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.

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.

How it works

Choose your ML framework

Choose and optimize your ML algorithm

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

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

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

Execute the model for inference

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

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

Choose a flexible and easy-to-use solution

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

Multiple machine learning models supported

Multiple data types supported

Achieve optimized throughput and latency

High-throughput and low-latency canonen because fewer code changes are needed.

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

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.

Up to 4x 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

Up to 2000 TRILLION

2000

"4X"

Large on-chip memory

Supports widely-used frameworks

with few, if any, code changes.

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/