
AWS
re:Invent

CMP401

Machine learning inference with Amazon EC2 Inf1 instances

John Barrus

Global EC2 Business Development Manager

AWS

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS helps customers of all sizes deploy machine learning at scale in every industry

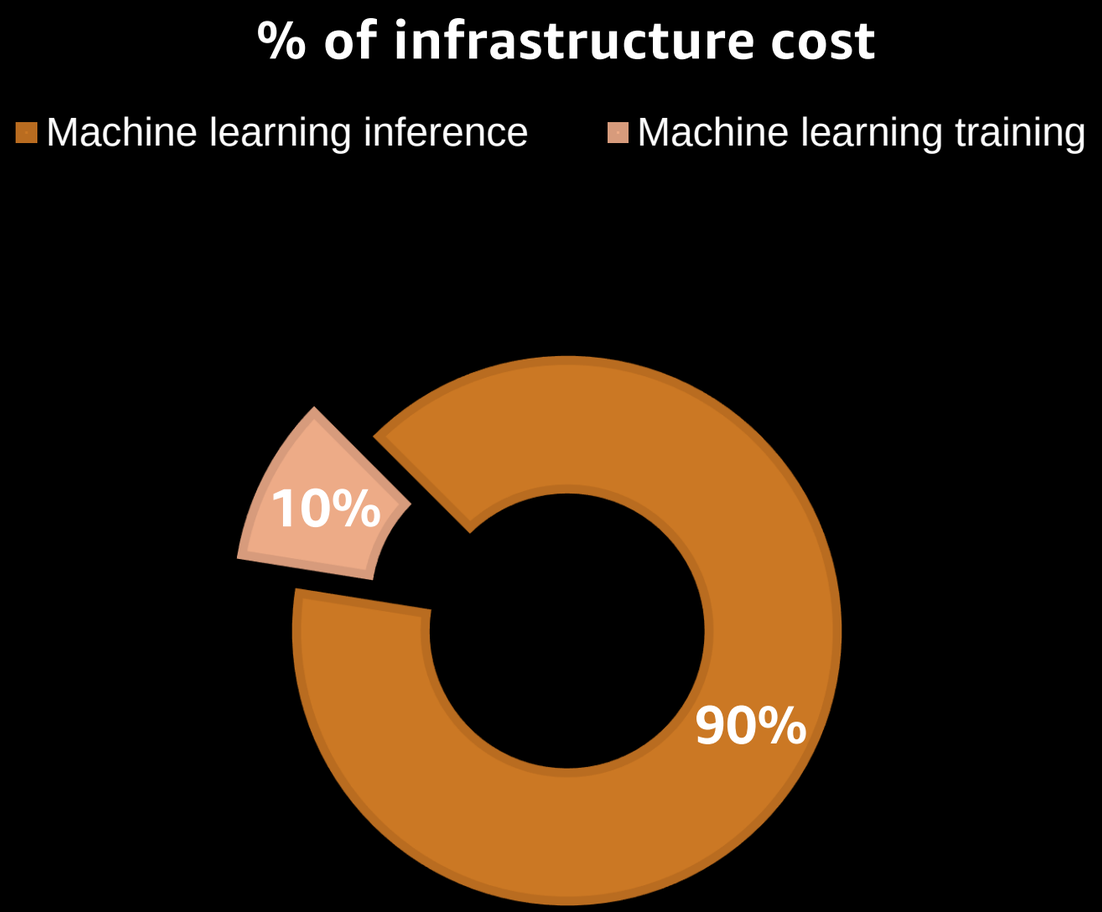
Tens of thousands of customers have chosen AWS for their machine learning workloads
More than twice as many customers use ML on AWS than on any other cloud provider



Inference dominates ML infrastructure costs

1. ML inference in production can consume up to 90% of total infrastructure costs

2. The infrastructure can become prohibitively more expensive to run as scientists strive to improve predictions by using larger, more complex models



AWS Inferentia



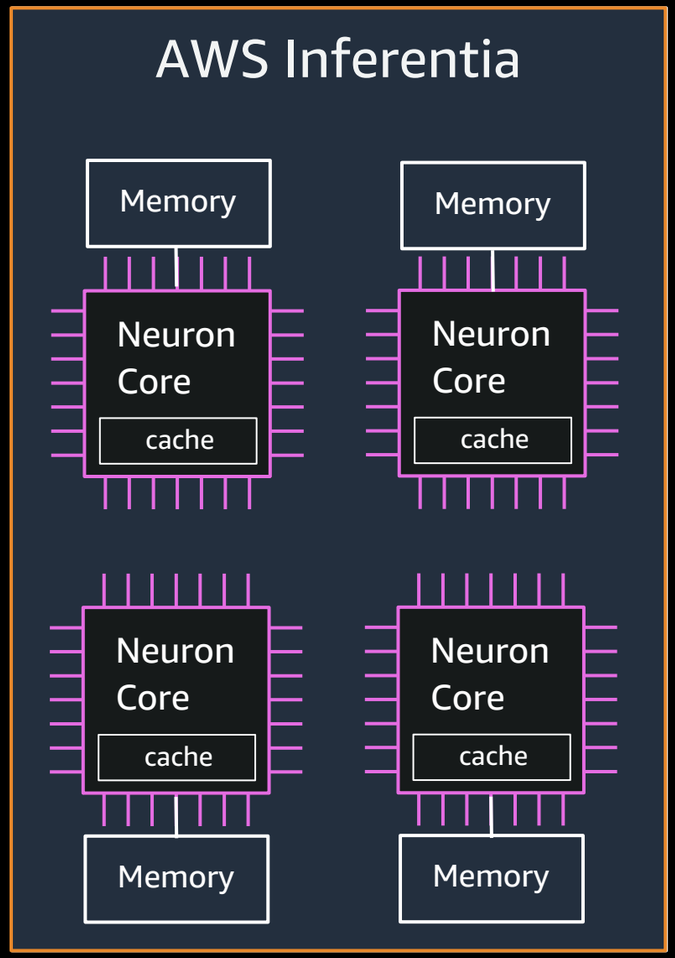
AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Inferentia overview

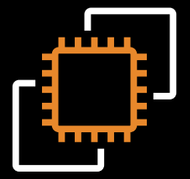
Inferentia is a custom ML chip designed and built by Amazon

- 4 NeuronCores
- Up to 128 TOPS per chip
- 2-stage memory hierarchy
  - Large on-chip cache and commodity DRAM
- Supports FP16, BF16, and INT8 data types with mixed precision
- Fast chip-to-chip interconnect

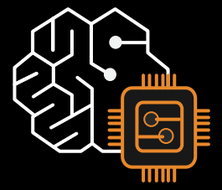


Inf1 instance purpose-built for ML inference

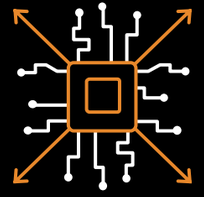
Instances are designed and built from the ground up by AWS



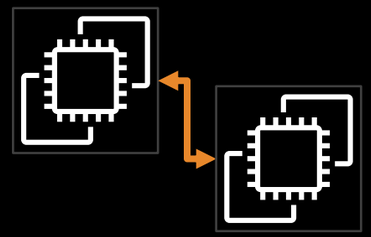
AWS Nitro



AWS Inferentia



AWS Custom 2nd Gen Intel Xeon Scalable processors



100Gbps networking



High performance

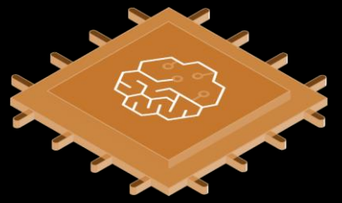


Low cost

Amazon EC2 Inf1 instances

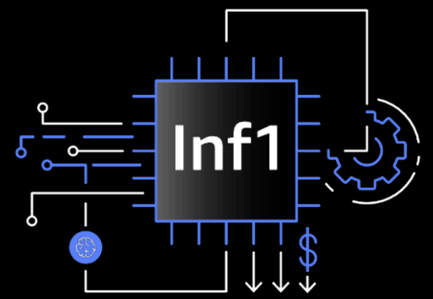
High-performance, lowest-cost AWS instance for machine learning inference

- Most affordable instance in the cloud for running deep learning models – up to 45% cheaper than GPU instances
- Get started quickly with minimal code changes – software seamlessly integrated with ML frameworks including TensorFlow, PyTorch, and MXNet
- Available in VMs, containers (Amazon ECS and Amazon EKS), and AWS SageMaker



AWS Inferentia

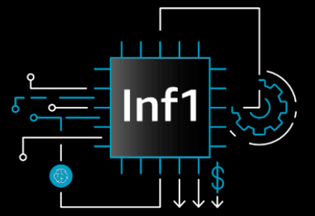
High performance machine learning inference chip, custom designed by AWS

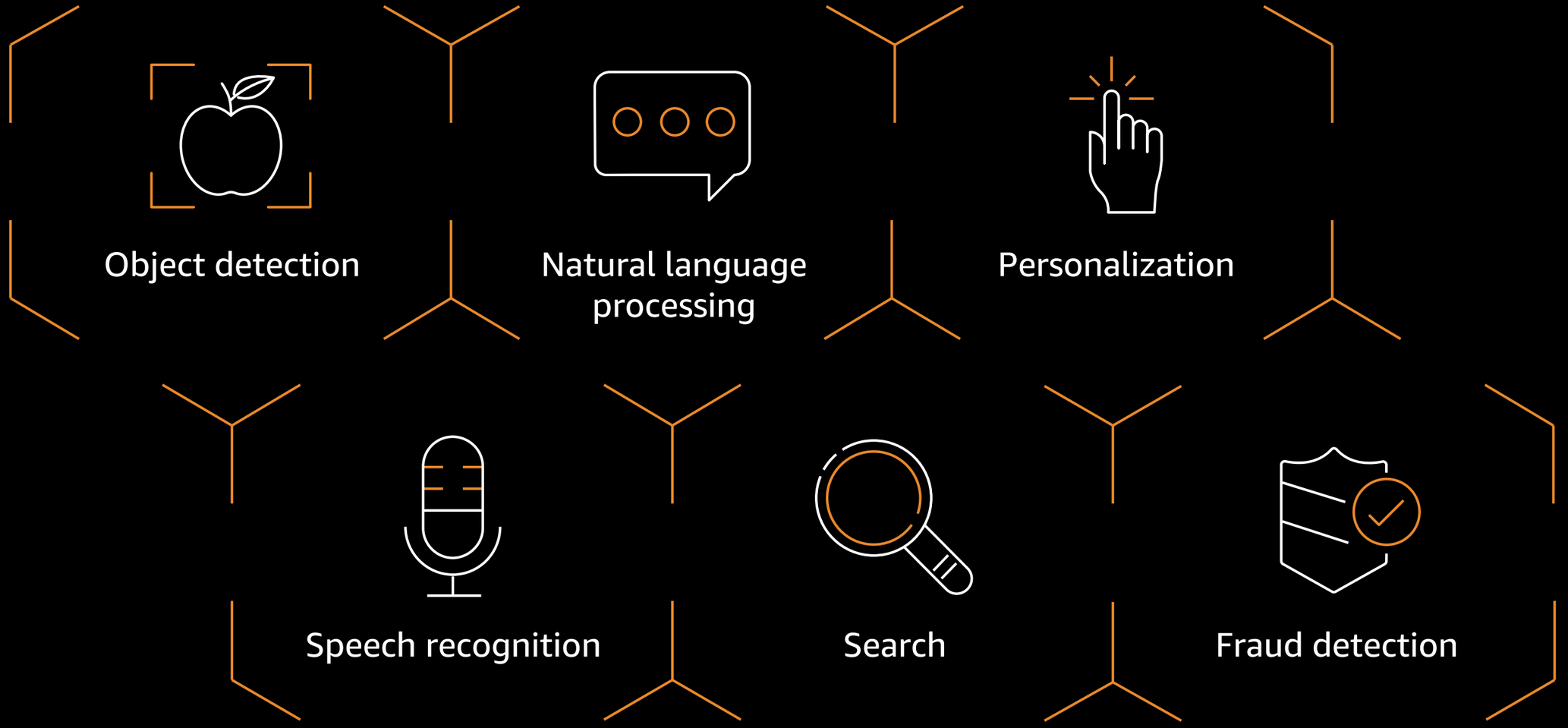


EC2 Inf1 instances

Fastest and lowest cost inference in the cloud

Broad set of supported use cases





Inf1 instances are available in four sizes

	vCPUs	VM memory (GIB)	Inferentia chips	Storage	Network B/W	EBS B/W
inf1.xlarge	4	8	1	EBS	Up to 25 Gbps	Up to 3.5 Gbps
inf1.2xlarge	8	16	1	EBS	Up to 25 Gbps	Up to 3.5 Gbps
inf1.6xlarge	24	48	4	EBS	25 Gbps	3.5 Gbps
inf1.24xlarge	96	192	16	EBS	100 Gbps	19 Gbps

- 1 to 16 Inferentia processors per instance
- Up to 192 GiB of memory

- AWS 2nd Gen Intel Xeon host processors
- Up to 100Gbps networking bandwidth

AWS Neuron SDK



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Neuron

High-performance software development kit (SDK)



Neuron Compiler



Neuron Runtime



Profiling tools



Easy to get started

Integrated with major frameworks

TensorFlow mxnet PyTorch

Flexibility of choice

Deploy existing models with minimal code changes.
Maintain hardware portability without dependence on AWS software

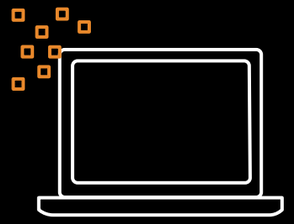


Documentation, examples, and support

github.com/aws/aws-neuron-sdk

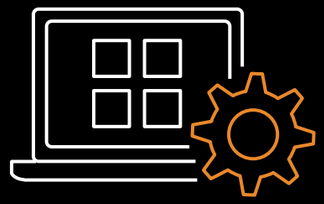
TensorFlow, the TensorFlow logo and any related marks are trademarks of Google Inc.

AWS Neuron key features



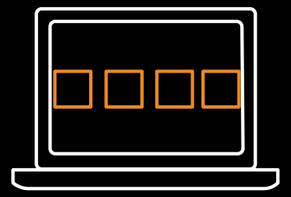
Smart partitioning

Automatically optimize neural-net compute



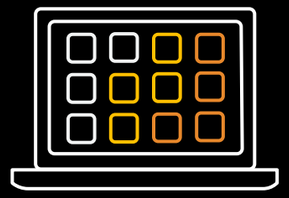
FP32 auto casting

Ingest FP32 trained models, and Neuron auto casts to BF16



NeuronCore pipeline

Super low-latency full bandwidth



NeuronCore groups

Concurrently run multiple models

Build and deploy using popular ML frameworks

AWS Neuron works with common ML frameworks

- Develop and optimize algorithms in your preferred framework
- Maintain portability between hardware platforms
- Compile pre-trained models for deployment on Inferentia chips via the Neuron SDK with just a few lines of code

```
# Import the AWS Neuron SDK
>> import tensorflow.neuron as tfn
# Compile the model using Neuron
>> library.tfn.saved_model.compile(model_dir, compiled_model_dir)
```



AWS Neuron is integrated in Amazon EKS, Amazon ECS, Kubernetes, and AWS Deep Learning AMIs. You can also install AWS Neuron in your own AMI or container

TensorFlow, the TensorFlow logo and any related marks are trademarks of Google Inc.

Inferentia support in AWS Managed Services

Amazon SageMaker is a managed service for building, training, and deploying machine learning models

- SageMaker is the fastest way to get started with Inf1 instances
- Integration of Inf1 instances and Neuron SDK provides one-click deployment of models onto Inf1



Amazon EKS and Amazon ECS

- Inf1 instances are available for use in Amazon EKS and Amazon ECS
- Ideal for deploying models on Inf1 instances using managed container orchestration or Kubernetes services.





Best price / performance for DL inference in the cloud

Object detection:

	Throughput (images/second)	Cost-per-1M inferences	Throughput: Inf1 vs. G4	Cost-per-inference: Inf1 vs. G4
inf1.2xlarge	118	$1.362	87% higher	37% lower
g4dn.xlarge	65	$2.163		

Results based on running YoloV4 model end-to-end with TensorFlow

Natural language:

	Throughput (sequences/second)	Cost-per-1M inferences	Throughput: Inf1 vs. G4	Cost-per-inference: Inf1 vs. G4
inf1.xlarge	360	$0.284	38% higher	49% lower
g4dn.xlarge	260	$0.562		

Results based on running BERT-Large model end-to-end with TensorFlow, seqlen=128

Image classification:

	Throughput (images/second)	Cost-per-1M inferences	Throughput: Inf1 vs. G4	Cost-per-inference: Inf1 vs. G4
inf1.xlarge	2,226	$0.045	24% higher	45% lower
g4dn.xlarge	1,792	$0.082		

Results based on running ResNet-50 model end-to-end with TensorFlow

Users achieve higher performance with Inf1



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Alexa Neural text-to-speech challenges

Responsiveness (low latency)

- React immediately to requests from users of 100M+ devices

High memory requirements

- Context generation is a memory heavy, sequence-to-sequence model
- Inference is bound by processor-to-memory bandwidth
- Model processes an entire sentence at once

Immense compute requirements

- 90 billion calculations used to generate each second of speech (90 GFLOPS)

Must scale at reasonable cost

- System scales to support requests from over 100M devices
- 2x performance improvement can cut by half the required number of compute instances

Alexa service overview



Utterance

Text to speech (TTS)

Alexa cloud service

Recognize

Recognition result

Automatic speech recognition

Recognition result

Intent

Natural language understanding

Intent

“Play” directive

Skill

Text

Output speech

Text to speech

Alexa Neural text-to-speech overview

Linguistic input representation

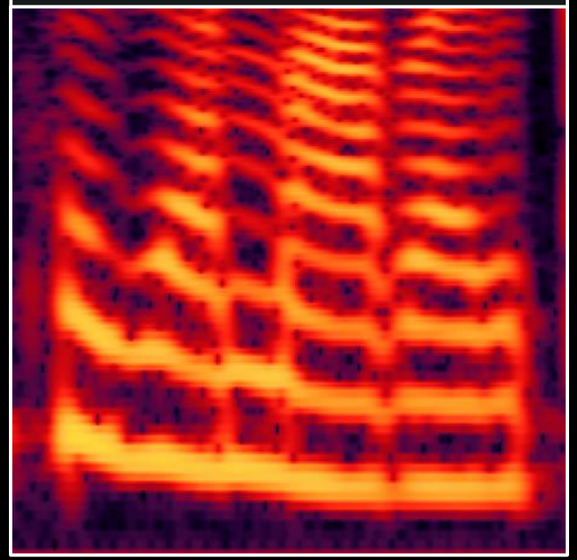
Mel spectrograms



Speech

Context generation

(Sequence-to-sequence network)



Speech production

(Neural vocoder)

- Focused on speech representation
- Speaker and style specific

- Focused on Hi-Fi speech signal
- Independent of speaker or style



Alexa text-to-speech

Alexa has deployed a highly complex text-to-speech model that generates human-like speech to support over 100 million Alexa devices globally. With Inf1 instances, the Alexa team has been able to lower their operating costs by about 65% over GPU instances, while achieving better inference latency.

Users selecting Inf1 instances for production









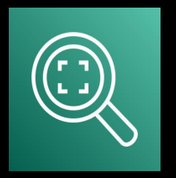












Amazon Rekognition







User momentum



"We incorporate machine learning (ML) into many aspects of Snapchat, and exploring innovation in this field is a key priority. Once we heard about Inferentia we started collaborating with AWS to adopt Inf1/Inferentia instances to help us with ML deployment, including **around performance and cost**. We started with our recommendation models, and look forward to adopting more models with the Inf1 instances in the future."

- Nima Khajehnouri, VP Engineering, Snap Inc.



"Autodesk is advancing the cognitive technology of our AI-powered virtual assistant, Autodesk Virtual Agent (AVA) by using Inferentia. AVA answers over 100,000 customer questions per month by applying natural language understanding (NLU) and deep learning techniques to extract the context, intent, and meaning behind inquiries. Piloting Inferentia, **we are able to obtain a 4.9x higher throughput over G4dn for our NLU models**, and look forward to running more workloads on the Inferentia-based Inf1 instances."

- Binghui Ouyang, Sr Data Scientist, Autodesk

User momentum



"The market of digital health platforms is growing at a remarkable rate. Gathering intelligence on this market is a challenging task due to the vast amounts of customer opinions data and its unstructured nature. Our application automates the generation of actionable insights from customer opinions via deep learning natural language models (Transformers). Our application is computationally intensive and needs to be deployed in a highly performant manner. We seamlessly deployed our deep learning inferencing workload onto Amazon EC2 Inf1 instances powered by the AWS Inferentia processor. The new **Inf1 instances provide 2X higher throughput compared to GPU-based instances** and allowed us to streamline our inference workloads."

– Numan Laanait PhD, Principal AI/Data Scientist
Miro Mihaylov PhD, Principal AI/Data Scientist



"Condé Nast's global portfolio encompasses over 20 leading media brands, including Wired, Vogue, and Vanity Fair. Within a few weeks, our team was able to integrate our recommendation engine with AWS Inferentia chips. This union enables multiple runtime optimizations for state-of-the-art natural language models on **SageMaker's** Inf1 instances. As a result**, we observed a 72% reduction in cost than the previously deployed GPU instances**."

– Paul Fryzel, Principal Engineer, AI Infrastructure

Amazon

amazon alexa

“Amazon Alexa’s AI and ML-based intelligence, powered by Amazon Web Services, is available on more than 100 million devices today – and our promise to customers is that Alexa is always becoming smarter, more conversational, more proactive, and even more delightful. Delivering on that promise requires continuous improvements in response times and machine learning infrastructure costs, which is why we are excited to use Amazon EC2 Inf1 to lower inference latency and cost-per-inference on Alexa text-to-speech. **With Inf1 instances, they have been able to lower their operating costs by about 65% over GPU instances**, while achieving better inference latency.”

– Tom Taylor, Senior Vice President, Amazon Alexa



Amazon Rekognition

“Amazon Rekognition is a simple and easy image and video analysis application that helps customer identify objects, people, text, and activities. Amazon Rekognition needs high-performance deep learning infrastructure that can analyze billions of images and videos daily for our customers. With AWS Inferentia-based Inf1 instances, running **Rekognition models such as object classification, resulted in 8X lower latency, and 2X the throughput compared to running these models on GPUs**. Based on these results we are moving Rekognition to Inf1, enabling our customers to get accurate results, faster.”

– Rajneesh Singh, Director, SW Engineering, Rekognition and Video

Try EC2 Inf1 instances today!

- Amazon EC2 Inf1 instances are generally available in 11 AWS Regions
- Available in VMs, containers (Amazon ECS and Amazon EKS), and AWS SageMaker
- Amazon EC2 Inf1 instances available with On-Demand, Savings Plans, and Spot pricing options

For more information, visit:

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

Thank you!



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Please complete the session survey

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.