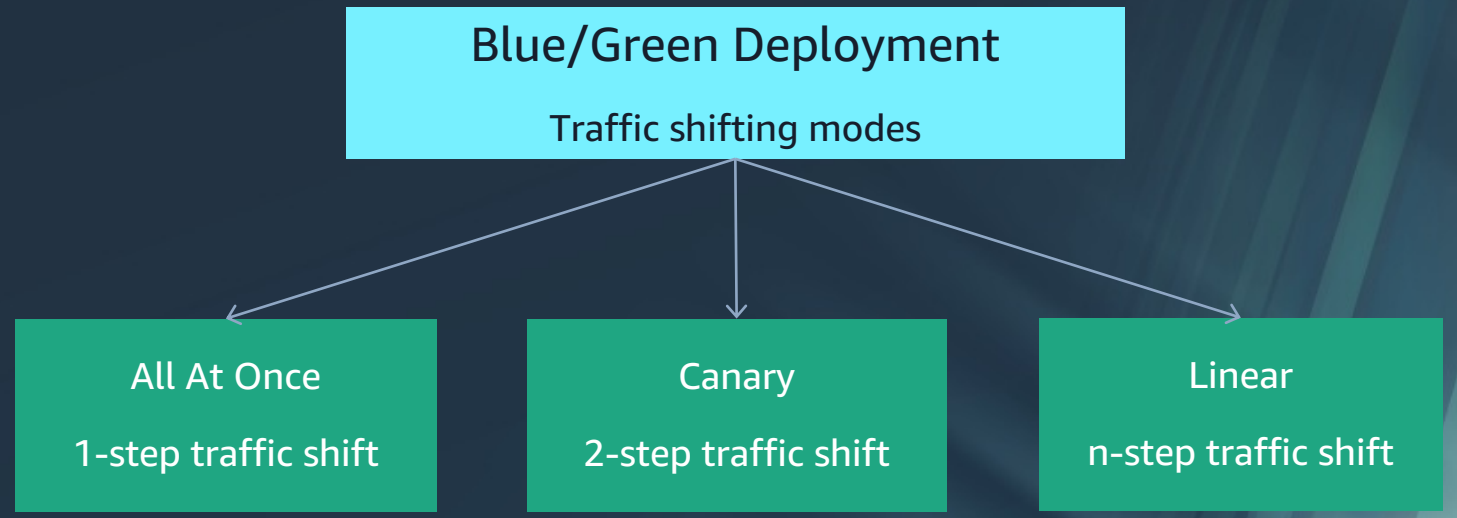


Deployment complexity

Deployment Guardrails

Minimize deployment risk

- Traffic shifting modes, such as canary and linear
- Deployment safety while updating production environments
- Built-in safeguards such as auto-rollbacks
- Fully managed deployment
- Visibility: Track the progress of your deployment



Performance

Inference performance

ML-BASED APPLICATIONS HAVE DIVERSE AND STRINGENT PERFORMANCE REQUIREMENTS

Ad serving

Personalized recommendations

Fraud detection

Object detection

Document analysis

• • •

Latency

Ultra-low latency for real-time interactions

Availability

Minimize downtime and improve reliability

Throughput

Serve millions of transactions per second

Scale

Thousands of models – one per person

SageMaker for best inference performance

Under 5 ms for small payloads

70+ instances to fit performance requirements
Ability to co-locate endpoint with application

10M+ transactions per second

Built-in routing layer to distribute your traffic and prevent overloading

Deploy thousands of models

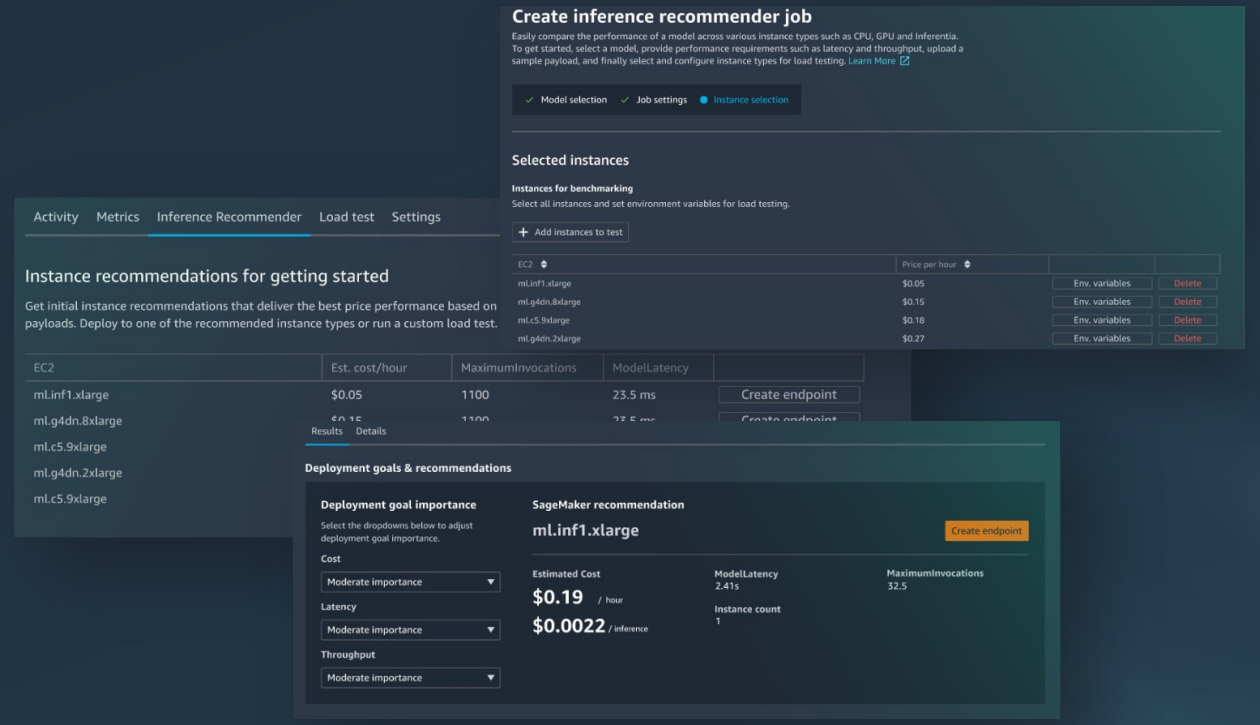
Use multi-model endpoints

99.95% availability SLA

- Automatically deploy across multiple AZs
- Routine health checks and replacement of unhealthy instances

SageMaker Inference Recommender

- Get instance recommendation in just a few clicks
- Run extensive load tests to optimize cost, latency, and throughput
- Deploy models with confidence

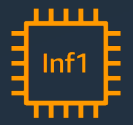


Cost

- Infrastructure costs
- Operations and maintenance costs

Infrastructure cost

Choose the right instances



Inferentia: 2.3x throughput with 70% lower cost compared to GPU instances

Graviton 2 and 3: 40% better price performance over comparable current generation x86-based instances

Autoscaling: Provision instances dynamically to meet traffic pattern; auto scales in a few minutes

SageMaker Serverless Inference

1 

ECR image location for inference code

2 

S3 location for model artifacts

3 

Choose a memory size

Serverless Inference endpoint

Automatically spins up and manages compute resources

Automatic scaling based on demand

Managed logging and monitoring

Sends inference requests

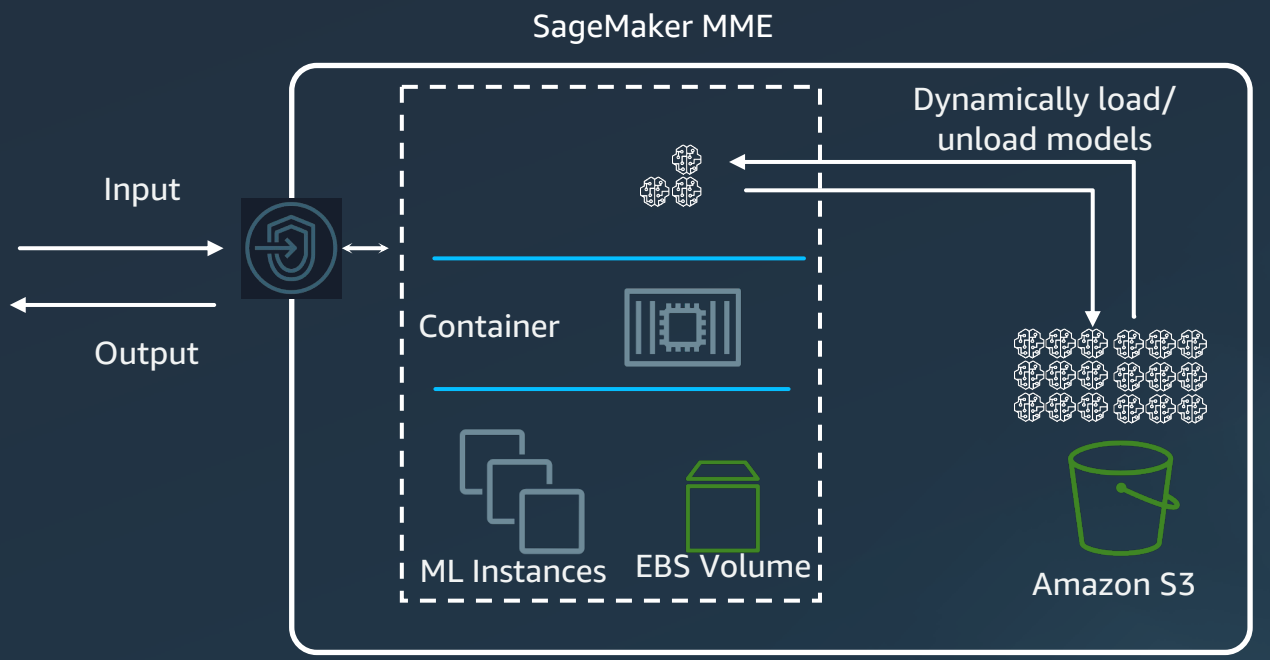
Inference results

Trigger from client application or other AWS services

End user

SageMaker multi-model endpoints (MME)

- Deploy thousands of models to a single endpoint
- Save 90% of cost deploying personalization models



Should I use Amazon SageMaker or build my own deployment platform?

Why run inference on Amazon SageMaker?

Infrastructure Management & Monitoring

- Pre-built optimized deep learning & ML framework containers
- Logging and Monitoring (pre-built CloudWatch metrics)
- Request routing + Smart routing (multi-model endpoints)
- Storage provisioning
- VPC, NAT/IGW provisioning

Model Release Management

- A/B testing
- Blue/green deployments
- Canary & Linear traffic shifting
- Update endpoints w/o availability loss
- Auto-rollback protection

Security & Compliance

- VPC endpoint support
- Authentication (Sigv4) / Authorization
- Secure connection (TLS 1.2)
- Support for CMK & KMS
- Security compliance validation

ML-Specific Capabilities

- High level Python SDK, AWS SDKs, CLIs, APIs
- Multiple deployment modes
- Inference pipelines
- Pre/post data processing
- Request & response capturing
- Model version management
- Model monitoring
- Inference load testing
- Explainability
- Lineage tracking
- Model registry
- Model caching (with multi-model endpoints)

Large generative AI model hosting

- SageMaker Jumpstart (Foundation models)
- Large model Partitioning across GPUs (DeepSpeed, Hugging Face Accelerate)
- Large model compression
- SageMaker compatible pre-built Large Model Inference container
- Low-code/No-code setup

High Availability

- Automatic multi-AZ provisioning (provides 99.95% SLA for real-time inference)
- Automatic monitoring and patching of underlying instances
- Automatic storage monitoring
- Automatic bad instance replacement
- Autoscaling & Continuous health checks

Cost Optimization Tooling

- Multi-model endpoints
- Multi-container endpoints
- Instance rightsizing
- Serverless Inference (zero cost if idle)
- Model compilation

Survey!

aws machine learning