Amazon EC2 Inf1 Instances: High Performance with the Lowest Cost Machine Learning Inference in the Cloud

With Amazon EC2 Inf1 instances powered by AWS Inferentia chips, you can optimize the deployment of your machine learning applications with high throughput, low latency, at the lowest cost per inference in the cloud.

Achieve optimized throughput and latency

High throughput and low latency mean you can achieve faster processing without compromise.

Up to 30% higher throughput

- Compared to Amazon EC2 G4 instances

Can scale up to 2000 Tera (Trillion) Operations per Second (TOPS)

- With 1 to 16 AWS Inferentia chips per instance

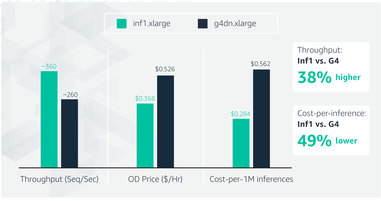
Large on-chip memory

- Allows caching of machine learning models directly on the chip instead of having to access external memory, resulting in low latency

Enable the lowest cost machine learning inference in the cloud

The cost of deploying a machine learning model can have a significant impact on budgets. Inf1 instances outperform other instances with the lowest cost per inference in the cloud.

Up to 90% of the cost of machine learning is incurred by inference in deployment



Throughput:
Inf1 vs. G4
38% higher

Cost-per-inference:
Inf1 vs. G4
49% lower

Amazon EC2 Inf1 instances deliver the lowest cost machine learning inference in the cloud

Up to 45% lower than the next lowest-cost instance (Amazon EC2 G4 instances)

Choose a flexible and easy-to-use solution

Inf1 instances support multiple machine learning models and data types, requiring few code changes to support models trained on the most popular frameworks.

Multiple machine learning models supported

Image and video recognition and classification: identify objects, people, text, scenes and activities to

- Detect inappropriate content
- Verify users
- Count people

Natural language processing and translation: turn text into lifelike speech and translate content to

- Synthesize speech for virtual assistants
- Create speech-enabled products
- Localize content for international users

Supports widely-used frameworks with few, if any, code changes

Multiple data types supported

The AWS Neuron SDK can automatically convert FP32 trained models to BF16

Deploy using popular AWS services

AWS BATCH

Amazon EKS

Amazon ECS

Amazon SageMaker

How it works

Choose ML Framework

Choose and optimize your ML algorithm

Build Model

You can build your model by using Jupyter Notebooks hosted on EC2 or within Amazon SageMaker, a fully managed service

Train Model

You can train your model using EC2 P3/P3dn or use Amazon SageMaker for automated workflow

Deploy Model on Inferentia-based Inf1 instances

Distribute the compiled model to an EC2 Inf1 instance or fleet of instances

Execute the model for inference

Compile Model Using AWS Neuron

Take your trained model and invoke AWS Neuron through the ML framework's API

Compile your trained model so that it is optimized for use with AWS Inferentia

Save the output model to an S3 bucket

With Amazon EC2 Inf1 instances, you can run a variety of large-scale ML inference applications at high throughput, low latency, at the lowest cost in the cloud.

Learn more at https://aws.amazon.com/ec2/instance-types/inf1/