

Amazon SageMaker serverless inference (Preview)

Michael Lin

Senior Solutions Architect
AWS



SageMaker inference options

NEW

Real-time inference

Low latency
Ultra high throughput
Multi-model endpoints
A/B testing

Batch transform

Process large datasets
Job-based system

Asynchronous inference

Near real-time
Large payloads (1 GB)
Long timeouts (15 mins)

Current customer challenges with ML inference



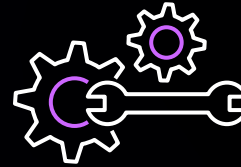
Workloads are intermittent

Some ML workloads have less predictable usage patterns and long periods of inactivity



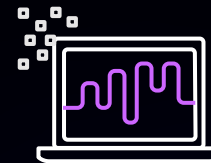
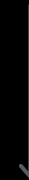
End up over-provisioning capacity

Utilization is low and costs are high regardless of number of requests



Challenging to provision capacity

Data scientists are challenged with selecting optimal instance types and managing autoscaling policies



Increases TCO

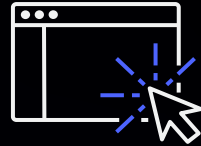
Spend lot of time in provisioning and managing servers

PREVIEW

Introducing Amazon SageMaker Serverless Inference



First purpose built serverless
ML inference in cloud



Fully managed



Pay only for what you use,
billed in milliseconds

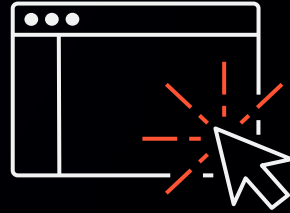
What are the key benefits?

1. Fully managed offering



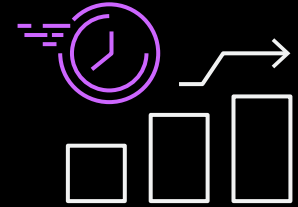
Managed infrastructure

Security
Monitoring
Logging
Built in availability
and fault tolerance



Serverless

No need to select instance
types or provision capacity
Choose memory options based
on inference processing needs



Automatically scale out, in, and down to 0

No need to set
scaling policies

2. Pay only for what you use

INFERENCE DURATION PRICING

Memory (MB)	Price per sec
1024	\$0.000020
2048	\$0.000040
3072	\$0.000060
4096	\$0.000080
5120	\$0.000100
6144	\$0.000120

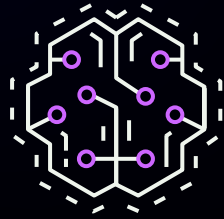
Pay only for the duration to process each request and for data processing

No charge for idle periods

Billed in milli-seconds granularity

Free tier: 150k seconds free inference duration/month for the first 2 months

3. Purpose built for ML



ML science

1P algorithms

Optimized ML frameworks

BYO



**Seamlessly move from
serverless to real-time**

One-click update

No changes required to container

How to deploy on SageMaker Serverless Inference?

Simple one-click deployment

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

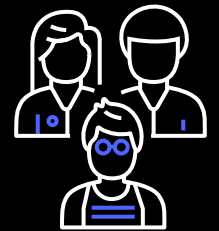
Managed logging
and monitoring

Sends
inference
requests



Trigger from
client application
or other
AWS services

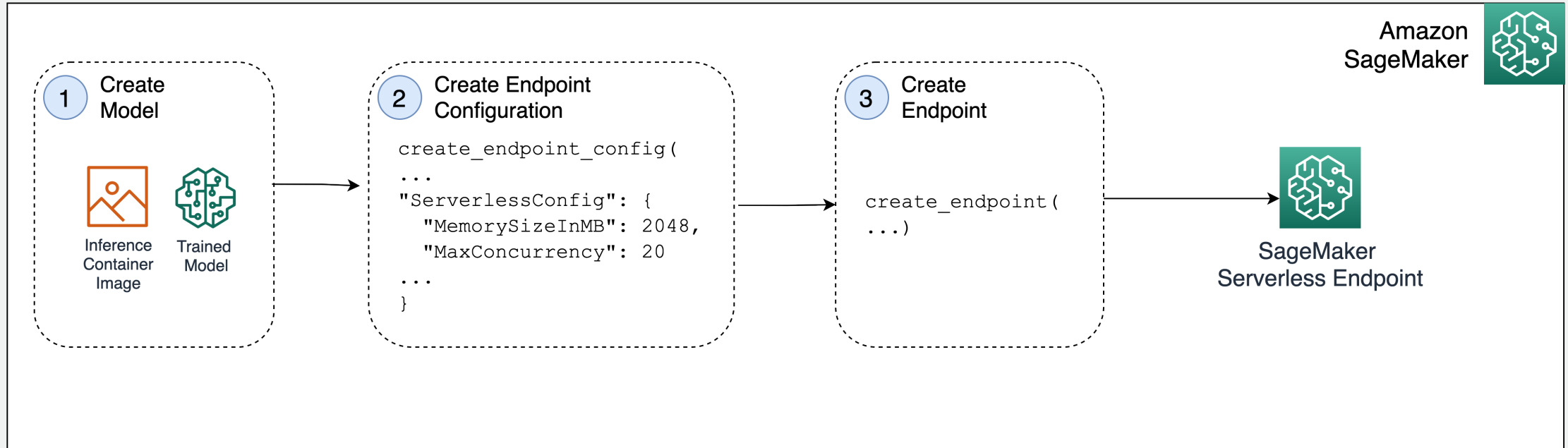
Inference
results



End user

How it works

Creating a Serverless Endpoint



Console experience

Container definition 1

▼ Container input options

- ☒ **Provide model artifacts and inference image location**
Use this for models trained using built-in algorithms, BYO algorithms, or models trained outside Amazon SageMaker.
- ☐ **Use a model package resource**
Use this for model packages that contain inference images and artifacts from AWS Marketplace subscribed algorithms.
- ☐ **Use a model package subscription from AWS Marketplace**
Use this for model packages published by vendors from AWS Marketplace.

▼ Provide model artifacts and inference image options

- ☒ **Use a single model**
Use this to host a single model in this container.
- ☐ **Use multiple models**
Use this to host multiple models in this container.

Location of inference code image

Type the registry path where the inference code image is stored in Amazon ECR.

`aws_account_id.dkr.ecr.region.domain/repository[:tag] or [@digest]`

Location of model artifacts - *optional*

Type the URL where model artifacts are stored in S3.

`s3://bucket/path-to-your-data/`

The path must point to a single gzip compressed tar archive (.tar.gz suffix).

First create a model by providing ECR image location and S3 location for model artifacts

No changes to current experience

Console experience

THEN CREATE ENDPOINT CONFIGURATION BY SELECTING SERVERLESS OPTION

Amazon SageMaker > Endpoint configuration > Create endpoint configuration

Create endpoint configuration

To deploy models to Amazon SageMaker, first create an endpoint configuration. In the configuration, specify which models to deploy, and the relative traffic weighting and hardware requirements for each. See [Deploying a Model on Amazon SageMaker Hosting Services](#) [Learn more about the API](#)

Endpoint configuration

Endpoint configuration name

Maximum of 63 alphanumeric characters. Can include hyphens (-), but not spaces. Must be unique within your account in an AWS Region.

Type of endpoint

☐ Provisioned

☒ Serverless (In Preview)

Production variants

Model name	Training job	Variant name	Memory Size	Max Concurrency	Actions
Beta-Model	-	variant-name-1	1 GB	20	Edit Remove

▼ Tags - optional

Key	Value	
<input type="text"/>	<input type="text"/>	Remove

[Add tag](#)

[Cancel](#) [Create endpoint configuration](#)

Search for services, features, marketplace products, and docs [Option 5]

Edit Production Variant

Model name

Beta-Model

Variant name

variant-name-1

Memory Size

1 GB

1 GB

2 GB

3 GB

4 GB

5 GB

6 GB

[Cancel](#) [Save](#)

Console experience

Amazon SageMaker > Endpoints > Create and configure endpoint

Create and configure endpoint

To deploy models to Amazon SageMaker, first create an endpoint. Provide an endpoint configuration to specify which models to deploy and the hardware requirements for each. See [Deploying a Model on Amazon SageMaker Hosting Services](#) [Learn more about the API](#)

Endpoint

Endpoint name
Your application uses this name to access this endpoint.

Maximum of 63 alphanumeric characters. Can include hyphens (-), but not spaces. Must be unique within your account in an AWS Region.

Attach endpoint configuration

☒ **Use an existing endpoint configuration**
Use an existing endpoint configuration or clone an endpoint configuration.

☐ **Create a new endpoint configuration**
Add models and configure the instance and initial weight for each model.

New endpoint configuration

Change Clone

Endpoint configuration name
ABCD23232

Encryption key
-

Production variants

Model name	Training job	Variant name	Instance type	Elastic Inference	Initial instance count	Initial weight
Beta-Model	-	variant-name-1	-	-	-	1

▼ **Tags - optional**

Key	Value	
<input type="text"/>	<input type="text"/>	Remove

[Add tag](#)

Cancel **Create endpoint**

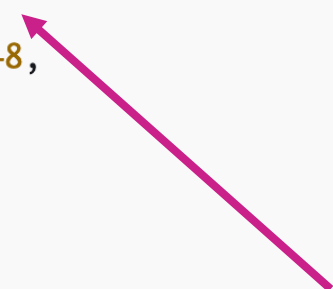
Last step is to create an endpoint using the configuration you chose earlier

No changes to current experience

Code Experience Overview

New Config in existing create endpoint config API

```
response = client.create_endpoint_config(  
    EndpointConfigName="<your-endpoint-configuration>",  
    ProductionVariants=[  
        {  
            "ModelName": "<your-model-name>",  
            "VariantName": "AllTraffic",  
            "ServerlessConfig": {  
                "MemorySizeInMB": 2048,  
                "MaxConcurrency": 20  
            }  
        }  
    ]  
)
```



New ServerlessConfig option

Link to documentation:

<https://docs.aws.amazon.com/sagemaker/latest/dg/serverless-endpoints.html>

Best Practices

for Serverless Inference



Use only for workloads that can tolerate cold starts

Use only for workloads that can tolerate cold starts as pre-warming is not recommended.



Convert from serverless to real-time as needed

Serverless Inference provides the capability to update a serverless endpoint to a real-time endpoint if traffic patterns change. Updating from a real-time endpoint to serverless is not supported at launch.



Choose serverless endpoint memory size based on model size

Memory size should be at least as large as your model for optimal usage. Benchmark to measure performance.



BYOC: Create only one worker & load one copy of model

Serverless Inference supports SageMaker-provided containers and Bring-Your-Own Container (BYOC)

Thank you!

Michael Lin

linmicht@amazon.com

