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**

1. **Choose ML Framework**
   - You can build your model by using Jupyter Notebooks hosted on EC2 or within Amazon SageMaker, a fully managed service.

2. **Build Model**
   - Take your trained model and invoke AWS Neuron through the ML framework’s API.

3. **Compile Model**
   - Compile your trained model so that it is optimized for use with AWS Inferentia.

4. **Save the output model to an S3 bucket**

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

6. **Execute the model for inference**

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

**Multiple machine learning models supported**

- Transformer
- BERT
- Single shot detector (SSD)
- ResNet

**Image and video recognition and classification**

- Identify objects, people, text, scenes and activities
- Synthesize speech for virtual assistants
- Create speech-enabled products
- Localize content for international users
- Detect inappropriate content
- Verify users
- Count people

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

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

<table>
<thead>
<tr>
<th>INF1 Instance Type</th>
<th>Throughput (Seq/Sec)</th>
<th>Price ($/Hr)</th>
<th>Cost-per-inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>inf1.xlarge</td>
<td>~360</td>
<td>~260</td>
<td>~0.368</td>
</tr>
<tr>
<td>g4dn.xlarge</td>
<td>~260</td>
<td>~0.526</td>
<td>~0.284</td>
</tr>
<tr>
<td>inf1.2xlarge</td>
<td>~720</td>
<td>~0.562</td>
<td>~0.142</td>
</tr>
<tr>
<td>inf1.4xlarge</td>
<td>~1440</td>
<td>~0.562</td>
<td>~0.071</td>
</tr>
<tr>
<td>inf1.8xlarge</td>
<td>~2880</td>
<td>~0.562</td>
<td>~0.044</td>
</tr>
</tbody>
</table>

**Cost-per-inference: Inf1 vs. G4**

- Lower 49%: Inf1
-higher 38%

**Multiple data types supported**

- Supports multiple data types: 32-bit, 16-bit, 8-bit, and 4-bit
- Displays popular AWS services: S3, Batch, Deep Learning

**Enable the lowest cost machine learning inference in the cloud**

Amazon EC2 Inf1 instances deliver 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.

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

**Achieve optimized throughput and latency**

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

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

**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 90% of the cost of machine learning is incurred by inference in deployment.**

**How Amazon EC2 Inf1 Instances Enable the Lowest Cost Machine Learning Inference in the Cloud**

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

- Achieve optimized throughput and latency
- Enable the lowest cost machine learning inference in the cloud
- Choose a flexible and easy-to-use solution

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.