AWS re:Invent

NOV. 29 – DEC. 3, 2021 | LAS VEGAS, NV

CMP304-R2

Amazon EC2 Inf1: High-performance and low-cost ML inference

Alex Iankoulski
Principal Solutions Architect, Autonomous Computing
AWS

Mahadevan Balasubramaniam
Principal Solutions Architect, Autonomous Computing
AWS

Hanish Garg
Senior Solutions Architect, State and Local Governments
AWS

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Optimize and deploy ML models on Amazon EC2 Inf1 instances at scale

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

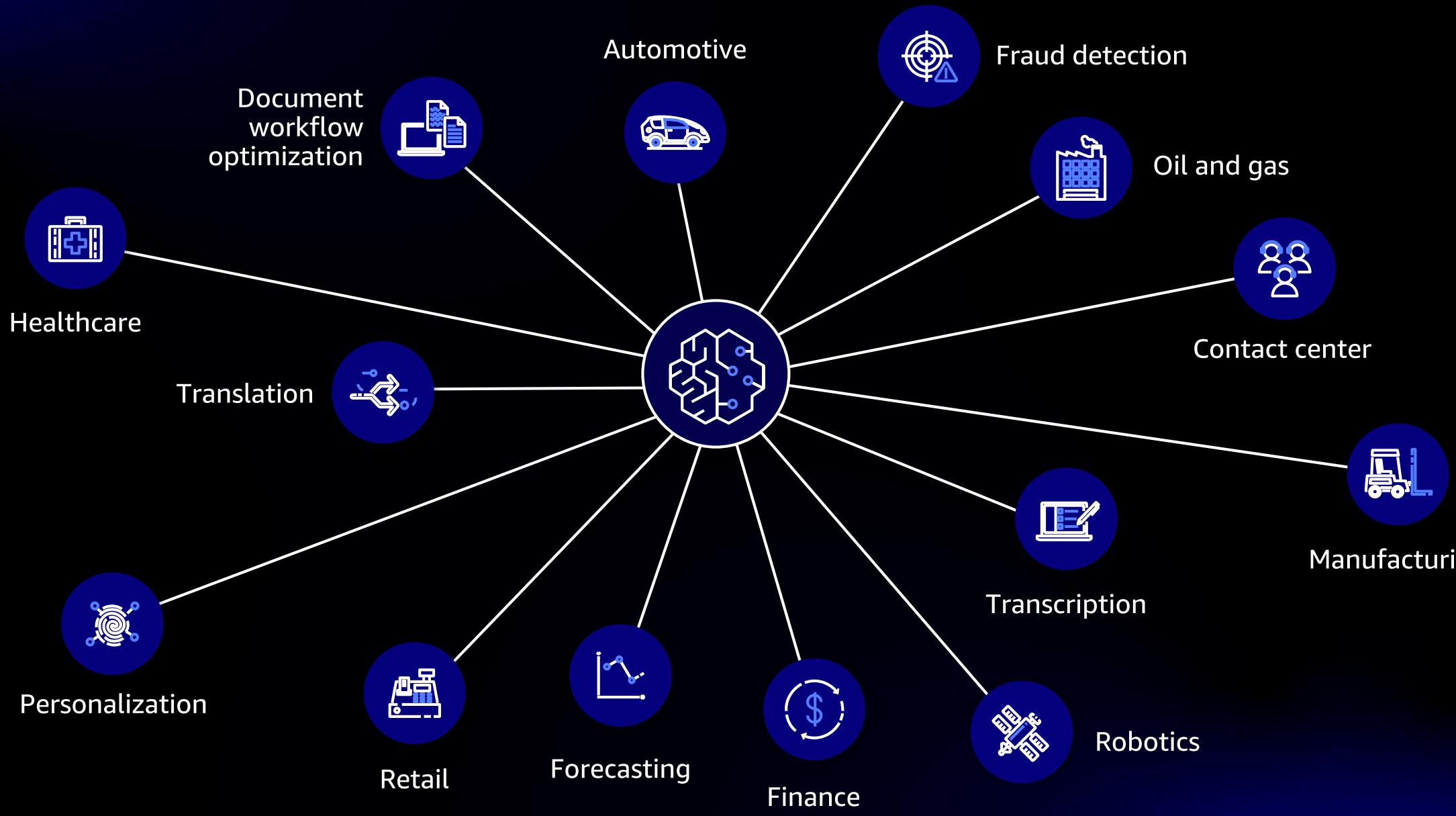
1. Focus - Amazon EC2 Inf1 Instances Overview (10 min)
2. Walk - Build and deploy single model on Inf1 EC2 instance (1 hr)
3. Run - Optimize model deployment using AWS Batch (1 hr)
4. Relax – Conclusion and awards (5 min)

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Focus: Amazon EC2 Inf1 Instances
Featuring Amazon Inferentia Chips

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The reach of ML is rapidly growing



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

ML is happening in companies of every size and in every industry

10,000+ CUSTOMERS | MORE ML HAPPENS ON AWS THAN ANYWHERE ELSE

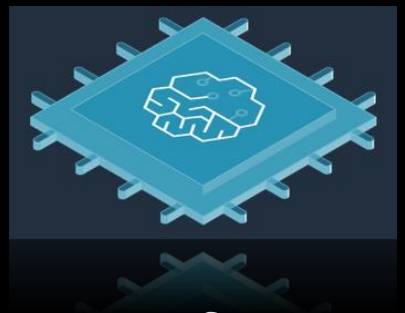


© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon EC2 Inf1 Instances

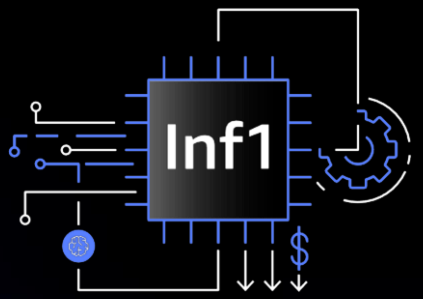
High-performance, lowest-cost Machine Learning Inference

- Featuring AWS Inferentia, the first ML chip designed by Amazon
- Lowest in the cloud for running Deep Learning Models
  - Up to 80% lower cost than GPU instances
- Seamless software Integration with ML frameworks like TensorFlow, PyTorch, and MXNet for quickly getting started & with minimal code changes
- Available through EC2, ECS, EKS, and Amazon SageMaker



AWS Inferentia

High performance machine learning inference chip, custom designed by AWS



EC2 Inf1 Instances

Fastest and lowest Cost Inference in the cloud

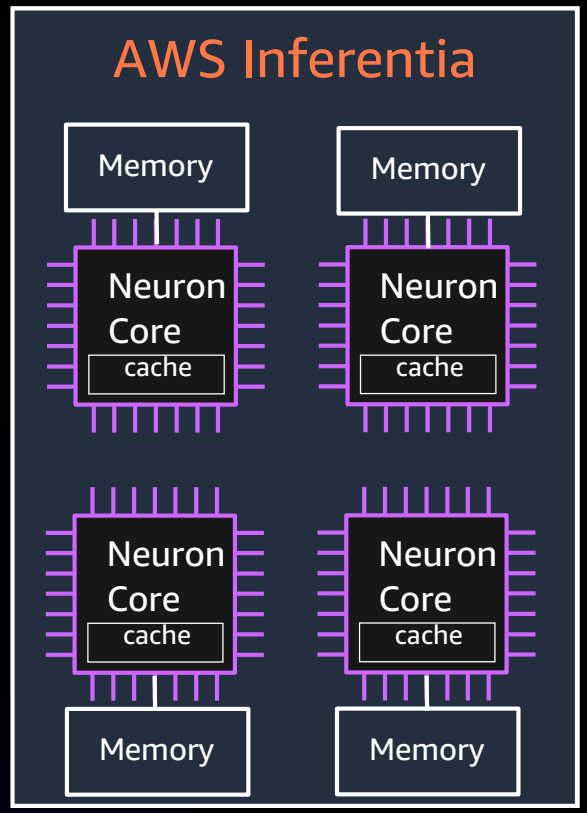
aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Inferentia quick tour

FIRST ML CHIP DESIGNED BY AMAZON

- 4 NeuronCores with up to 128 TOPS
- 2-stage memory hierarchy: large on-chip cache + 8 GB DRAM
- Supports FP16, BF16, INT8 data types with mixed precision
- 1 to 16 AWS Inferentia cores per instance with high-speed interconnect
- Optimized for high throughput and real-time low latency



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Inf1 Instance choices

Instance size	vCPUs	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 chips per instance
- Up to 192 GB of Memory

- AWS 2nd Gen Intel Xeon Host Processors
- Up to 100Gbps networking bandwidth

aws

© 2021, 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

PyTorch TensorFlow mxnet

Minimal Code Change

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



Documentation, Examples & Support

github.com/aws/aws-neuron-sdk

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Native integration with minimum code change

Run HuggingFace Transformers models on Inferentia with single line of code added

```
from transformers import BertTokenizer, BertModel
import torch
import torch_neuron

tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')
model = BertModel.from_pretrained('bert-base-uncased',return_dict=False)

inputs = tokenizer("Hello, my dog is cute", return_tensors="pt")

neuron_model = torch.neuron.trace(model,
                                  example_inputs = (inputs['input_ids'],inputs['attention_mask']),
                                  verbose=1)

outputs = neuron_model(*(inputs['input_ids'],inputs['attention_mask']))
```

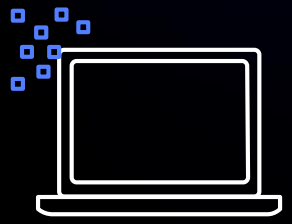


Sample code taken from docstring for BERT Base model: https://huggingface.co/transformers/model_doc/bert.html#bertmodel

aws

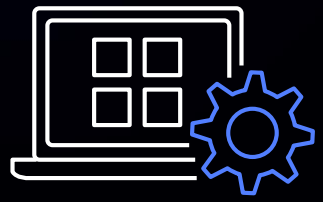
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Neuron highlights



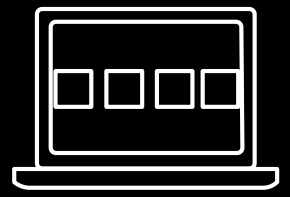
Smart partitioning

Automatically optimize neural-net compute



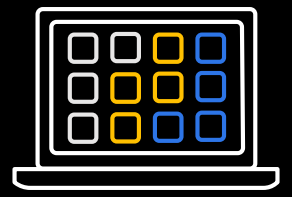
FP32 auto casting

Ingest FP32 trained models, and Neuron auto casts to BF16



NeuronCore pipeline

Real-time apps with bounded latency requirements at full bandwidth



NeuronCore groups

Concurrently run multiple models

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

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	185	$0.544	2.85x higher	76% lower
g4dn.xlarge	65	$2.248		

Neuron v1.16, fp32, PT 1.8, b1

Natural language:

PyTorch + 🤗	Throughput (sequences/second)	Cost-per-1M inferences	Throughput: Inf1 vs. G4	Cost-per-inference: Inf1 vs. G4
inf1.xlarge	1155	$0.055	2.74x higher	84% lower
g4dn.xlarge	422	$0.346		

Neuron v1.16, fp32, bert-base-uncased, sequence-length=128, PT 1.8, b6

Image classification:

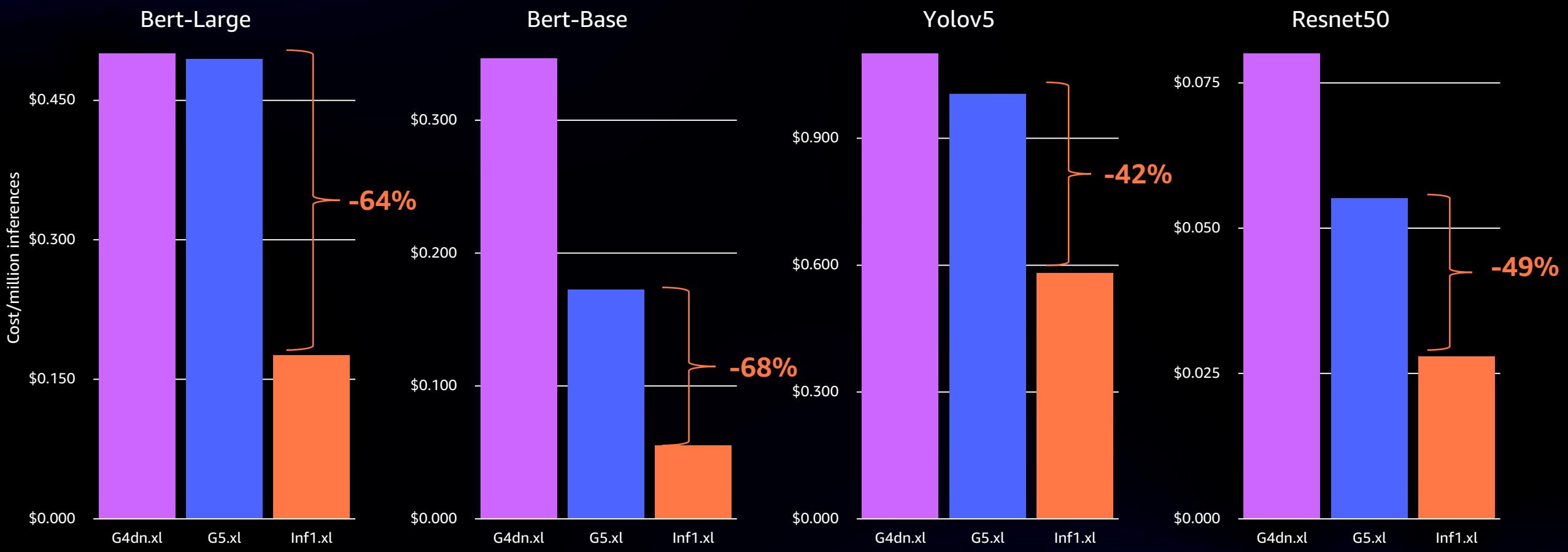
	Throughput (images/second)	Cost-per-1M inferences	Throughput: Inf1 vs. G4	Cost-per-inference: Inf1 vs. G4
inf1.xlarge	2,270	$0.028	27% higher	66% lower
g4dn.xlarge	1,792	$0.082		

Neuron v1.16, fp16, PT 1.8, b10

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Inf1 up to 68% lower cost than G5



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Customer momentum with AWS Inferentia

airbnb

AUTODESK

sprinklr

NTT PC COMMUNICATIONS

amazon robotics

amazon advertising

amazon alexa

The Asahi Shimbun

CONDÉ NAST

Talroo™

Anthem

DISCO

SKYWATCH

Hotpot.ai

Amazon Rekognition

DMP DIGITAL MEDIA PROFESSIONALS

BUNGEE

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Get started quickly with Inf1



- Amazon Sagemaker
  - Easiest and quickest way to get started with Inf1 instances
  - Amazon SageMaker is a fully managed service for building, training and deploying ML models quickly
  - Inf1 instances and Neuron are integrated in SageMaker to provide one-click deployment of models onto Inf1
- Amazon EKS & ECS
  - Inf1 instances are also available for use via Amazon EKS or ECS
  - Ideal for deploying models on Inf1 instances via managed containerized pipelines
- Amazon DLAMI & DL Containers
  - Neuron comes pre-installed in AWS Deep Learning AMIs, as well as in AWS Deep Learning Containers.
- Try Inf1 Today!
  - Available in 23 Regions today!
  - EC2 Inf1 instances support On-Demand, Reserved and Spot purchasing options; also available as part of Savings Plan
  - Inf1/Neuron Quick Start Guide
  - Inf1 roadmap & supported models

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Walk: Deploy ML model on EC2 Inf1 instance

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Run: Optimize ML model on EC2 Inf1 instances

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Relax: Conclusion

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Takeaways

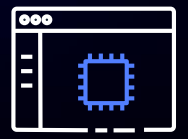
1. EC2 Inf1 instances offer the best price/performance for ML inference in the cloud

2. To run a model on Inferentia, trace it using the Neuron SDK, then build a container and deploy it

3. To maximize model performance and minimize cost, use AWS Batch to experiment with different instance types and batch sizes

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Resources



AWS Neuron SDK Documentation

Neuron Containers Documentation

Github Tutorials/Project

PyTorch

Getting Started with Torch Neuron



AWS Neuron forum & blogs

TensorFlow

Getting Started with TF Neuron

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Awards

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Alex Iankoulski

https://github.com/iankouls-aws

https://www.linkedin.com/in/alex-iankoulski/

Mahadevan Balasubramaniam

https://github.com/mahadevan-aws

https://www.linkedin.com/in/bmdevan/

Hanish Garg

https://github.com/hanish-garg

https://www.linkedin.com/in/hanishgarg/

Fabio Nonato de Paula

https://github.com/nonatofabio

https://www.linkedin.com/in/fabiononato

Scott Perry

https://github.com/5cp

https://www.linkedin.com/in/scott-perry-yyc

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.