

SageMaker Endpoints

<https://github.com/ChandraLingam/AmazonSageMakerCourse>

SageMaker SDK

```
estimator.fit(...)
```

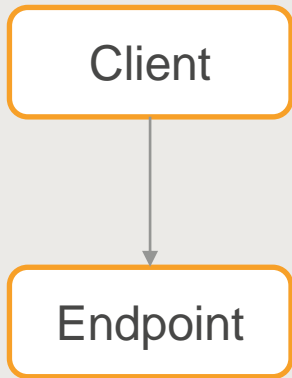
- Stores model artifacts in S3

```
estimator.deploy(...)
```

- Create Model
- Create Endpoint Configuration
- Create Endpoint

Deployment Flexibility!

Motivation

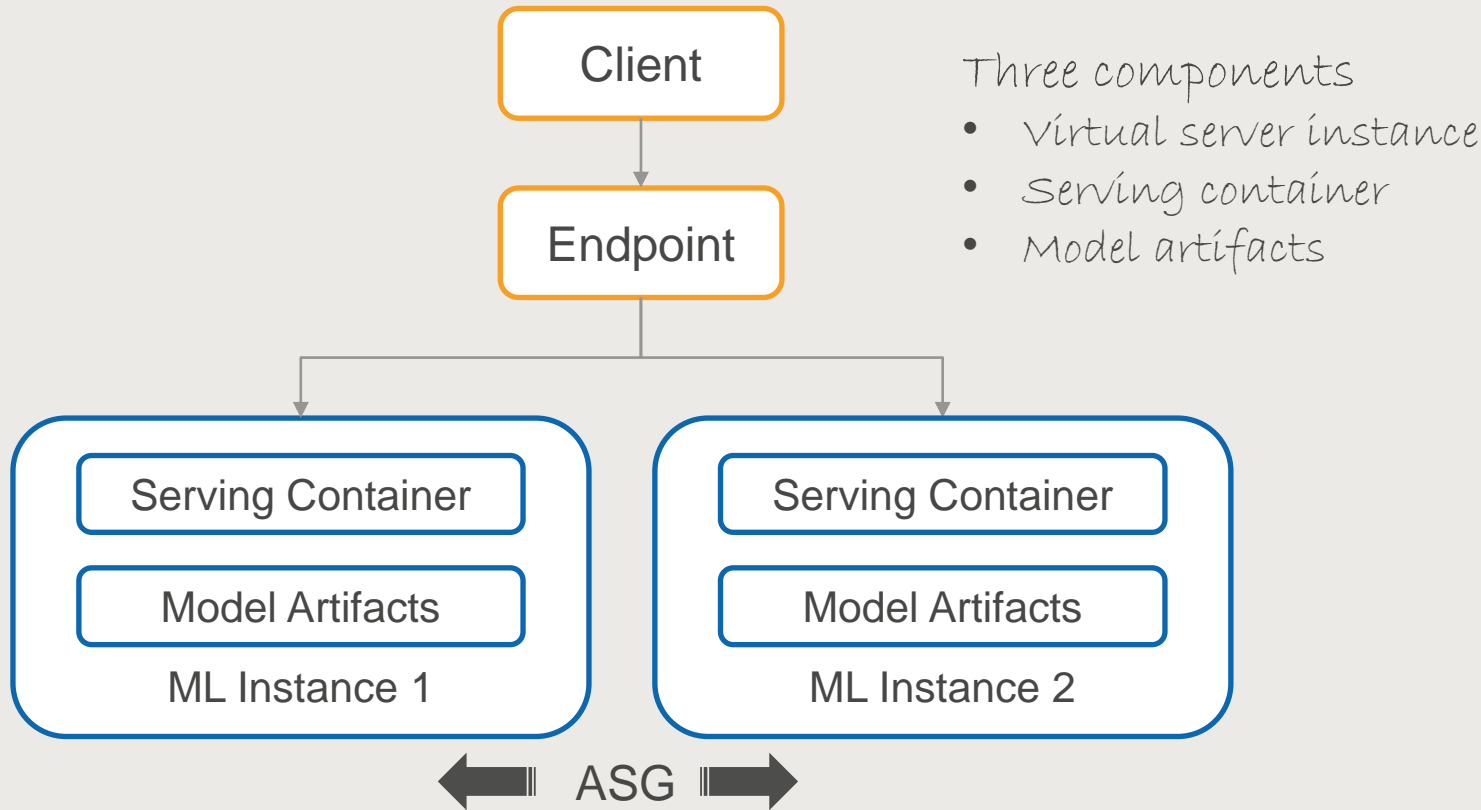


Continuous Improvement

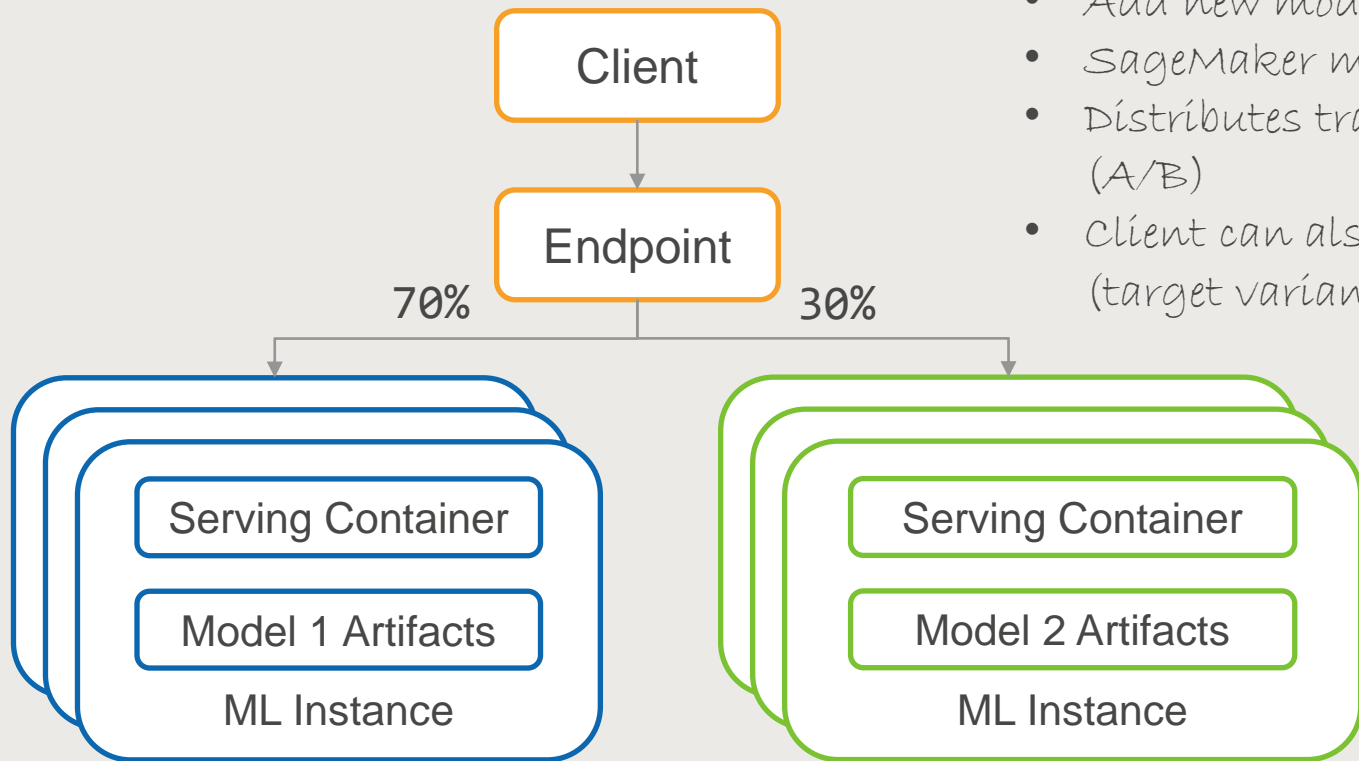
- New model versions
- New algorithm versions
- Different algorithms

How to incorporate these changes in a live system with zero-downtime?

Single-model Endpoint



Multiple Production Variants



- Add new model variants
- SageMaker manages deployment
- Distributes traffic based on weight (A/B)
- Client can also call a specific variant (target variant parameter)

Multiple Production Variants (with single-model endpoint)

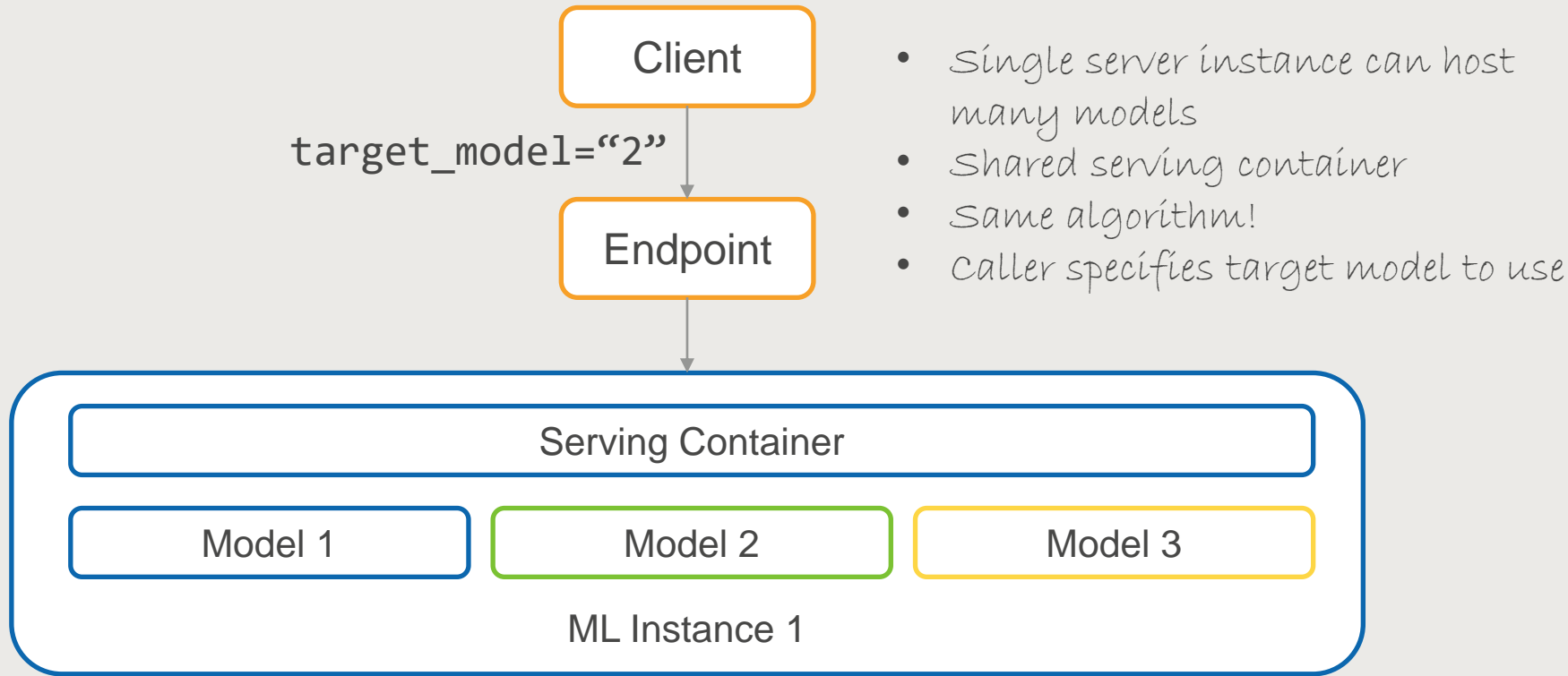
Advantages

- Make model changes with zero-downtime
- Distribute traffic based on weight
- Client can also use a specific variant
- AutoScaling rules by variant
- Mix and Match Algorithms

Disadvantages

- Each model is hosted in a separate server instance
- One serving container and model artifact per instance
- Too many servers!

Multi-model Endpoint



Multi-model Endpoint

Advantages

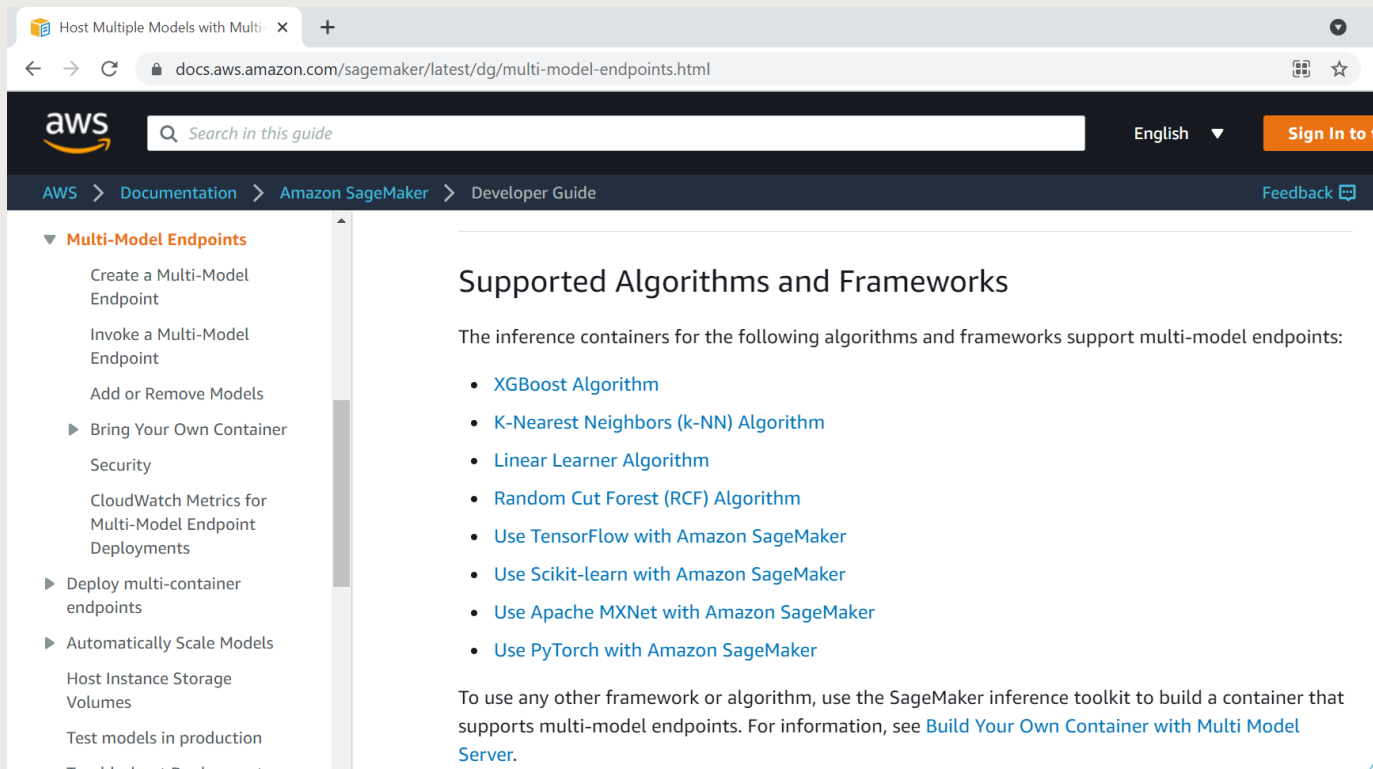
- Single serving container to host many models
- Reduce infrastructure costs
- Automatically host new models when you add new model artifacts in S3

Disadvantages

- Cold start delay when a model is invoked for the first time
- Container needs to download artifacts from S3 and host
- Models need to be algorithm compatible
- Caller must keep track of models and their S3 location
- Only some of the algorithms [support multi-model hosting](#)

Multi-model algorithm support (2021)

*** Check AWS documentation for updated list ***



The screenshot shows the AWS SageMaker Developer Guide page for Multi-Model Endpoints. The browser address bar shows the URL: docs.aws.amazon.com/sagemaker/latest/dg/multi-model-endpoints.html. The page header includes the AWS logo, a search bar, and a 'Sign In to t' button. The breadcrumb trail is: AWS > Documentation > Amazon SageMaker > Developer Guide. The left sidebar shows the 'Multi-Model Endpoints' section expanded, with sub-items: Create a Multi-Model Endpoint, Invoke a Multi-Model Endpoint, Add or Remove Models, Bring Your Own Container, Security, CloudWatch Metrics for Multi-Model Endpoint Deployments, Deploy multi-container endpoints, Automatically Scale Models, Host Instance Storage Volumes, Test models in production, and Troubleshoot Deployments. The main content area is titled 'Supported Algorithms and Frameworks' and contains the following text: 'The inference containers for the following algorithms and frameworks support multi-model endpoints:'. Below this text is a bulleted list of supported algorithms and frameworks: XGBoost Algorithm, K-Nearest Neighbors (k-NN) Algorithm, Linear Learner Algorithm, Random Cut Forest (RCF) Algorithm, Use TensorFlow with Amazon SageMaker, Use Scikit-learn with Amazon SageMaker, Use Apache MXNet with Amazon SageMaker, and Use PyTorch with Amazon SageMaker. At the bottom of the main content area, there is a paragraph: 'To use any other framework or algorithm, use the SageMaker inference toolkit to build a container that supports multi-model endpoints. For information, see Build Your Own Container with Multi Model Server.'

Host Multiple Models with Multi- x +

docs.aws.amazon.com/sagemaker/latest/dg/multi-model-endpoints.html

aws Search in this guide English Sign In to t

AWS > Documentation > Amazon SageMaker > Developer Guide Feedback

Multi-Model Endpoints

- Create a Multi-Model Endpoint
- Invoke a Multi-Model Endpoint
- Add or Remove Models
- Bring Your Own Container
 - Security
 - CloudWatch Metrics for Multi-Model Endpoint Deployments
- Deploy multi-container endpoints
- Automatically Scale Models
 - Host Instance Storage Volumes
 - Test models in production
 - Troubleshoot Deployments

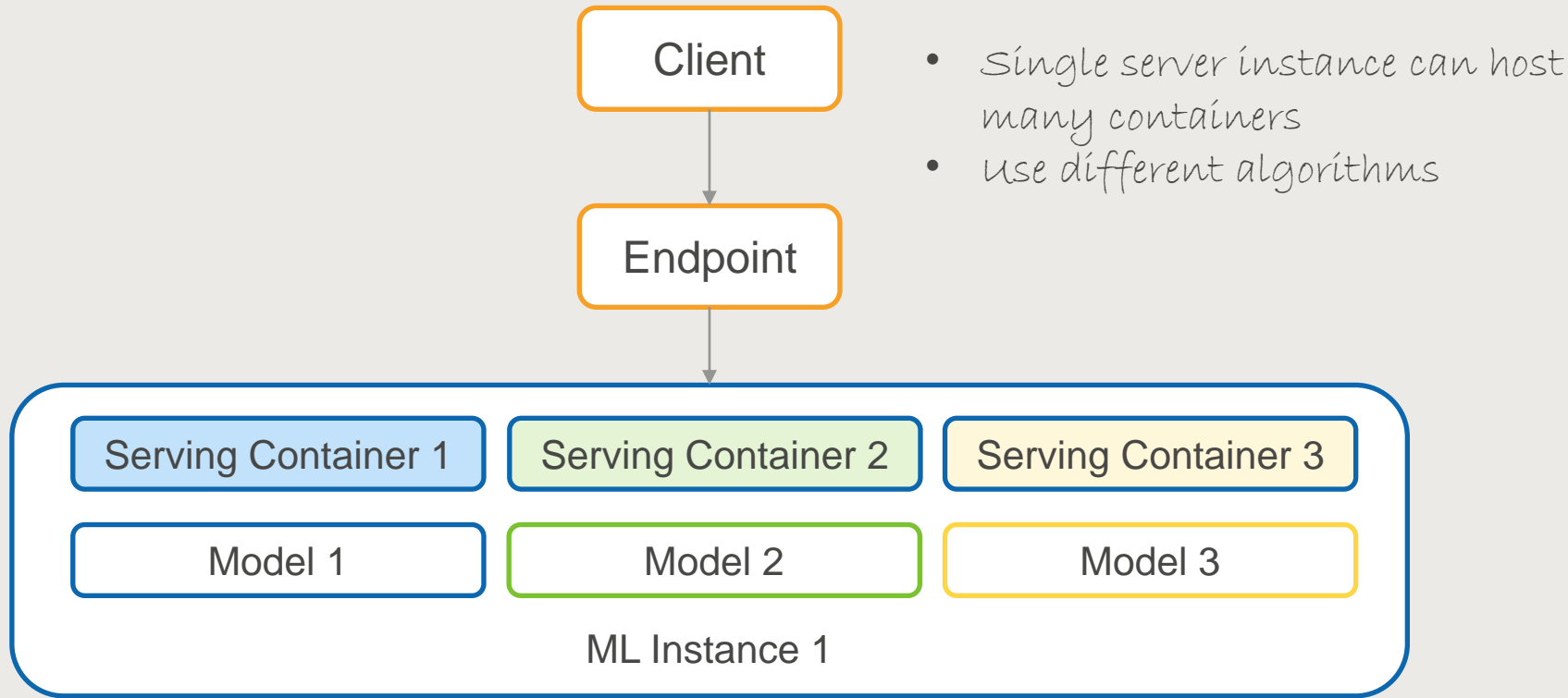
Supported Algorithms and Frameworks

The inference containers for the following algorithms and frameworks support multi-model endpoints:

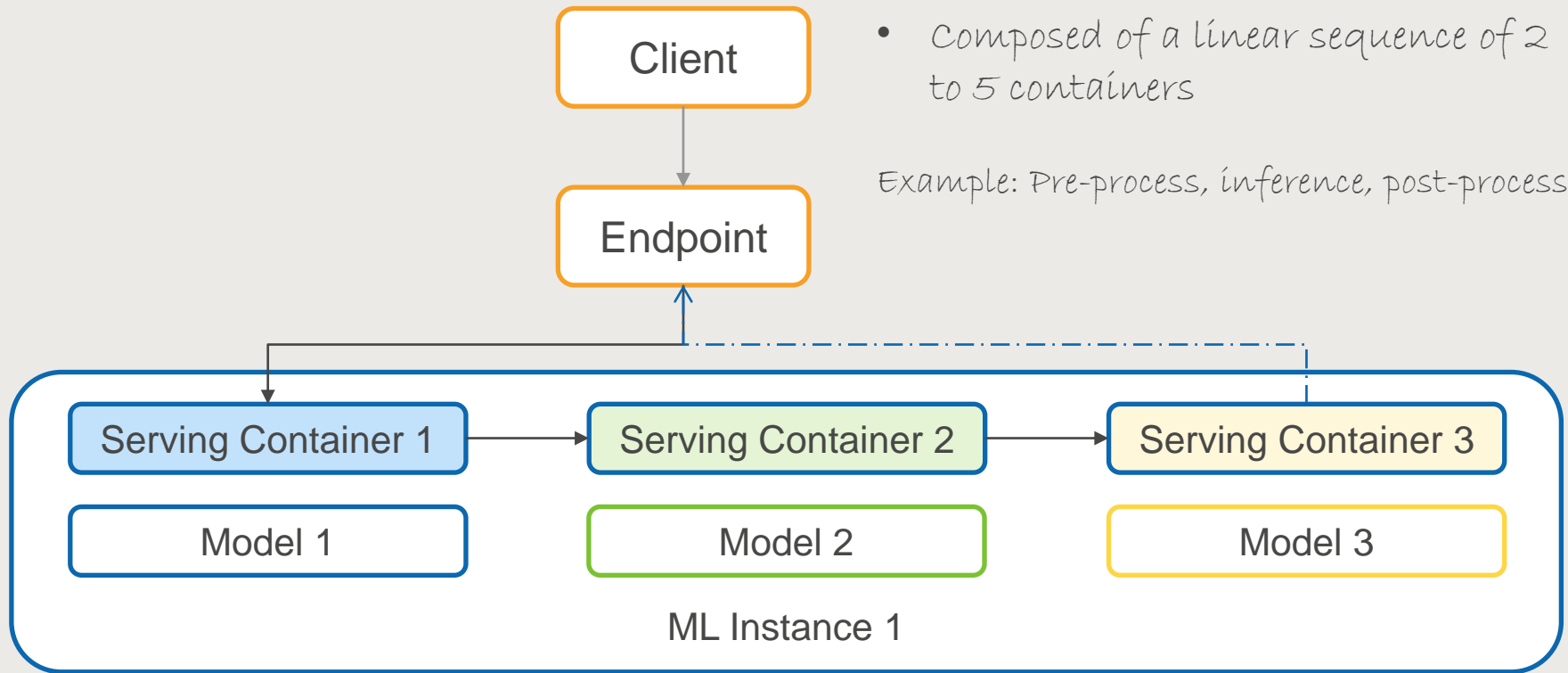
- XGBoost Algorithm
- K-Nearest Neighbors (k-NN) Algorithm
- Linear Learner Algorithm
- Random Cut Forest (RCF) Algorithm
- Use TensorFlow with Amazon SageMaker
- Use Scikit-learn with Amazon SageMaker
- Use Apache MXNet with Amazon SageMaker
- Use PyTorch with Amazon SageMaker

To use any other framework or algorithm, use the SageMaker inference toolkit to build a container that supports multi-model endpoints. For information, see [Build Your Own Container with Multi Model Server](#).

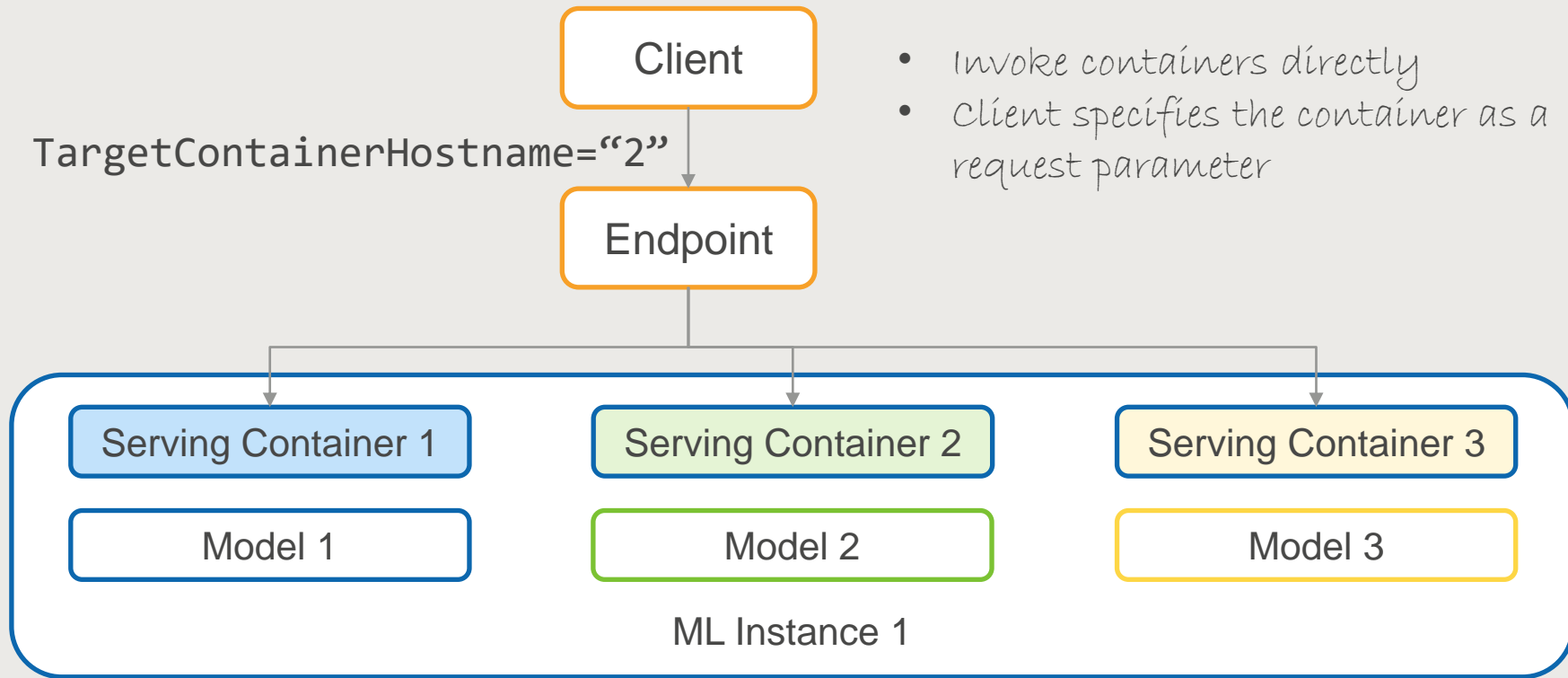
Multi-container Endpoint



Inference Pipeline Mode



Direct Mode



Multi-container Endpoint

Advantages

- Host multiple models and algorithms in an instance
- Chain containers as an inference pipeline
- Invoke containers directly
- Reduce infrastructure costs

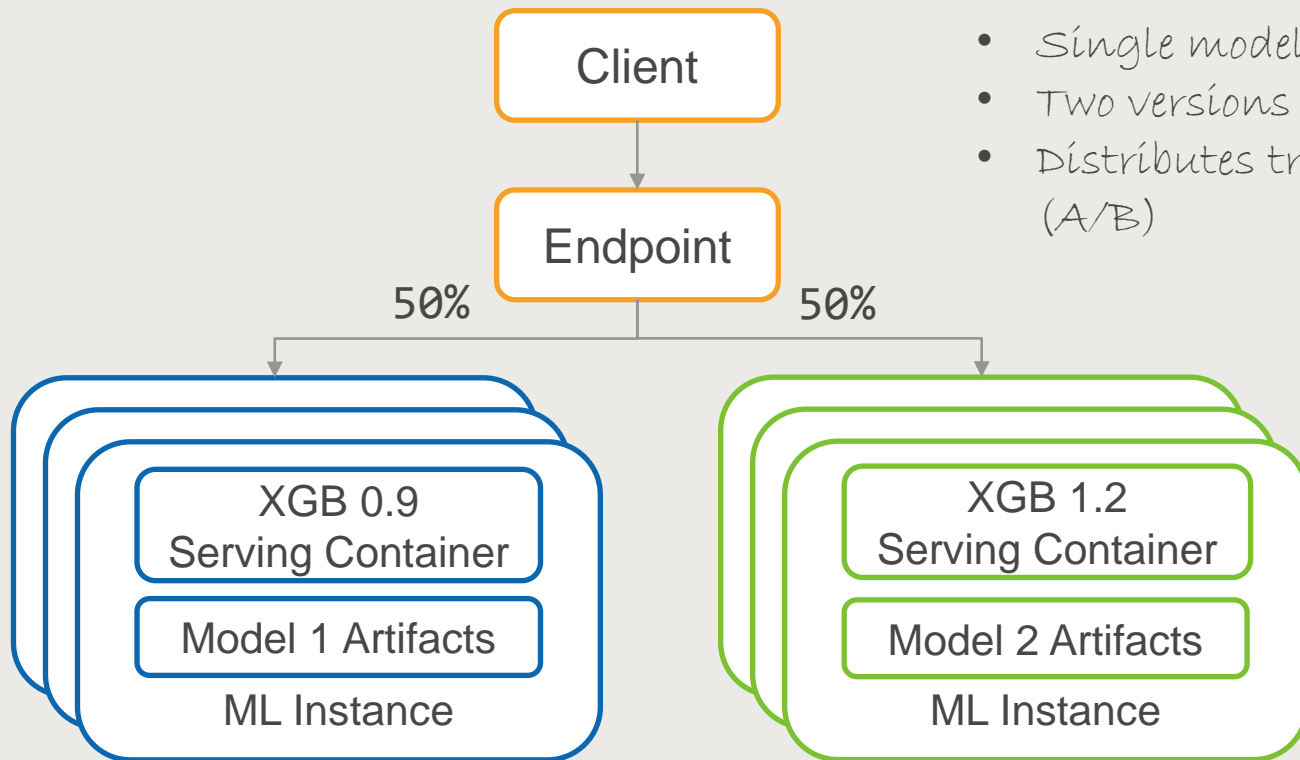
Disadvantages

- Memory and storage are shared among containers
- Cross-interference, one bad serving container can affect stability
- May need more powerful instances
- Limit of 5 containers that can be co-hosted

Summary

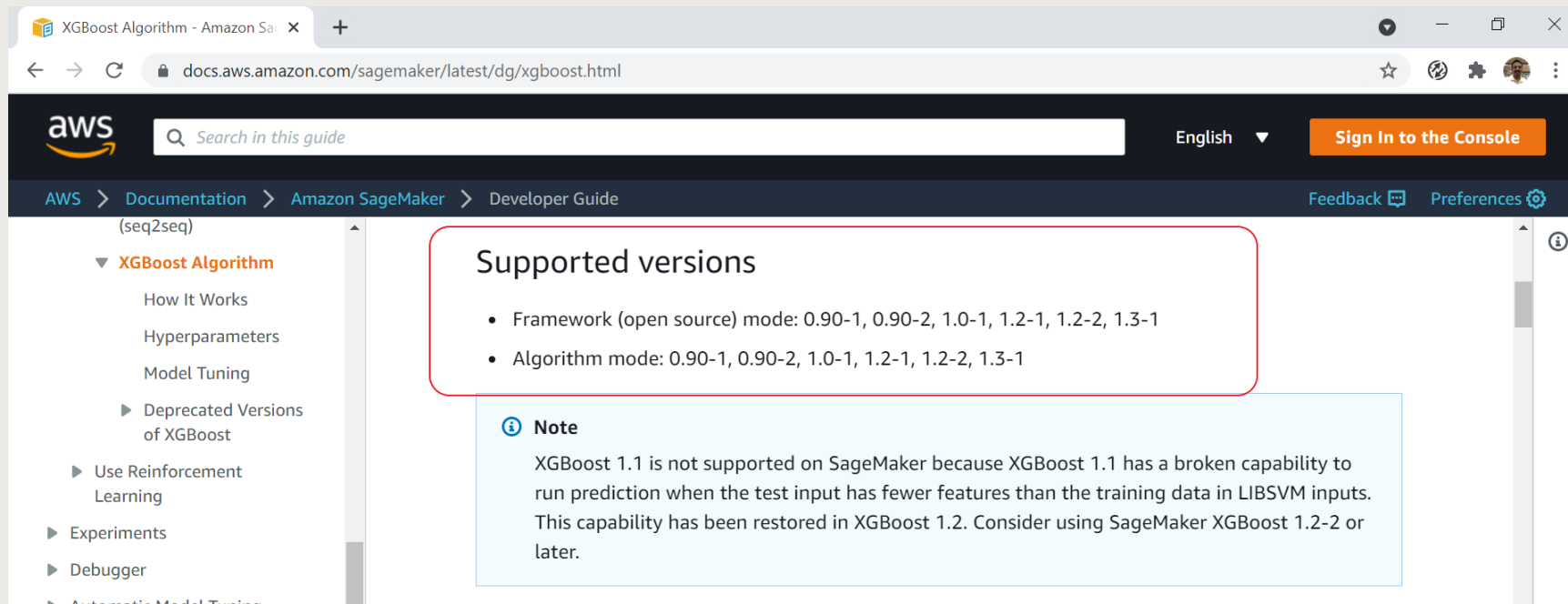
- 1 Single-model endpoint
- 2 Multiple production variants
- 3 Multi-model endpoint
- 4 Multi-container endpoint

Lab – A/B Testing Multiple Production Variants



- Single model endpoints
- Two versions of XGBoost 0.9 and 1.2
- Distributes traffic based on weight (A/B)

XGBoost Versions

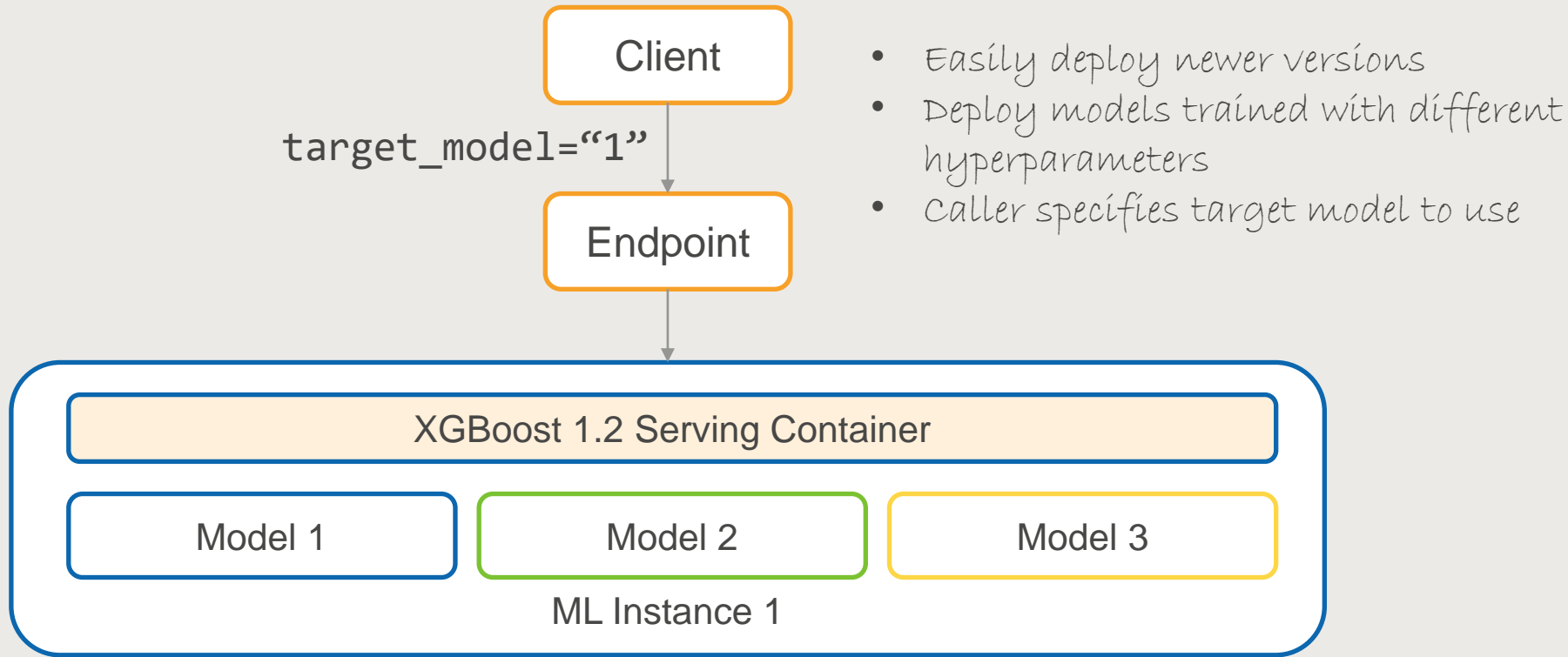


The screenshot shows a web browser window displaying the AWS SageMaker documentation for XGBoost. The browser's address bar shows the URL `docs.aws.amazon.com/sagemaker/latest/dg/xgboost.html`. The AWS logo and a search bar are visible in the top navigation bar. The left sidebar contains a navigation menu with the following items: **XGBoost Algorithm** (expanded), How It Works, Hyperparameters, Model Tuning, Deprecated Versions of XGBoost, Use Reinforcement Learning, Experiments, Debugger, and Automatic Model Tuning. The main content area is titled "Supported versions" and lists the following supported versions:

- Framework (open source) mode: 0.90-1, 0.90-2, 1.0-1, 1.2-1, 1.2-2, 1.3-1
- Algorithm mode: 0.90-1, 0.90-2, 1.0-1, 1.2-1, 1.2-2, 1.3-1

A "Note" box below the list states: "XGBoost 1.1 is not supported on SageMaker because XGBoost 1.1 has a broken capability to run prediction when the test input has fewer features than the training data in LIBSVM inputs. This capability has been restored in XGBoost 1.2. Consider using SageMaker XGBoost 1.2-2 or later."

Lab – Multi-model Endpoint





Chandra Lingam

70,000+ Students



AWS Certified Machine Learning Specialty

