Lower the Cost of Running ML Applications with New Amazon EC2 Inf1 Instances

Gadi Hutt, Senior Director of Business Development and Product, AWS
Agenda

- Machine learning at AWS
- AWS Inferentia overview (HW and SW)
- Amazon EC2 Inf1 instance
- Customer story: Alexa Text-to-Speech (TTS)
Tens of thousands customers have chosen AWS for their ML workloads | More than twice as many customers using ML than any other cloud provider
Machine learning Infrastructure costs

- Machine learning training (<10%)
- Machine learning inference (>90%)
ML inference deployment options on Amazon EC2

**Custom chip**
EC2 Inf1 instances
Applications that leverage common ML frameworks
Powered by AWS Inferentia
Best price/performance for ML inferencing in the cloud
Up to 40% lower inference than EC2 G4 instances

**GPU based**
EC2 G4 instances
Applications that require access to CUDA, CuDNN or TensorRT libraries
Amazon EC2 G4 instances based on NVIDIA T4 GPUs

**CPU based**
EC2 C5 instances
Low sensitivity to inference performance
Intel Skylake CPUs
Support for AVX-512/VNNI instruction set

Featuring AWS Inferentia
Powered by AWS Inferentia
AWS Inferentia quick tour

AWS custom built: chip, software, and server

- 4 NeuronCores
- Up to 128 TOPS
- 2-stage memory hierarchy
  - Large on-chip cache and commodity DRAM
- Supports FP16, BF16, INT8 data types
- Fast chip-to-chip interconnect
Inf1 instances for ML inferencing

- Object detection
- Natural language processing
- Personalization
- Speech recognition
- Search
- Fraud detection
Inf1 instances are built from the ground up by AWS

- AWS Nitro
- AWS Custom 2nd Gen Intel Xeon Scalable Processors
- AWS Neuron
- AWS Inferentia

High performance
Low cost
## Inf1 instance sizes

<table>
<thead>
<tr>
<th>Instance size</th>
<th>vCPUs</th>
<th>Memory (GIB)</th>
<th>Storage</th>
<th>Inferentia Chips</th>
<th>Network B/W</th>
<th>EBS B/W</th>
</tr>
</thead>
<tbody>
<tr>
<td>inf1.xlarge</td>
<td>4</td>
<td>8</td>
<td>EBS only</td>
<td>1</td>
<td>Up to 25 Gbps</td>
<td>Up to 3.5 Gbps</td>
</tr>
<tr>
<td>inf1.2xlarge</td>
<td>8</td>
<td>16</td>
<td>EBS only</td>
<td>1</td>
<td>Up to 25 Gbps</td>
<td>Up to 3.5 Gbps</td>
</tr>
<tr>
<td>inf1.6xlarge</td>
<td>24</td>
<td>48</td>
<td>EBS only</td>
<td>4</td>
<td>25 Gbps</td>
<td>3.5 Gbps</td>
</tr>
<tr>
<td>inf1.24xlarge</td>
<td>96</td>
<td>192</td>
<td>EBS only</td>
<td>16</td>
<td>100 Gbps</td>
<td>14 Gbps</td>
</tr>
</tbody>
</table>

- Available in 4 sizes
- Single and multi chip instances
- AWS 2\(^{nd}\) Gen Intel Xeon Scalable Processors
- Up to 100Gbps networking bandwidth
- Amazon SageMaker, Amazon EKS, and Amazon ECS support coming soon
**Best price/performance for ML inference in the cloud**

<table>
<thead>
<tr>
<th>Instance type</th>
<th>Throughput (Seq/Sec)</th>
<th>OD Price ($/Hr)</th>
<th>Cost-per-1M inferences</th>
<th>Throughput: Inf1 vs. G4</th>
<th>Cost-per-inference: Inf1 vs. G4</th>
</tr>
</thead>
<tbody>
<tr>
<td>inf1.24xlarge</td>
<td>26,100</td>
<td>$7.619</td>
<td>$0.081</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G4.12xlarge</td>
<td>6,156</td>
<td>$3.912</td>
<td>$0.1765</td>
<td>4.24x</td>
<td>54.1%</td>
</tr>
</tbody>
</table>

Results based on running BERT-base model end-to-end with TensorFlow

EC2 Inf1 offers up to 4x higher throughput\(^\d\) and 54% lower cost per inference\(^\d\) than EC2 G4 instances

Highest price to performance in the cloud for ML inference
AWS Neuron
high-performance deep learning inference SDK

Neuron Compiler
Neuron Runtime
Profiling tools

Supports all major frameworks
TensorFlow
mxnet
PyTorch

github.com/aws/aws-neuron-sdk

AWS Neuron support forum
AWS Neuron quick tour

**Compile**
- Neuron Compiler (NEFF output)

**Deploy**
- Neuron Runtime (NRT)
- Neuron Binary (NEFF)

**Profile**
- Neuron Tools

C:\>code --version
1.1.1
AWS Neuron Compiler highlights

Smart partitioning

Automatically optimize neural-net compute
AWS Neuron Compiler highlights

Smart partitioning

Automatically optimize neural-net compute

FP32 auto casting

Ingest FP32 trained models, and Neuron auto casts to BF16
AWS Neuron Compiler highlights

- **Smart partitioning**: Automatically optimize neural-net compute
- **FP32 auto casting**: Ingest FP32 trained models, and Neuron auto casts to BF16
- **NeuronCore pipeline**: Super low-latency full bandwidth
NeuronCore Pipeline

Enabling high-throughput for latency sensitive applications
AWS Neuron Compiler highlights

- **Smart partitioning**: Automatically optimize neural-net compute
- **FP32 auto casting**: Ingest FP32 trained models, and Neuron auto casts to BF16
- **NeuronCore pipeline**: Super low-latency full bandwidth
Ingest FP32 trained models, and Neuron auto casts to BF16.

Automatically optimize neural-net compute.

Ingest FP32 trained models, and Neuron auto casts to BF16.

Super low-latency full bandwidth.

Concurrently run multiple models.
Neuron runtime

Deploy

App container
- Customer app
- Framework

Container runtime
- Neuron runtime

AWS Inferentia
Profiling with Neuron tools
Neuron Performance since Dec-4 Launch

- Results based on running BERT-base and Resnet-50 model end-to-end with TensorFlow
- BERTbase measured with inf1.24xlarge / ResNet50 measured with inf1.1xlarge
AWS Neuron Roadmap

Updated on Dec 23, 2019

12. We are working on it
   - Frameworks and deployments: add Sagemaker Neuron support
     #27 opened by AWSGH
   - Frameworks and deployments: add ECS Neuron support
     #28 opened by AWSGH
   - Operators: add non-Max-suppression operator support
     #39 opened by AWSGH
   - Frameworks and deployments: add EKS Neuron support
     #29 opened by AWSGH
   - Performance: batching performance updates
     #30 opened by AWSGH

4. Coming soon
   - Models: add CV Object detection (Mask R-CNN) support
     #40 opened by AWSGH
   - Frameworks and deployments: provide source code of the Neuron TensorFlow framework integration
     #41 opened by AWSGH
   - Frameworks and deployments: provide source code of the Neuron MXNet framework integration
     #43 opened by AWSGH
   - Performance: Neuron Runtime performance improvements and CPU load reduction
     #38 opened by AWSGH

1. Completed
   - Frameworks and deployments: Neuron Pytorch framework general availability
     #41 opened by AWSGH
Alexa usage of Inf1 instances
Neural text-to-speech challenges

- Low-latency requirement for dialog system
- High throughput requirement implied by streaming of the output speech
- Context generation is a sequence-to-sequence auto-regressive model
- Memory-bandwidth bound inference
- High temporal density of speech production model resulting with 90GFLOPs to generate 1s of output is Compute-bound inference
- Using EC2 GPU instances to meet requirements results in high operational cost
Alexa TTS migration to EC2 Inf1—Ease of integration

- Alexa TTS uses MXNet that is supported natively by AWS Neuron
- Support for C and Python APIs
- Options to migrate original FP32 models to FP16 or Bfloat16
Alexa TTS migration to EC2 Inf1—Architecture

Inferentia1
Running as NeuronCore Pipeline

Inferentia2
Running as NeuronCore Group

TTS Process

Context Generation

Vocoder 1

Vocoder 3

Vocoder 2

Vocoder 4
Alexa TTS migration to EC2 Inf1—Gains

Alexa generic traffic is able to get the lowest cost at the same latency budget.
Alexa TTS migration to EC2 Inf1—Long texts

Alexa long text traffic (ex: books, news) has even higher gains

Speech Generation Cost

- P3: 100%
- G4: 74%
- Inf1: 45%

Lowest cost
Alexa TTS migration to EC2 Inf1—Projected gains

With software optimizations Alexa can take full advantage of Inf1 for significant gains
Try Inf1 instances today!

Inf1 instances are available in US-East-1 (N. Virginia) and US-West-2 (Oregon) today. More regions coming soon!

EC2 Inf1 instances support On-Demand, Reserved and Spot purchasing options; also available as part of Savings Plan.

<table>
<thead>
<tr>
<th>Instance size</th>
<th>On-Demand</th>
<th>1YR Standard RI (40% discount)</th>
<th>3YR Standard RI (60% discount)</th>
</tr>
</thead>
<tbody>
<tr>
<td>inf1.xlarge</td>
<td>$0.368/Hr</td>
<td>$0.221/Hr</td>
<td>$0.147/Hr</td>
</tr>
<tr>
<td>inf1.2xlarge</td>
<td>$0.584/Hr</td>
<td>$0.351/Hr</td>
<td>$0.234/Hr</td>
</tr>
<tr>
<td>inf1.6xlarge</td>
<td>$1.905/Hr</td>
<td>$1.143/Hr</td>
<td>$0.762/Hr</td>
</tr>
<tr>
<td>inf1.24xlarge</td>
<td>$7.619/Hr</td>
<td>$4.572/Hr</td>
<td>$3.048/Hr</td>
</tr>
</tbody>
</table>
Thank you!