# INNOVATE

FOR EVERY APPLICATION EDITION

25 August, 2022

# AWS Silicon Innovations

## Pushing price performance boundaries and providing more choices for compute

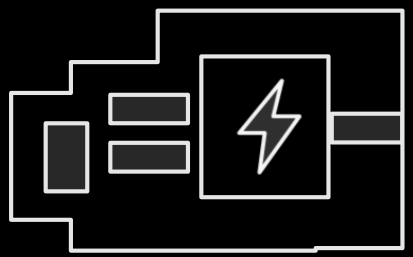
Bhushan Desam

Principal Compute Specialist
Amazon Web Services

# Today's agenda

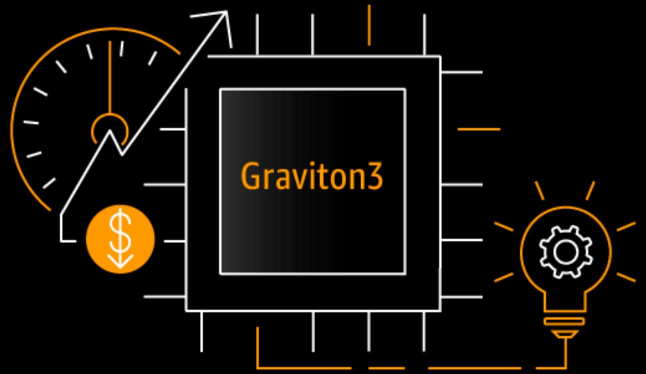
- AWS silicon innovations
- AWS Graviton processor & Amazon EC2 Graviton instances
- Customer adoption, workloads, and software ecosystem for Graviton
- AWS Inferentia and AWS Trainium for ML workloads
- Amazon EC2 Inf1 instances and customer adoption
- High-perf ML training with Amazon EC2 Trn1 instances
- Key takeaways and resources

# Silicon innovation at AWS



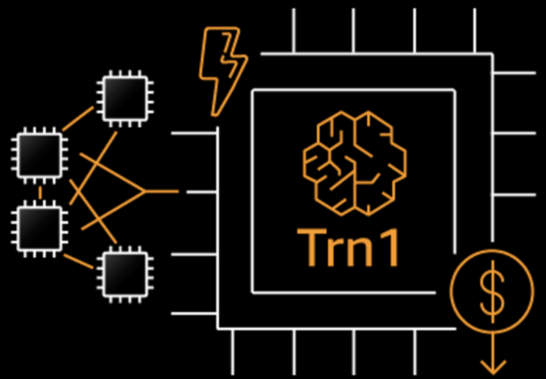
**AWS Nitro System**
**AWS Nitro SSD**

Hypervisor, network, storage, and security




**AWS Graviton2**
**AWS Graviton3**

Powerful and efficient, modern applications




**AWS Inferentia**
**AWS Trainium**

Machine learning acceleration

# Why build our own chips?

Innovation

Specialization

Speed

Operations

# AWS Graviton

Enabling the best price performance in Amazon EC2

# AWS Graviton processors

BEST PRICE PERFORMANCE IN AMAZON EC2

**Graviton**

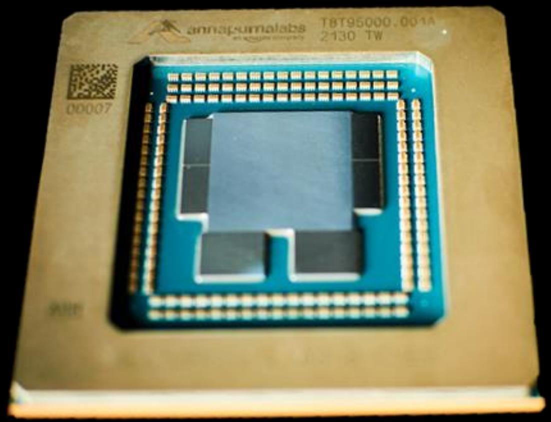
**2018**

**Graviton2**

**2019**


**Graviton3**

**2021**




aws

# AWS Graviton: Broad workload applicability

Web and gaming servers

Open-source databases

High performance computing

In-memory caches

Media encoding

Electronic design automation

Analytics

Microservices

# AWS Graviton2-based Amazon EC2 instances

BEST PRICE-PERFORMANCE IN EC2

## M6g, M6gd

General purpose workloads

## T4g

Burstable general purpose workloads

## C6g, C6gd, C6gn

Compute-intensive workloads

## R6g, R6gd, X2gd

Memory-intensive workloads

## Im4gn, Is4gen

Storage-intensive workloads

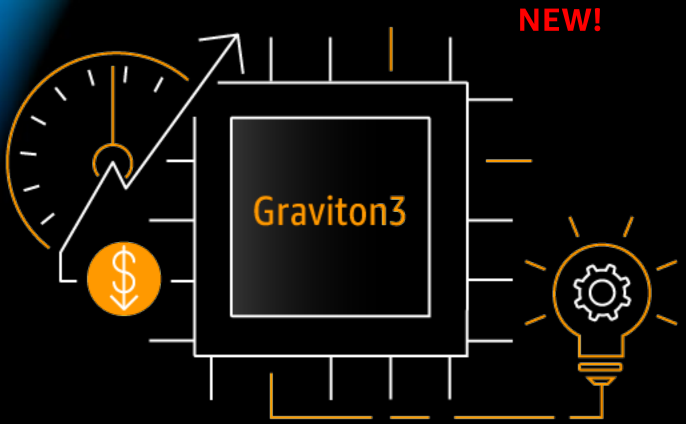
## G5g

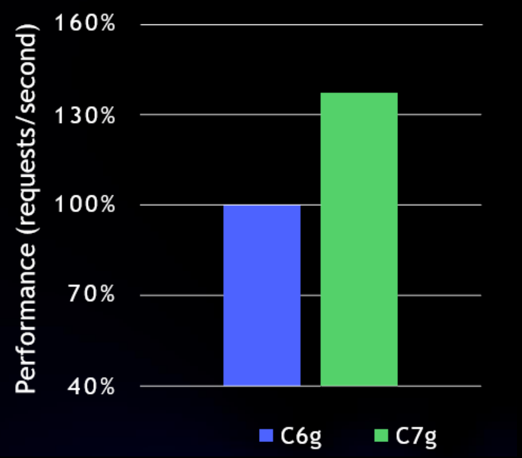
GPU-based graphics and machine learning workloads

AVAILABLE ACROSS 23 AWS REGIONS GLOBALLY

# Customers adopting AWS Graviton2

# AWS Graviton3 and Amazon EC2 C7g instances

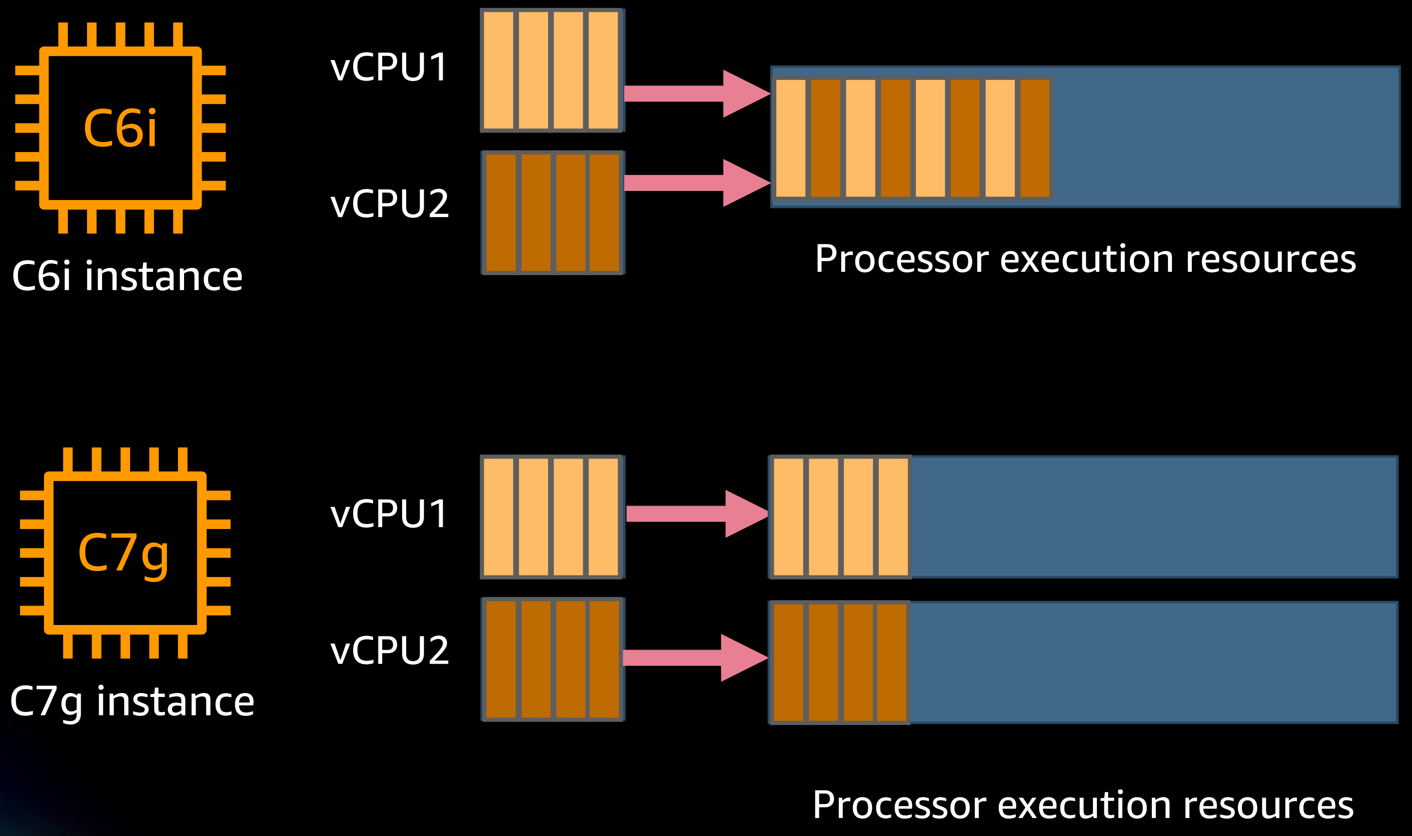



- Up to 25% better performance compared to Graviton2
- Up to 2x higher floating-point performance, up to 2x faster cryptographic workload performance, and up to 3x better machine learning performance compared to Graviton2
- First in the cloud to feature DDR5 memory
- 60% more energy efficient over comparable EC2 instances
- C7g instances will provide the best price performance for compute-intensive workloads in Amazon EC2

Node.JS 16.7.0, AcmeAir test application with one process per vCPU, JMeter load generator on c6g.4xlarge in a cluster placement group with NGiNX as reverse proxy, HTTP connections, connection count varied to control load

# AWS Graviton3 – vCPU




Every vCPU is a physical core

No simultaneous multithreading (SMT)

# What customers are saying about Amazon EC2 C7g instances

"We found Graviton3-based C7g instances deliver 20-80% higher performance vs. Graviton2 based C6g instances, while also reducing tail latencies by as much as 35%. "

"We were able to run 30% fewer instances of C7g than C6g serving the same workload, and with 30% reduced latency."

"They are suitable for even the most demanding latency sensitive workloads while providing significant price performance benefits."

"We have now found Graviton3 C7g instances to be 40% faster than the Graviton2 C6gn instances for those same simulations."

https://aws.amazon.com/ec2/instance-types/c7g/

# AWS managed services supporting Graviton

EXTENDING THE GRAVITON2 PRICE PERFORMANCE TO MANAGED SERVICES

## Databases

- Amazon DocumentDB
- Amazon Aurora
- Amazon RDS
- Amazon Elasticache
- Amazon MemoryDB
- Amazon Neptune

## Analytics

- Amazon OpenSearch
- Amazon EMR

## Compute

- AWS Lambda
- AWS Fargate
- AWS Elastic Beanstalk

## Storage

- Amazon FSx for Lustre, Amazon FSx for OpenZFS

https://github.com/aws/aws-graviton-getting-started/blob/main/managed_services.md

# Use AWS Graviton directly on AWS

**Operating Systems**

- Amazon Linux 2
- Red Hat Enterprise Linux 8.2+
- SLES 15 SP2+
- ubuntu 18.04LTS, 20.04LTS
- CentOS
- Debian
- fedora
- FreeBSD
- NetBSD

**Containers**

- Amazon Elastic Container Service (Amazon ECS)
- AWS Fargate
- Amazon Elastic Container Service for Kubernetes (Amazon EKS)
- Docker
- Kubernetes

# AWS Graviton Ready

CERTIFIED PARTNER SOLUTIONS FOR GRAVITON CUSTOMERS

aqua
aviatrix
Buildkite
CANONICAL
circleci
Cribl
DATADOG
docker
dynatrace
elastic
Genymotion Device image
GitLab
GRANULATE
honeycomb.io
InterSystems Creative data technology
Lacework
New Relic
Qualys.
Red Hat
SUSE
Starburst
tenable
TREND MICRO

https://aws.amazon.com/ec2/graviton/partners/

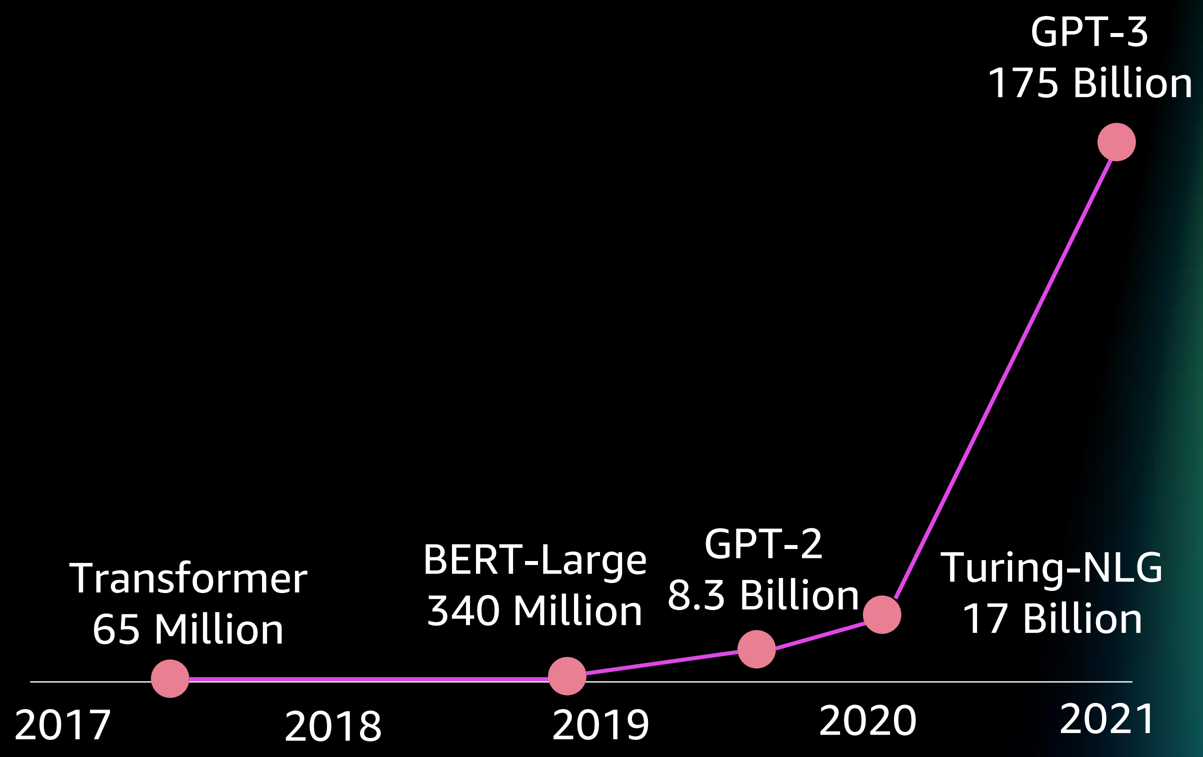
aws

# AWS Inferentia and AWS Trainium

## High performance machine learning chips

# Key trends in AI/ML infrastructure

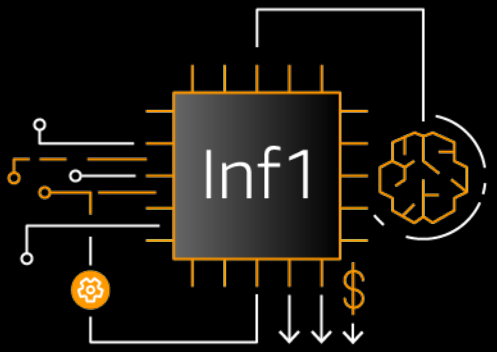
- Models are becoming more complex, with end users moving from classical ML to deep learning
- Time and cost to train these models have exploded from days to weeks
- Similar challenges with running these models in production, serving customer traffic




**ML model size**

# AWS chips optimized for deep learning

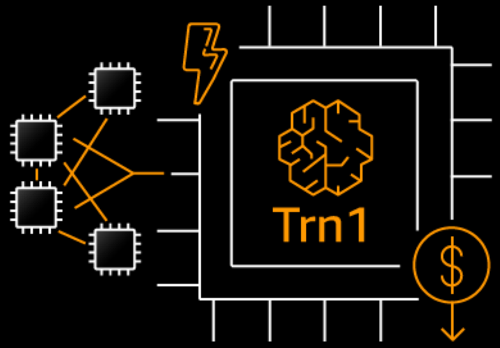
## AWS Inferentia




**Lowest cost in the cloud** for running deep learning models – up to 80% lower cost than GPU instances


Seamless **integration with ML frameworks** like TensorFlow and PyTorch with minimal code changes

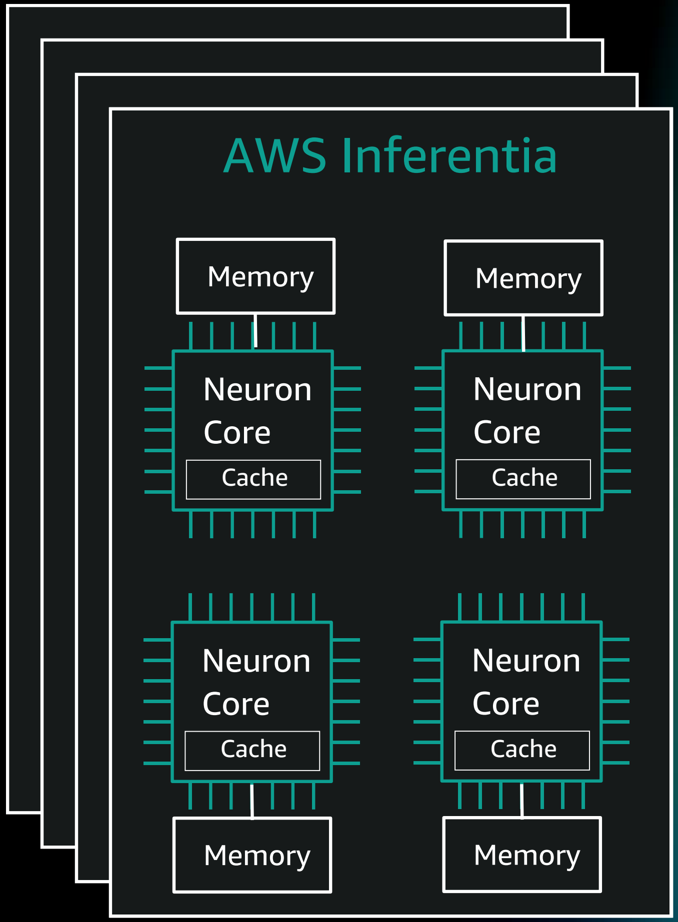
## AWS Trainium




Purpose-built for the **most cost-efficient and fast DL training in the cloud** for a broad spectrum of applications

# AWS Inferentia

- 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 Inferentia chips per instance with high-speed interconnect
- Optimized for high throughput and real-time low latency

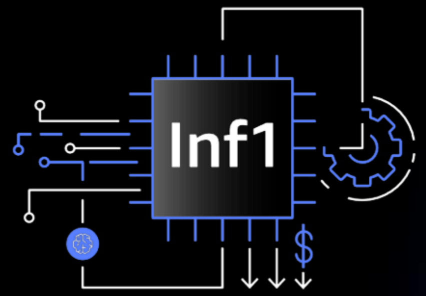


# Amazon EC2 Inf1 Instances

## HIGH-PERFORMANCE, LOWEST-COST MACHINE LEARNING INFERENCE

- Featuring AWS Inferentia, the first ML chip designed by Amazon
- Lowest cost 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 VMs, Containers, Kubernetes, 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

# Customer momentum with AWS Inferentia

airbnb
AUTODESK
Finch COMPUTING
sprinklr

NTT PC COMMUNICATIONS
prime video
amazon advertising
amazon alexa
SCREENING EAGLE

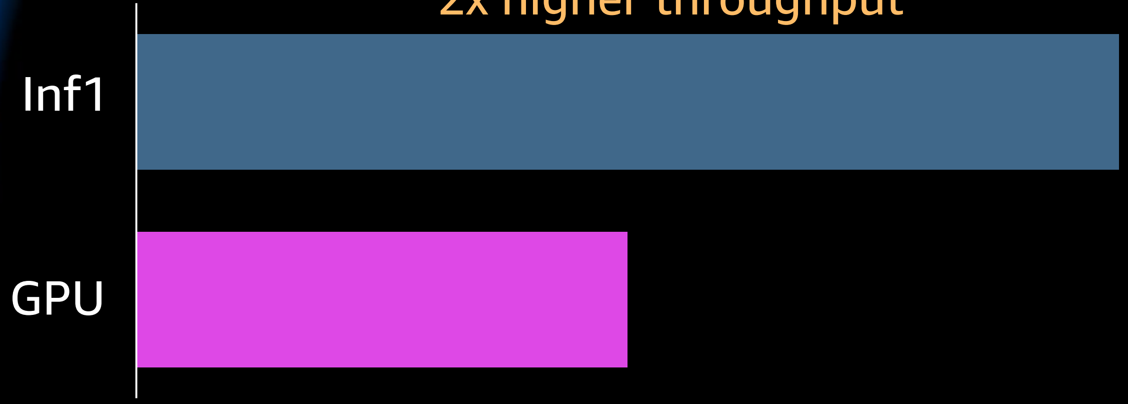
The Asahi Shimbun
CONDÉ NAST
Talroo™
Anthem.
DISCO

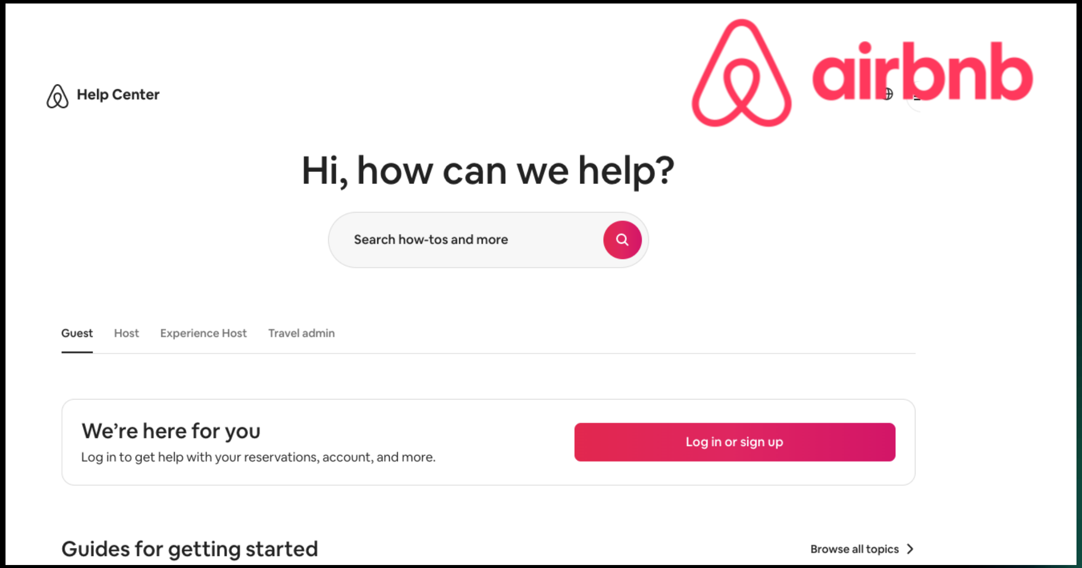
SKYWATCH
Hotpot.ai
Amazon Rekognition
DMP DIGITAL MEDIA PROFESSIONALS
robotics

# 2x higher throughput over GPU instances

**PyTorch BERT throughput**

2x higher throughput




"Airbnb's Community Support Platform enables intelligent, scalable, and exceptional service experiences to our community of millions of guests and hosts around the world. We are constantly looking for ways to improve the performance of our Natural Language Processing models that our support chatbot applications use. With Amazon EC2 Inf1 instances powered by AWS Inferentia, we see a 2x improvement in throughput out of the box, over GPU-based instances for our PyTorch based BERT models. We look forward to leveraging Inf1 instances for other models and use cases in the future."

Bo Zeng, Engineering Manager, Airbnb

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

aws

# 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

AWS Neuron

## Minimal Code Change

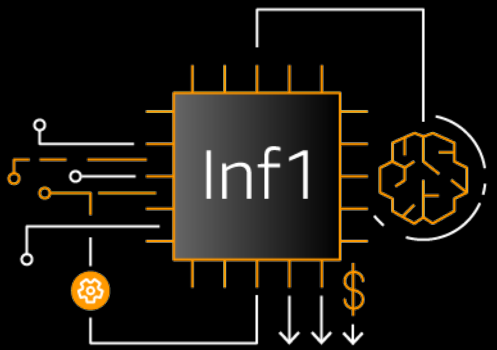
Deploy existing models with minimal code changes.
Maintain hardware portability without dependency on AWS software

Documentation, Examples & Support

github.com/aws/aws-neuron-sdk

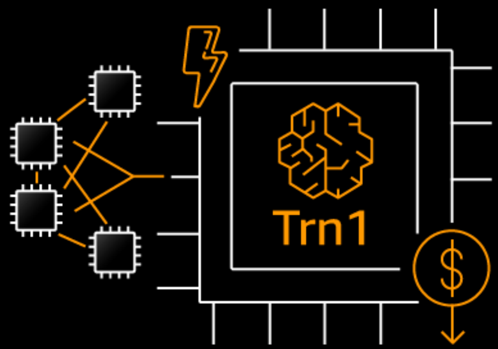
# AWS chips optimized for deep learning

## AWS Inferentia





## AWS Trainium




**Lowest cost in the cloud** for running deep learning models – up to 80% lower cost than GPU instances

Seamless **integration with ML frameworks** like TensorFlow and PyTorch with minimal code changes

Purpose-built for the **most cost-efficient and fast DL training in the cloud** for a broad spectrum of applications

# Amazon EC2 Trn1 instances

## THE MOST COST-EFFICIENT DL INSTANCE IN THE CLOUD

**Trn1**

MATH ENGINE FREQUENCY
**3 GHz**

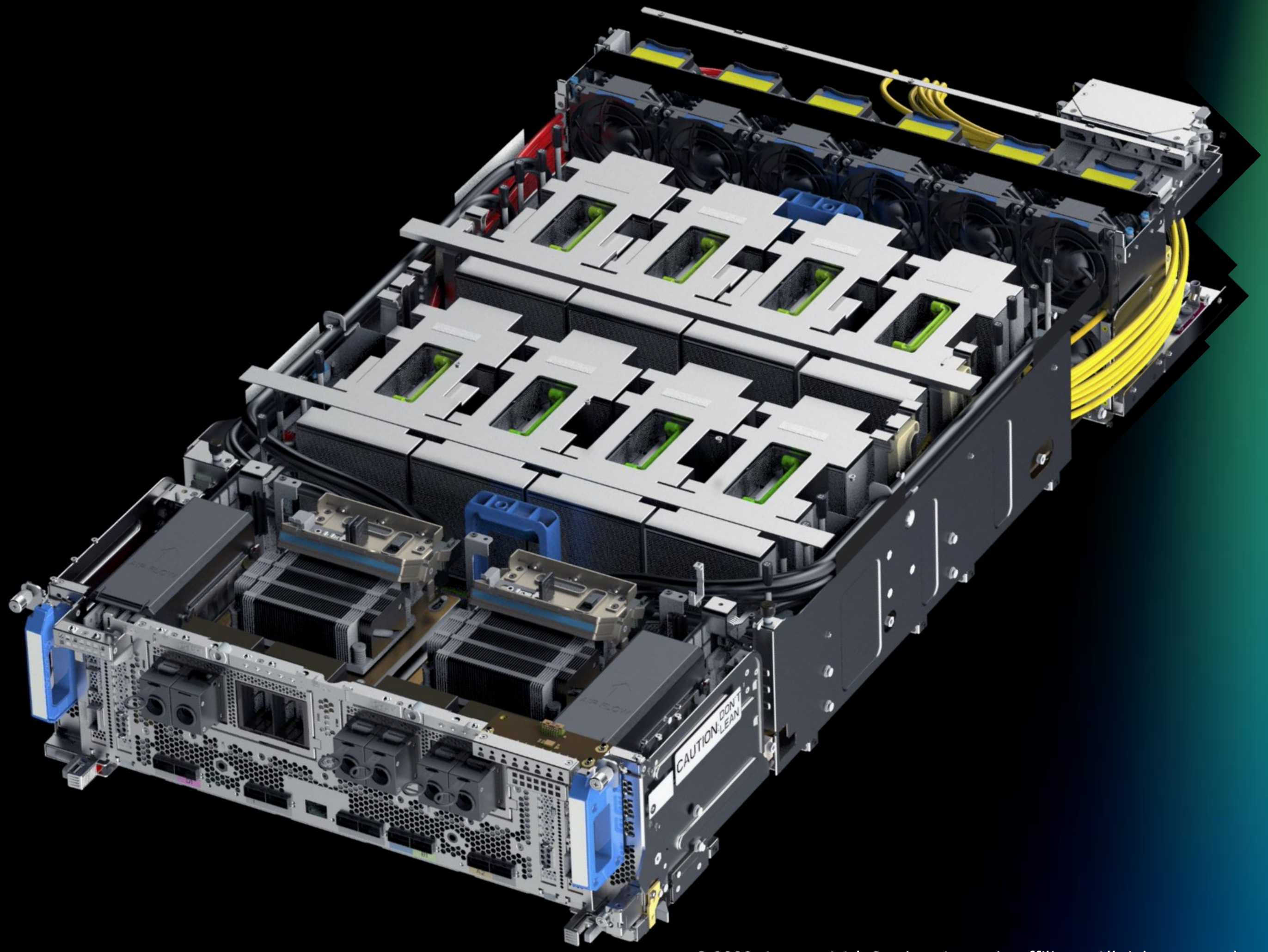
| BF16/FP16 | TF32 | FP32 |
|---|---|---|
| **3.4 PFLOPS** | **3.4 PFLOPS** | **840 TFLOPS** |

AGGREGATE ACCELERATOR MEMORY
**512 GB**

PEAK MEMORY BANDWIDTH
**13.1 TB/sec**

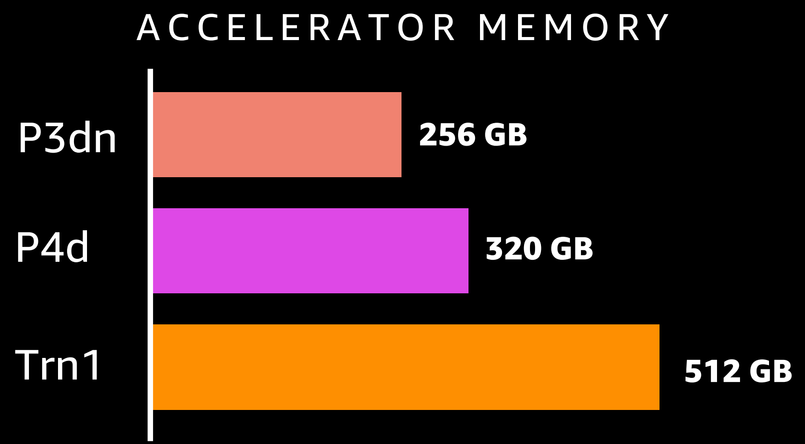
NEURONLINK BANDWIDTH BETWEEN CHIPS
**768 GB/sec**

NETWORK CONNECTIVITY
**800 Gbps EFA**

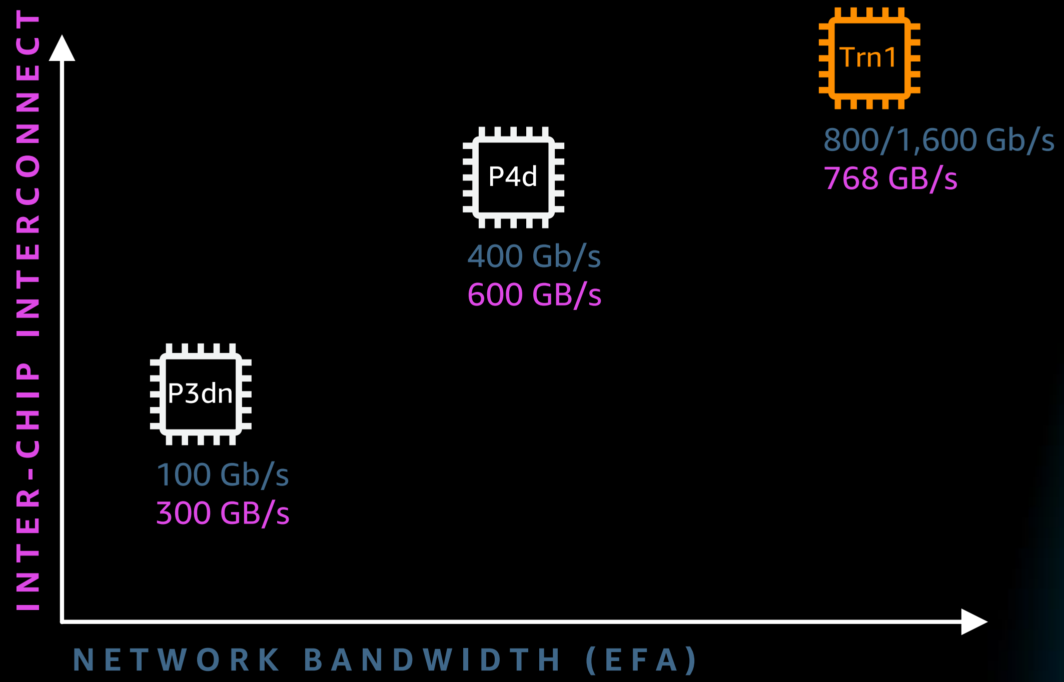


aws

# Trn1 built for scale-out




Large in-server accelerator memory




High-bandwidth and low-latency interconnect

"A major key to our success is access to modern infrastructure that allows us to spin up very large fleets of high-performance deep learning accelerators. We are looking forward to using Amazon EC2 Trn1 instances powered by AWS Trainium, as their unprecedented ability to scale to tens of thousands of nodes and higher network bandwidth will enable us to iterate faster while keeping our costs under control."

**Tom Brown**
Co-Founder at Anthropic

ANTHROP\C

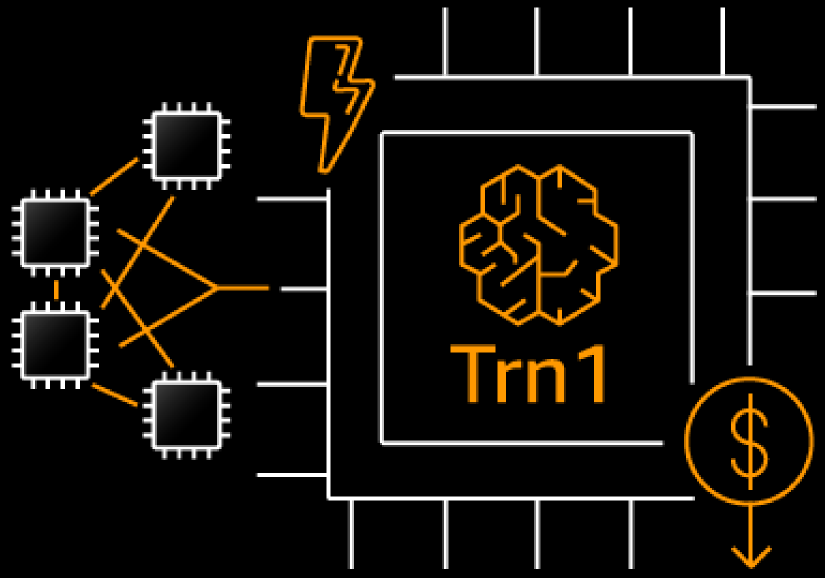
https://aws.amazon.com/ec2/instance-types/trn1/

aws

# Amazon EC2 Trn1 instances

THE MOST COST-EFFICIENT DL TRAINING IN THE CLOUD

In preview **now** –
general availability in 2022




60% higher accelerator memory (vs. P4d)

2x network bandwidth (vs. P4d)

Native support for PyTorch and TensorFlow

Train on Trn1 and deploy anywhere

| Instance size* | vCPUs | AWS Trainium chips | Accelerator memory | NeuronLink | Instance memory | Instance networking | Local instance storage |
|---|---|---|---|---|---|---|---|
| Trn1.2xlarge | 8 | 1 | 32 GB | N/A | 32 GB | Up to 10Gbps | 500 GB NVMe |
| Trn1.32xlarge | 128 | 16 | 512 GB | **768 GB/sec** | 512 GB | **800 Gbps** | 8 TB NVMe |

# AWS portfolio integration

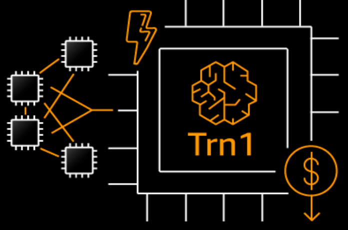
**Frameworks & workflow services**

- Pytorch
- TensorFlow
- Amazon SageMaker
- AWS Deep Learning AMIs
- Amazon EKS
- Amazon ECS
- AWS Deep Learning Containers

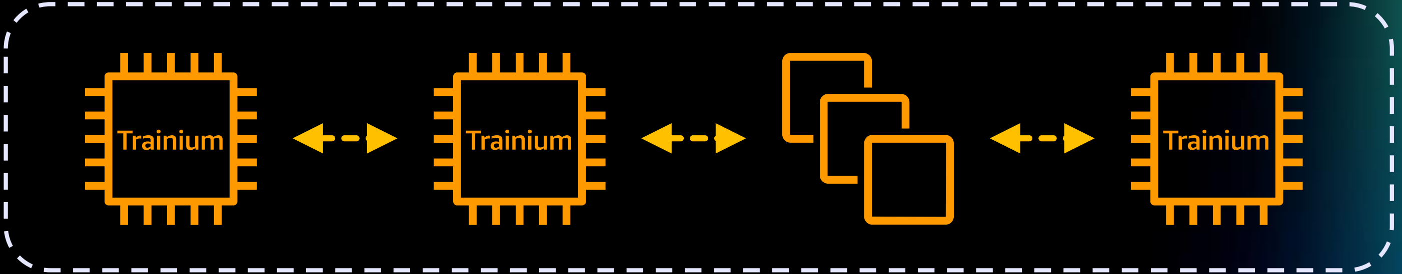
**Networking & storage**

- Elastic Fabric Adapter
- Amazon S3
- Amazon EBS
- Amazon FSx for Lustre
- Amazon EFS

**Compute acceleration**




**Amazon EC2 Trn1**




**EC2 Trn1 UltraCluster**

# Recap

- Amazon EC2 Graviton instances provide **better price-performance** over comparable x86-based instances and support **a broad spectrum of workloads**

# Recap

- Amazon EC2 Graviton instances provide **better price-performance** over comparable x86-based instances and support **a broad spectrum of workloads**
- Amazon EC2 Inf1 instances, powered by AWS Inferentia processors, **are the lowest cost in the cloud** for running deep learning inference

# Recap

- Amazon EC2 Graviton instances provide **better price-performance** over comparable x86-based instances and support **a broad spectrum of workloads**
- Amazon EC2 Inf1 instances, powered by AWS Inferentia processors, **are the lowest cost in the cloud** for running deep learning inference
- Amazon EC2 Trn1 instance are purposely built for the **most cost-efficient and fast deep learning training in the cloud** for a broad spectrum of applications

# Other resources

## AWS Graviton

**Getting started guide**
https://github.com/aws/aws-graviton-getting-started

**Performance runbook**
https://github.com/aws/aws-graviton-getting-started/blob/main/perfrunbook/graviton_perfrunbook.md

**Blogs, testimonials, and other resources**
https://aws.amazon.com/ec2/graviton/

## AWS Inferentia

**AWS Neuron SDK documentation**
https://awsdocs-neuron.readthedocs-hosted.com/en/latest/neuron-intro/get-started.html

**Github tutorials/projects**
https://github.com/aws/aws-neuron-sdk

# Visit the AWS resource hub

Start building upon a scalable, reliable, and globally available infrastructure so that you can focus on innovation and bringing new applications to market. Dive deeper with these resources today.

- Accelerate innovation with AWS
- Get more performance for your applications at lower costs with AWS
- Global-scale solutions
- How startups succeed with AWS

https://tinyurl.com/for-every-app-hub-aws

**Visit resource hub**

# AWS Training and Certification

## Self – Paced Digital Training on AWS

Explore learning plans and 500+ digital courses from our new learning center, AWS Skill Builder, to help you achieve your goals on your schedule.

bit.ly/3lzVj0g

## AWS Certification

Validate technical skills and cloud expertise to grow your career and business.

go.aws/3PwN3ff

# Thank you for attending AWS Innovate – For Every Application Edition

We hope you found it interesting! A kind reminder to **complete the survey.**
Let us know what you thought of today's event and how we can improve the event experience for you in the future.

aws-apj-marketing@amazon.com

twitter.com/AWSCloud

facebook.com/AmazonWebServices

youtube.com/user/AmazonWebServices

slideshare.net/AmazonWebServices

twitch.tv/aws

# Thank you!