
AWS
re:Invent
NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

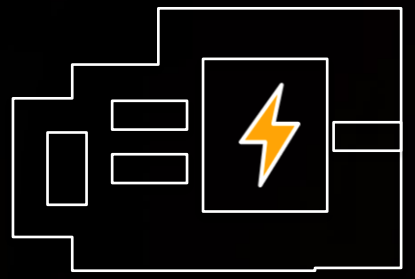
CMP201

Silicon innovation at AWS

Ali Saidi (he/him)

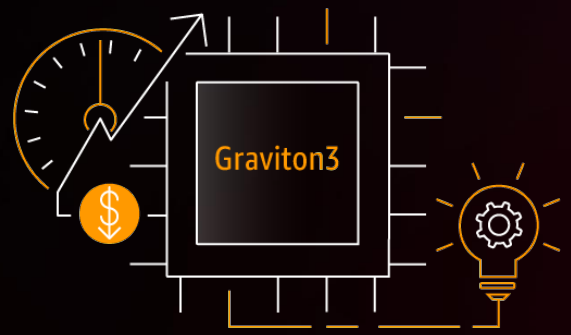
Sr. Principal Engineer
AWS

Journey of silicon innovation at AWS



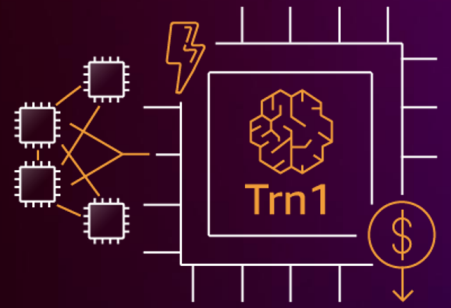
AWS Nitro System

Hypervisor, Nitro Cards, network, storage, SSD, and security



AWS Graviton

Powerful and efficient compute

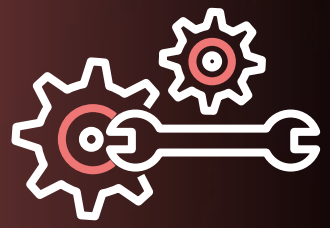


AWS Inferentia and AWS Trainium

Machine learning acceleration

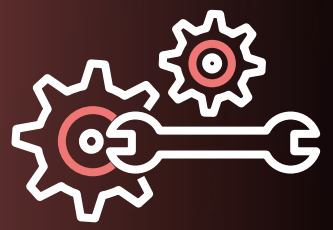
Why build our own chips?

Why build our own chips?



Specialization

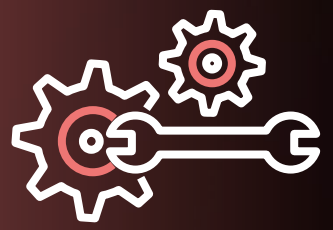
Why build our own chips?



Specialization

Speed

Why build our own chips?

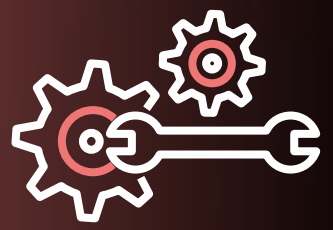


Specialization

Speed

Innovation

Why build our own chips?



Specialization

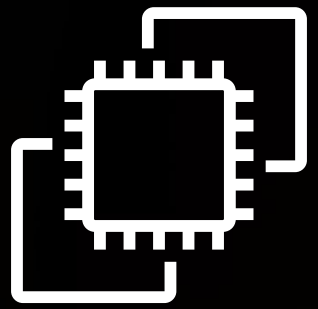
Speed

Innovation

Security

AWS Nitro System

AWS Nitro System



AWS Nitro System

If we applied all of our learnings, how would we change our server platforms?

Improve throughput

Simplify hypervisor

Reduce latency & jitter

Bare-metal instances

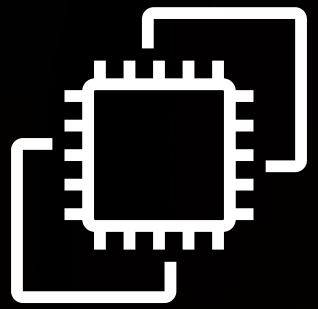
Transparent encryption

Hardware root of trust

No operator access

Narrow & auditable APIs

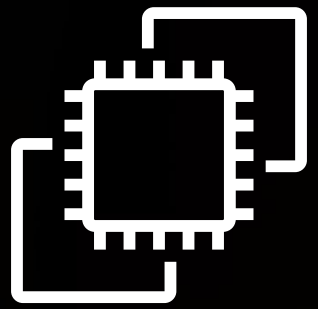
AWS Nitro System



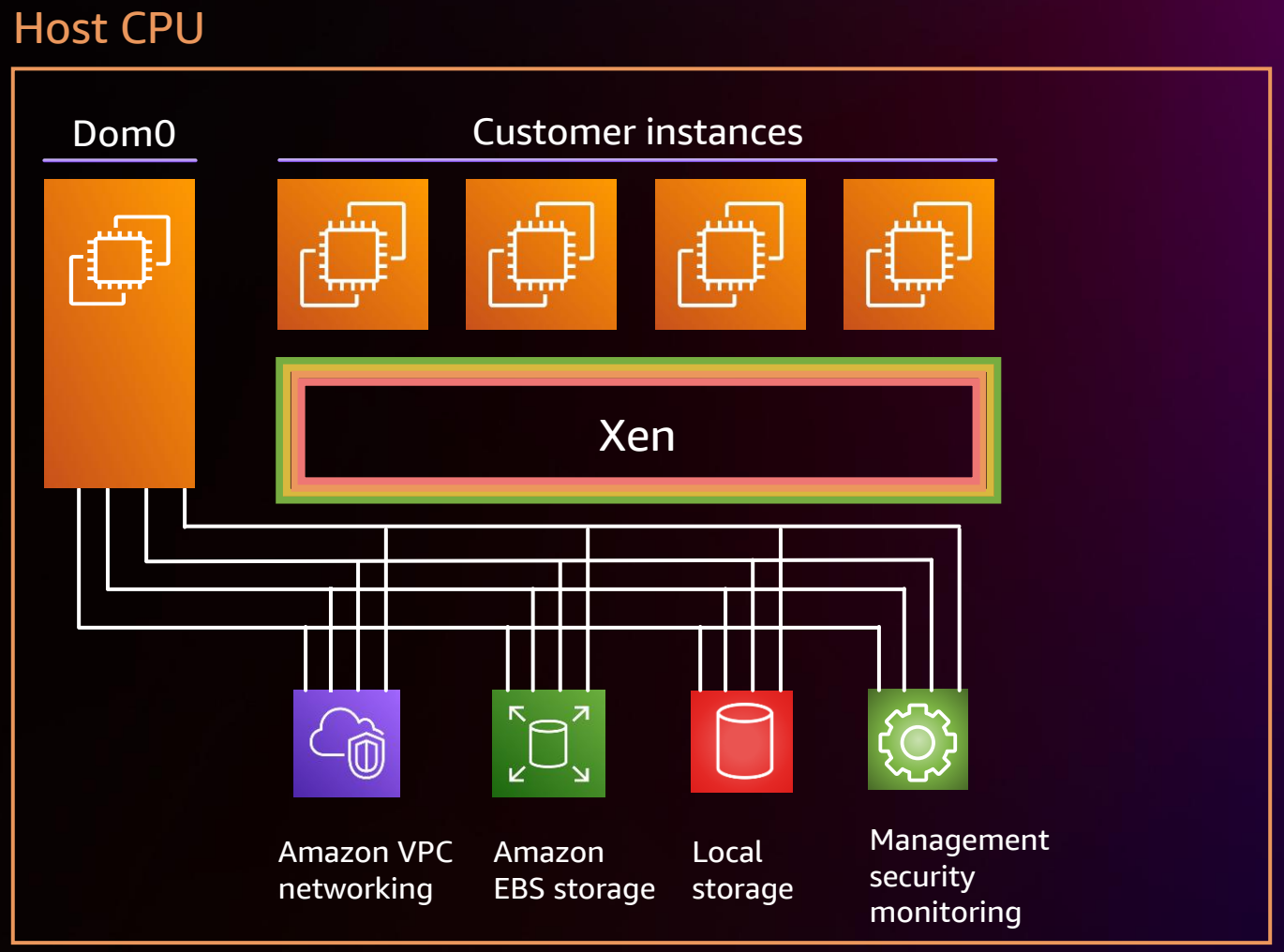
AWS Nitro System

- Purpose-built hardware/ software
- 5 generations of custom chips
- Hypervisor built for AWS
- Powering over 500 instance types

Before Nitro...

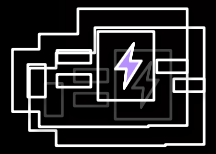


AWS Nitro System



AWS Nitro System

Nitro Cards



VPC networking

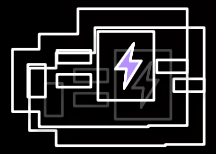
Amazon EBS

Nitro SSDs

System controller

AWS Nitro System – VPC Networking

Nitro Cards



VPC networking

Amazon EBS

Nitro SSDs

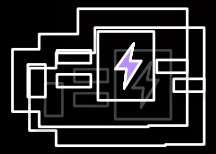
System controller

VPC data-plane offload

- ENI attachment, security groups, flow logs, routing, port mirroring, DHCP, DNS

AWS Nitro System – VPC Networking

Nitro Cards



VPC networking

Amazon EBS

Nitro SSDs

System controller

VPC data-plane offload

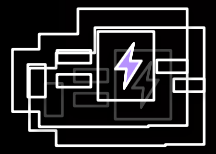
- ENI attachment, security groups, flow logs, routing, port mirroring, DHCP, DNS

VPC encryption

- Authentication and transparent end-to-end 256-bit encryption

AWS Nitro System – VPC Networking

Nitro Cards



VPC networking

Amazon EBS

Nitro SSDs

System controller

VPC data-plane offload

- ENI attachment, security groups, flow logs, routing, port mirroring, DHCP, DNS

VPC encryption

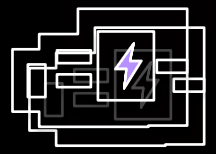
- Authentication and Transparent end-to-end 256-bit encryption

Elastic Network Adapter (ENA)

- Extendible host interface
- ENA Express

AWS Nitro System – VPC Networking

Nitro Cards



VPC networking

Amazon EBS

Nitro SSDs

System controller

VPC data-plane offload

- ENI attachment, security groups, flow logs, routing, port mirroring, DHCP, DNS

VPC encryption

- Authentication and transparent end-to-end 256-bit encryption

Elastic Network Adapter (ENA)

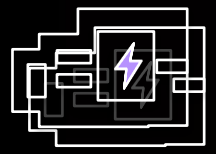
- Extensible host interface
- ENA Express

Elastic Fabric Adapter (EFA)

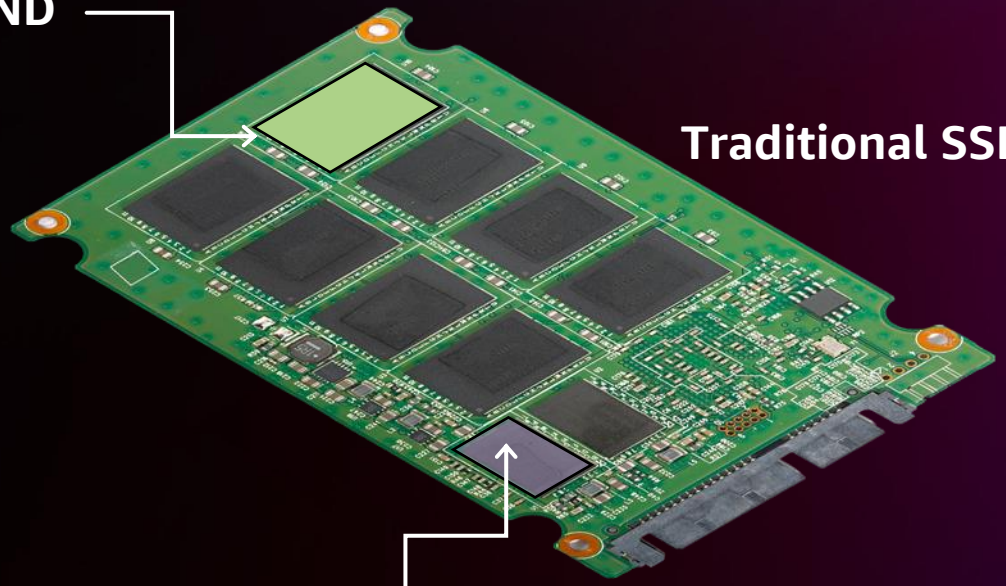
- ML and HPC UltraClusters, RDMA and GPU-RDMA
- Low latency at scale
- Multipathing (SRD)

AWS Nitro System – Nitro SSDs

Nitro Cards



VPC networking
Amazon EBS
Nitro SSDs
System controller



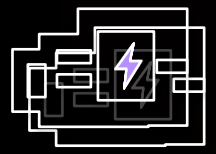
Traditional SSD design

Flash translation layer (FTL)

- Maps logical to physical addresses
- Performs garbage collection
- Manages NAND wear leveling

AWS Nitro System – Nitro SSDs

Nitro Cards

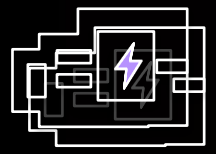


VPC networking
Amazon EBS
Nitro SSDs
System controller

Each vendor has their own FTL and differing performance

AWS Nitro System – Nitro SSDs

Nitro Cards



VPC networking

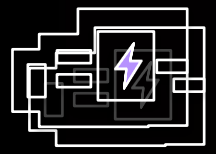
Amazon EBS

Nitro SSDs

System controller

AWS Nitro System – Nitro SSDs

Nitro Cards

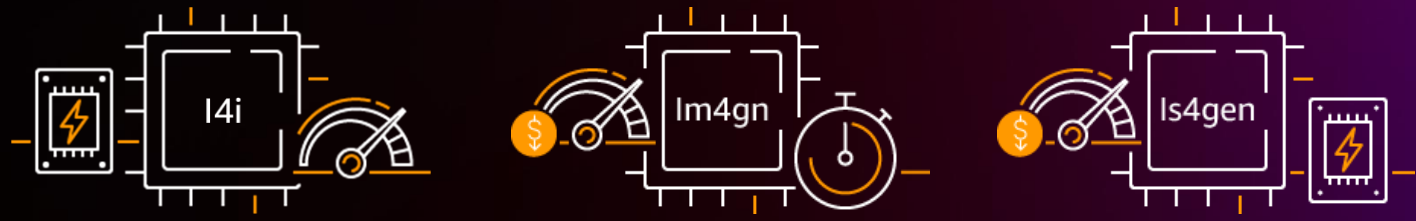


VPC networking

Amazon EBS

Nitro SSDs

System controller



60%

Lower average I/O latency

75%

Lower tail latency

Reducing storage costs

UP TO 50% DECREASE IN TCO

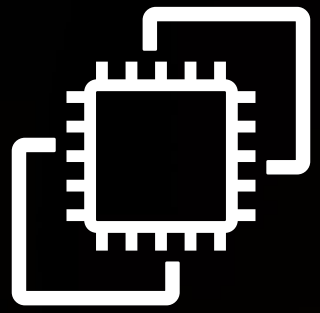
splunk>

Splunk is a leading data platform provider, and is designed to investigate, monitor, analyze and act on data at any scale

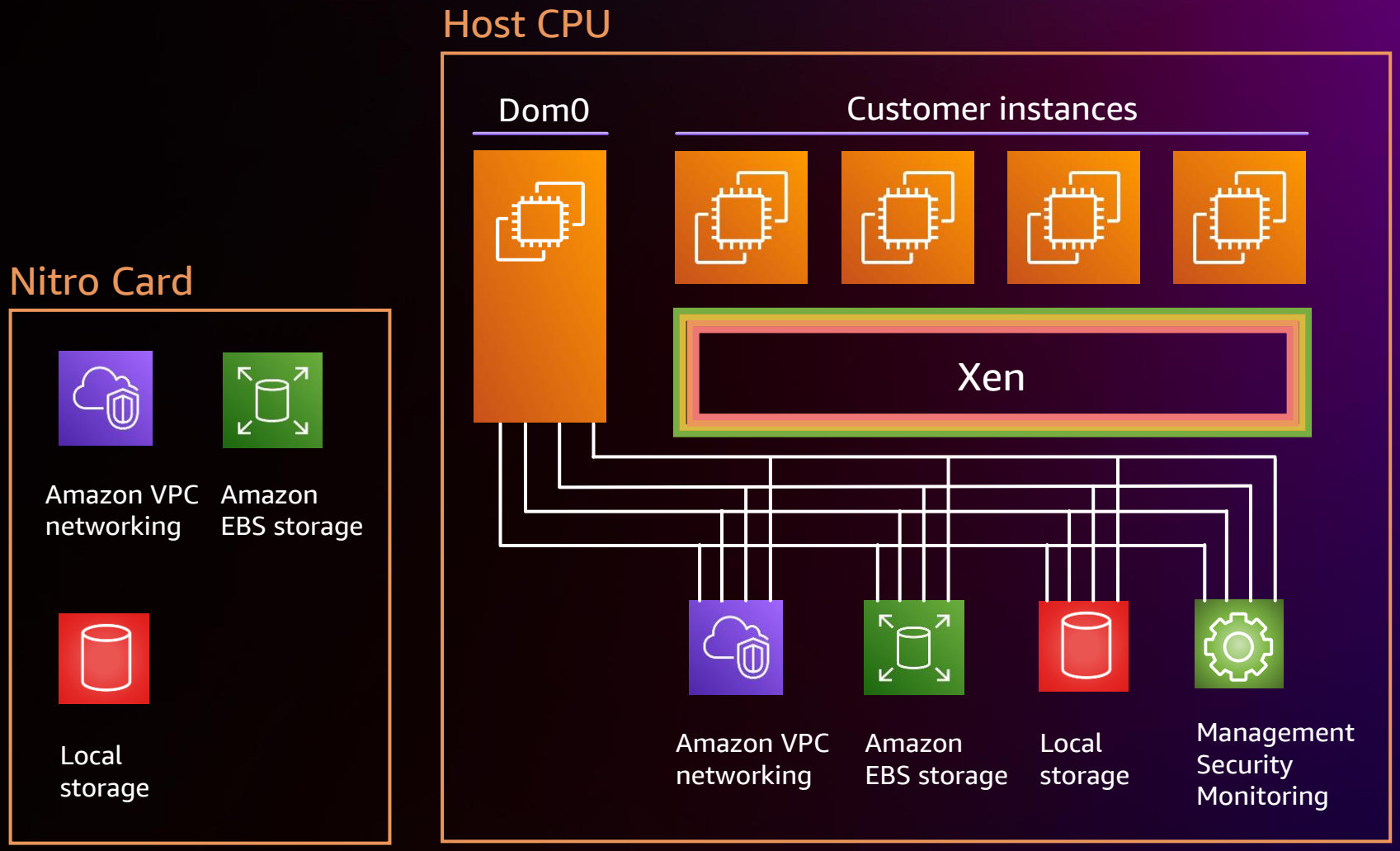
"When evaluating the new Im4gn/Is4gen instances powered by AWS Graviton2, we observed an **up to 50% decrease in search runtime** compared to I3/I3en instances, which we currently use."

—Brad Murphy, VP of Engineering, Splunk

AWS Nitro System

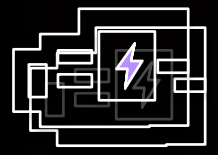


AWS Nitro System



AWS Nitro System

Nitro Cards



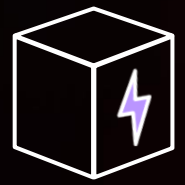
VPC networking

Amazon EBS

Nitro SSDs

System controller

Nitro Hypervisor



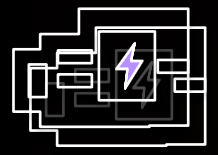
Lightweight hypervisor

Memory and CPU allocation

Bare metal–like performance

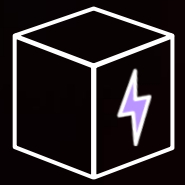
AWS Nitro System

Nitro Cards



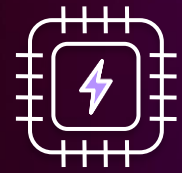
VPC networking
Amazon EBS
Nitro SSDs
System controller

Nitro Hypervisor



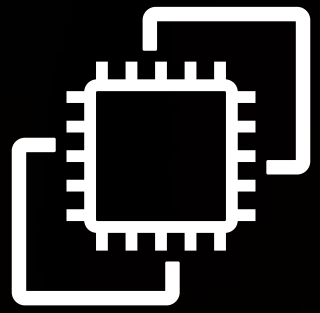
Lightweight hypervisor
Memory and CPU allocation
Bare metal–like performance

Nitro Security Chip



Integrated into motherboard
Traps I/O to nonvolatile storage
Hardware root of trust

Before Nitro...



AWS Nitro System

Nitro Card

Amazon VPC networking

Amazon EBS storage

Local storage

Management Security Monitoring

Host CPU

Dom0

Customer instances

Host CPU

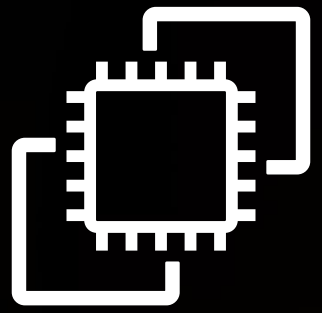
Xen

Customer instances

Nitro Hypervisor

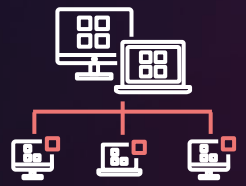
nagement
urity
Monitoring

Nitro security



AWS Nitro System

All interactions with the AWS Nitro System are through narrow, authorized, and authenticated APIs



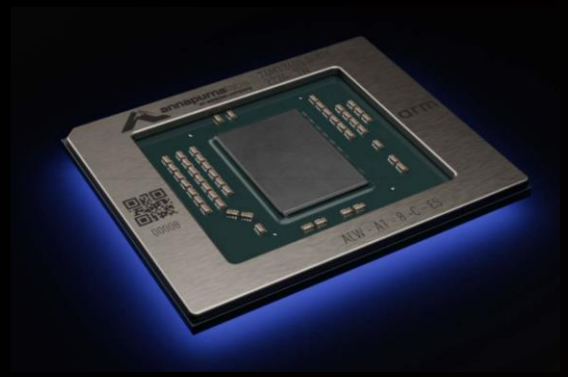
There is no mechanism for any system or person to log in to the underlying EC2 host

There is no interactive access

The Security Overview of the AWS Nitro System whitepaper

https://a.co/hYWhsH9

5th generation Nitro Card



AWS Nitro System

Encryption support on all interfaces including Network, DRAM and PCIe

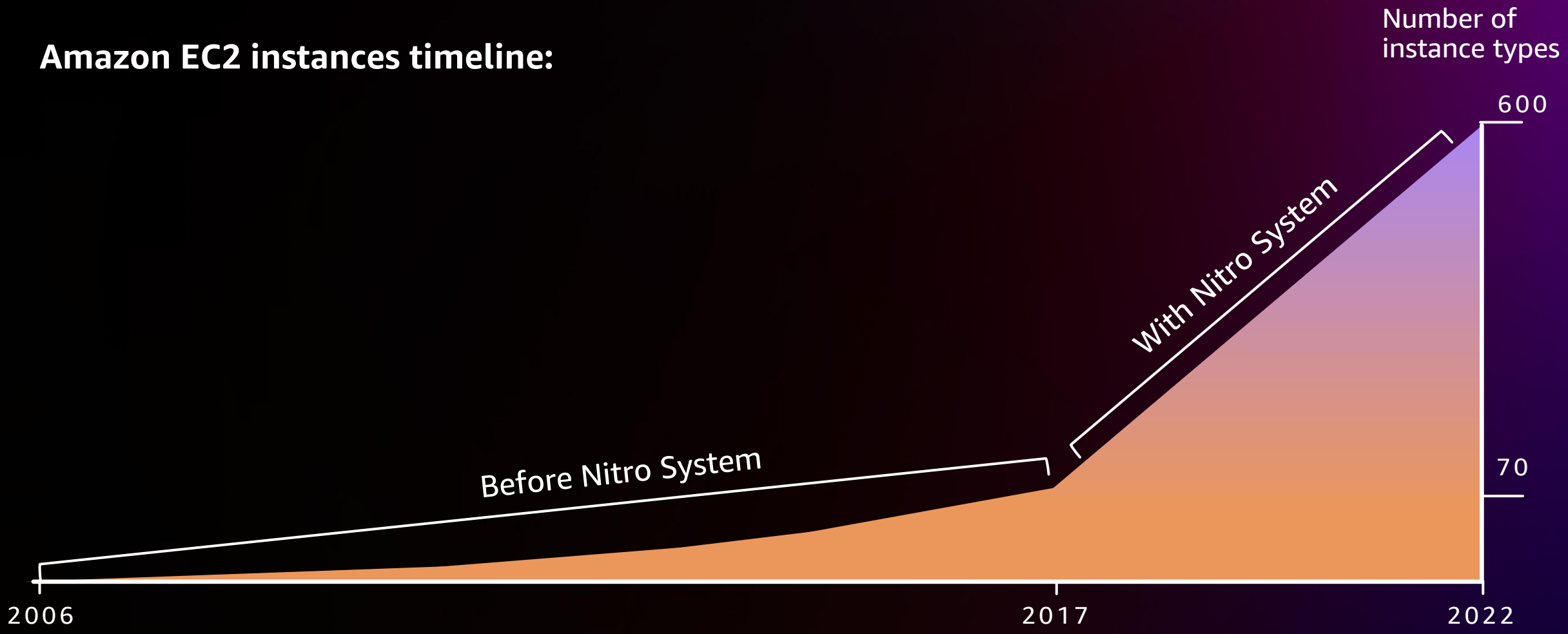
22B transistors providing enhanced performance

Latest generation interfaces including PCIe gen5 & DDR5

up to 60% packet rate, up to 30% lower latency, lower power consumption

Increased pace of innovation

Amazon EC2 instances timeline:



AWS Graviton

AWS Graviton

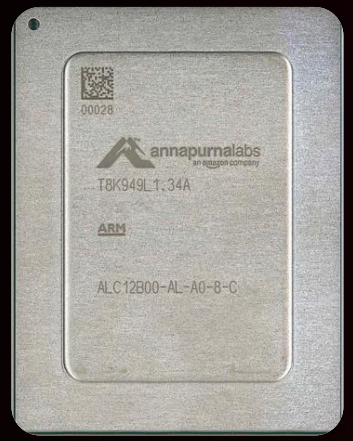
BEST PRICE PERFORMANCE IN AMAZON EC2

Graviton

2018

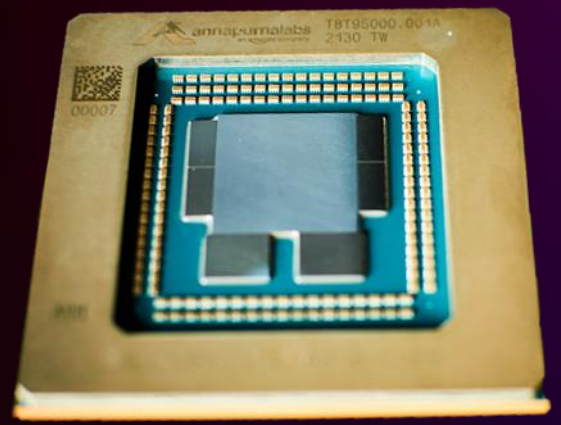
Graviton2

2019



Graviton3

2021



AWS Graviton2-based Amazon EC2 instances

UP TO 40% BETTER PRICE PERFORMANCE OVER COMPARABLE X86-BASED INSTANCES

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 28 AWS REGIONS GLOBALLY

AWS managed services supporting Graviton2

EXTENDING THE GRAVITON PRICE PERFORMANCE TO MANAGED SERVICES

Databases

Amazon DocumentDB

Amazon Aurora

Amazon RDS

Amazon ElastiCache

Amazon MemoryDB

Amazon Neptune

Analytics

Amazon OpenSearch Service

Amazon EMR

Compute

AWS Lambda

AWS Fargate

AWS Elastic Beanstalk

Machine Learning

Amazon SageMaker

Customer momentum with Graviton




Graviton Energy Efficiency

\Orchestrating a brighter world

NEC

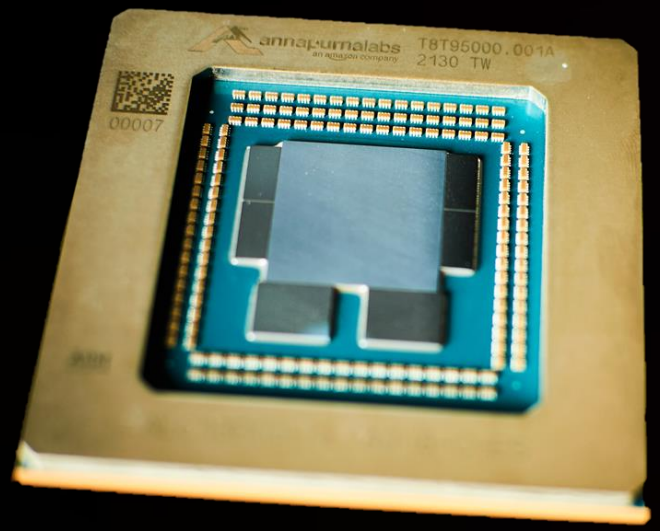
NTT docomo

NTT DOCOMO and NEC Reduce Power Consumption for 5G SA Core by an Average of 72% using AWS Graviton2

“We are delighted to announce that we achieved significant reduction of power consumption of 5GC thanks to NEC's advanced, cloud-native 5GC software and AWS's innovative and highly efficient Graviton2.”

Naoki Tani, Executive Vice President, Chief Technology Officer, Executive General Manager of the R&D Innovation Division of NTT DOCOMO

aws



AWS Graviton3

Third-generation Graviton processor

Chiplet-based design: seven silicon die

~55 billion transistors

First DDR5 system in our data centers – 50% more DDR bandwidth

Graviton3 CPU enhancements

PtrAuth
RNG Inst

AWS Graviton2

4- to 8- wide fetch

4-wide decode

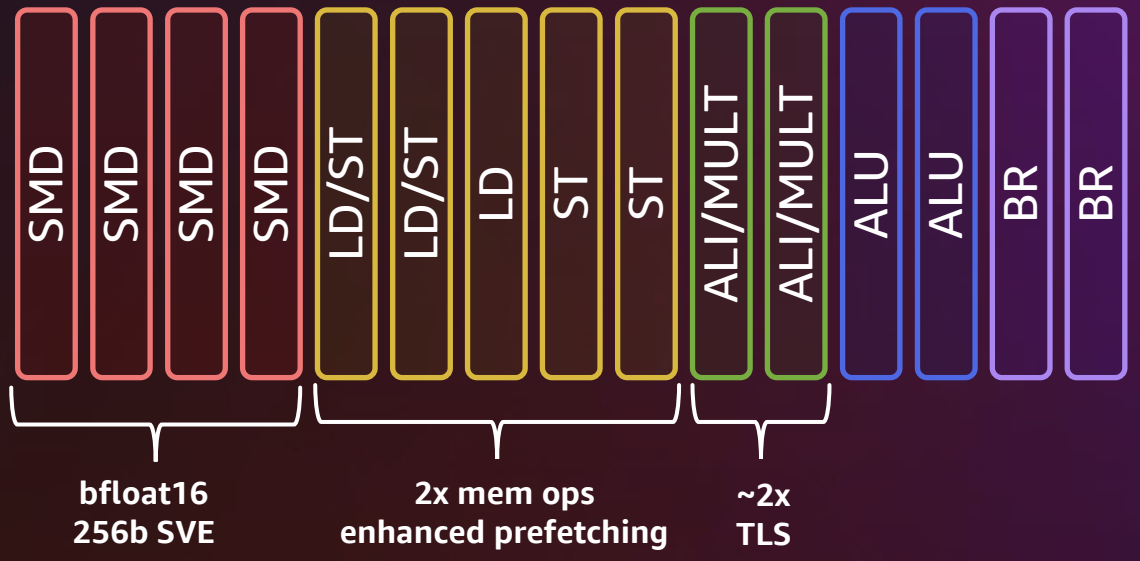
8-wide issue

AWS Graviton3

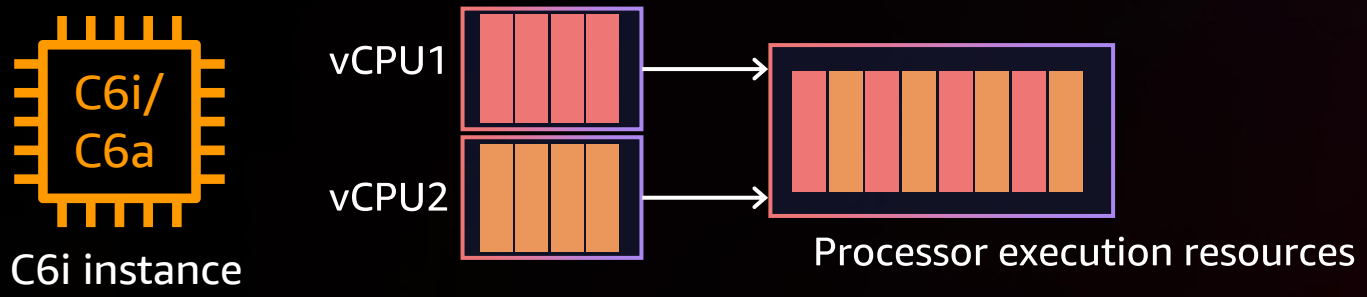
8-wide fetch

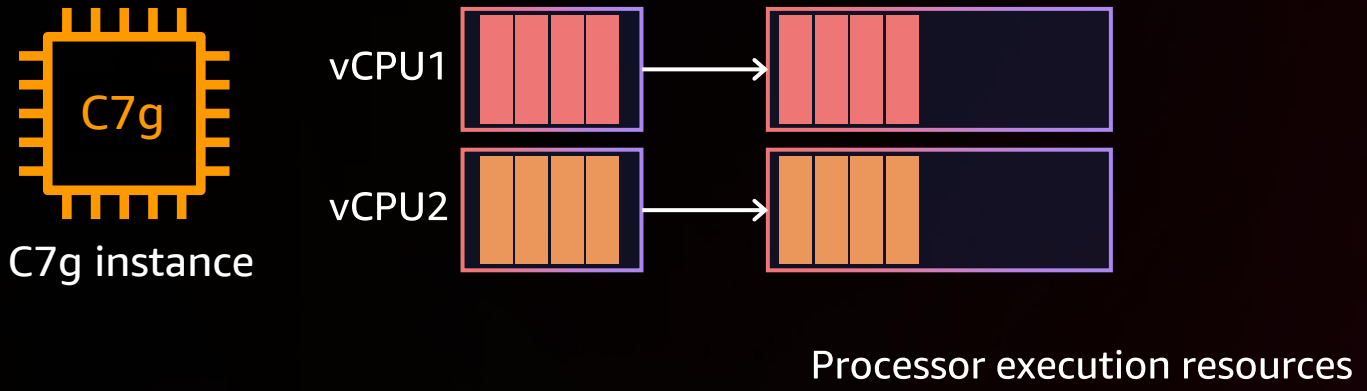
5- to 8-wide decode

15-wide issue & 2x larger instruction window



Graviton3 – vCPU

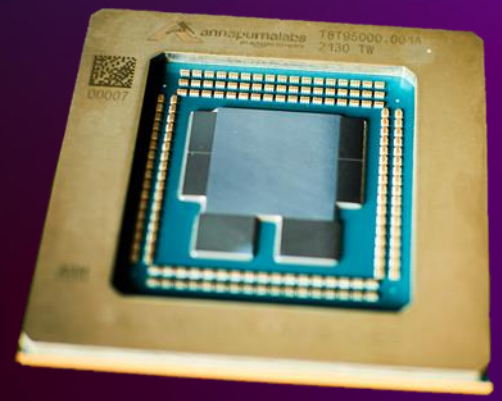


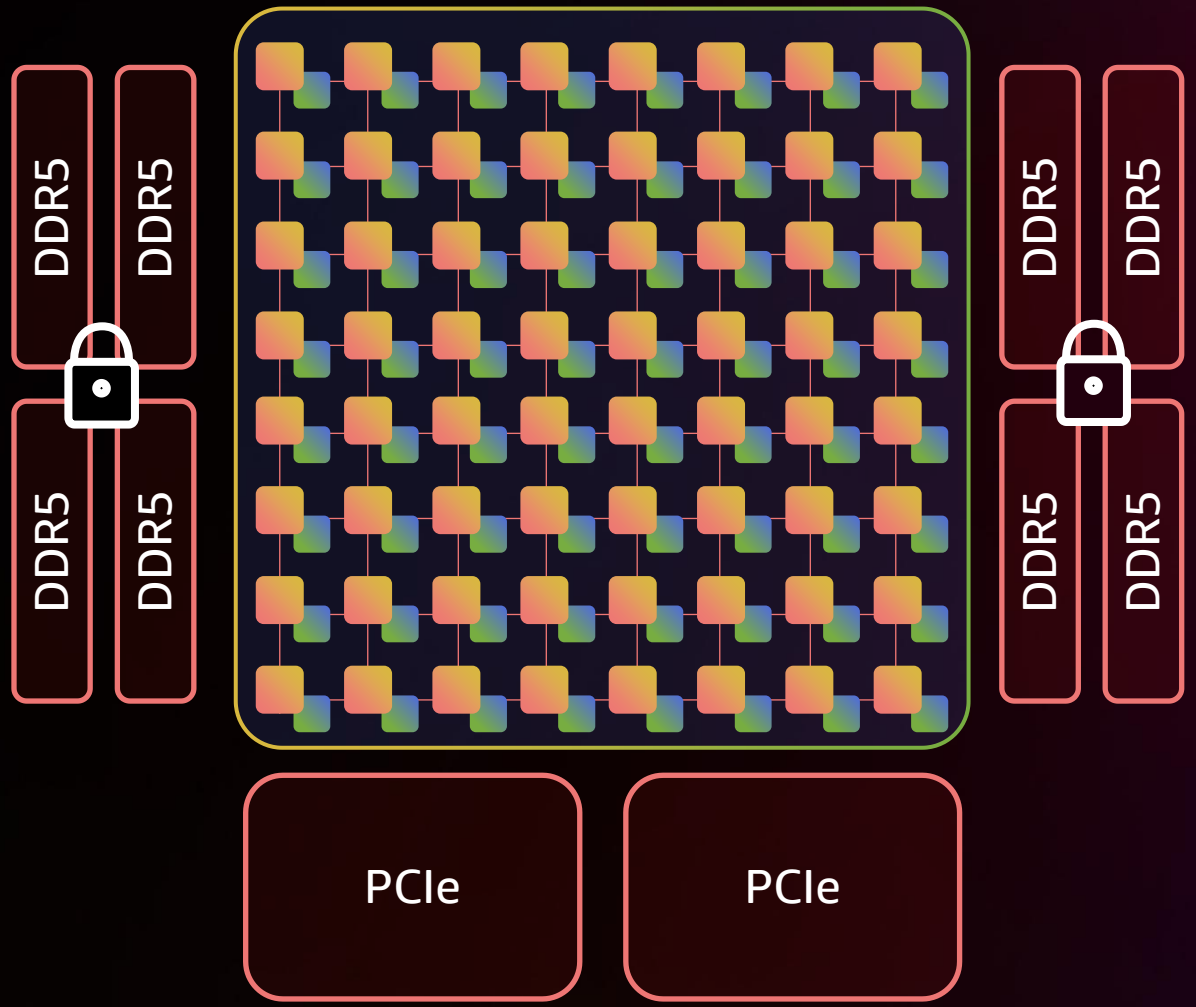


Every vCPU is a physical core

No simultaneous multithreading (SMT)

Graviton3 – chiplet design



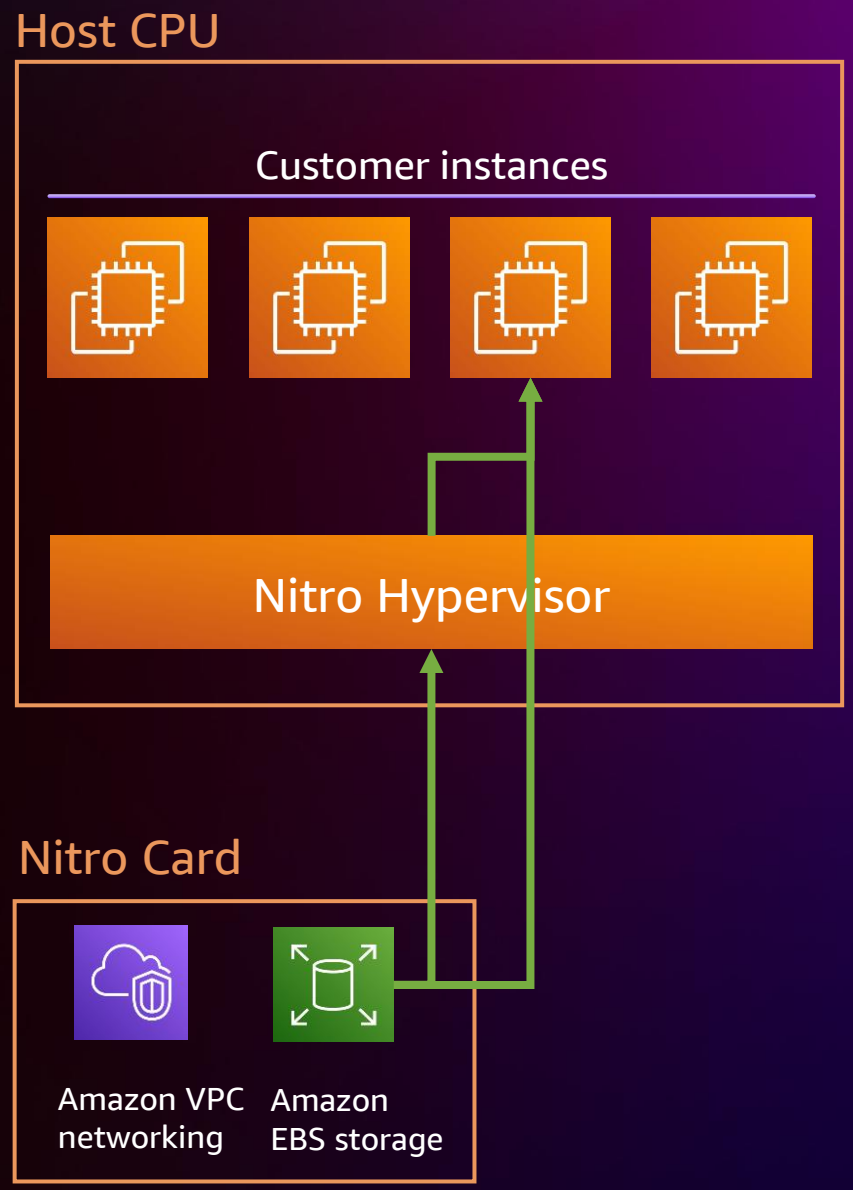


Direct interrupt injection

Traditionally interrupts go from IO-cards to customer VM via hypervisor

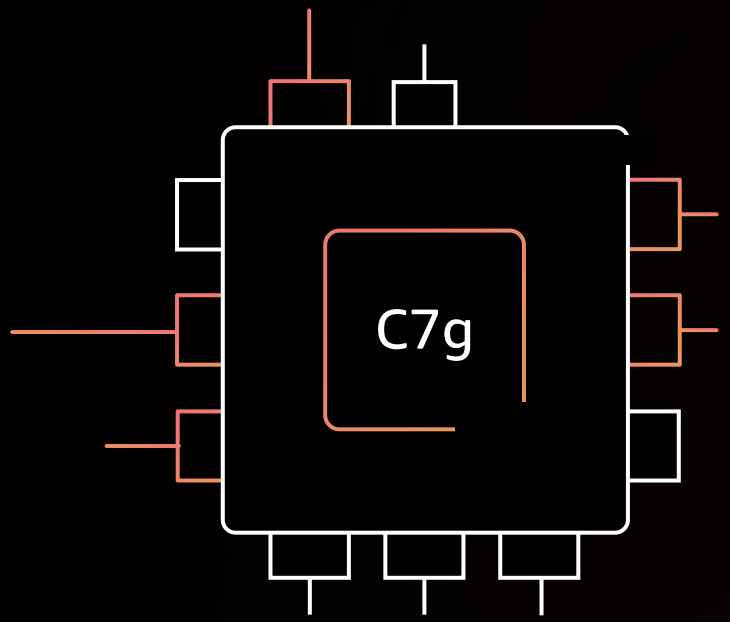
Graviton2 and Graviton3 instances directly inject into guests

Lower latency and higher throughput



Graviton3-based compute optimized instance

BEST PRICE PERFORMANCE FOR COMPUTE-INTENSIVE WORKLOADS ON AMAZON EC2



Up to 25% better performance compared to Graviton2

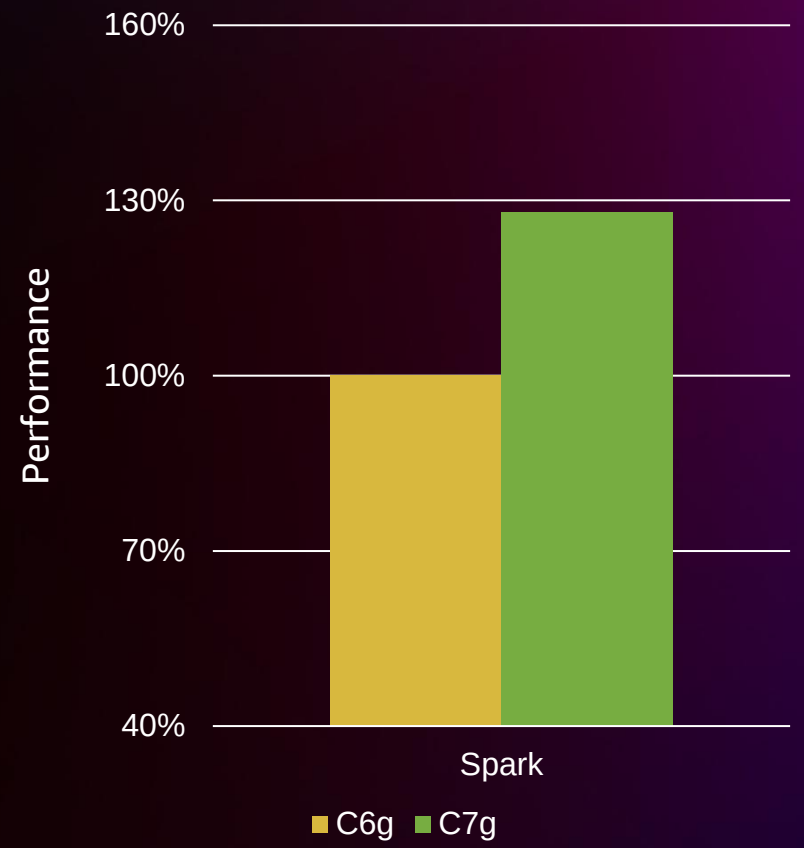
Up to 2x higher floating-point performance, up to 2x faster cryptographic workload performance, and up to 3x better machine learning performance compared to Graviton2

First in the cloud to feature DDR5 memory

Up to 60% more energy efficient vs comparable EC2 instances

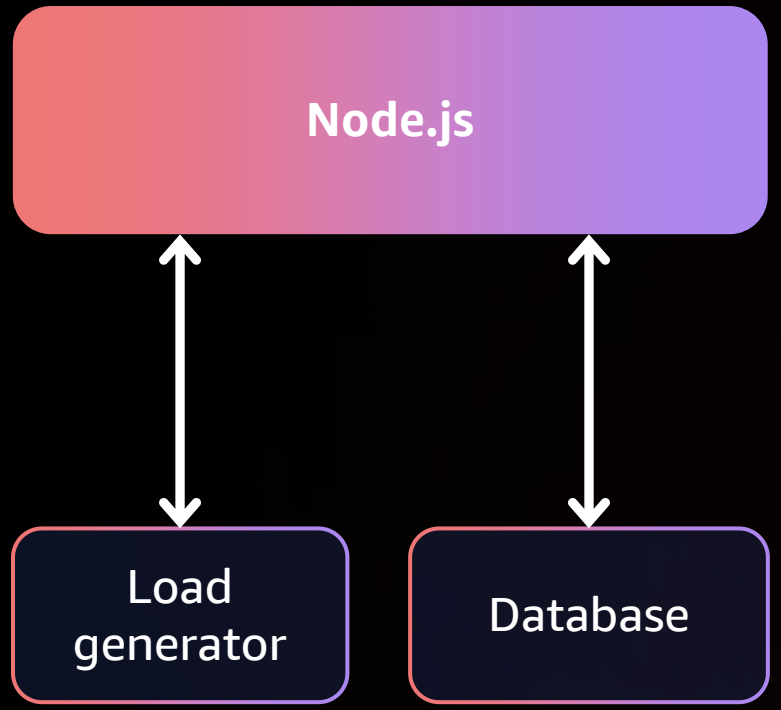
Spark SQL

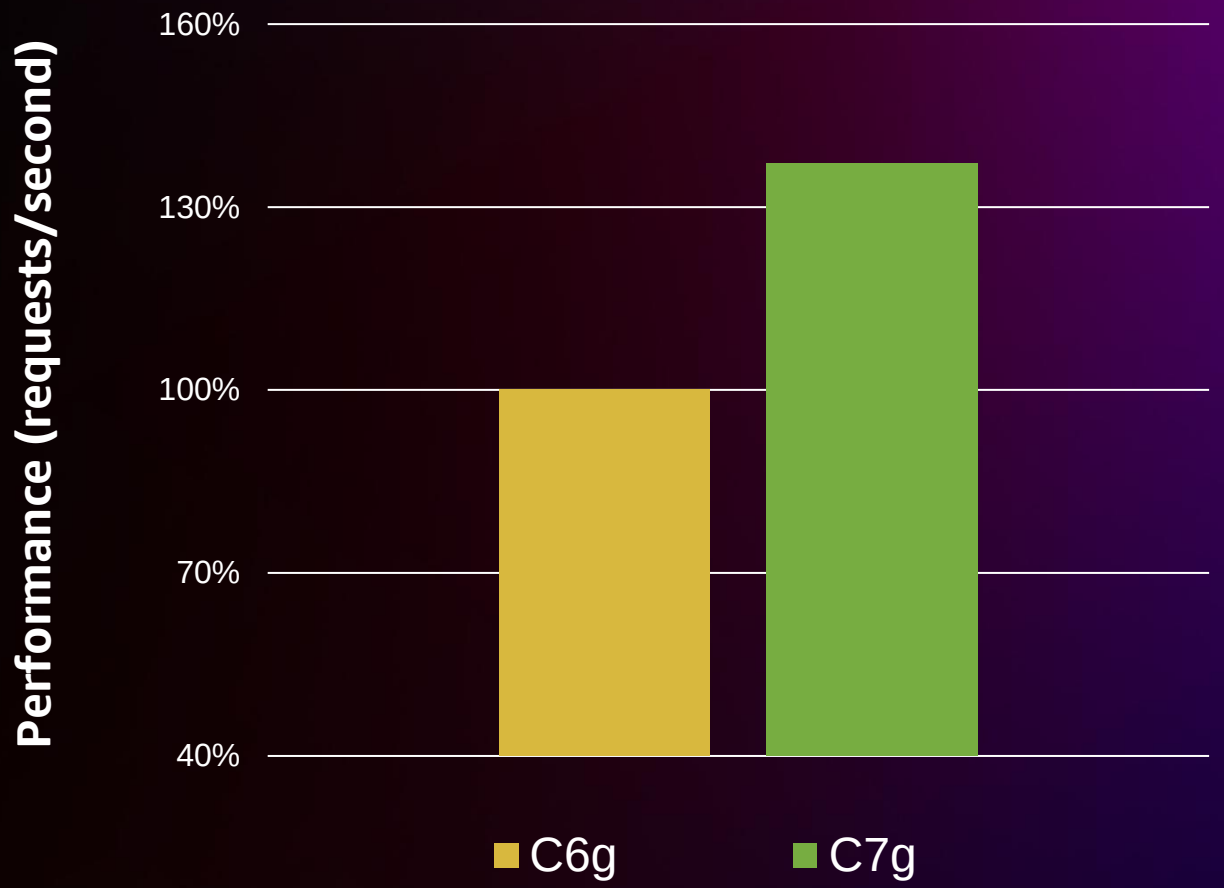
- Spark 3.3 with AWS Corretto 17
- 8-node cluster & 1TB of dataset
- Spark SQL performance is 28% faster



https://github.com/databricks/spark-sql-perf; 4xl large instances;

Node.js applications

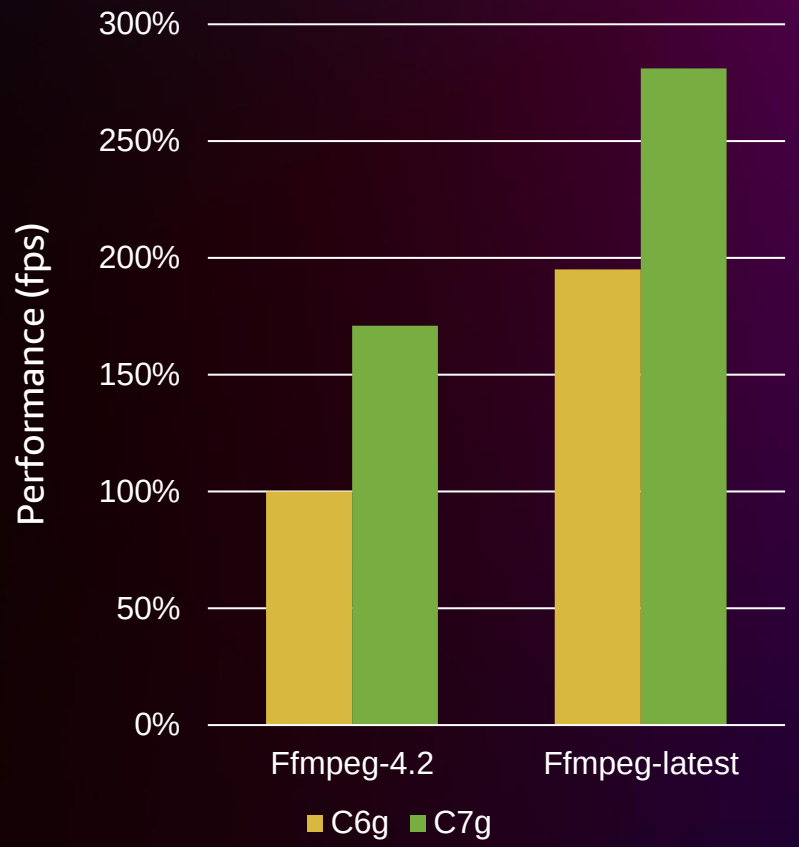




Node.JS 16.7.0, AcmeAir test application with one process per vCPU, JMeter load generator on c6g.4xlarge in a cluster placement group with nginx as reverse proxy, HTTP connections, connection count varied to control load

FFmpeg: Performance & software optimization

- Video accounts for over 60% of downstream traffic on the internet
- Encoding reduces the bandwidth to deliver and store video content
- AWS and open-source community have optimized video encoding on arm64
- Recent FFmpeg improvements (FFmpeg 4.2 → FFmpeg latest) have increased performance by 60%+ on Graviton3 processors
- Graviton3 is up to 50% faster than Graviton2



x265, 4xl instance size, mean frames per second

Machine learning performance

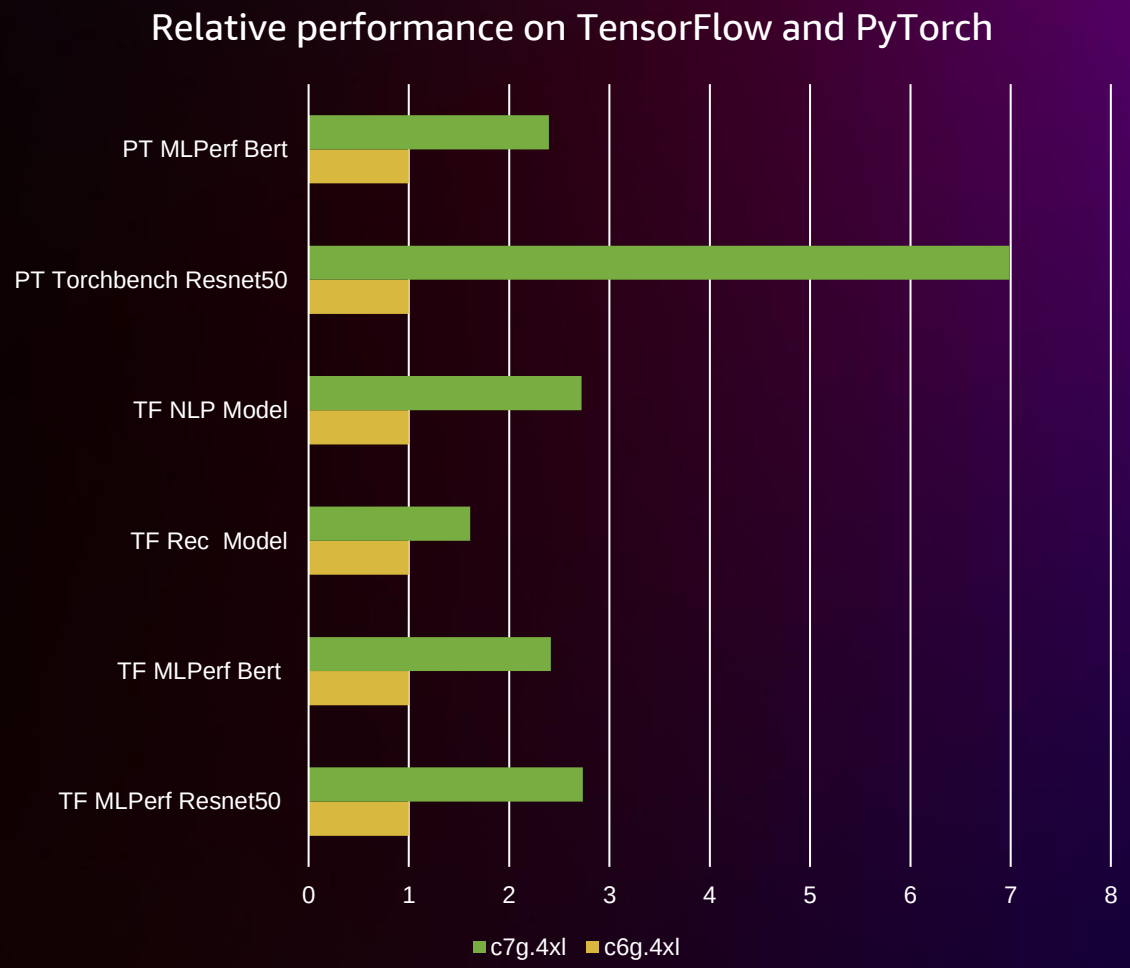
Leap in performance with Graviton3 vector width and bfloat16

+

Large amount of improvements in TensorFlow, PyTorch, OneDNN, and Arm Compute Library

=

Extremely fast CPU-based machine learning inference with TensorFlow and PyTorch across many models



AWS Graviton3–based Amazon EC2 instances

C7g

Compute-intensive workloads

Best price performance for compute-intensive workloads in EC2

C7gn

Network-intensive workloads

Up to 50% higher PPS & 2x bandwidth

Featuring 5th generation Nitro Card

HPC7g

High-performance computing workloads

Graviton3E

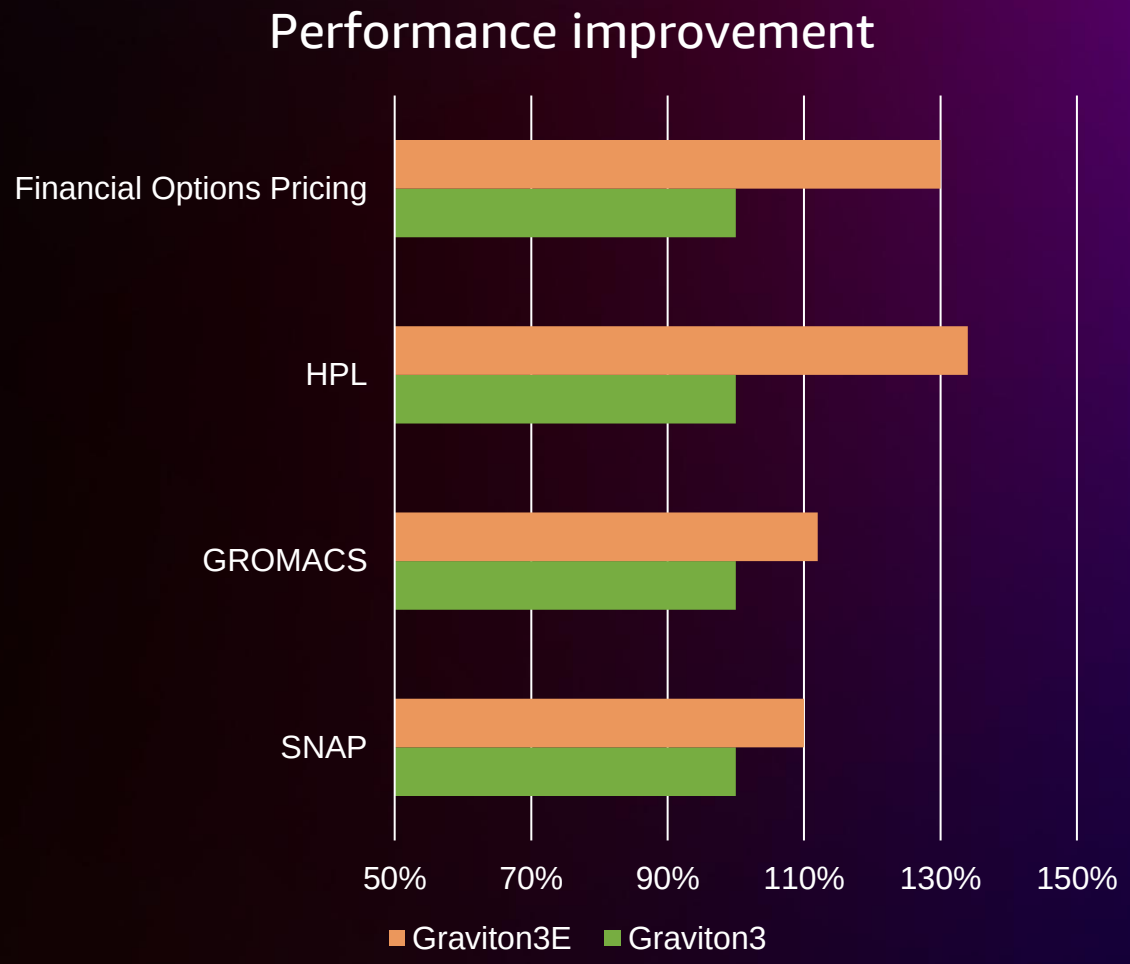
Featuring 5th generation Nitro Card

Graviton3E available in HPC7g

HPC-focused Graviton3 CPU

Increased SIMD performance

Up to 30% additional performance on HPC workloads



Innovation at the server level

- Typical servers
  - 2 sockets/server
  - 42U rack
- Traditionally run out of rack power before space
- Graviton2 runs out of space before power
- How can we put more sockets in a rack?

Graviton servers

Design the chip, package, and motherboard to fit…

Graviton servers

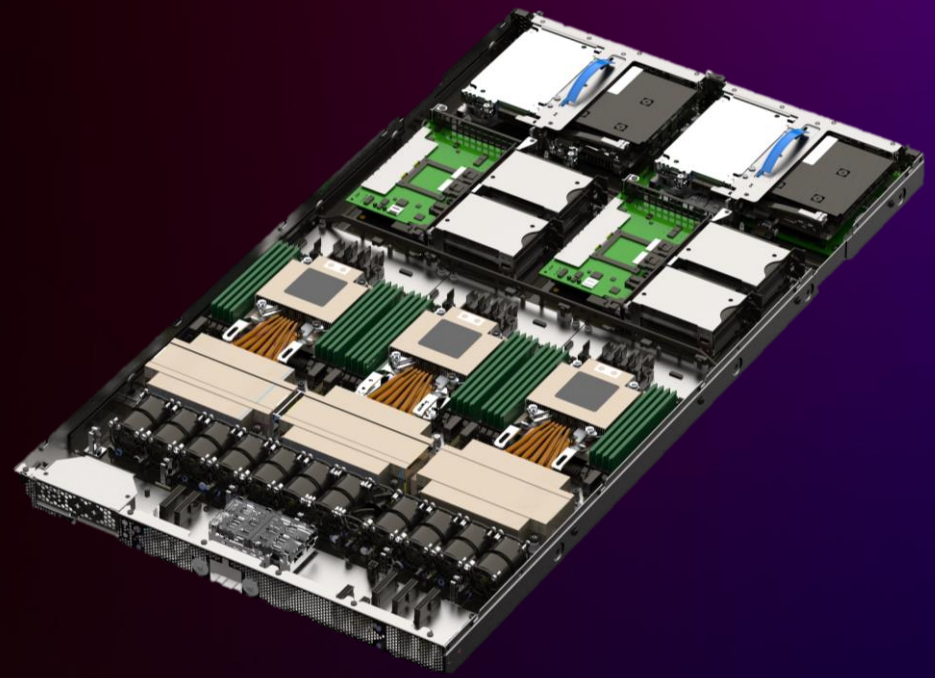
Design the chip, package, and motherboard to fit…

Three sockets per server

Increase the sockets/ rack 50%

Come closer to the power requirements of other vendor solutions

Nitro Card is capable of managing three sockets simultaneously



What customers are saying about Amazon EC2 C7g instances

"We are seeing about 15% more requests handled by C7g instances and up to 40% better latency compared to Graviton2-based C6g instances."

"We benchmarked our workloads on the new Amazon EC2 C7g instances and observed 27% better performance compared to the previous generation instances."

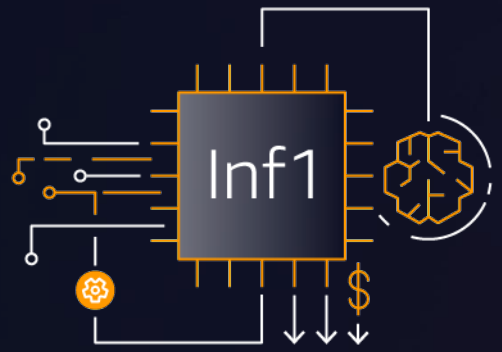
"We benchmarked 35% better performance on AWS Graviton3-based C7g instances compared to the previous generation instances."

"On production testing, we've seen ~45% performance improvement for heavy CPU workloads on AWS Graviton3-based Amazon EC2 C7g instances compared to AWS Graviton2-based C6g instances."

AWS Inferentia and AWS Trainium

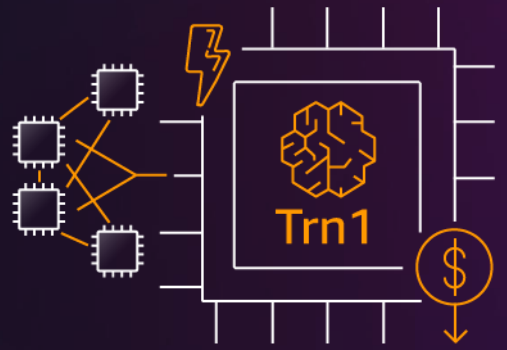
AWS chips optimized for deep learning

AWS Inferentia



Lowest cost inference in the cloud for running deep learning models—up to 70% lower cost than GPU instances

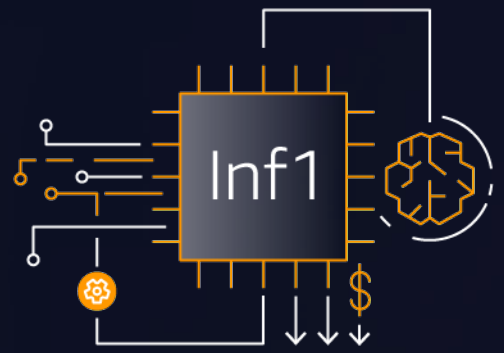
AWS Trainium



The most cost-efficient high-performance DL training instance in the AWS Cloud

AWS chips optimized for deep learning

AWS Inferentia

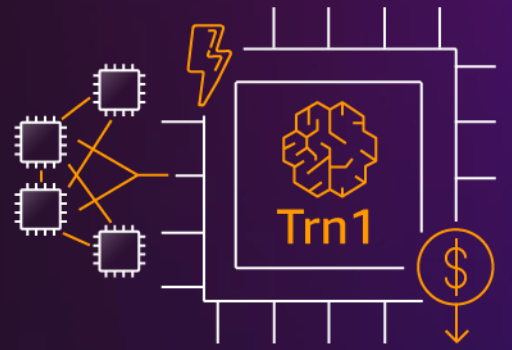


Lowest cost inference in the cloud for running deep learning models—up to 70% lower cost than GPU instances

AWS Neuron

Seamless integration with ML frameworks like PyTorch and TensorFlow

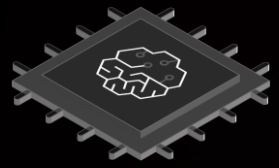
AWS Trainium



The most cost-efficient high-performance DL training instance in the AWS Cloud

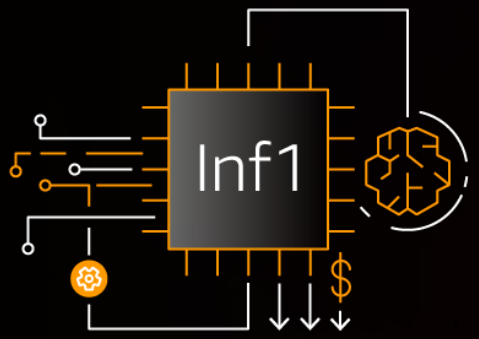
Amazon EC2 Inf1 instances

INTRODUCED IN 2019 BASED ON THE FIRST AWS-DESIGNED ML SILICON



AWS Inferentia

High-performance machine learning inference chip, purpose-built by AWS



EC2 Inf1 Instances

Inf1

BF16/FP16	INT8
1 Peta FLOPS	2 Peta OPS

AGGREGATE ACCELERATOR MEMORY	NETWORK CONNECTIVITY
128 GB	100 Gbps

NEURON-CORE V1 NEURON-LINK V1	PYTORCH & TENSORFLOW SUPPORTED

Amazon EC2 Inf1 instances

AMAZON EC2 INF1 INSTANCES DELIVER THE BEST INFERENCE PRICE PERFORMANCE

25%

higher throughput

70%

lower cost

* Measured on PyTorch BERT-Base

Customer momentum with Inferentia

airbnb

AUTODESK

Finch COMPUTING

sprinklr

NTTPC

amazonrobotics

amazonadvertising

amazon alexa

SCREENING EAGLE

The Asahi Shimbun

CONDÉ NAST

Talroo™

Anthem

DISCO

SKYWATCH

Money Forward

Amazon Rekognition

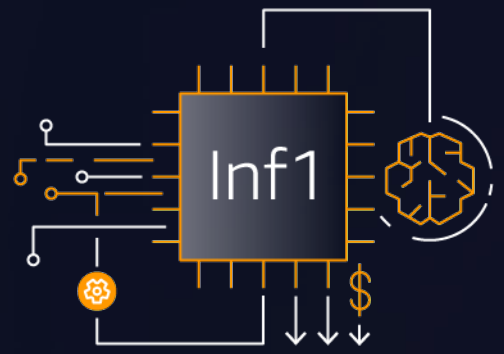
DMP DIGITAL MEDIA PROFESSIONALS

prime video

More customer testimonials

AWS chips optimized for deep learning

AWS Inferentia

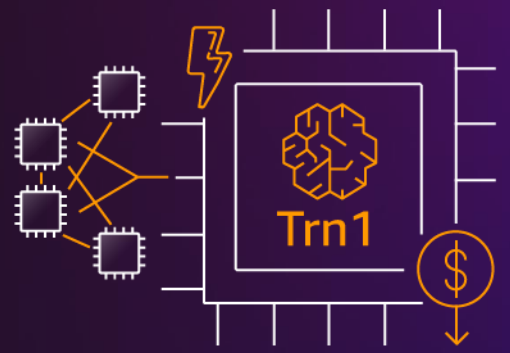


Lowest cost in the cloud for running deep learning models—up to 70% lower cost than GPU instances

AWS Neuron

Seamless Integration with ML frameworks like TensorFlow and PyTorch with minimal code changes

AWS Trainium



Purposely built for the most cost-efficient and fast DL training in the AWS Cloud for a broad spectrum of applications

Amazon EC2 Trn1/Trn1n instances

THE MOST COST-EFFICIENT HIGH-PERFORMANCE TRAINING INSTANCE IN THE AWS CLOUD

Trn1(n)

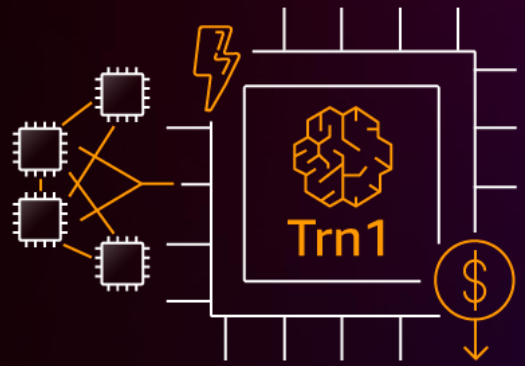
BF16/FP16	TF32	FP32
3.4 PFLOPS	3.4 PFLOPS	840 TFLOPS

AGGREGATE ACCELERATOR MEMORY	PEAK MEMORY BANDWIDTH
512 GB	13.1 TB/sec

EFA NETWORK CONNECTIVITY	NEURON-CORE V2 NEURON-LINK V2
800/1600 Gbps	

AWS Trainium

High-performance machine learning training chip, purpose-built by AWS



EC2 Trn1(n) Instances

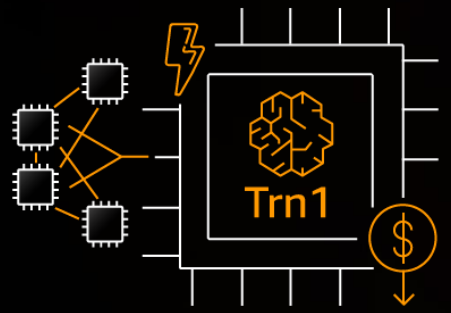
The most cost-efficient high-performance training instance in the AWS Cloud

Amazon EC2 Trn1/Trn1n instances

THE MOST COST-EFFICIENT HIGH-PERFORMANCE TRAINING INSTANCE IN THE AWS CLOUD

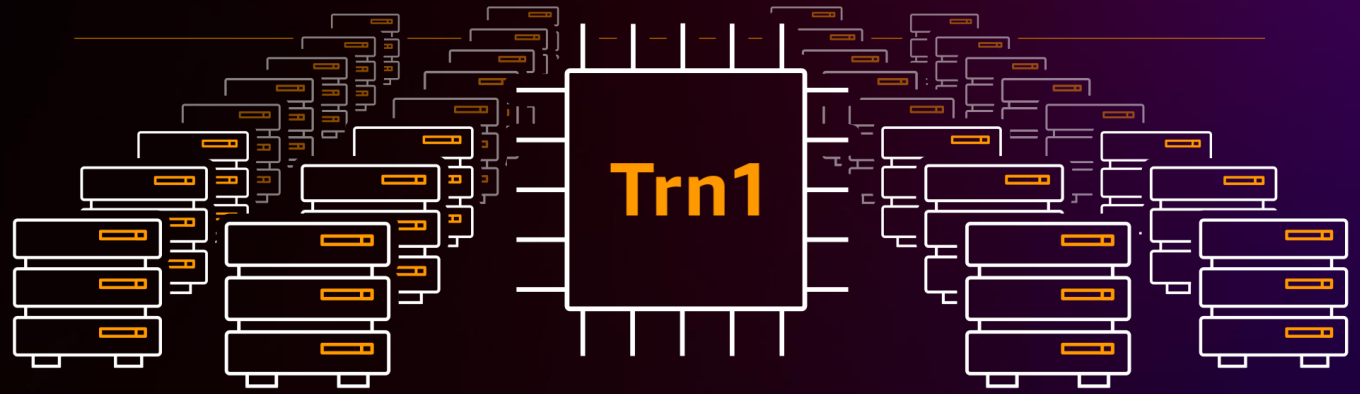
AWS Trainium

High-performance machine learning training chip, purpose-built by AWS



EC2 Trn1(n) Instances

The most cost-efficient high-performance training instance in the AWS Cloud

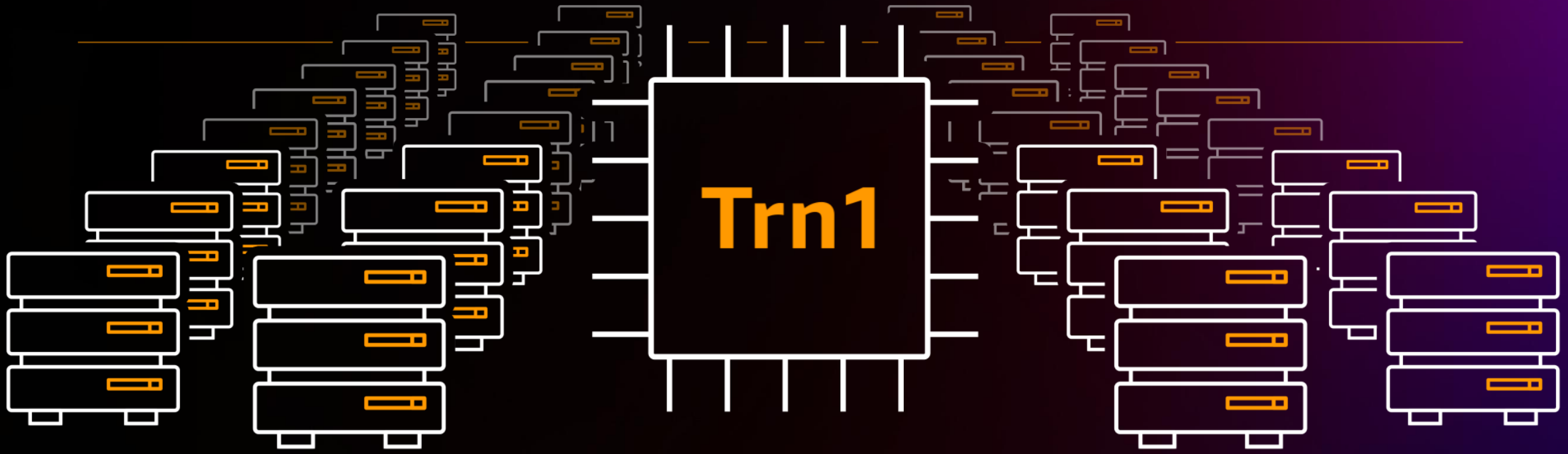


Trn1n UltraClusters

Train large models with 30K+ Trainium devices and non-blocking EFA network

UltraCluster for ultra-large models

30K+ TRAINIUM ACCELERATORS WITH NON-BLOCKING NETWORK CONNECTIVITY

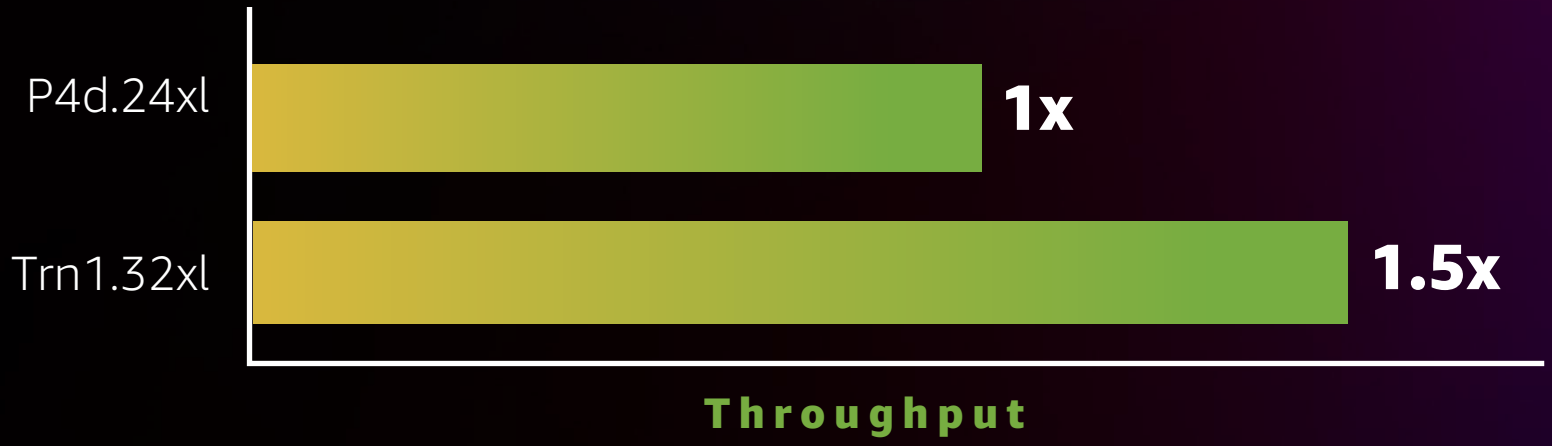


On-demand access to a world-class supercomputer,
6 ExaFLOPS of compute

Cost to train with Trn1

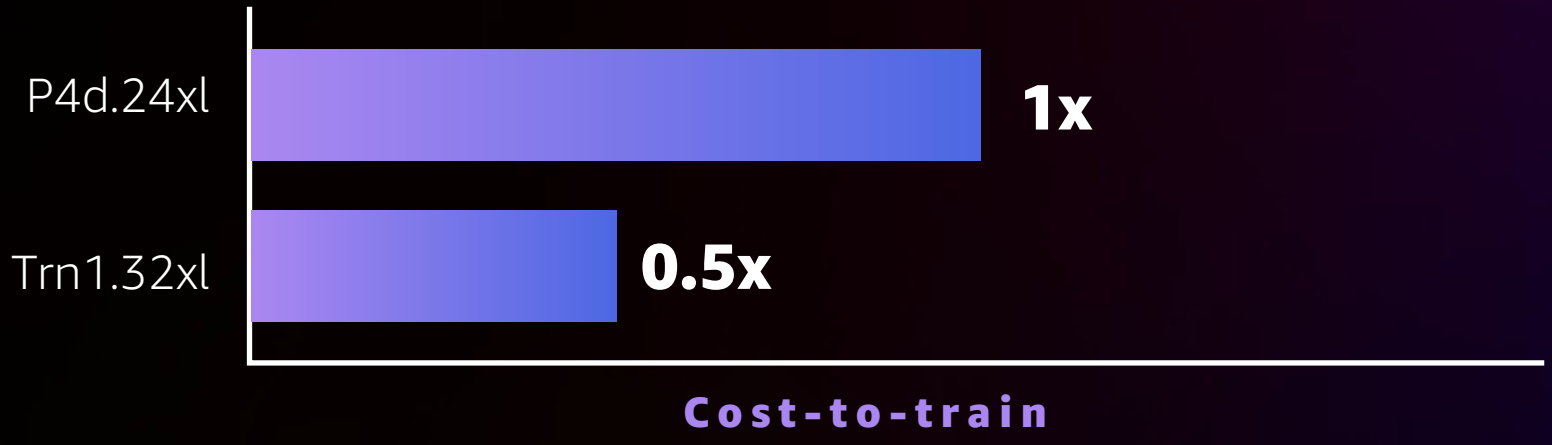
1.5x

higher throughput



50%

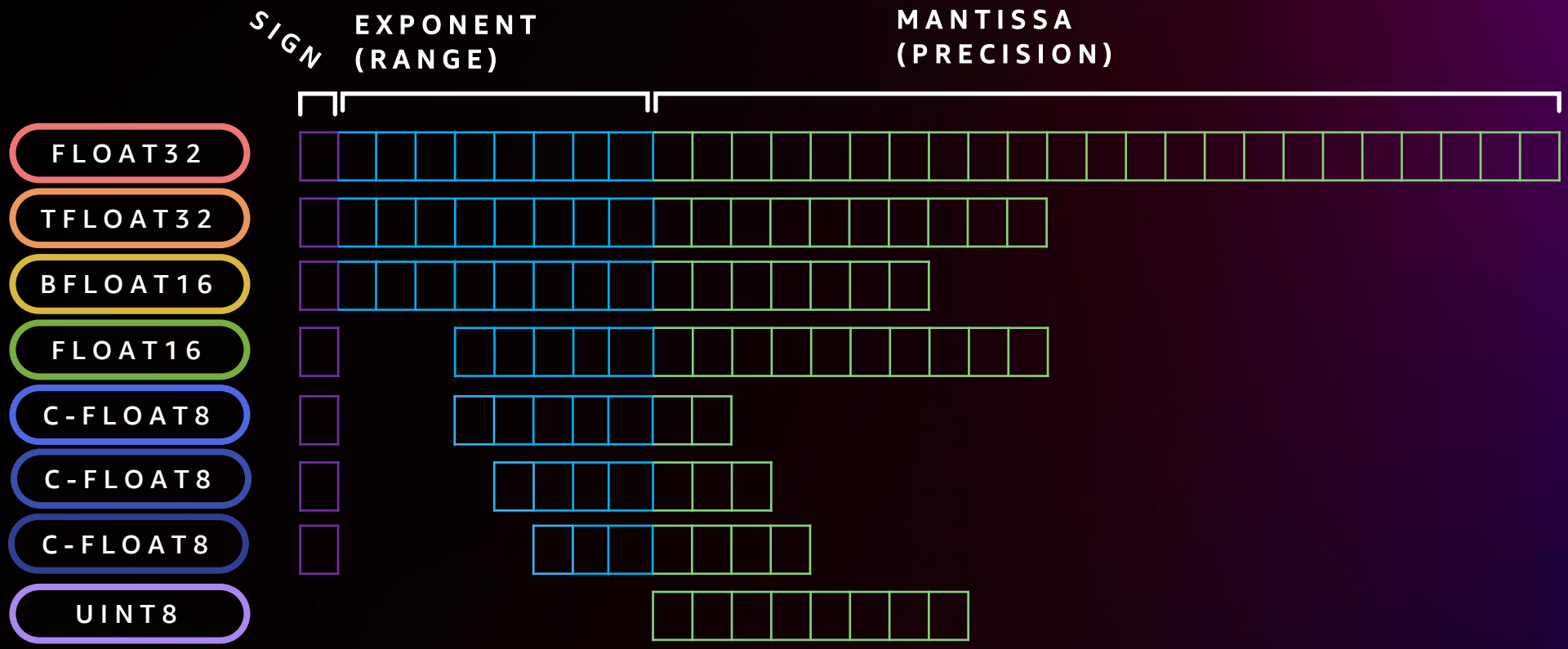
lower cost



* Measured on PyTorch BERT-Large

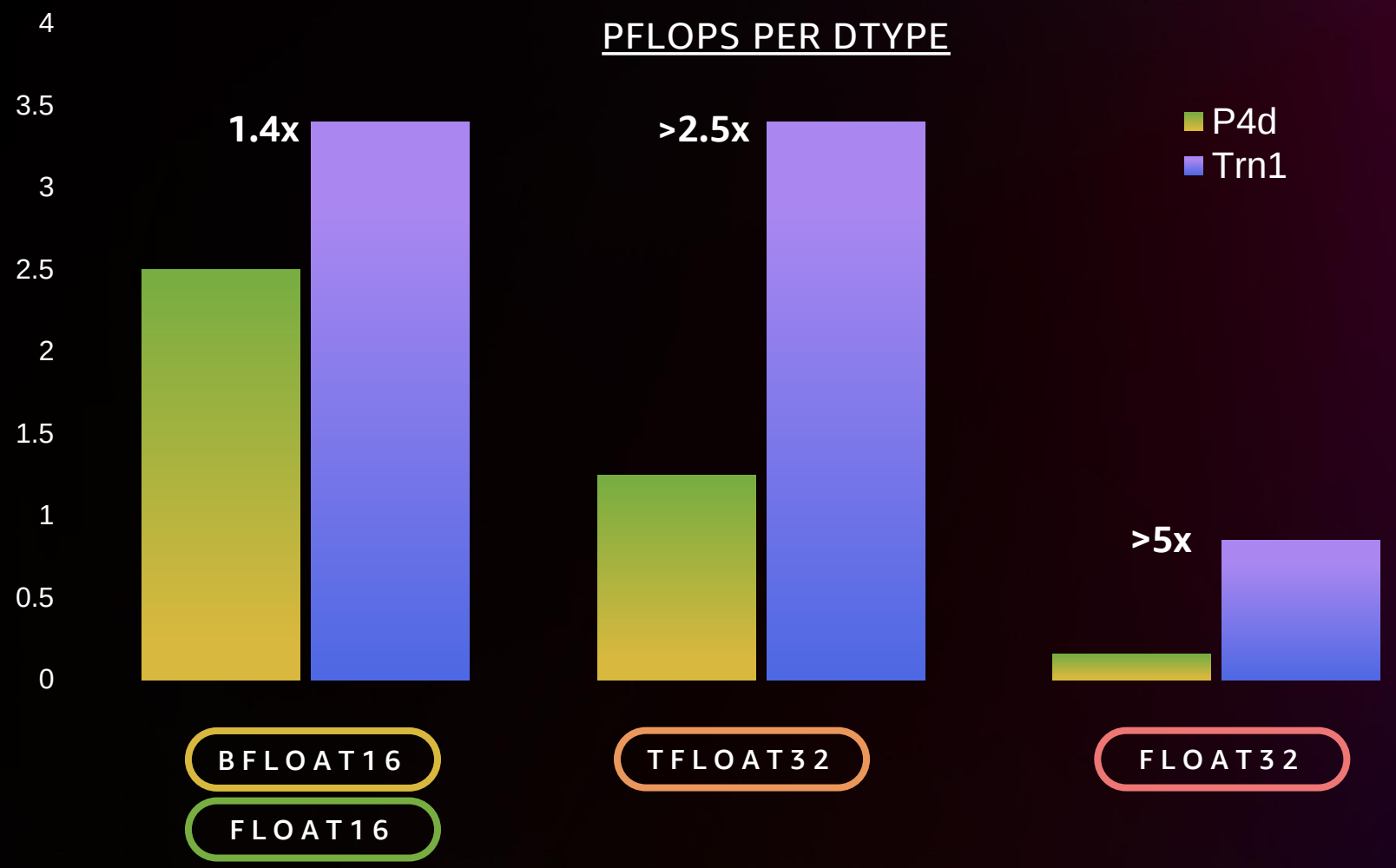
Rich data-type selection

CHOOSE THE RIGHT DATA-TYPE FOR YOUR WORKLOAD



Rich data-type selection

CHOOSE THE RIGHT DATA-TYPE FOR YOUR WORKLOAD



Stochastic rounding

IMPROVE TRAINING CONVERGENCE WITH NATIVE STOCHASTIC ROUNDING SUPPORT

Round nearest even

Round(0.9) = 1

PROBABILITY 100%

Round(0.2) = 0

PROBABILITY 100%

Stochastic rounding

IMPROVE TRAINING CONVERGENCE WITH NATIVE STOCHASTIC ROUNDING SUPPORT

Round nearest even

1.0

weight

0.2

gradient

Stochastic rounding

IMPROVE TRAINING CONVERGENCE WITH NATIVE STOCHASTIC ROUNDING SUPPORT

Round nearest even

Round(1.0 + 0.2) =

weight gradient

Stochastic rounding

IMPROVE TRAINING CONVERGENCE WITH NATIVE STOCHASTIC ROUNDING SUPPORT

Round nearest even

$$Round(1.0 + 0.2) = 1$$

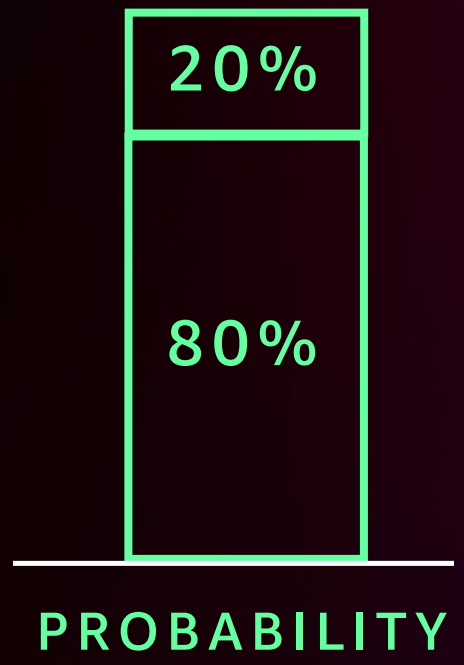
PROBABILITY 100%

Stochastic rounding

IMPROVE TRAINING CONVERGENCE WITH NATIVE STOCHASTIC ROUNDING SUPPORT

Stochastic rounding

$$\text{Round}(1.0+0.2) = \begin{cases} 2, \\ 1, \end{cases}$$



Stochastic rounding

IMPROVE TRAINING CONVERGENCE WITH NATIVE STOCHASTIC ROUNDING SUPPORT

Benefits of stochastic rounding

Customer momentum with AWS Trainium

Magic

ANTHROPIC

PyTorch

amazon

sprinklr

Money Forward

CACTUS

Helixon
华深智药

Wrap-up

Silicon innovation journey

BUILDING ON 15 YEARS OF INNOVATION

- Security
- Speed
- Innovation
- Price performance

Thank you!

Please complete the session survey in the **mobile app**