

The background features a dark blue gradient on the left, transitioning into a large, abstract, curved shape on the right. This shape is composed of various shades of purple and blue, with a bright orange-yellow highlight along its bottom edge. The overall design is modern and dynamic.

aws SUMMIT

ATLANTA | MAY 18-19, 2022

CMP201

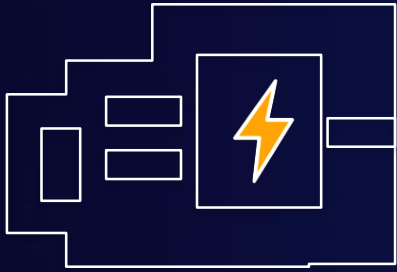
The journey of silicon innovation at AWS

Michael S. Fischer
Principal Specialist Solutions Architect
AWS



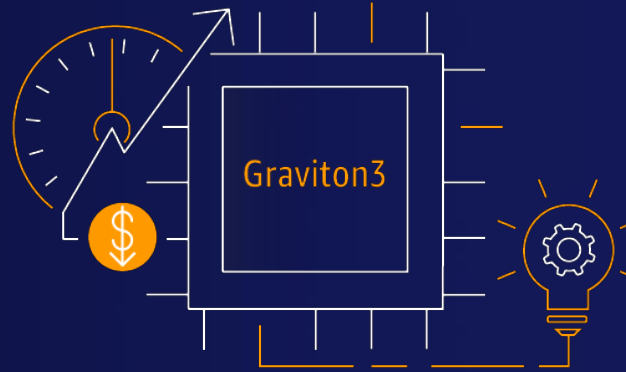
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Journey of silicon innovation at AWS



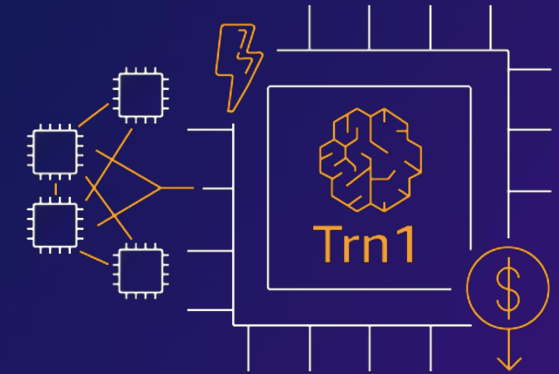
AWS Nitro System

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



AWS Graviton

Powerful and efficient, modern applications



AWS Inferentia and AWS Trainium

Machine learning acceleration

Why build our own chips?

Why build our own chips?

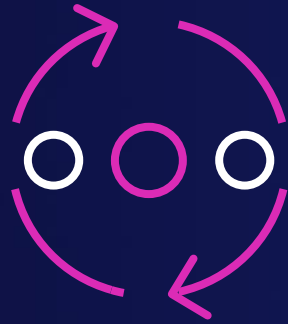


Specialization

Why build our own chips?



Specialization

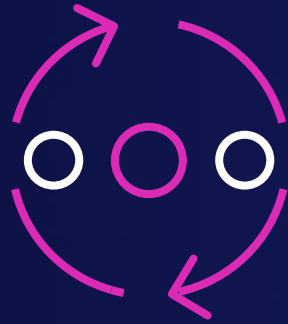


Speed

Why build our own chips?



Specialization



Speed

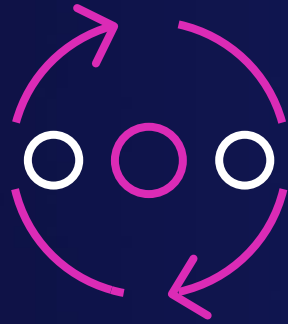


Innovation

Why build our own chips?



Specialization



Speed



Innovation



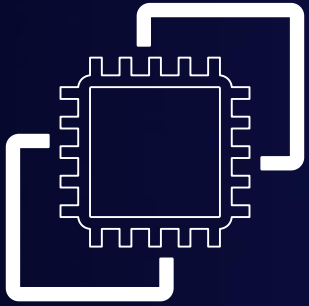
Security

AWS Nitro System

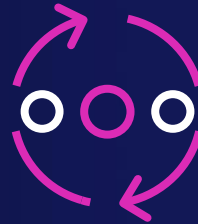


AWS Nitro System

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



AWS Nitro System



Improve throughput

Simplify hypervisor

Reduce latency & jitter

Bare-metal instances



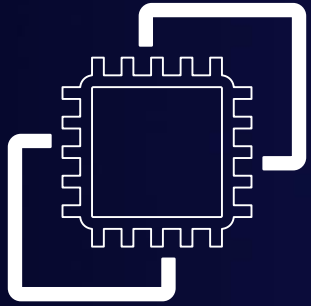
Transparent encryption

Hardware root of trust

No operator access

Narrow auditable APIs

AWS Nitro System



AWS Nitro System

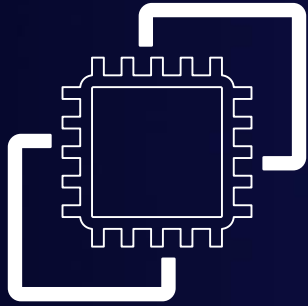
Purpose-built hardware/software

5 generations of custom chips

Hypervisor built for AWS

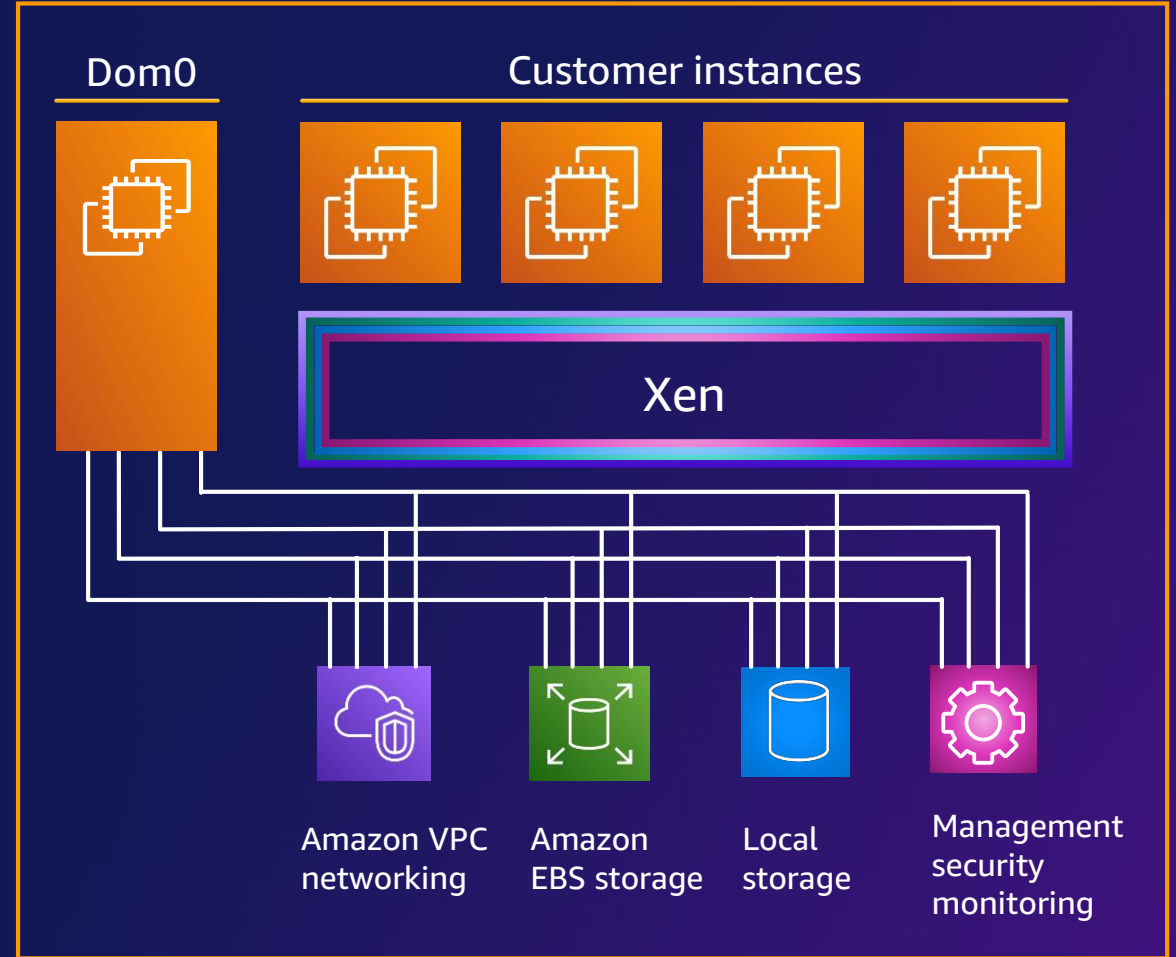
Powering over 500 instance types

Before Nitro . . .



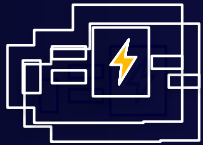
AWS Nitro System

Host CPU



AWS Nitro System

Nitro Cards



VPC networking

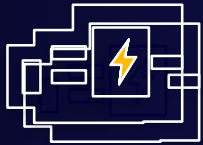
Amazon EBS

Nitro SSDs

System controller

Nitro Card for VPC networking

Nitro Cards



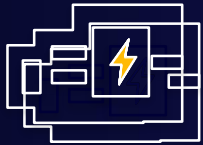
VPC networking
Amazon EBS
Nitro SSDs
System controller

VPC data-plane offload

Encapsulation, security groups, flow logs, routing, port mirroring, DHCP, DNS

Nitro Card for VPC networking

Nitro Cards



VPC networking

Amazon EBS

Nitro SSDs

System controller

VPC data-plane offload

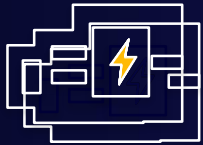
Encapsulation, security groups, flow logs, routing, port mirroring, DHCP, DNS

VPC encryption

Transparent end-to-end 256-bit encryption

Nitro Card for VPC networking

Nitro Cards



VPC networking

Amazon EBS

Nitro SSDs

System controller

VPC data-plane offload

Encapsulation, security groups, flow logs, routing, port mirroring, DHCP, DNS

VPC encryption

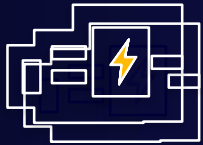
Transparent end-to-end 256-bit encryption

Elastic Network Adapter (ENA)

Extendible host interface

Nitro Card for VPC networking

Nitro Cards



VPC networking
Amazon EBS
Nitro SSDs
System controller

VPC data-plane offload

Encapsulation, security groups, flow logs, routing, port mirroring, DHCP, DNS

VPC encryption

Transparent end-to-end 256-bit encryption

Elastic Network Adapter (ENA)

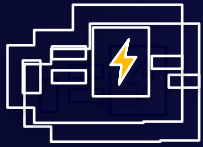
Extendible host interface

Elastic Fabric Adapter (EFA)

Kernel-bypass
Low latency at scale
Multi-pathing (SRD)

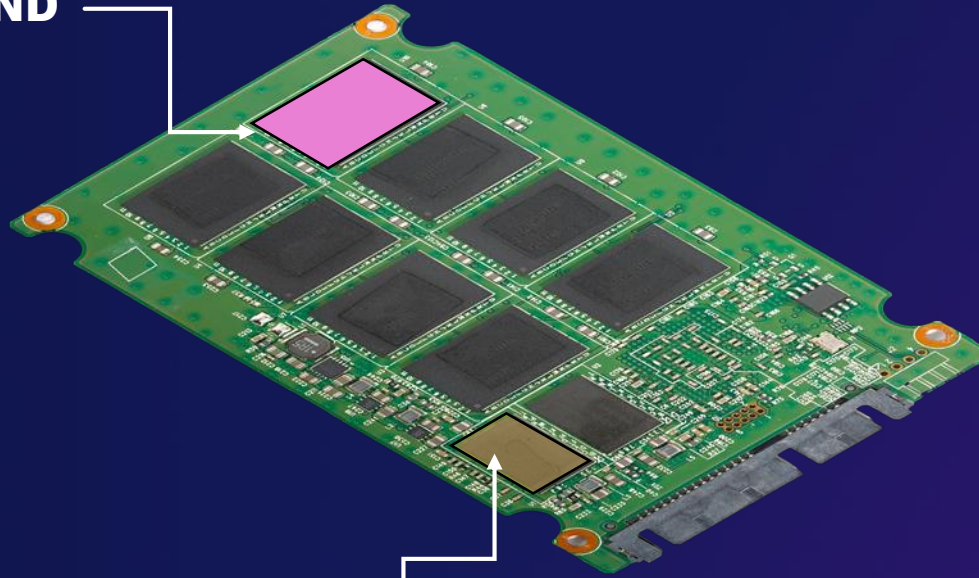
Nitro SSDs

Nitro Cards



VPC networking
Amazon EBS
Nitro SSDs
System controller

NAND

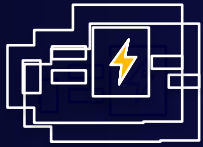


Flash Translation Layer (FTL)

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

Nitro SSDs

Nitro Cards



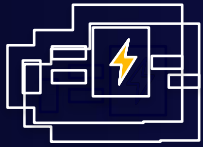
VPC networking
Amazon EBS
Nitro SSDs
System controller



Each vendor has their own FTL and differing performance

Nitro SSDs

Nitro Cards



VPC networking

Amazon EBS

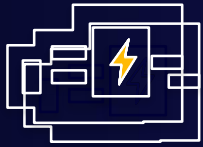
Nitro SSDs

System controller

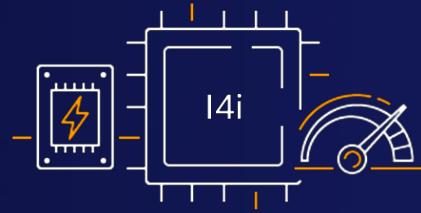


Nitro SSDs

Nitro Cards

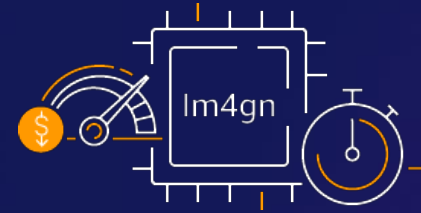


VPC networking
Amazon EBS
Nitro SSDs
System controller



60%

LOWER AVERAGE I/O LATENCY



75%

LOWER TAIL LATENCY





Reducing storage costs

UP TO 50% DECREASE IN TCO

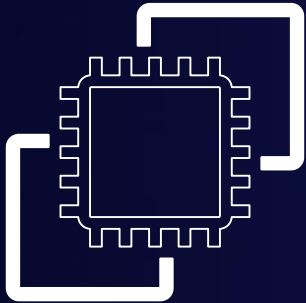


Splunk is a leading data platform provider and is designed to investigate, monitor, analyze, and act on data at any scale

"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

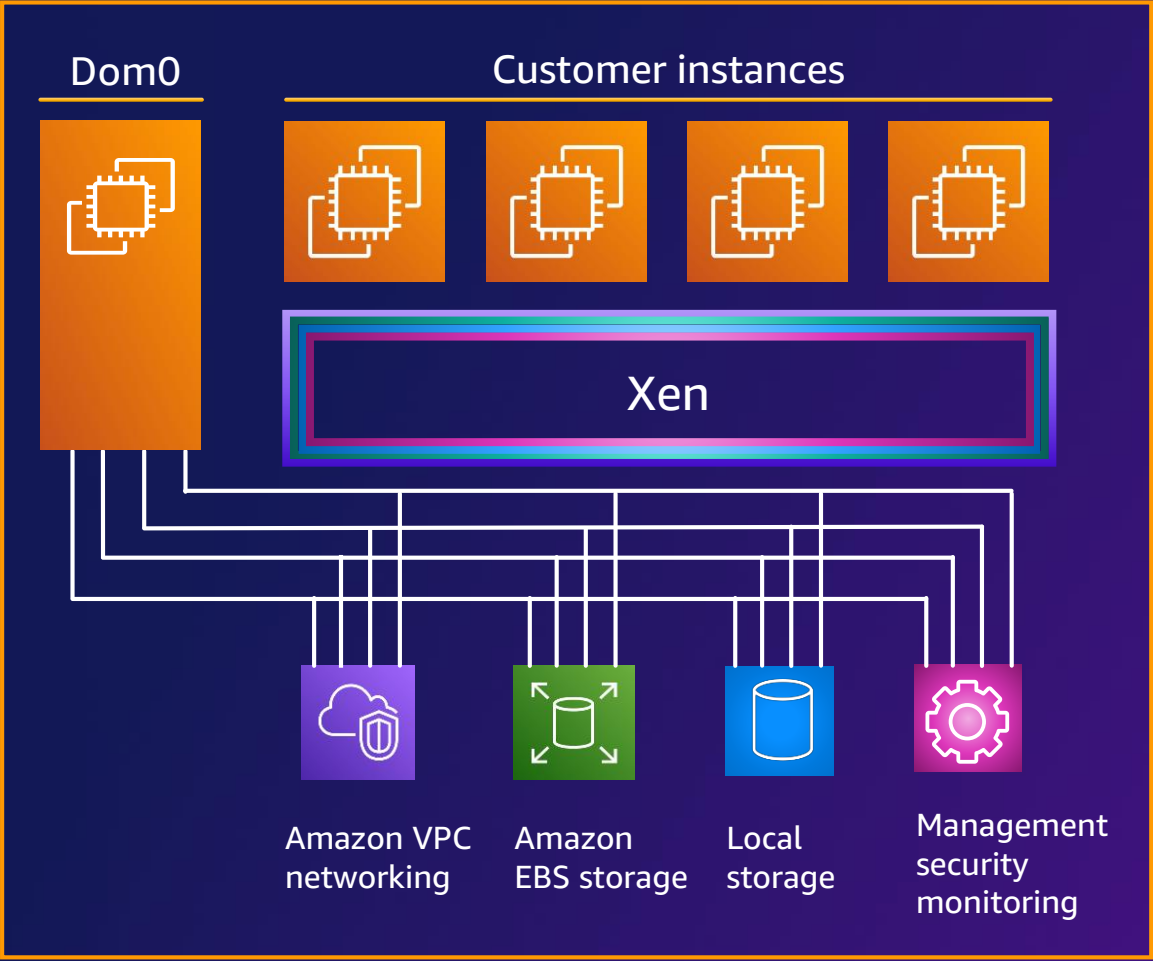


AWS Nitro System

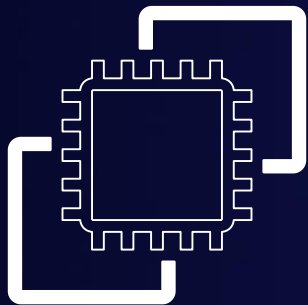
Nitro Card



Host CPU



AWS Nitro System



AWS Nitro System

Nitro Card



Amazon VPC
networking



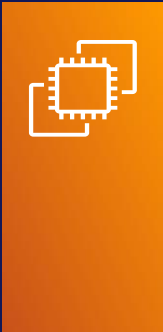
Amazon
EBS storage



Local
storage

Host CPU

Dom0



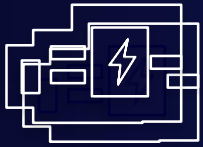
Customer instances



Management
security
monitoring

AWS Nitro System

Nitro Cards



VPC networking
Amazon EBS
Nitro SSDs
System controller

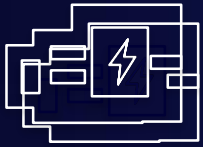
Nitro Hypervisor



Lightweight hypervisor
Memory and CPU allocation
Bare metal–like performance

AWS Nitro System

Nitro Cards



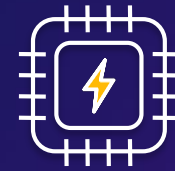
VPC networking
Amazon EBS
Nitro SSDs
System controller

Nitro Hypervisor



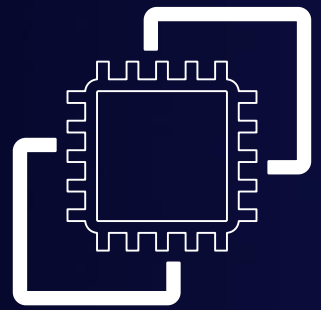
Lightweight hypervisor
Memory and CPU allocation
Bare metal–like performance

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



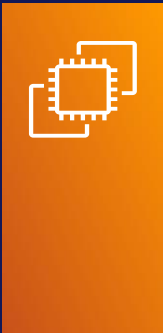
Local
storage



Management
security
monitoring

Host CPU

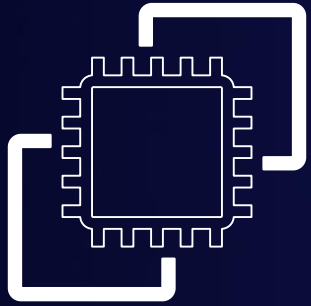
Dom0



Customer instances



Before Nitro . . .



AWS Nitro System

Nitro Card



Amazon VPC
networking



Amazon
EBS storage



Local
storage



Management
security
monitoring

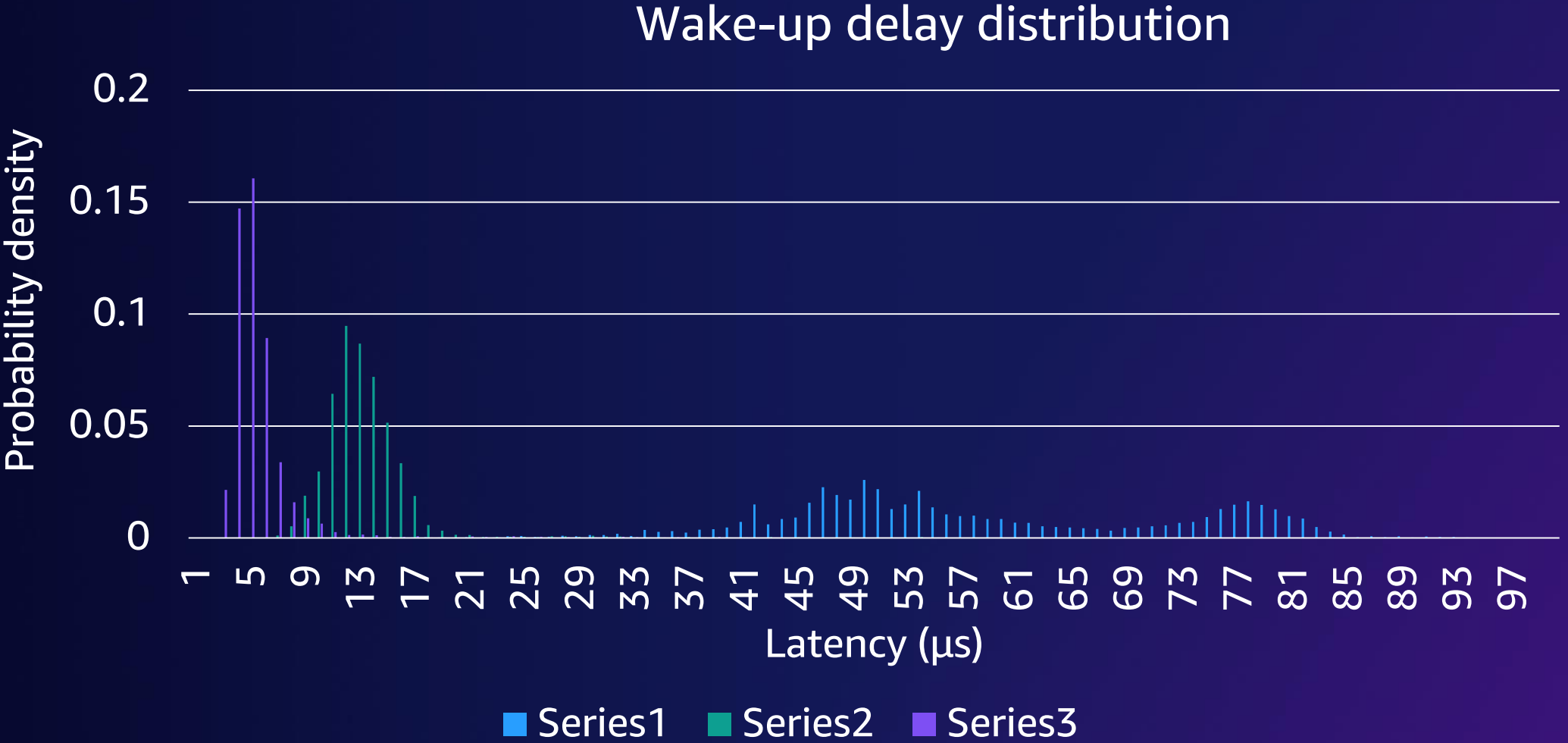
Host CPU

Customer instances

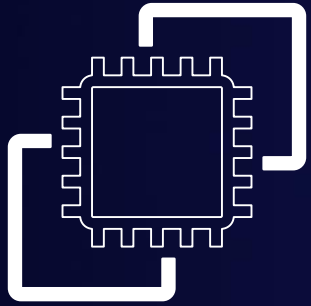


Nitro Hypervisor

AWS Nitro System – Reduced jitter



Nitro security – Confidential compute



AWS Nitro System

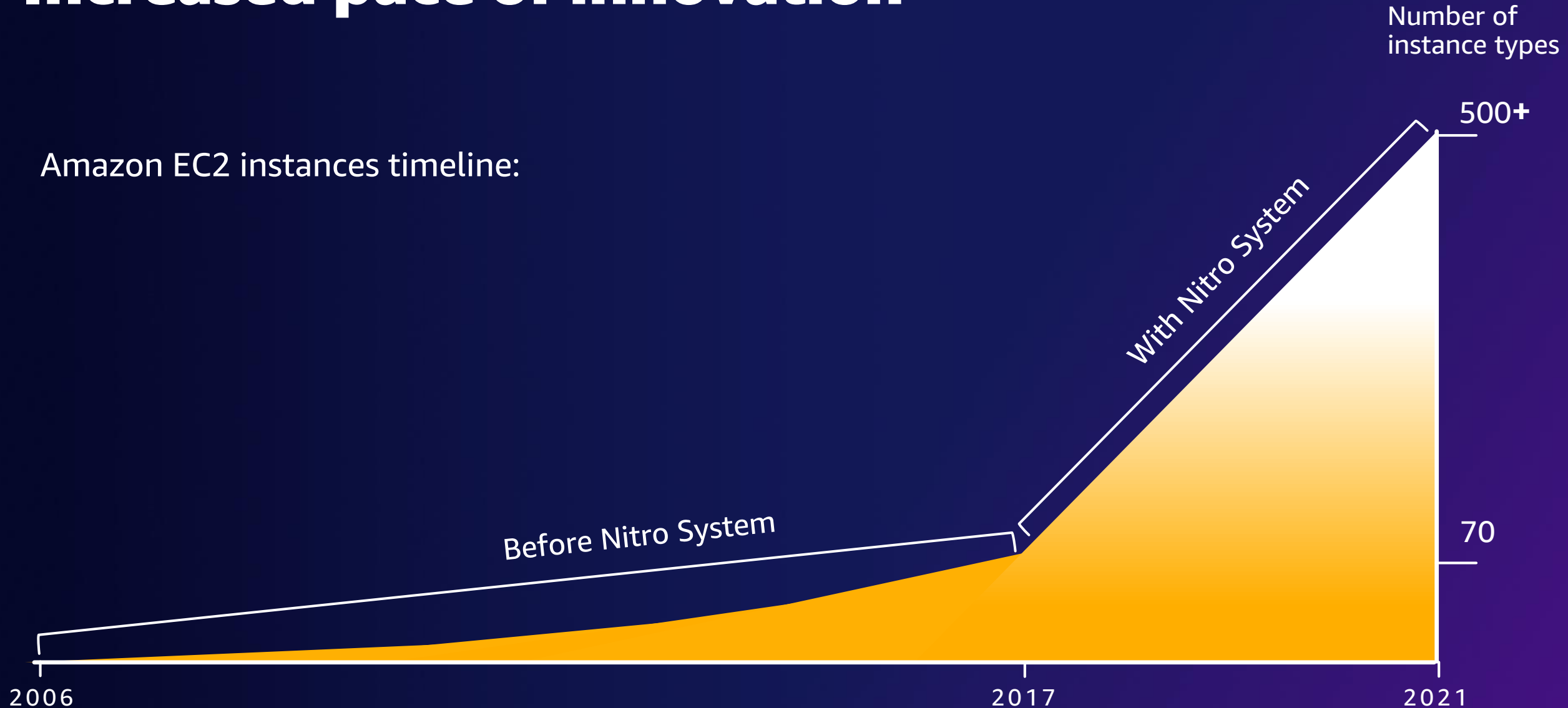
All interactions with the AWS Nitro System are through narrow, authorized, and authenticated APIs

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

There is no interactive access

Increased pace of innovation

Amazon EC2 instances timeline:



AWS Graviton



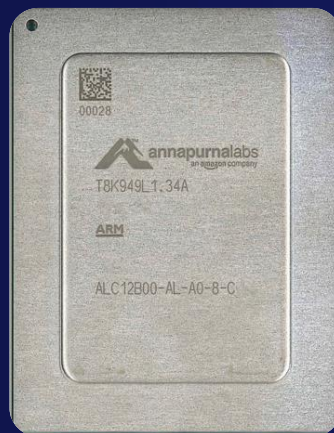
AWS Graviton

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 23 AWS REGIONS GLOBALLY



AWS managed services supporting Graviton2

EXTENDING THE GRAVITON2 PRICE PERFORMANCE TO MANAGED SERVICES

Databases



Amazon
DocumentDB



Amazon
Aurora



Amazon
RDS



Amazon
ElastiCache



Amazon
MemoryDB

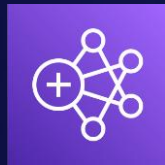


Amazon
Neptune

Analytics



Amazon
OpenSearch
Service



Amazon
EMR

Compute



AWS
Lambda



AWS Fargate



AWS Elastic
Beanstalk

Customers adopting Graviton2

DIRECTV
stream

 Discovery



intuit



lyft



SmugMug 

nielsen

 NextRoll

DOMO



hotelbeds

redbox

Lightyear 



 rinklr

splunk>



OneGame

AWS Graviton3

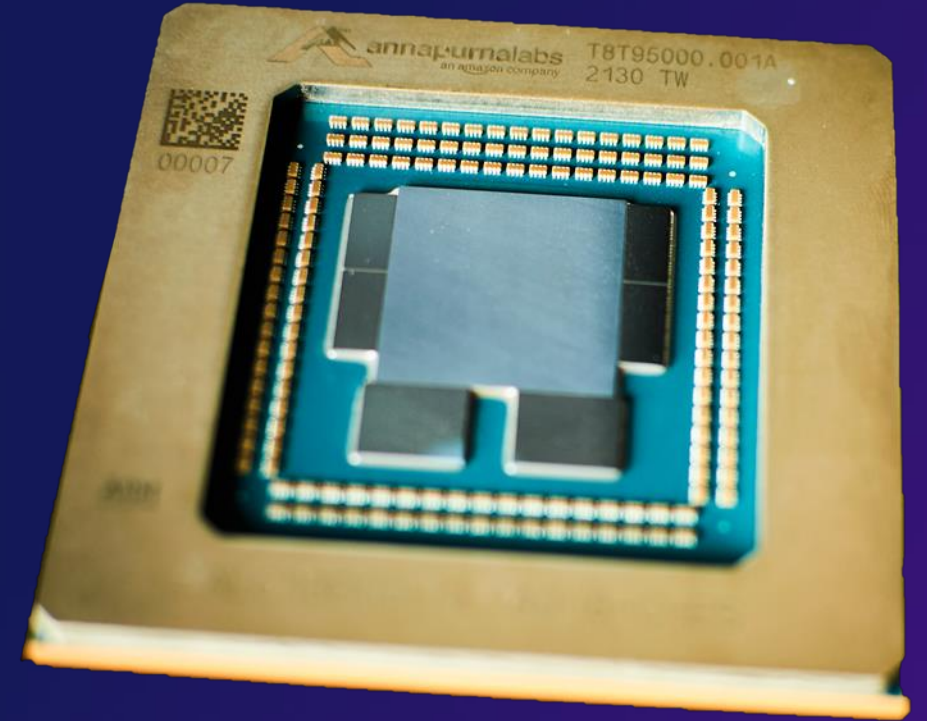
Third-generation Graviton processor

Chiplet-based design: Seven silicon dies

~55 billion transistors

25%+ higher performance/core vs. Graviton2

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



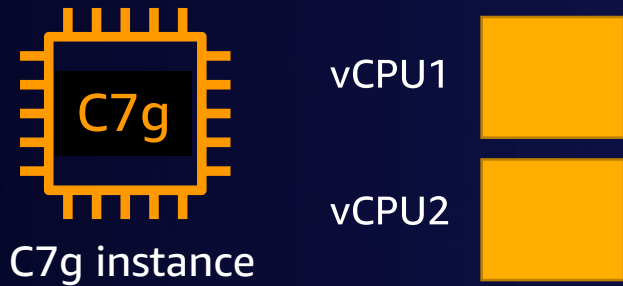
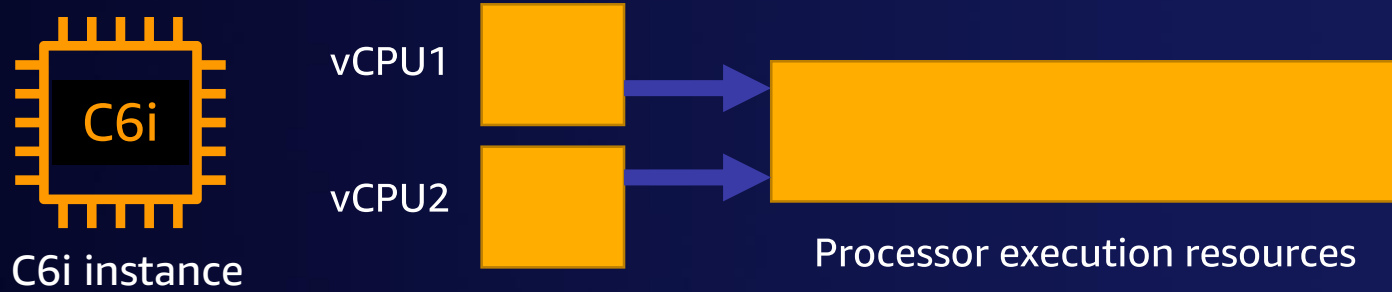
**bfloat16
256b SVE**

**2x mem ops
enhanced
prefetching**

**~2x
TLS**



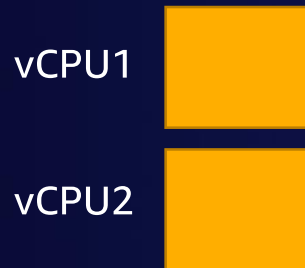
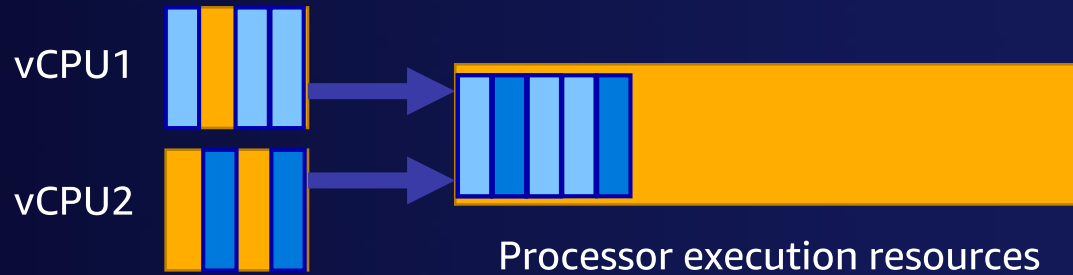
Graviton3 – vCPU



Every vCPU is a physical core

No simultaneous
multithreading (SMT)

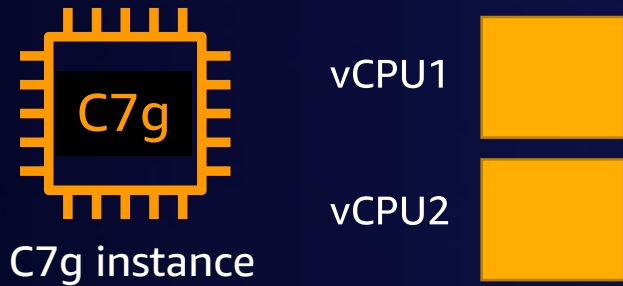
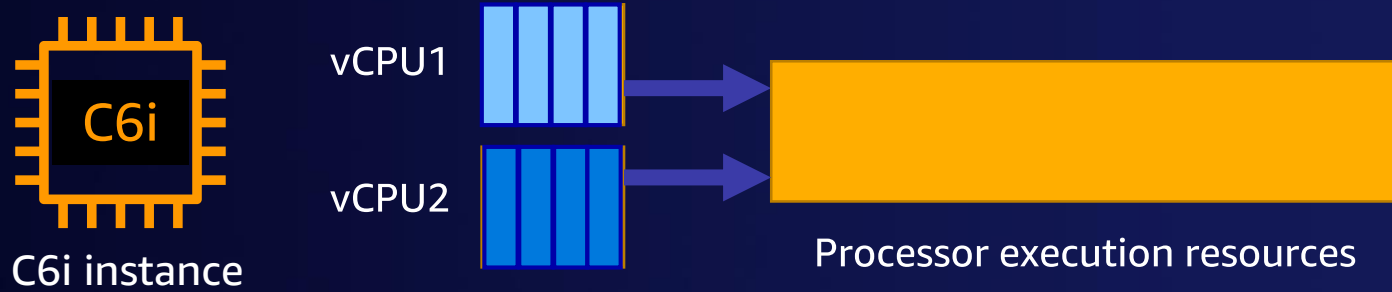
Graviton3 – vCPU



Every vCPU is a physical core

No simultaneous
multithreading (SMT)

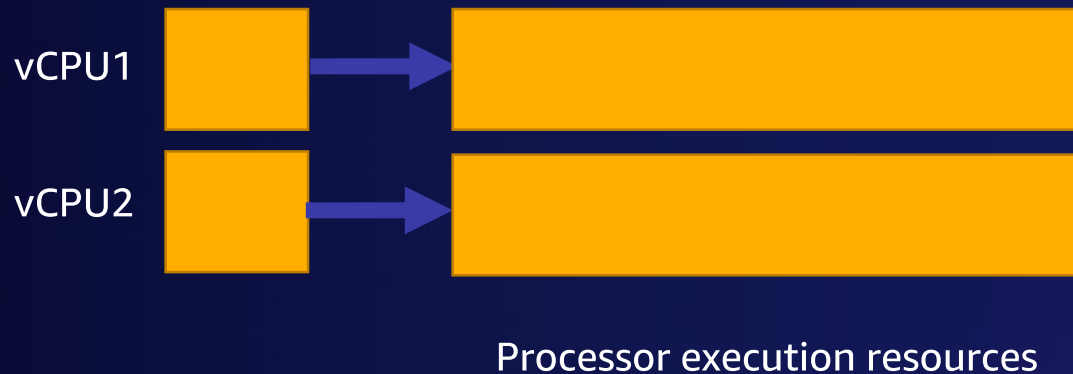
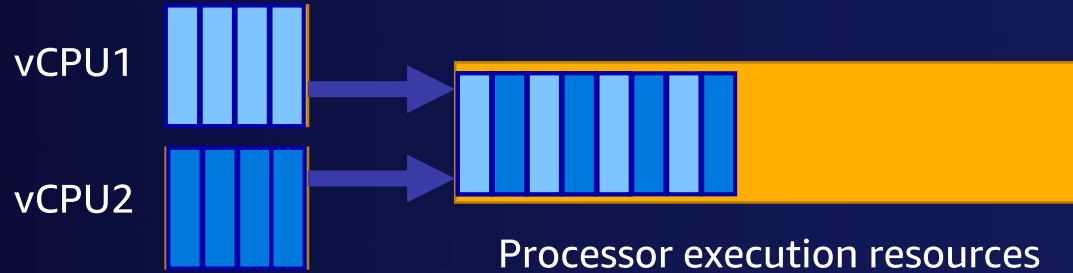
Graviton3 – vCPU



Every vCPU is a physical core

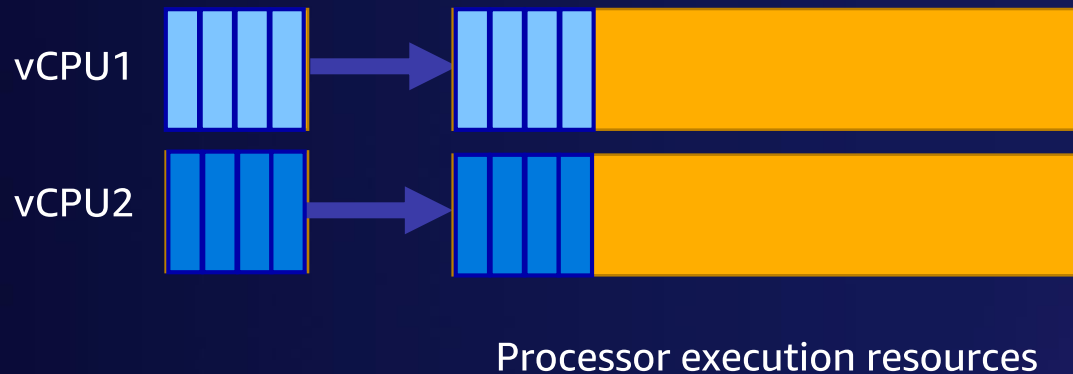
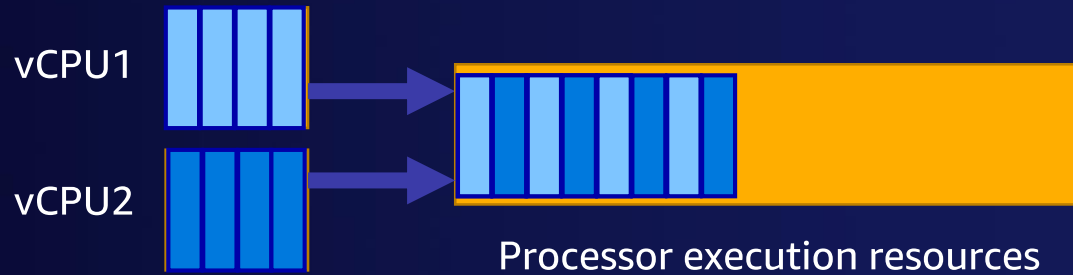
No simultaneous
multithreading (SMT)

Graviton3 – vCPU



Every vCPU is a physical core
No simultaneous
multithreading (SMT)

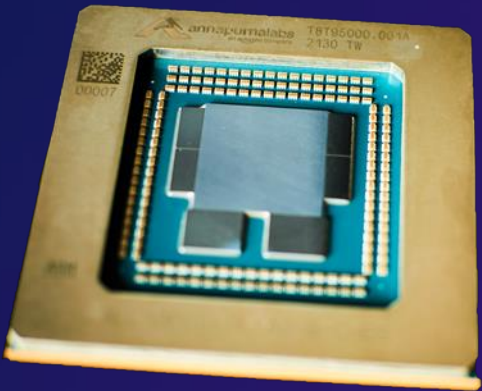
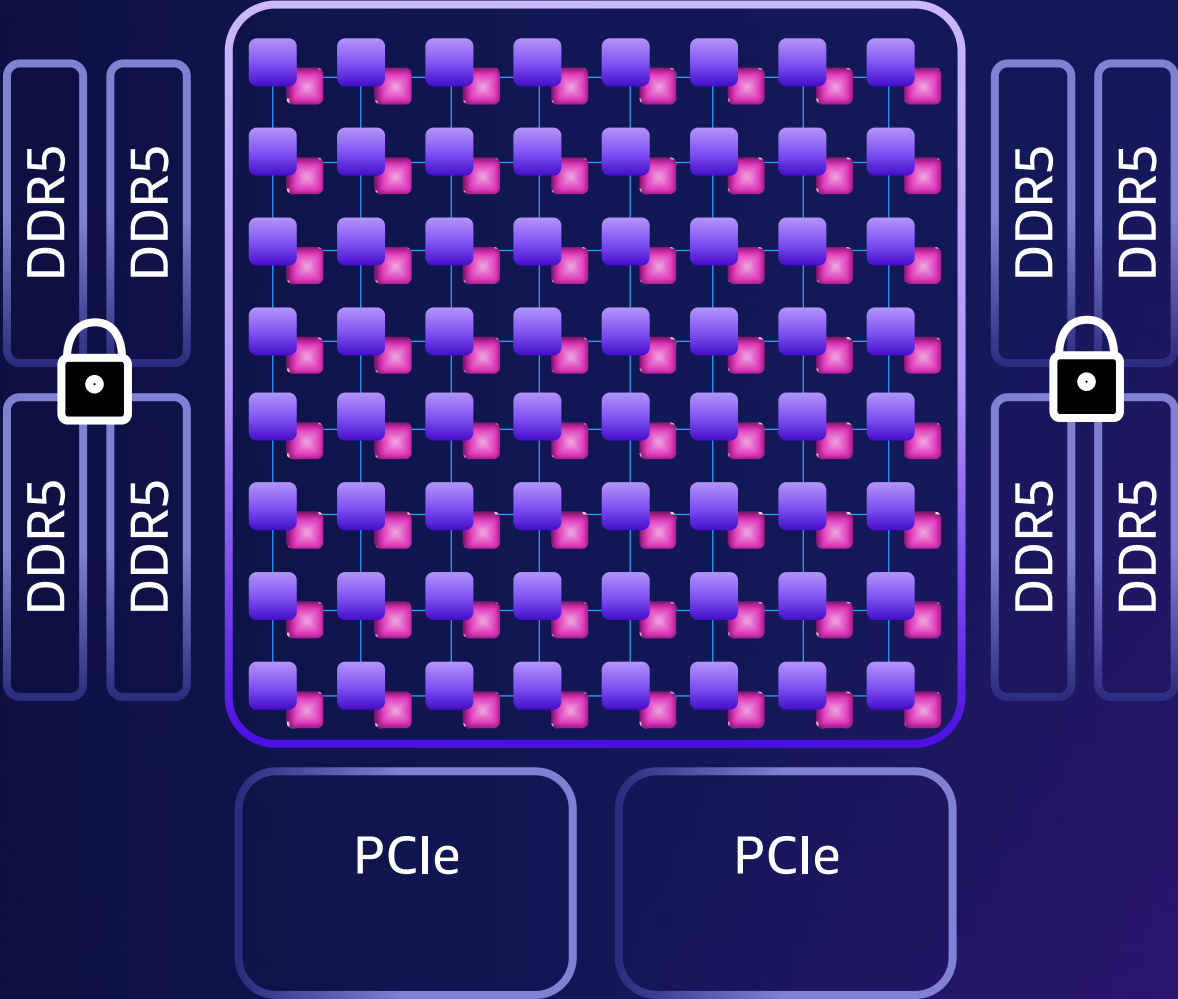
Graviton3 – vCPU



Every vCPU is a physical core

No simultaneous
multithreading (SMT)

Graviton3 – Chiplet design



Innovation at the server level

Typical servers

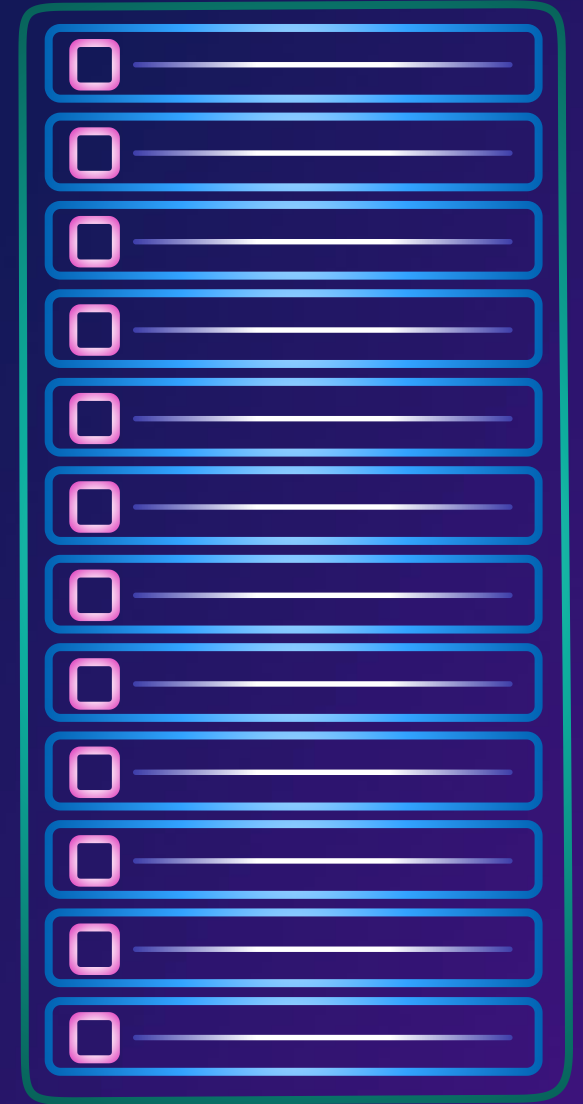
2 sockets/server

42U rack

Traditionally run out of rack power before space

Graviton 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

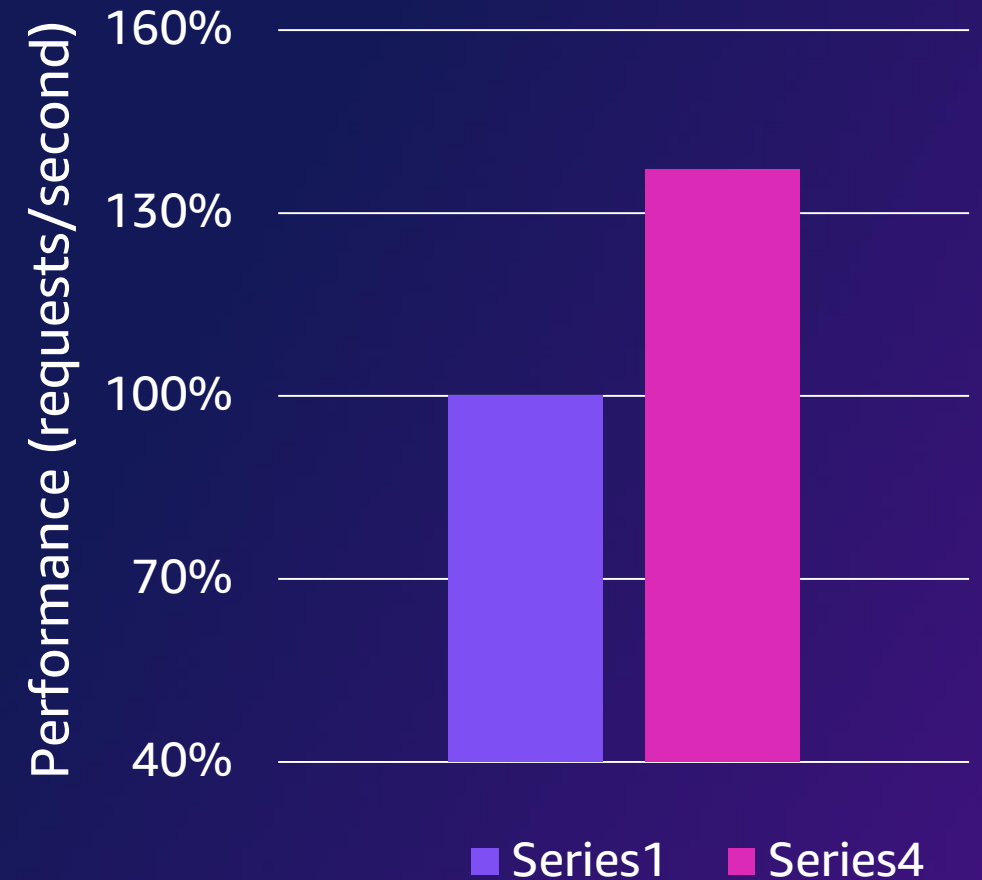
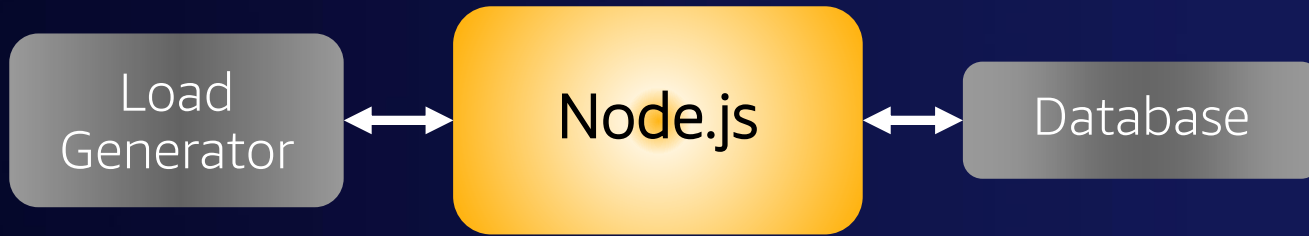
Increase the sockets/rack 50%

Come closer to the power requirements of other vendor solutions

Nitro Card is capable of managing three sockets simultaneously



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

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."



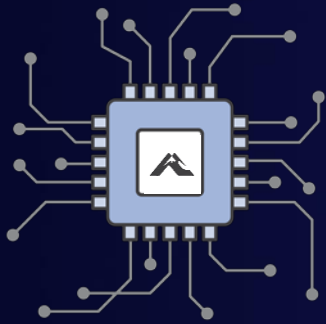
"They are suitable for even the most demanding latency sensitive workloads while providing significant price performance benefits."

Formula One
Group

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

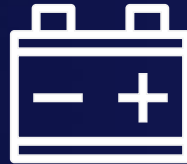
Graviton3-based seventh-generation EC2 instances

Graviton3



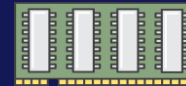
Industry-leading performance

Efficiency



Use up to 60% less energy for the same performance than comparable EC2 instances

DDR5



First DDR5 system in Amazon EC2

Best price performance



Best price performance in Amazon EC2

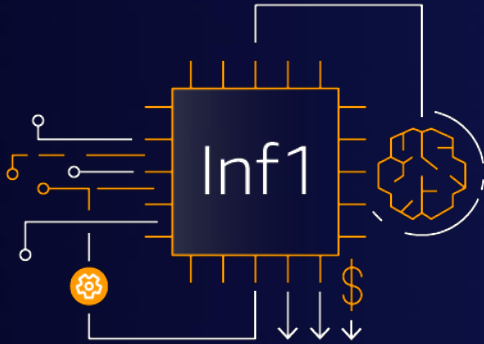
Sign up for the C7g preview at aws.amazon.com/ec2/instance-types/c7g

AWS Inference and AWS Trainium



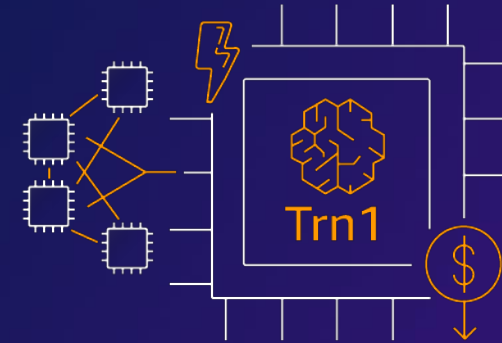
AWS chips optimized for deep learning

AWS Inferentia



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

AWS Trainium



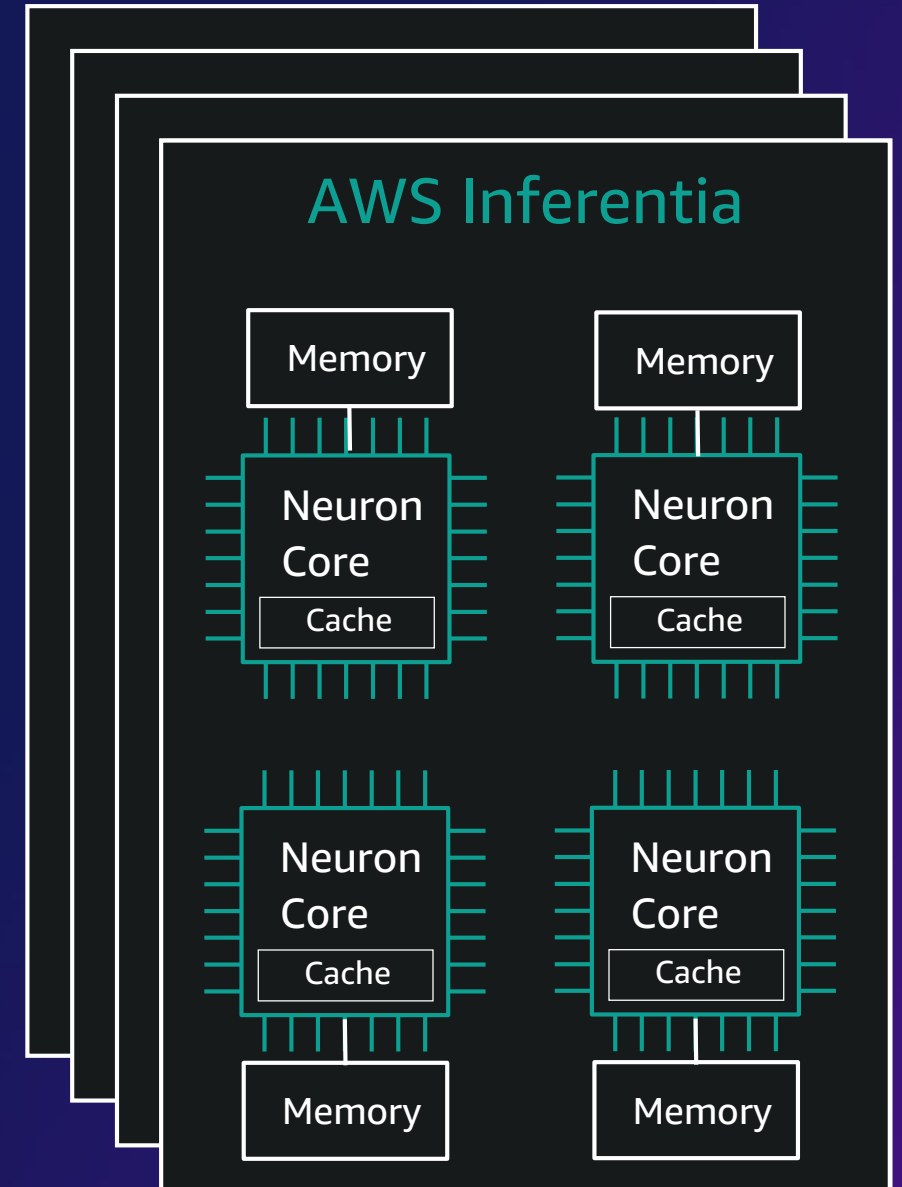
Purpose-built for the **most cost-efficient and fast DL training in the cloud** for a broad spectrum of applications



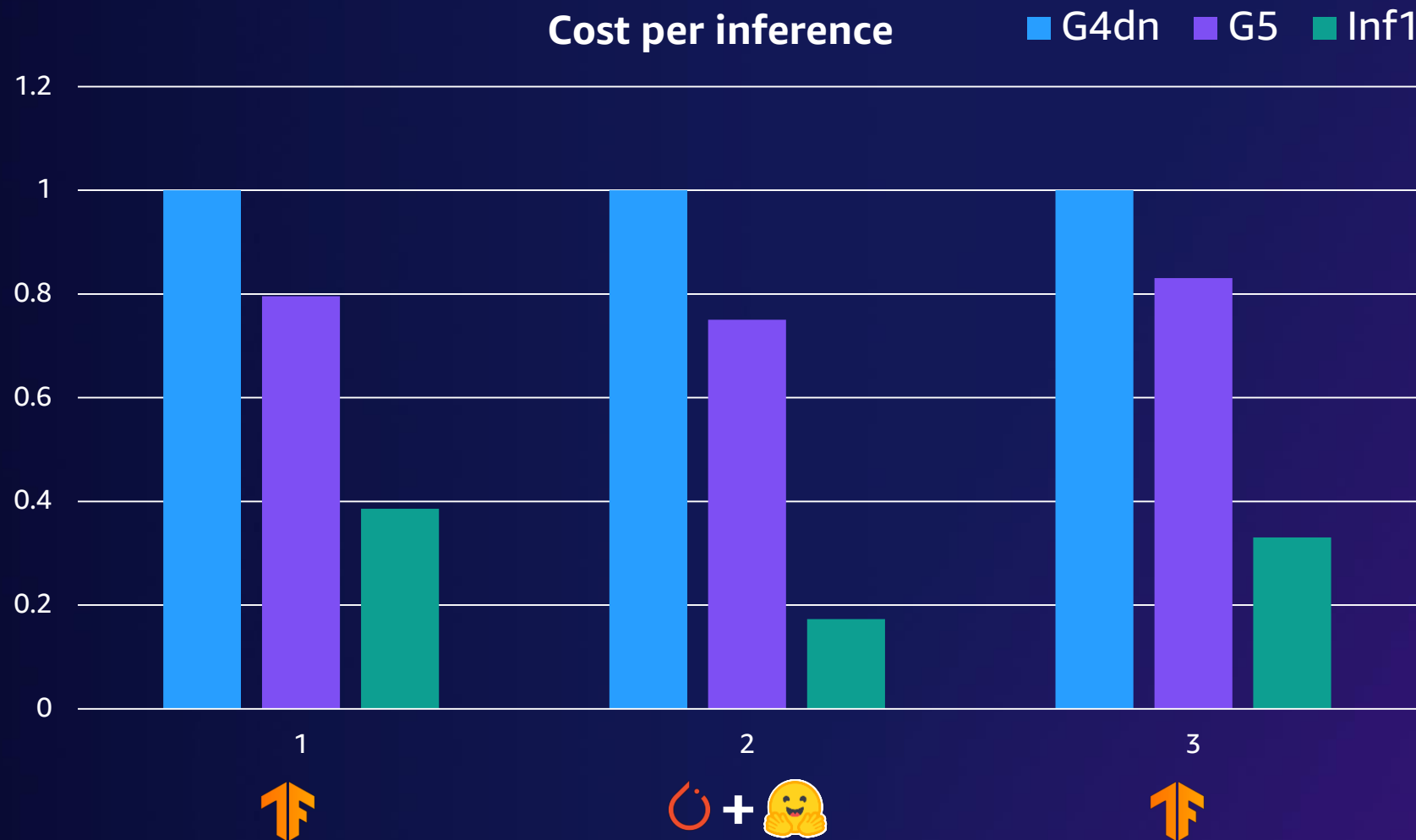
Seamless **integration with ML frameworks** like TensorFlow and PyTorch with minimal code changes

AWS Inferentia

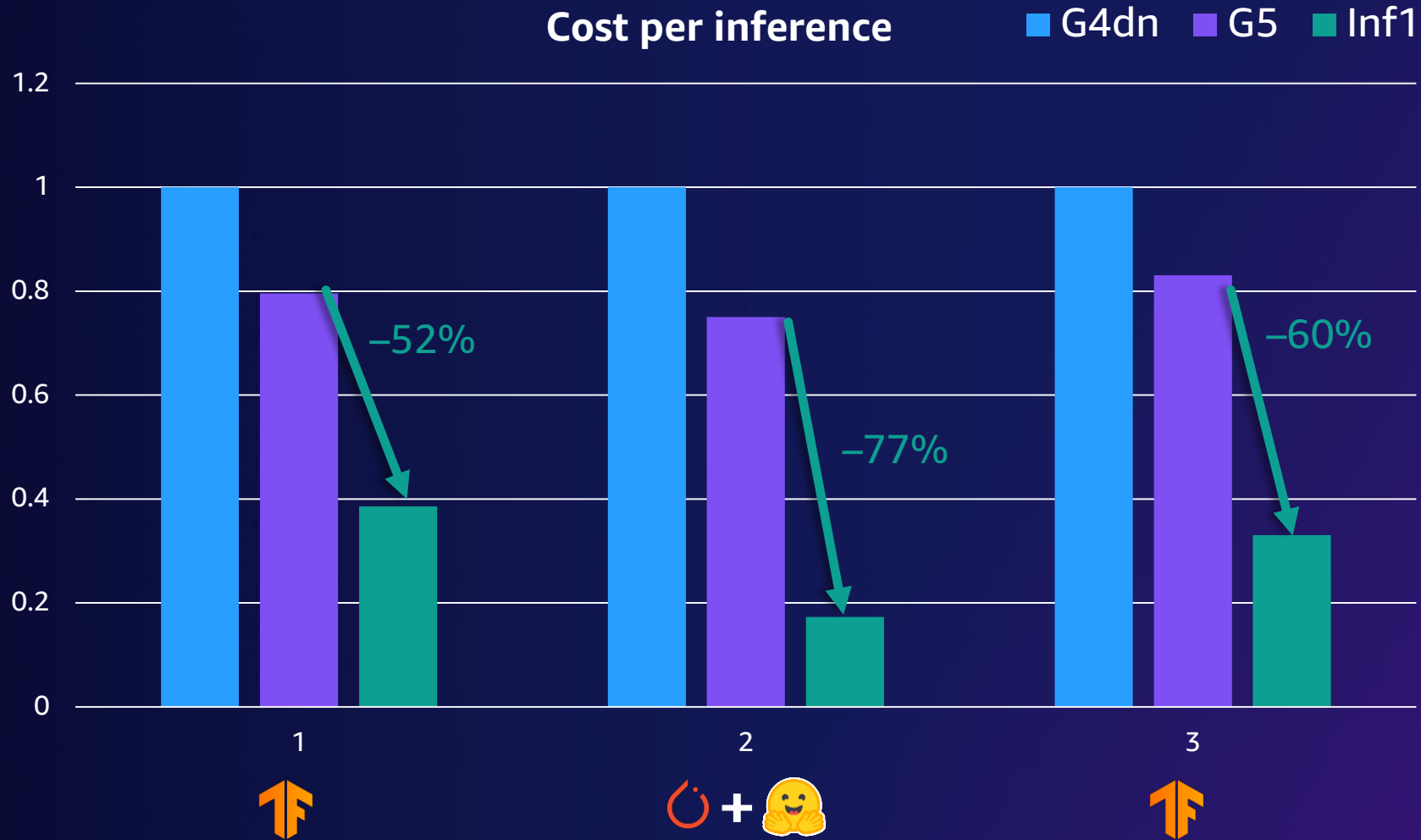
- 4 NeuronCores
- Up to 128 TOPs/chip
- Co-optimize throughput and latency
 - Large on-chip caches
 - Fast chip-to-chip interconnect
- Ease of use!
 - Supported in popular ML frameworks
 - FP16, BF16, INT8



AWS Inferentia – Cost savings

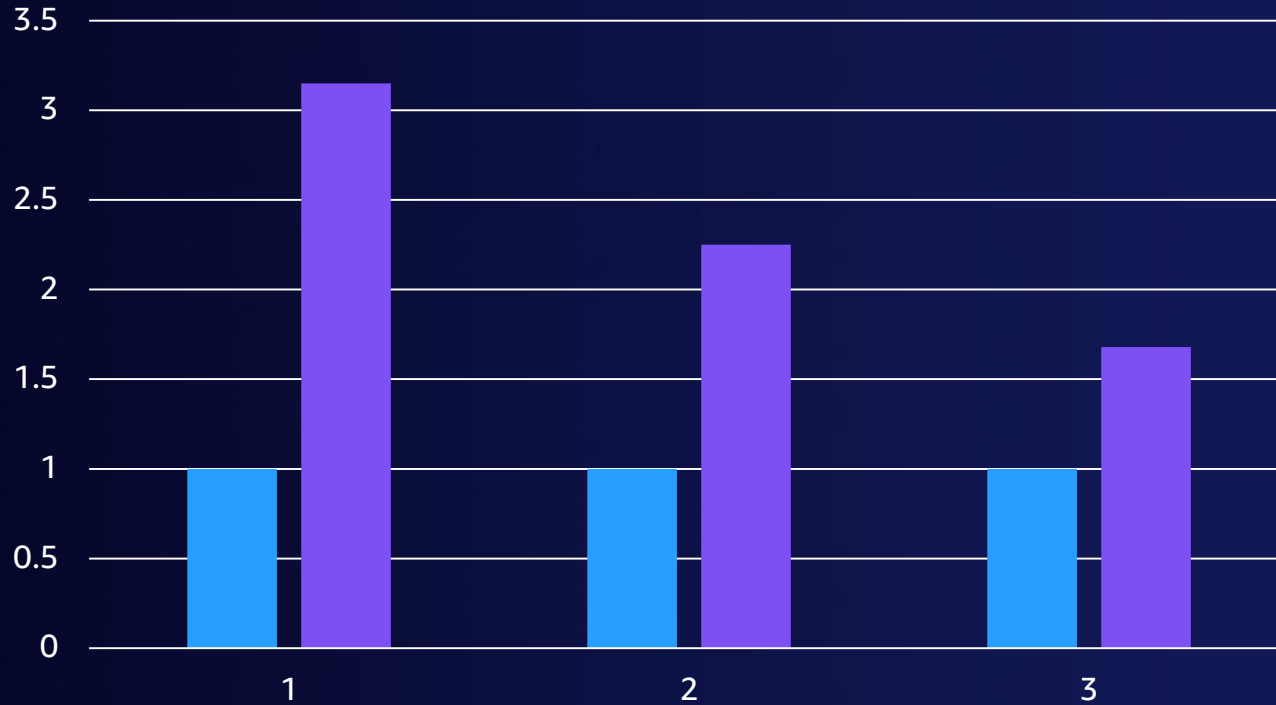


AWS Inferentia – Cost savings



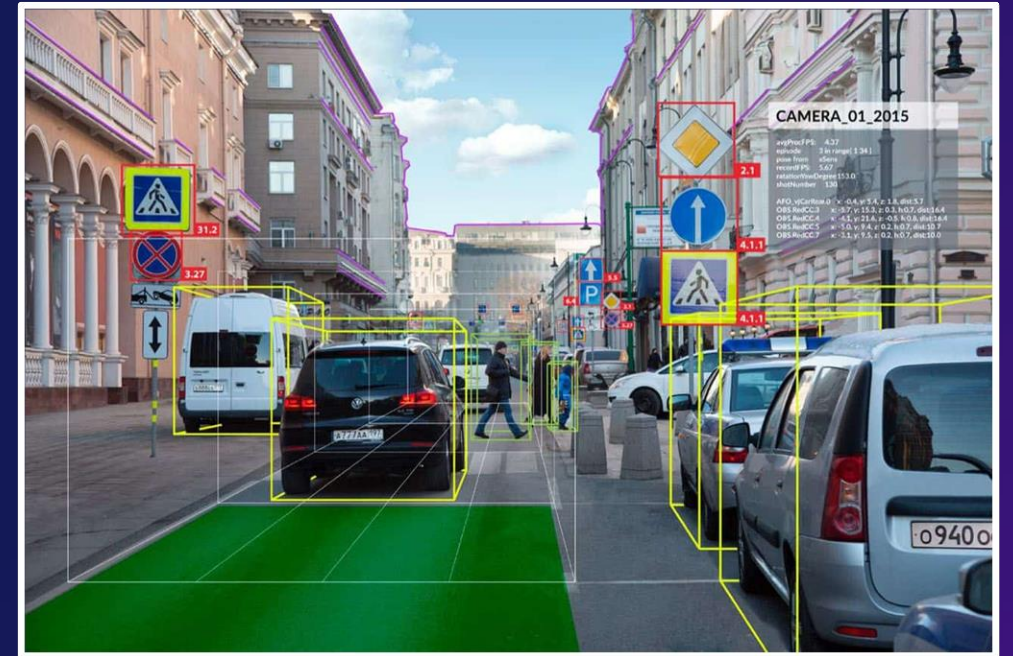
AWS Inferentia – Sustainable performance

Performance per watt



■ Series1

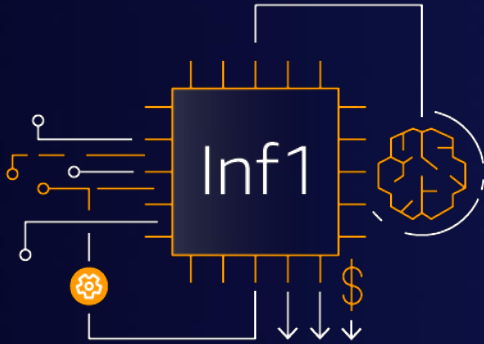
Measured 2-hour run, TensorFlow YOLOv4
(higher is better)



Objects in an image, as detected by YOLOv4

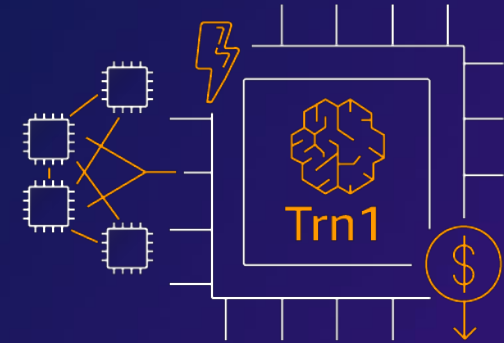
AWS chips optimized for deep learning

AWS Inferentia



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

AWS Trainium



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

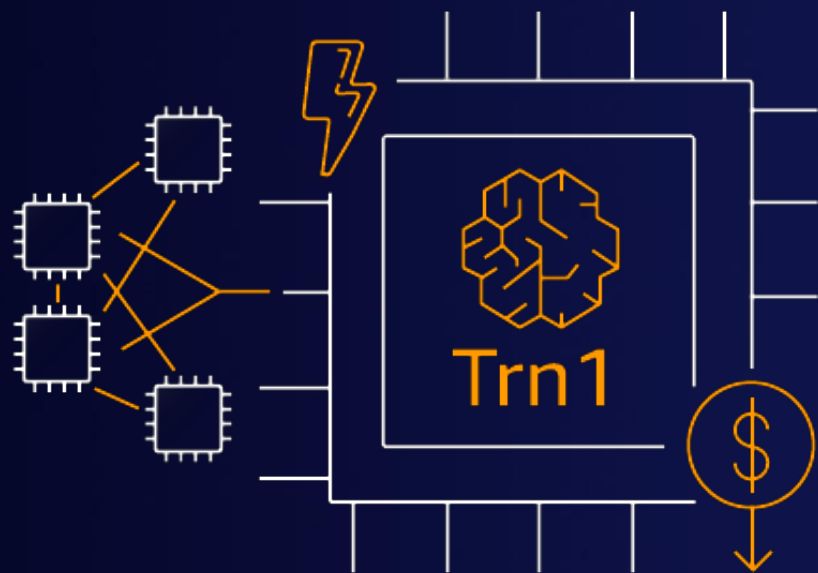


Seamless **integration with ML frameworks** like TensorFlow and PyTorch with minimal code changes

Amazon EC2 Trn1 instances

In preview

THE BEST PRICE PERFORMANCE FOR DL TRAINING IN THE CLOUD



60% higher accelerator memory (vs. P4d)

2x network BW (vs. P4d)

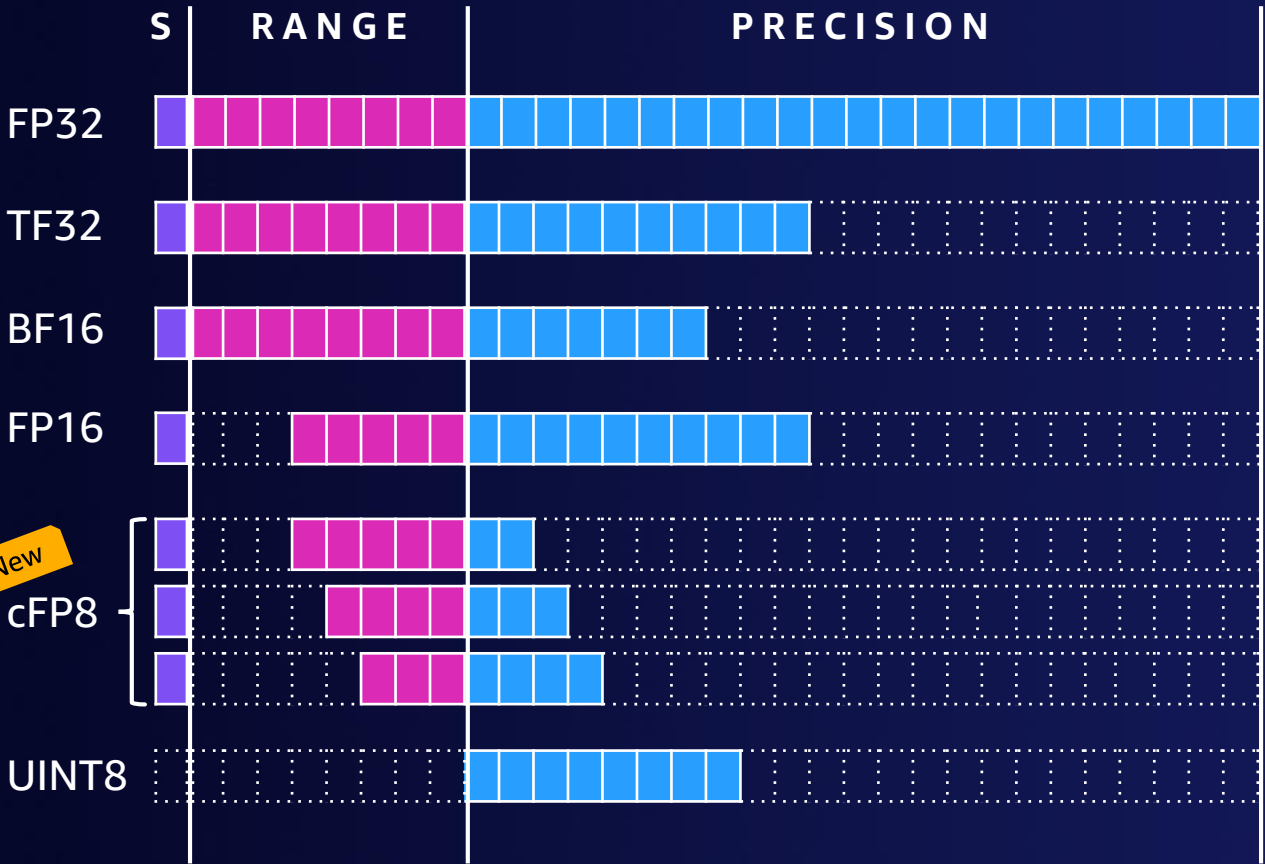
Native support for PyTorch and TensorFlow

Train on Trn1 and deploy anywhere

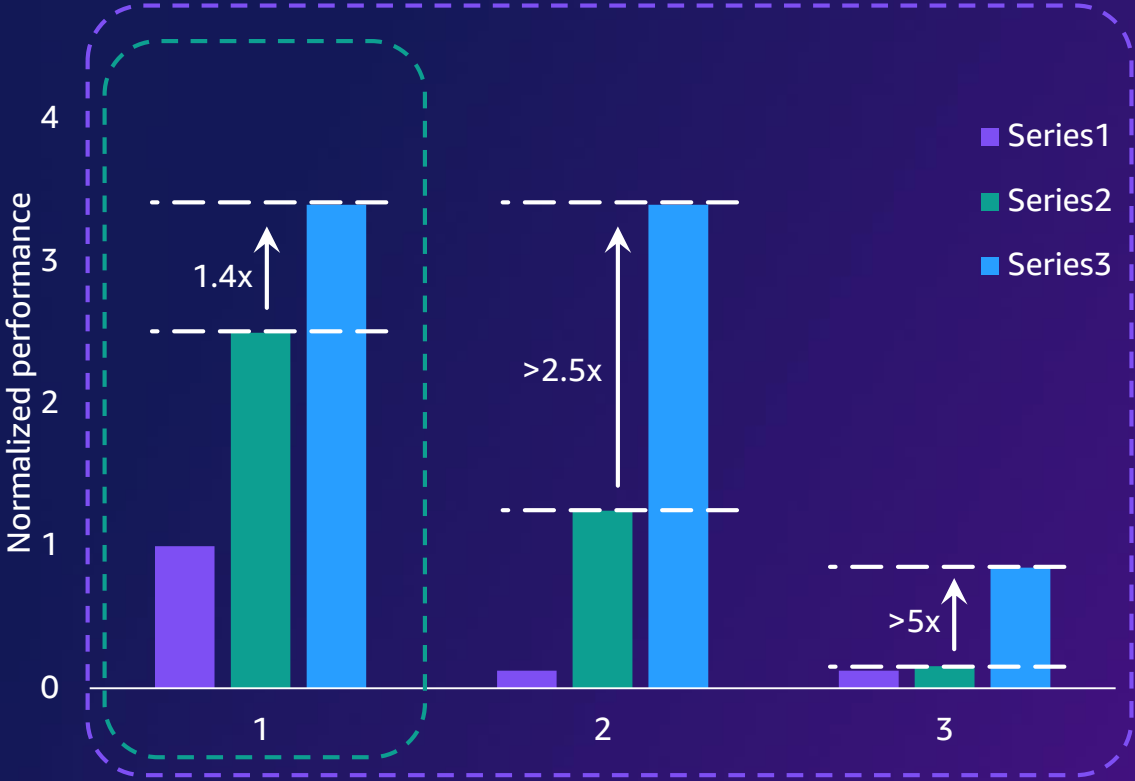
Instance size*	vCPUs	Trainium chips	Accelerator memory	NeuronLink	Instance memory	Instance networking	Local instance storage
Trn1.2xlarge	8	1	32 GB	N/A	32 GB	Up to 10Gbps	500 GB NVMe
Trn1.32xlarge	128	16	512 GB	768 GB/sec	512 GB	800 Gbps	8 TB NVMe

Trn1 – Fastest across data types

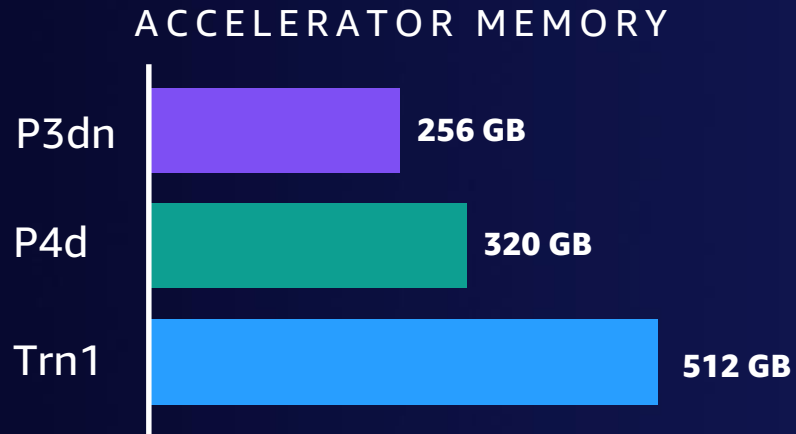
SUPPORTED DATA TYPES



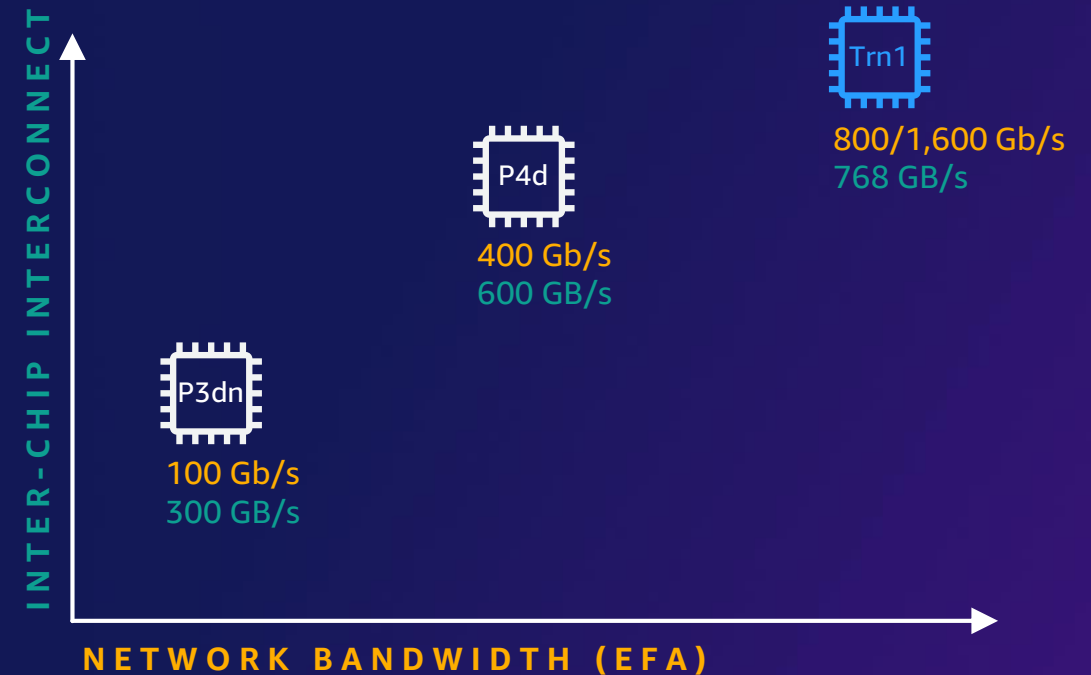
TFLOPS PER DATA TYPE



Trn1 built for scale-out



Large in-server accelerator memory



High-bandwidth and low-latency interconnect

Wrap-up

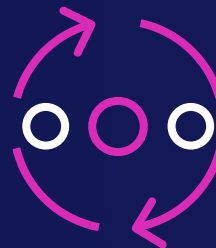


Silicon innovation journey

BUILDING ON 15 YEARS OF INNOVATION



Security



Speed



Innovation



Price performance

Learn in-demand AWS Cloud skills



AWS Skill Builder

Access **500+ free** digital courses and Learning Plans

Explore resources with a variety of skill levels and **16+** languages to meet your learning needs

Deepen your skills with digital learning on demand



Train now



AWS Certifications

Earn an industry-recognized credential

Receive Foundational, Associate, Professional, and Specialty certifications

Join the **AWS Certified community** and get exclusive benefits



Access **new** exam guides

Thank you!

