
AWS
re:Invent

CMP313

The Journey of silicon innovation at AWS

Ali Saidi

Sr. Principal Engineer

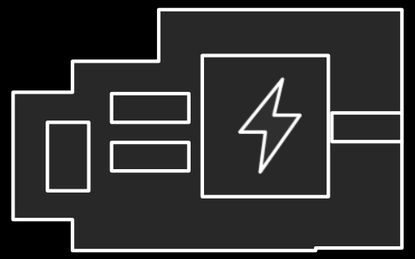
AWS

AWS re:Invent

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

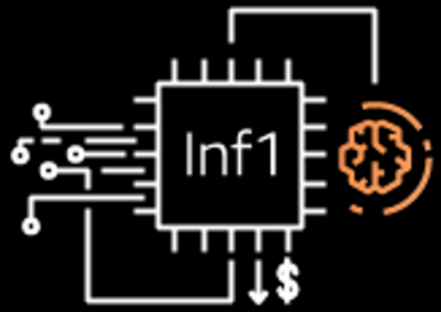
Journey of silicon innovation at AWS

- Why do we build our own chips?
- How does it help customers?
- Deep dive –



AWS Nitro System

Hypervisor, network, storage, and security



AWS Inferentia

Machine learning, hardware and software at scale



AWS Graviton2

Powerful and efficient, modern applications

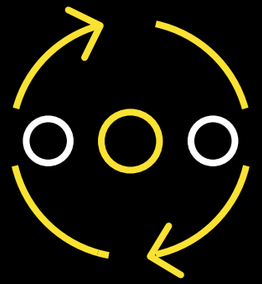
Why build our own chips?



Innovation



Specialization

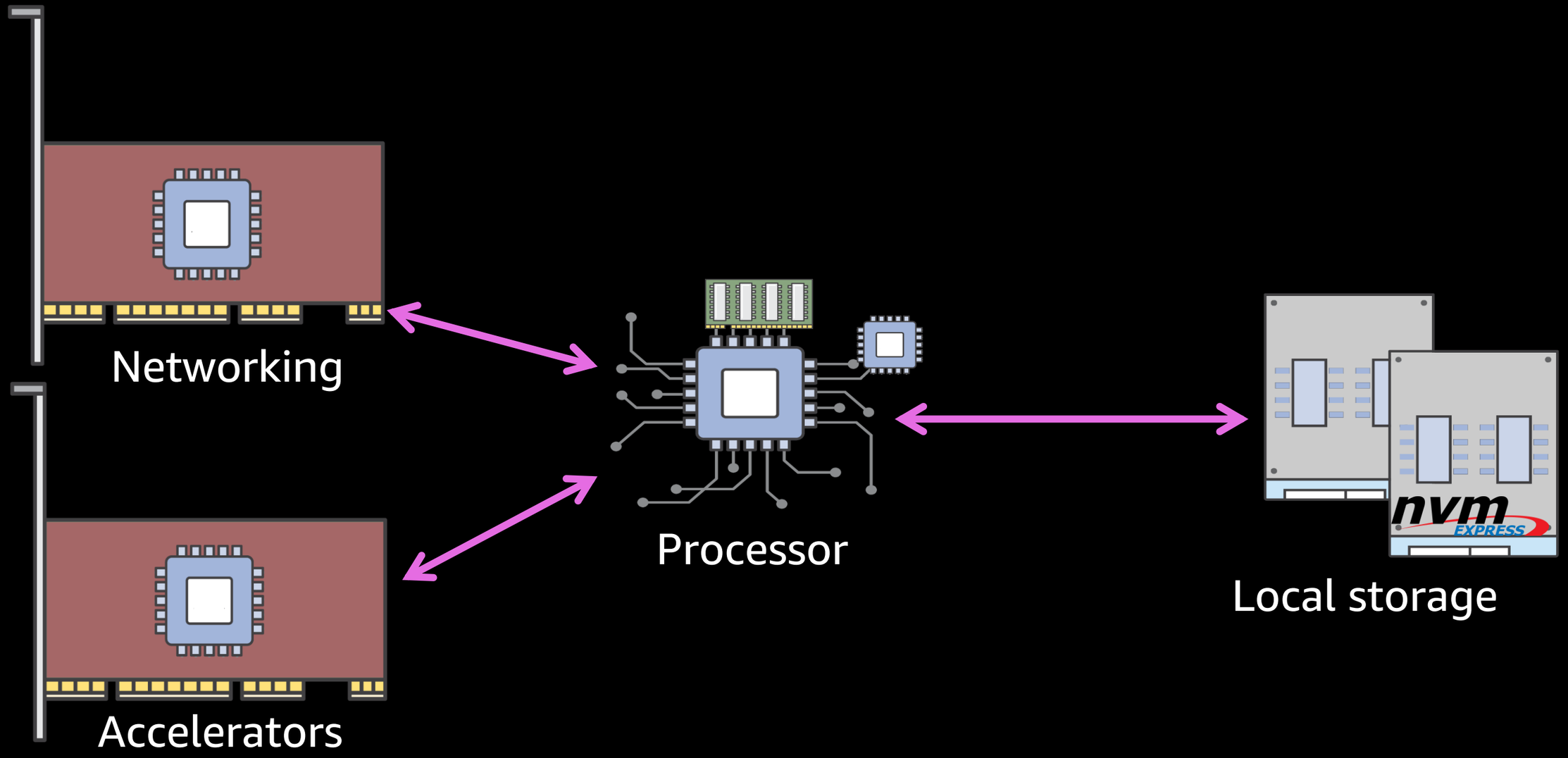


Speed

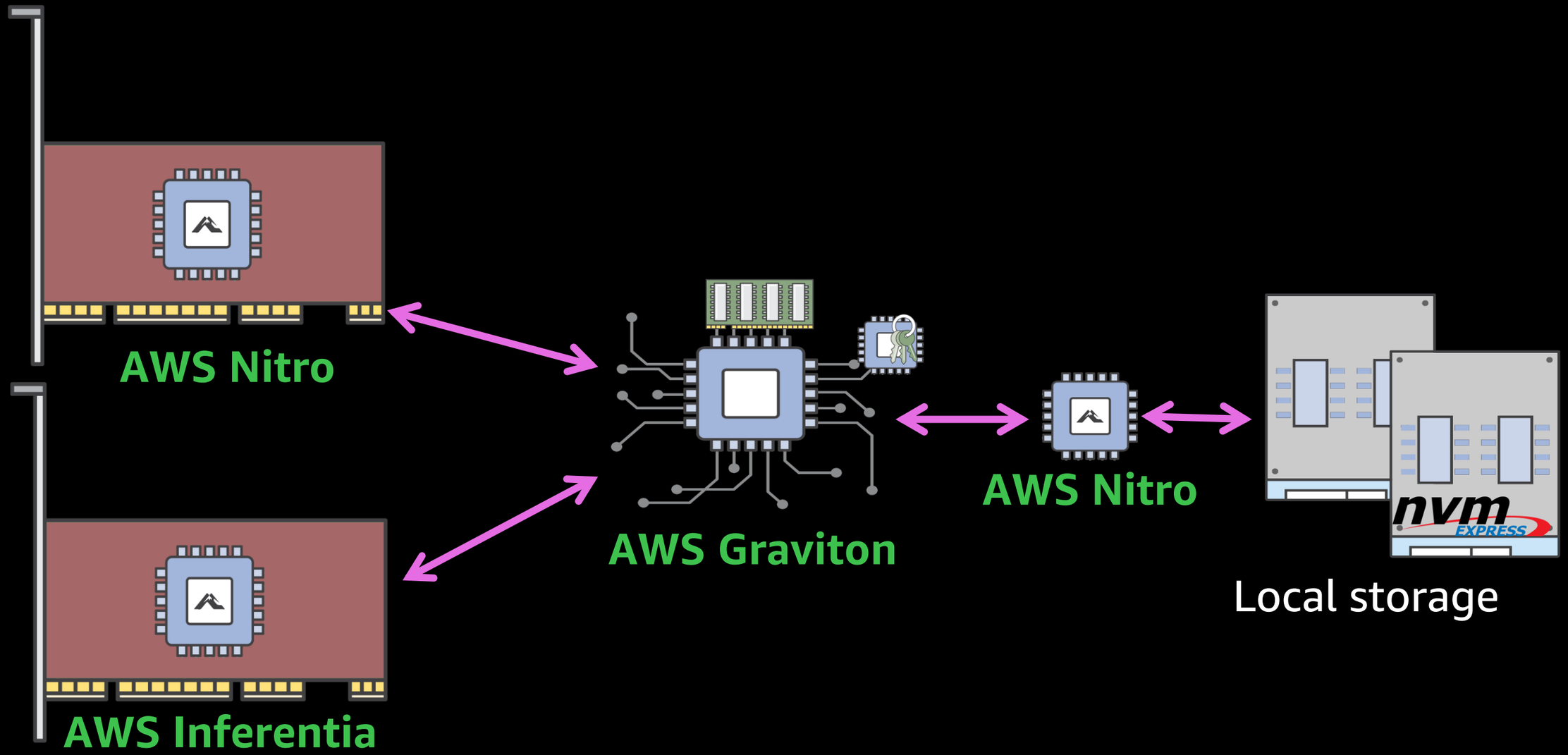


Operations

Typical server



Expanding opportunities for custom chips



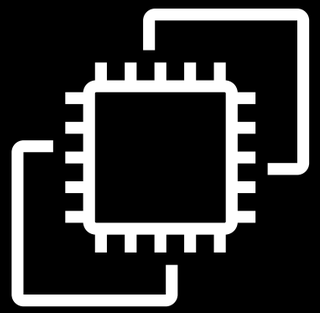
The Nitro System



AWS re:Invent

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

AWS Nitro System



AWS Nitro

Launched in November 2017

In development since 2013

Purpose-built hardware / software

Hypervisor built for AWS

All new instance launches use the Nitro system

Nitro in three parts

Nitro Hypervisor



Lightweight hypervisor
Memory and CPU allocation
Bare metal-like performance

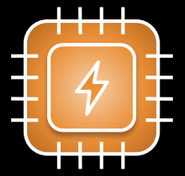
Nitro in three parts

Nitro Hypervisor



Lightweight hypervisor
Memory and CPU allocation
Bare metal-like performance

Nitro Security Chip



Integrated into motherboard
Protects hardware resources
Hardware Root of Trust

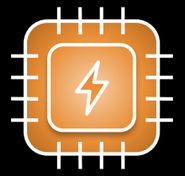
Nitro in three parts

Nitro Hypervisor



Lightweight hypervisor
Memory and CPU allocation
Bare metal-like performance

Nitro Security Chip



Integrated into motherboard
Protects hardware resources
Hardware Root of Trust

Nitro Cards



Amazon VPC networking
Amazon Elastic Block Store
Instance storage
System controller

Nitro Card for Amazon VPC



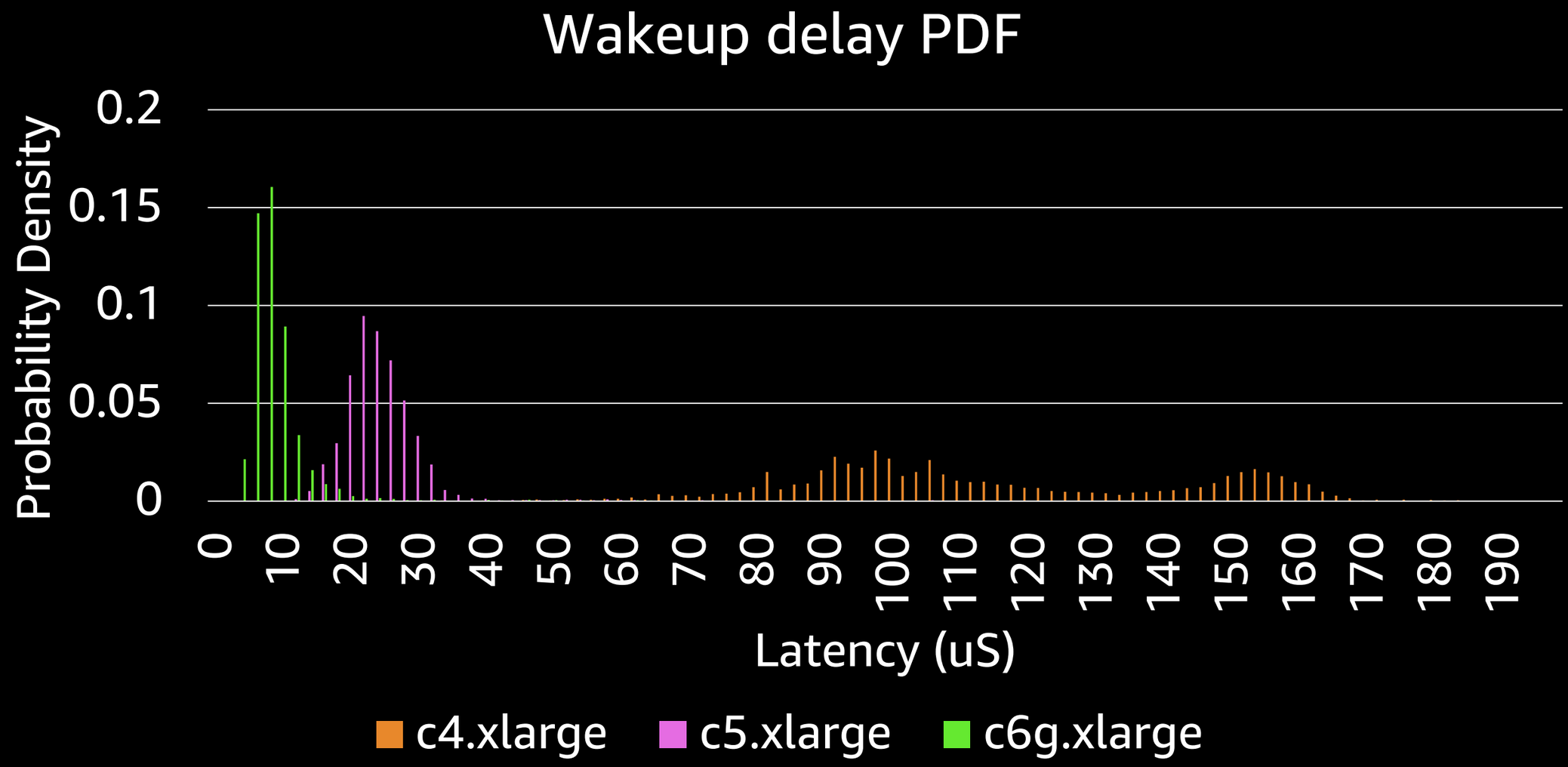
ENA Controller

- Drivers available for all major operating systems
- Independent of fabric

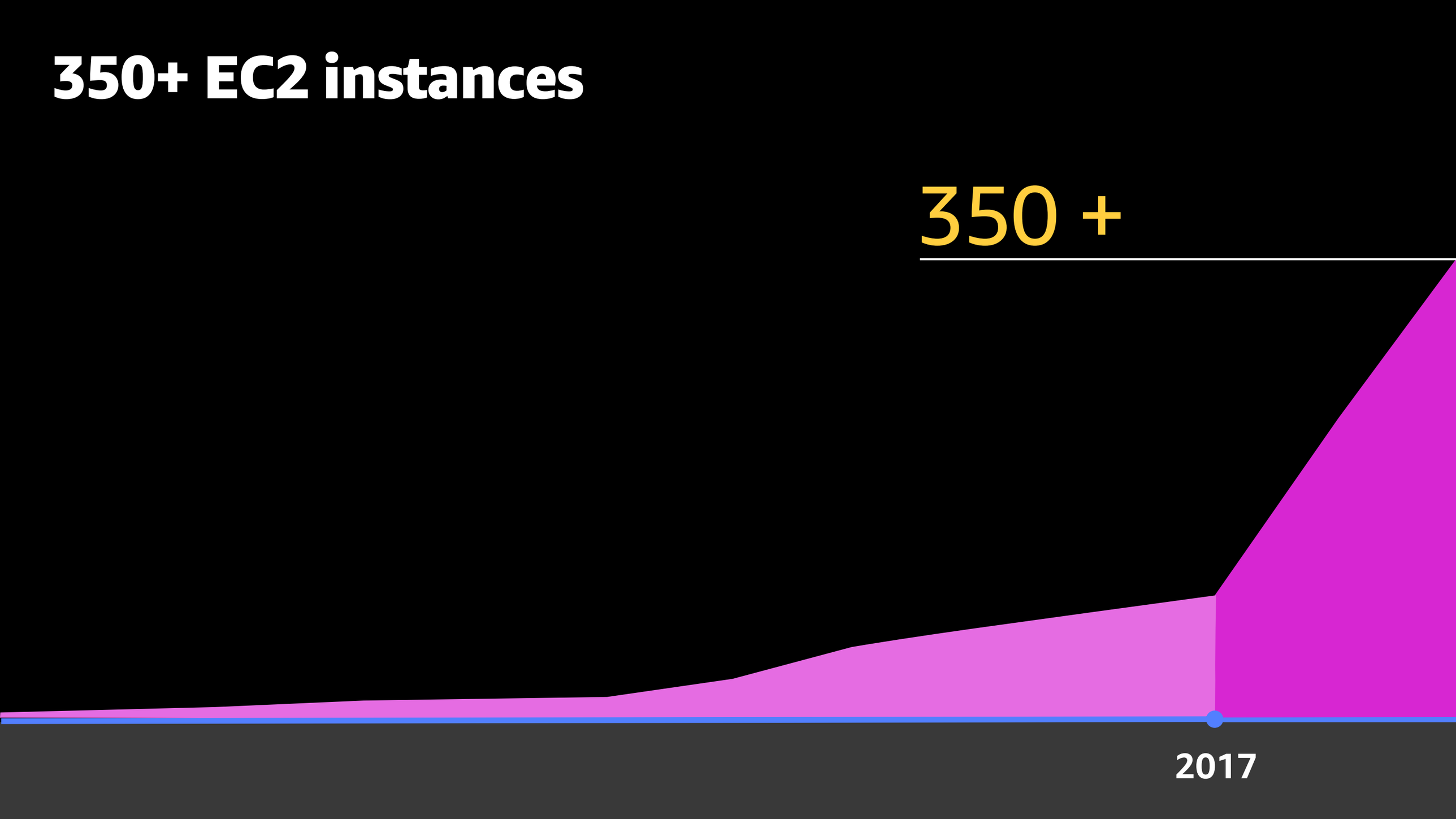
Amazon VPC Data Plane

- Encapsulation
- Security Groups
- Limiters
- Routing

Nitro Hypervisor – Reduced Jitter



350+ EC2 instances



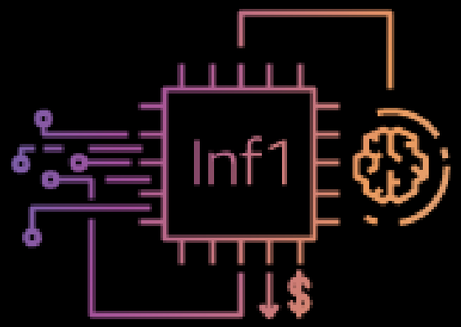
AWS Inferentia



AWS re:Invent

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

AWS Inferentia



AWS Inferentia

Purpose built chips for ML inference

Large on-chip caches

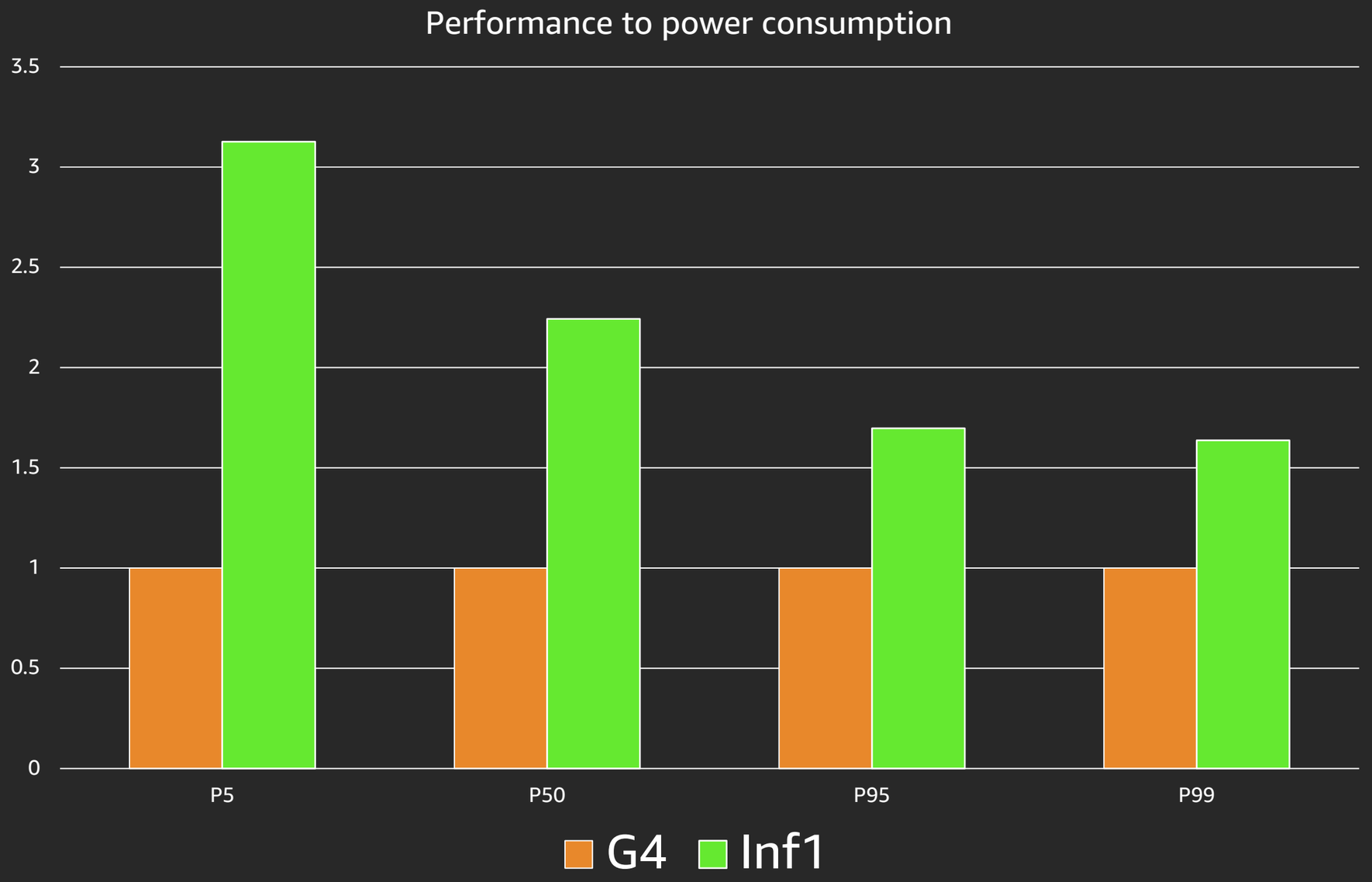
Optimized chip-to-chip interconnect

Support in popular ML frameworks

No quantization needed

Accessible via VMs, Containers, Amazon SageMaker

Sustainable high-performance



Measured 2hr run, TensorFlow YoloV4

Customers are using Inf1 instances in production

CONDÉ NAST

Anthem

SKYWATCH

Hotpot.ai

AUTODESK

INGA TECHNOLOGIES

DMP DIGITAL MEDIA PROFESSIONALS

The Asahi Shimbun

DISCO

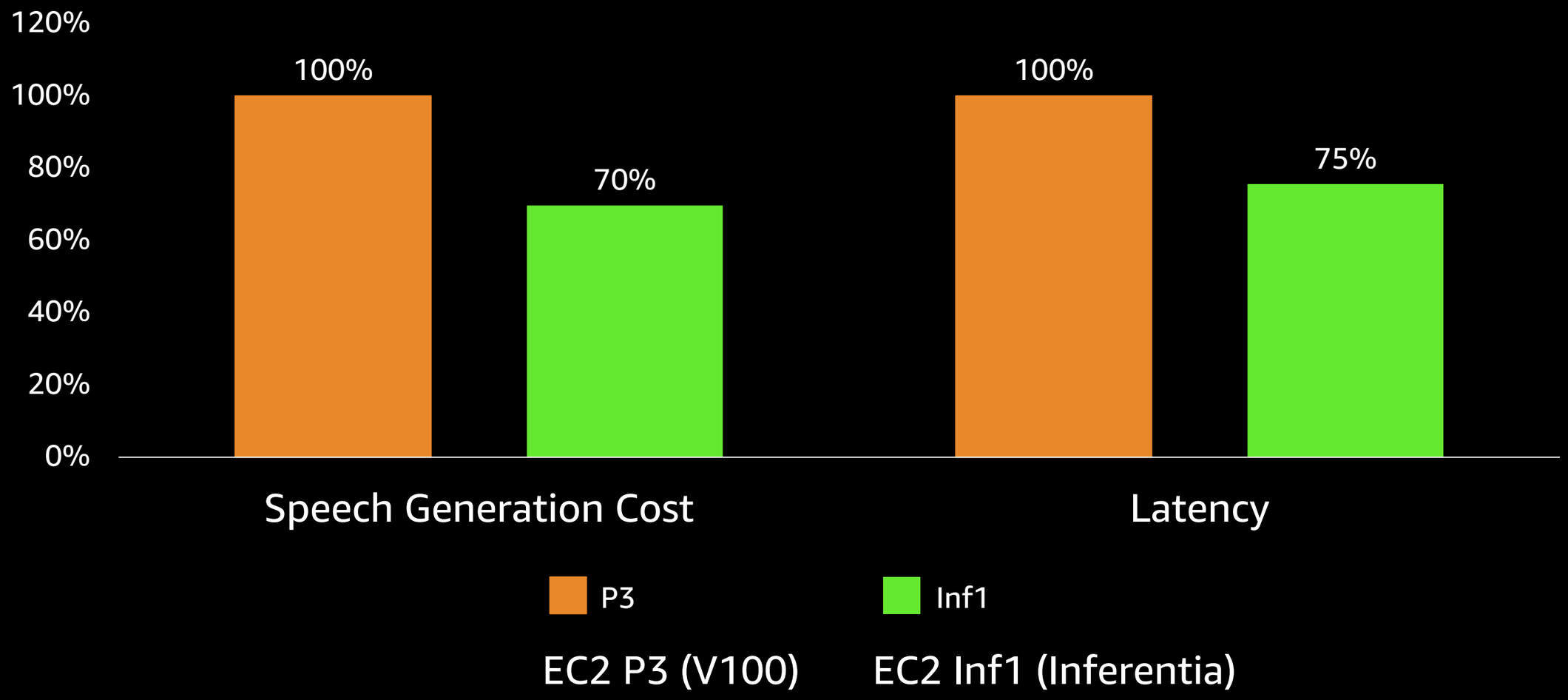
Talroo

Amazon Rekognition

Alexa TTS migration to EC2 Inf1



Alexa TTS migration to EC2 Inf1



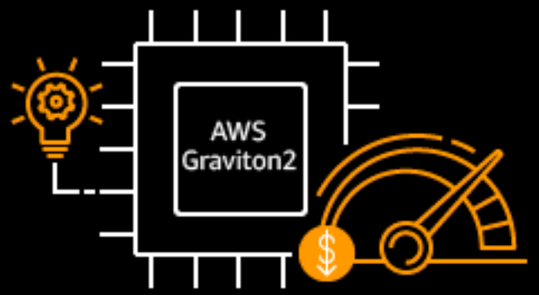
AWS Graviton



AWS re:Invent

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

AWS Graviton2



AWS Graviton

Custom AWS silicon with 64-bit Arm cores

Targeted optimizations to reduce costs

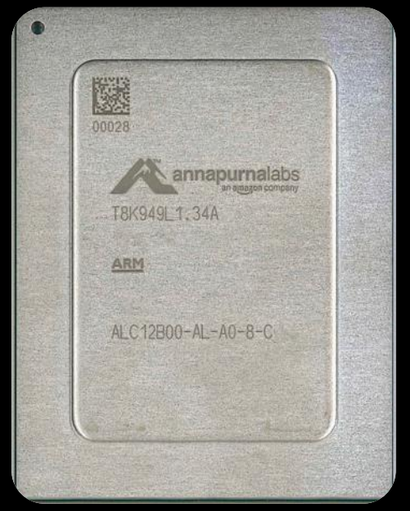
Rapidly innovate, build, and iterate

Range of instance types

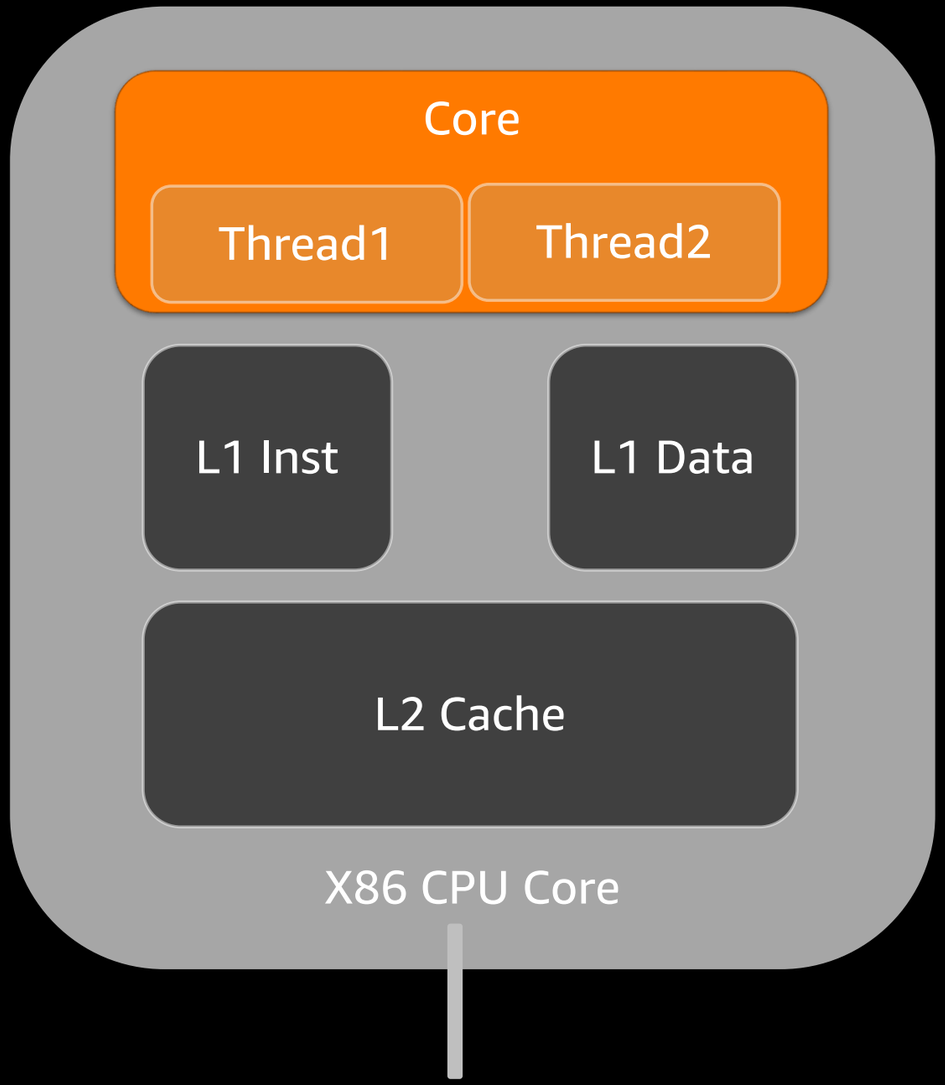
Same AWS Nitro building blocks

AWS Graviton2

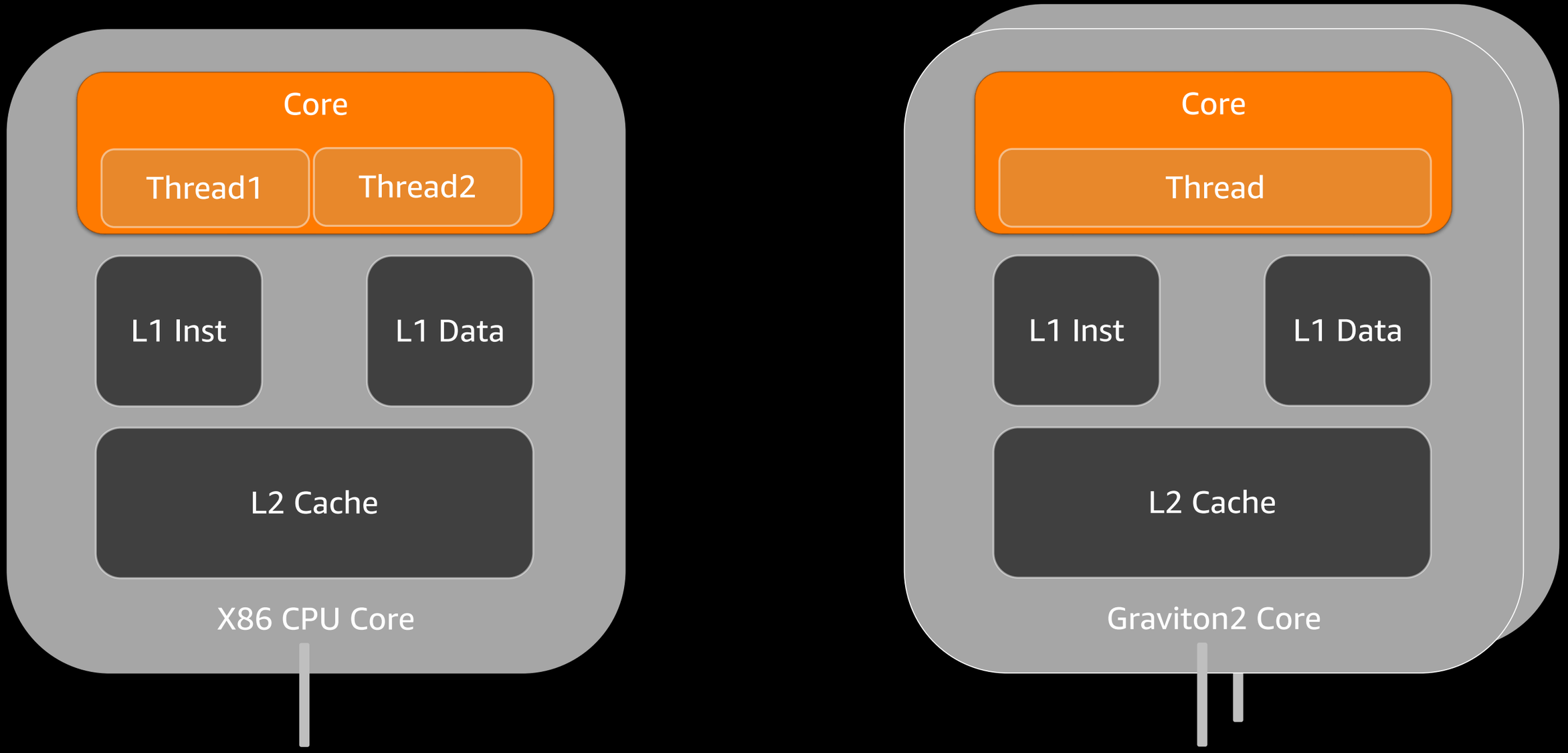
- Designed by AWS to provide the best value for our users
- First 7nm processor in AWS with ~30 billion transistors
- End-to-end ownership meant we could deliver this in a fraction of the time
- 20% less expensive than comparable equivalents
- >3x better performance/Watt



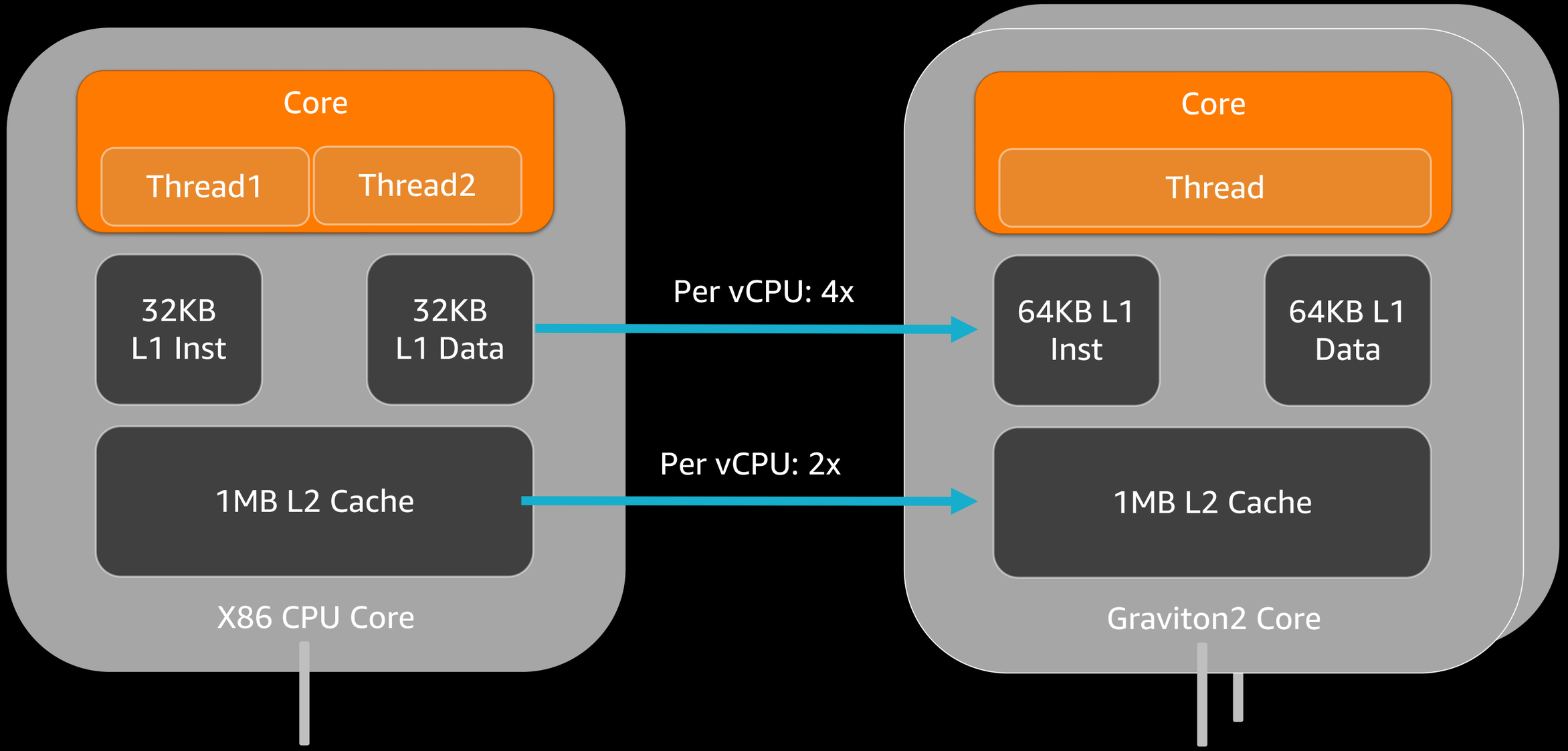
Graviton2 – Cores



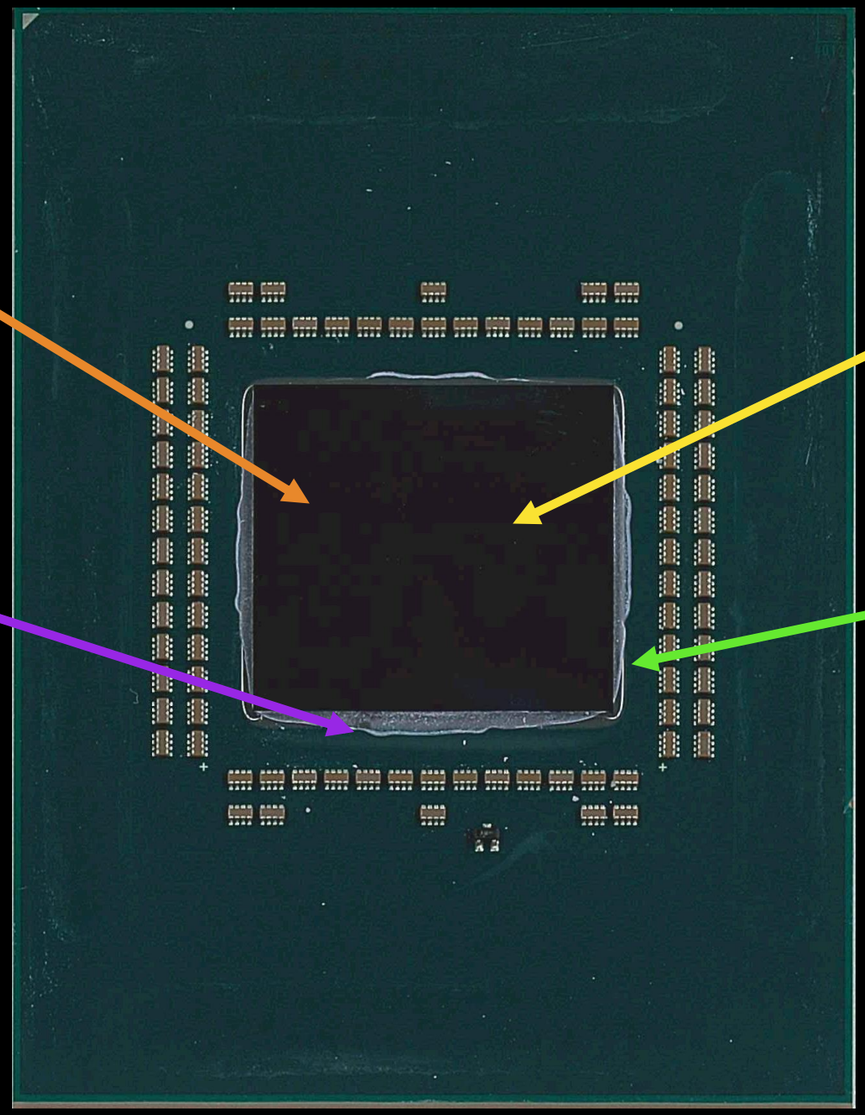
Graviton2 – Cores: No multi-threading!



Graviton2 – Cores: Large caches



Graviton2 - Interconnect



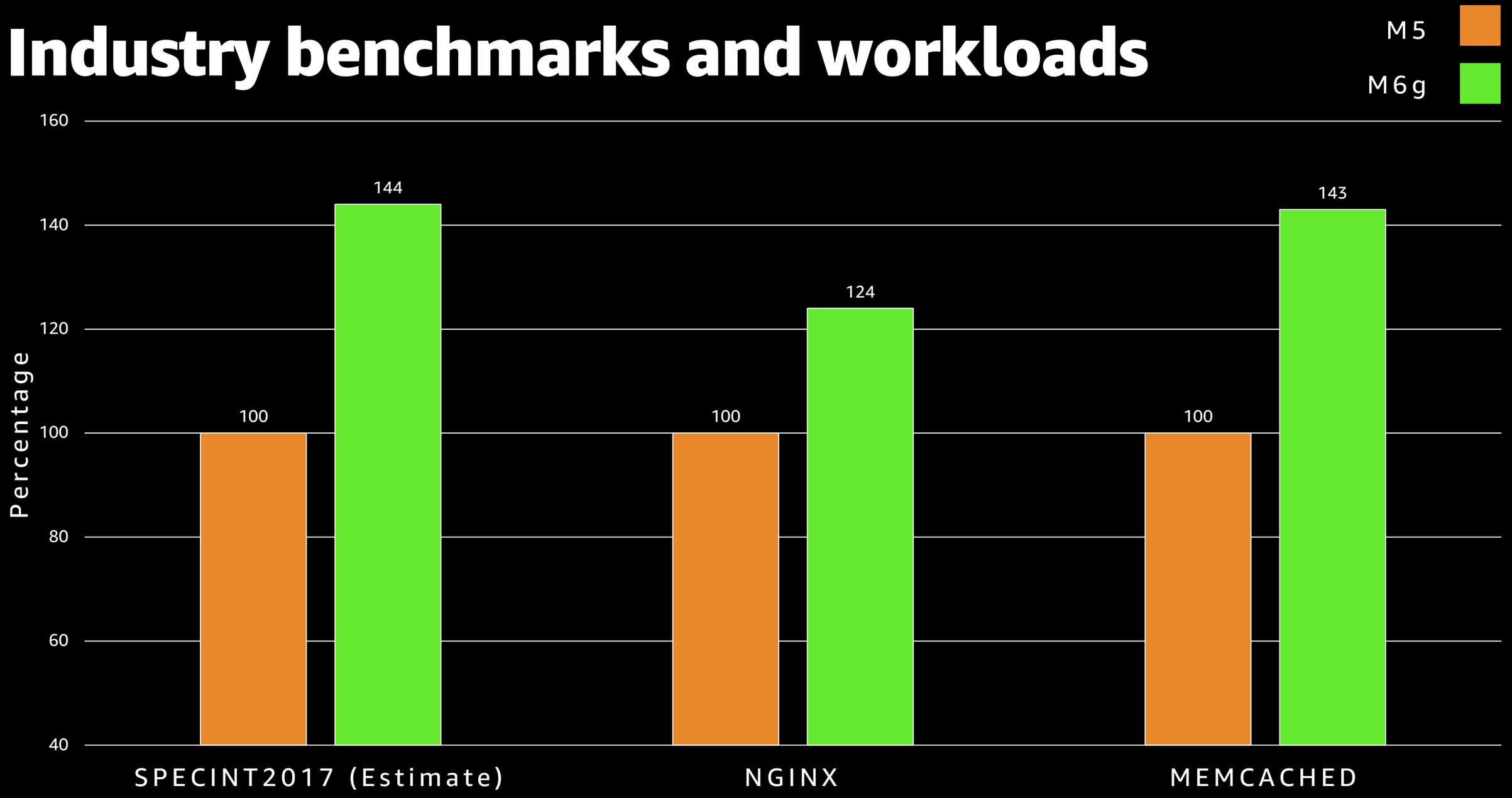
64 cores

64 PCIe

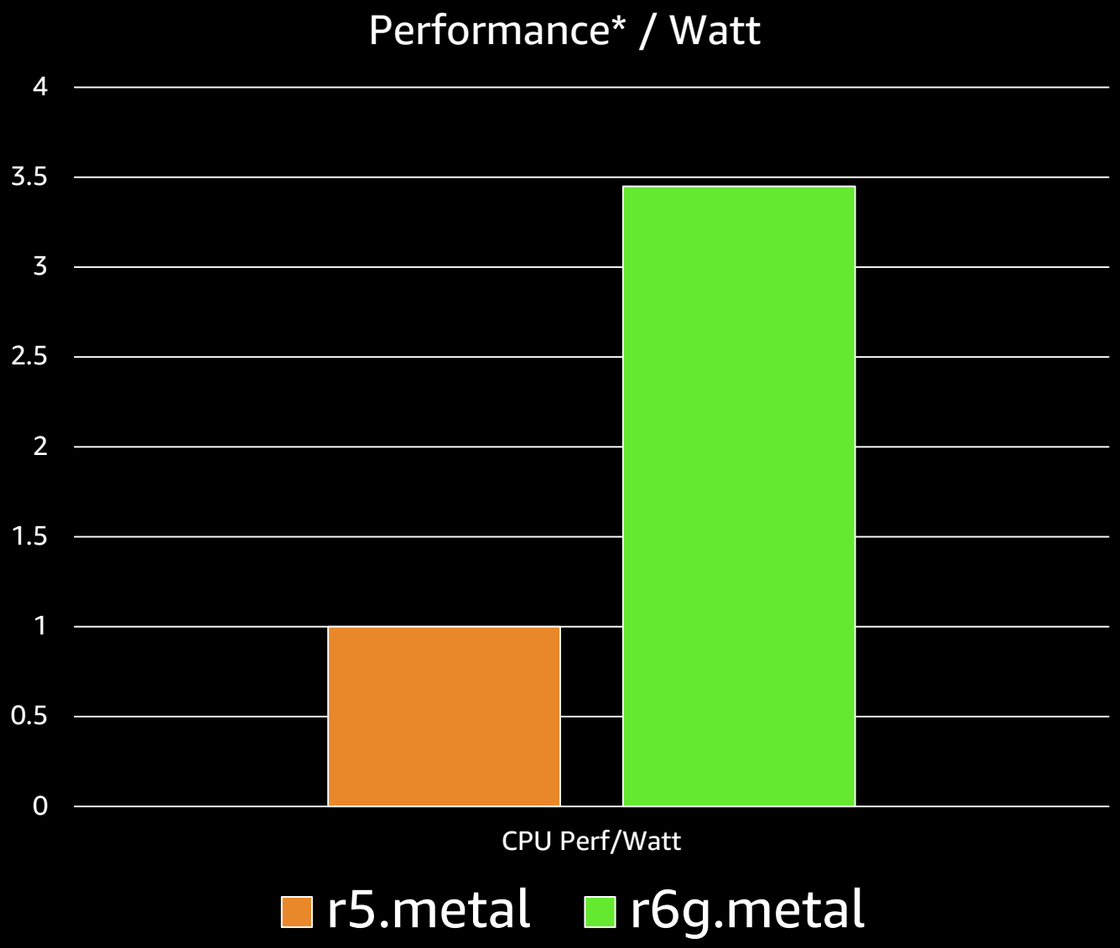
32 MB LLC

8 DDR channels

Industry benchmarks and workloads



CPU power efficiency



*Estimated SPECint2017 as show in previous slide

Most server power goes to CPUs

Lower power

→ Higher density

→ Lower costs

→ Lower carbon footprint

Next generation core EC2 instances

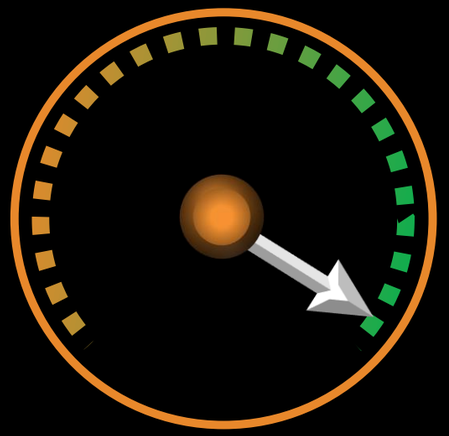
M6g	R6g	C6g	T4g	C6gn
General purpose 1:4 vCPU:memory	Memory optimized 1:8 vCPU:memory	Compute optimized 1:2 vCPU:memory	Burstable	Enhanced networking 100 Gbps, EFA

Also available – M6gd, R6gd, C6gd with NVMe local instance storage and metal instances for M, C, and R instance families

---

8 new EC2 instances powered by AWS Graviton2 processors

Lower TCO with Graviton2 powered instances



Highest performance
in their instance families



20% lower cost vs.
same-sized comparable instances



Up to **40%** better
price / performance vs
comparable instances

Best price performance within their instance families

Get started with the T4g free-trial today! *(until March 31st, 2021)*

https://aws.amazon.com/ec2/faqs/#t4g-instances

Graviton Customer feedback



"Our **Python and Go based services are compute intensive** and a major factor in the infrastructure cost of the platform. We benchmarked **more than 30% better price performance** for these services on AWS Graviton2 based Amazon EC2 C6g instances compared to the EC2 C5 instances we were previously using. **Moving to the C6g instances was reasonably straightforward**, and we are looking forward to running a majority of our service fleet on C6g instances for the equivalent performance and cost gains."



“We are excited to adopt AWS Graviton2 based M6g and C6g instances for our **web server** and **stream processing** workloads built on **Erlang, Go, and Java**. We have observed up to **50% total cost savings** over our previous generation EC2 instances for these workloads. All software packages that we needed to migrate were easily available and we plan to **switch everything we can to Graviton2** based instances.



“We have our **PHP-based photo-serving app** tier running on AWS Graviton2 based Amazon EC2 M6g and C6g instances. It's not very often that a piece of technology comes around that enables you to get **40% better price performance** just by recompiling. Moving to AWS Graviton2 based instances **required minimal code changes**; it just worked! ”



Snap uses Amazon EC2 C6g and M6g instances with **Amazon Elastic Kubernetes Service (EKS)** to run its Messaging Core Service. “Using Graviton2-based instances, we were able to **reduce the Snapchat messaging fleet size** and **significantly lower cost**. We are planning to move additional workloads to Graviton2 in the future.”

For more visit: https://aws.amazon.com/ec2/graviton/

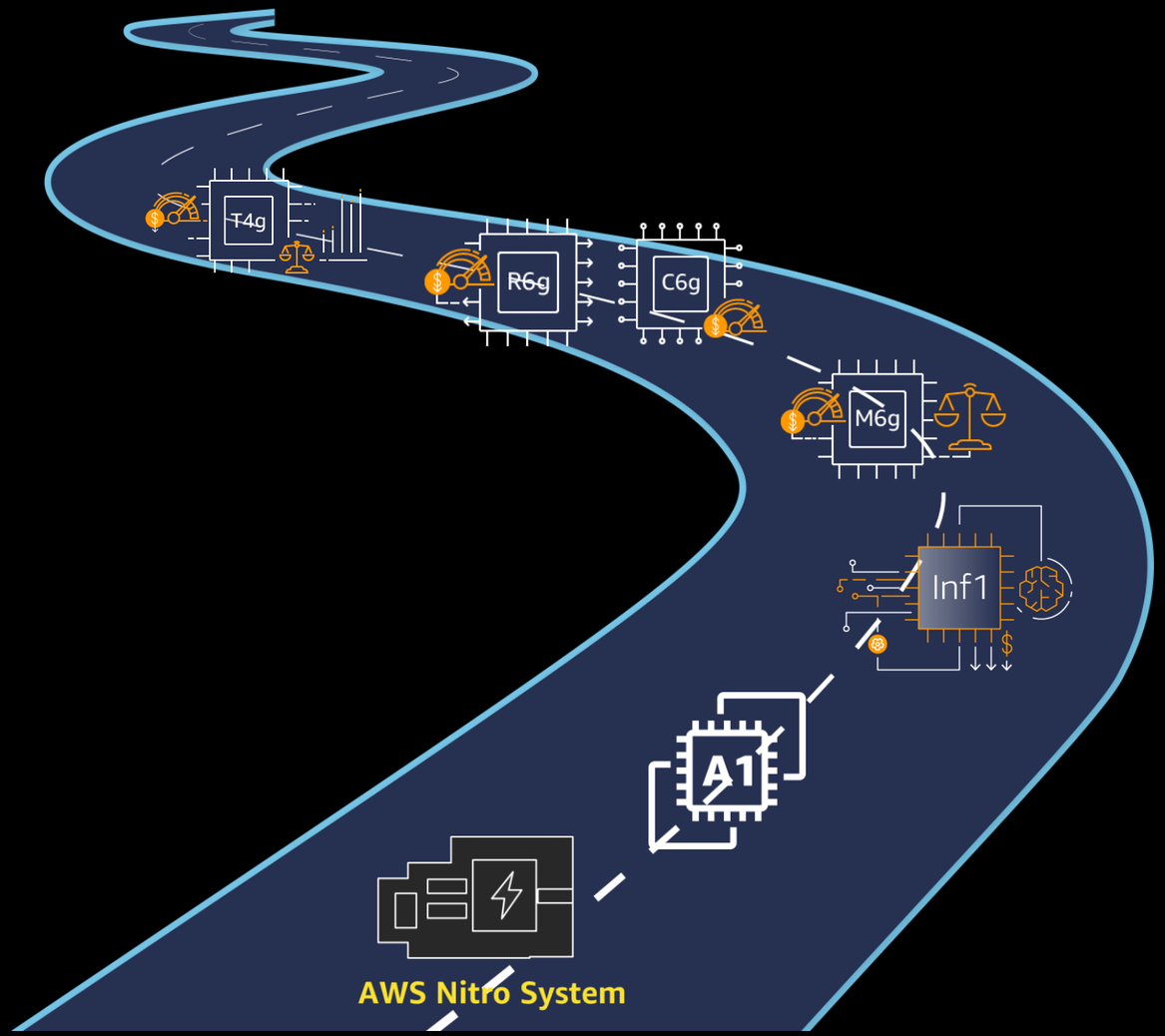
Wrapping Up



AWS re:Invent

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

More instances and more innovation



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.