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

CMP225

What’s new in Amazon EC2

Art Baudo (he/him)

Principal Product Marketing Manager, EC2
Amazon Web Services

Martin Yip (he/him)

Senior Manager, Product Marketing, EC2
Amazon Web Services

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



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

Over

30 billion

Instances launched since 2006

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

AWS Cloud Pillars

Providing customers with tools and services to securely and reliably run virtually any workload

Delivering more performance at lower cost to customers



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

“Sometimes small things make a big difference”

Dave Brown

Vice President, EC2

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

100 million

Instances launched every day

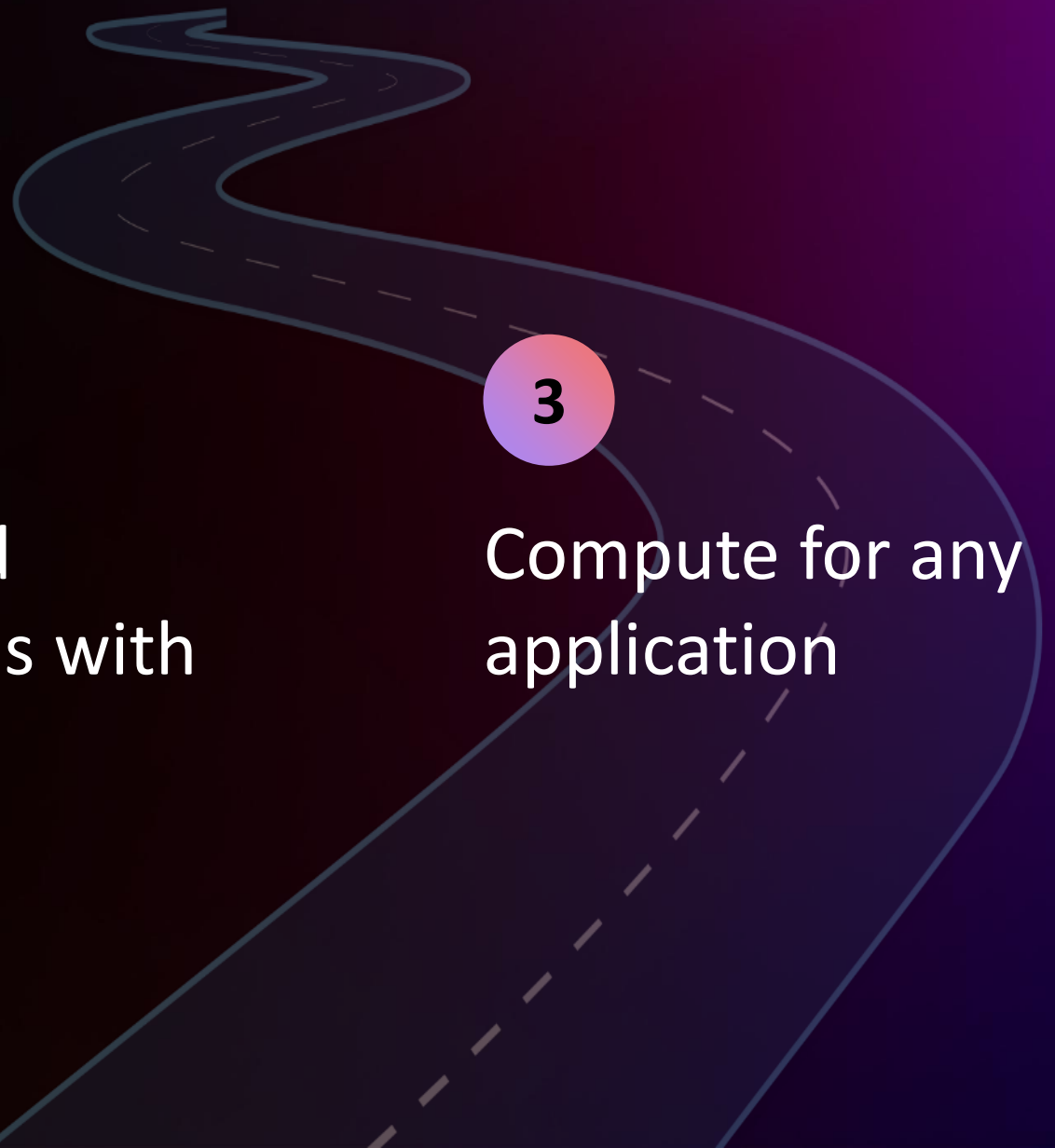
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

700 instances

Launched per second

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

Building on 16 years of innovation



1

Compute on a global scale

2

AWS Cloud Innovations with Nitro

3

Compute for any application

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

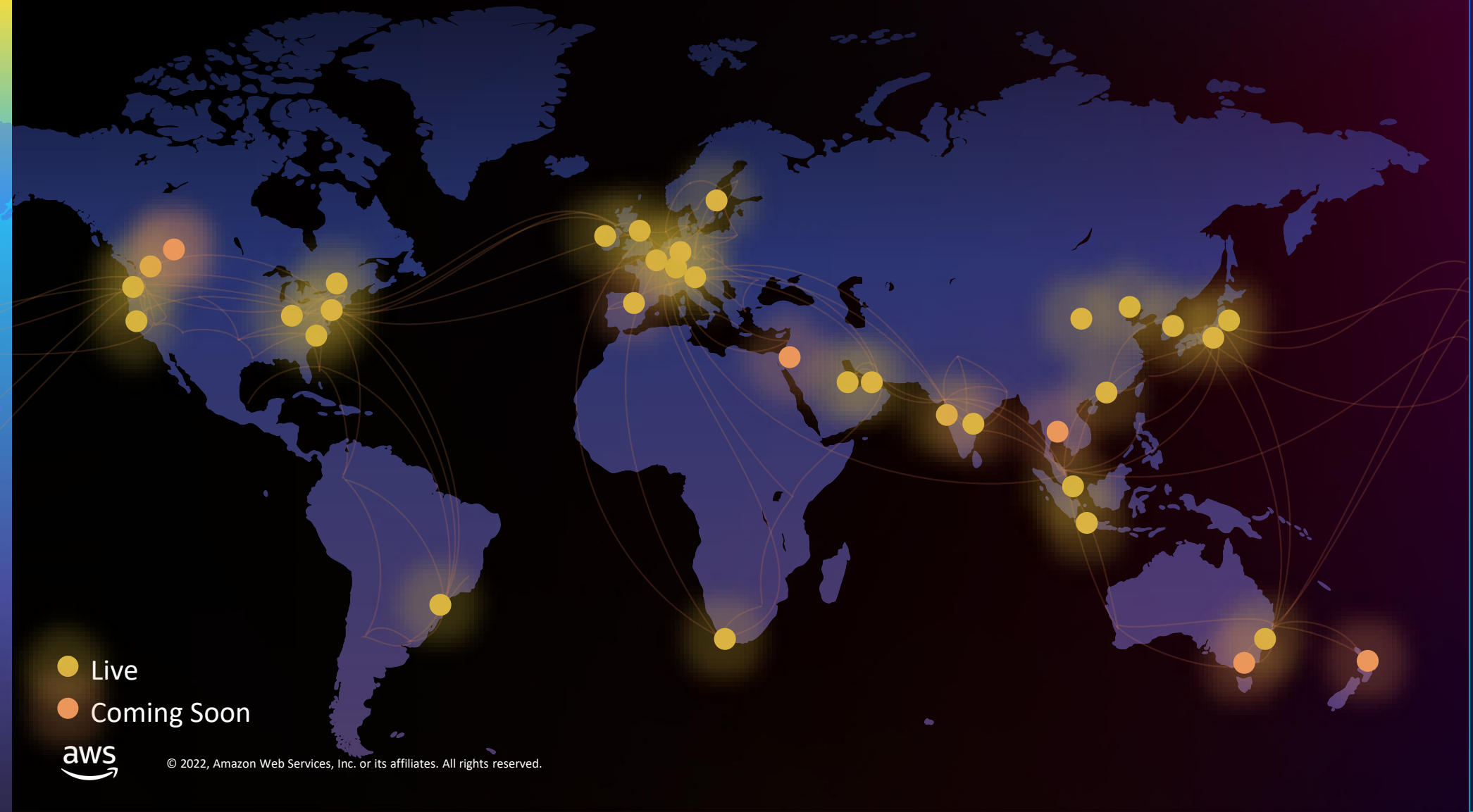
AWS Global Infrastructure



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

AWS Regions

The most secure, extensive, and reliable Global Cloud Infrastructure



30

Live Regions

5

Coming Soon

96

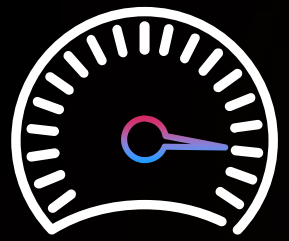
Availability Zones

aws

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

AWS Local Zones

RUN LATENCY-SENSITIVE APPLICATIONS AT THE EDGE USING AWS INFRASTRUCTURE AND SERVICES



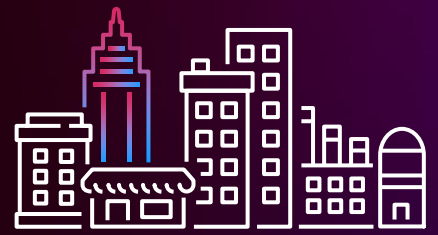
LOW LATENCY

Extends AWS infrastructure services, APIs, and tools to where customers need it to support low-latency applications



FULLY MANAGED

Fully owned, managed, and supported by AWS



CITIES

New type of AWS infrastructure that places AWS compute, storage, networking, and select AWS services closer to where your end users are located

aws

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

AWS Outposts

AWS INFRASTRUCTURE AND SERVICES IN YOUR ON-PREMISES LOCATION



AWS DESIGNED

Same AWS-designed infrastructure as in AWS data centers (built on AWS Nitro System)



FULLY MANAGED

Fully managed, monitored, and operated by AWS as if in AWS Regions



AWS API

Single pane of management in the cloud providing the same APIs and tools as in AWS Regions

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

AWS Outposts

AWS INFRASTRUCTURE AND SERVICES IN YOUR ON-PREMISES LOCATION



















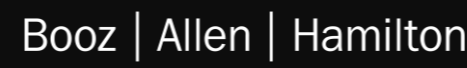












© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Why AWS infrastructure?



Routers

Load balancers

Custom compute servers and semiconductors

Custom software

Silicon

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

Broadest and deepest platform choice

CATEGORIES

General purpose

Burstable

Compute intensive

Memory intensive

Storage (High I/O)

Dense storage

GPU compute

Graphics intensive

CAPABILITIES

Choice of processor
(AWS, Intel, AMD, Apple)

Fast processors
(up to 4.5 GHz)

High memory footprint
(up to 24 TiB)

Instance storage
(HDD and NVMe)

Accelerated computing
(GPUs, ASICs, Video, FPGAs)

Networking
(up to 800 Gbps)

Bare metal

Size
(Nano to 112xlarge)

OPTIONS

Amazon EBS

Amazon Elastic Inference

MORE THAN

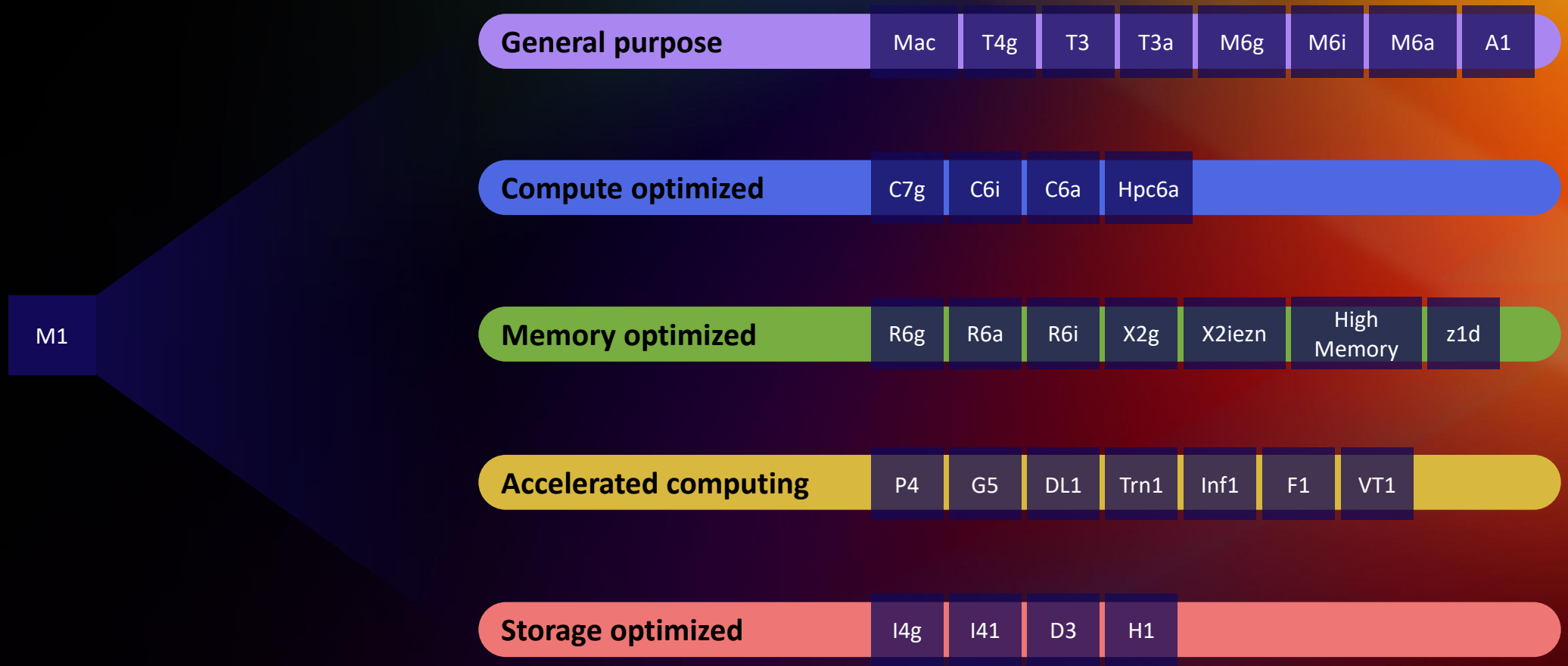
600

INSTANCE TYPES

for virtually every workload and business need

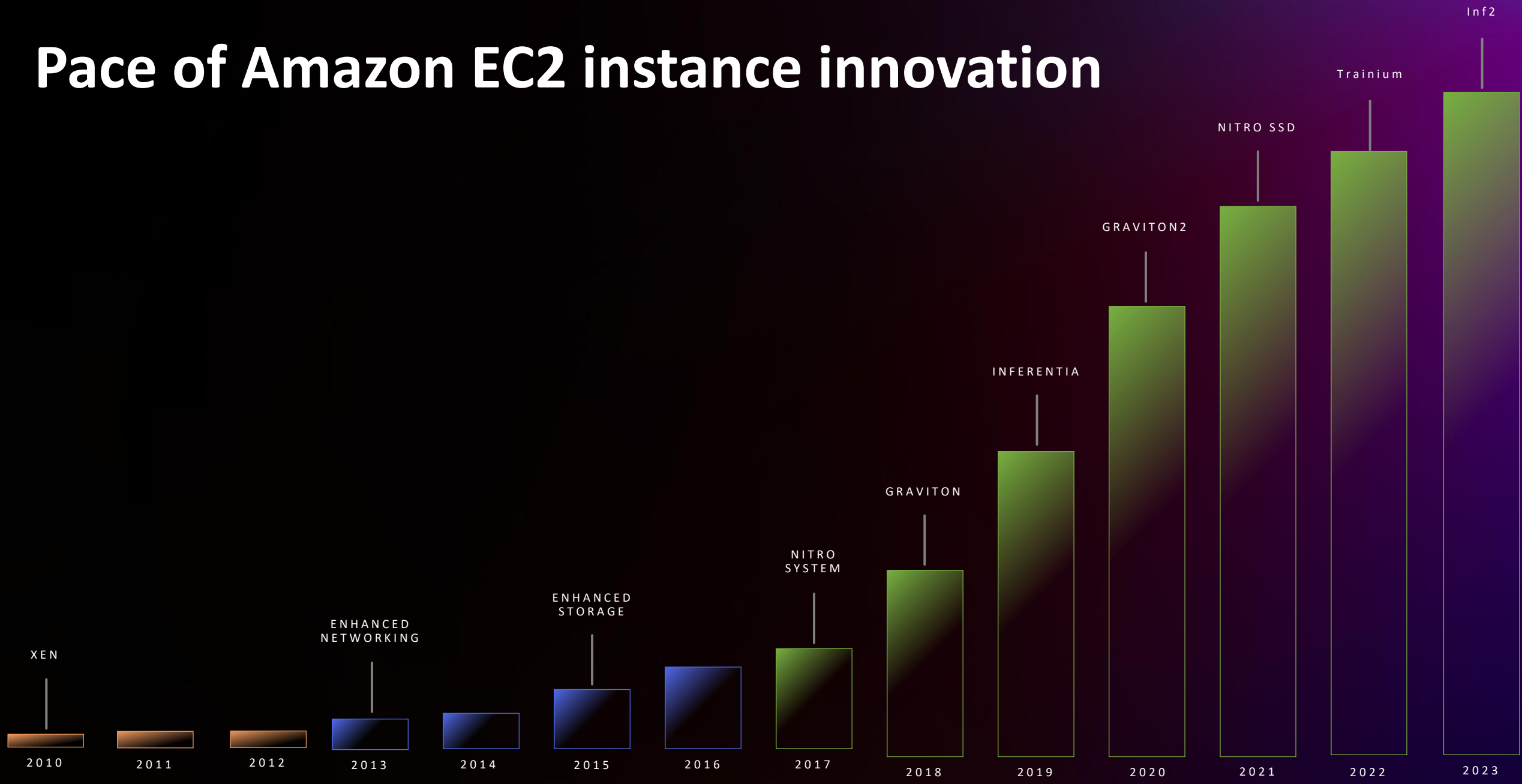
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

EC2 instances



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

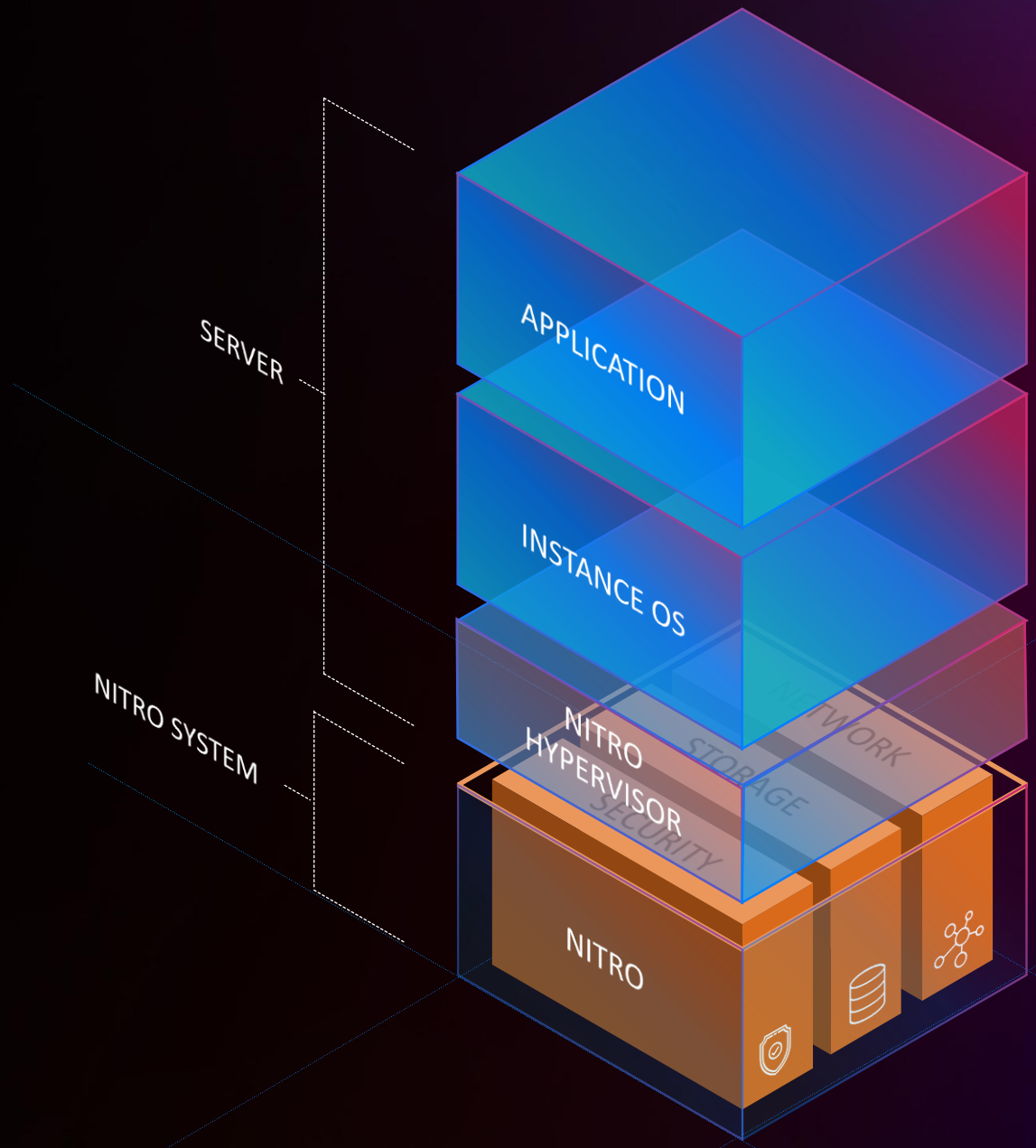
Pace of Amazon EC2 instance innovation



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

The AWS Nitro System architecture

Offering strong security, performance, and innovation in the cloud



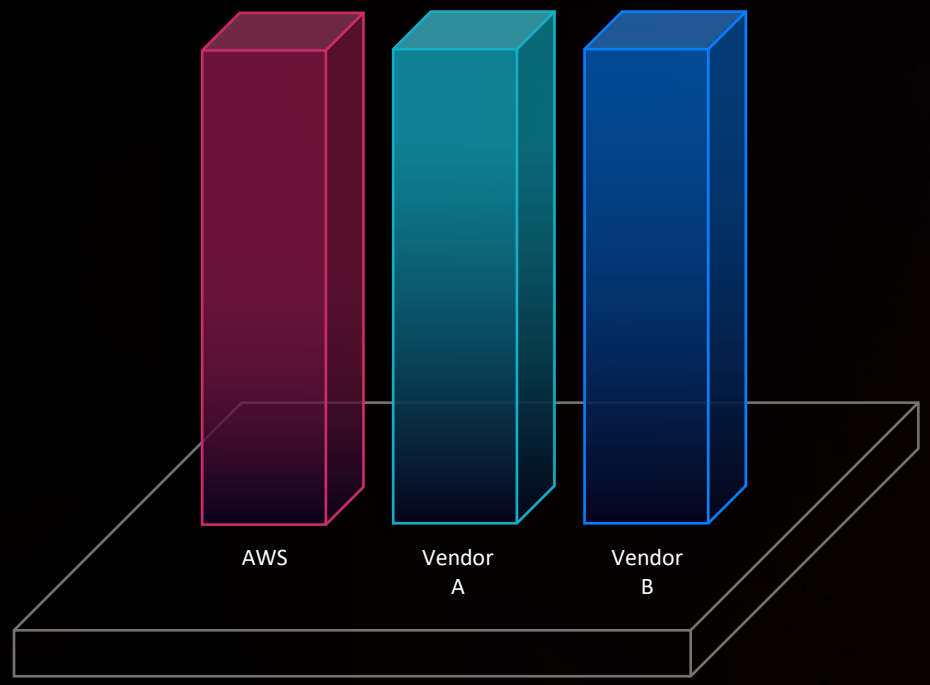
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

While the Nitro System provides an overall performance benefit, we also focus on component-level performance

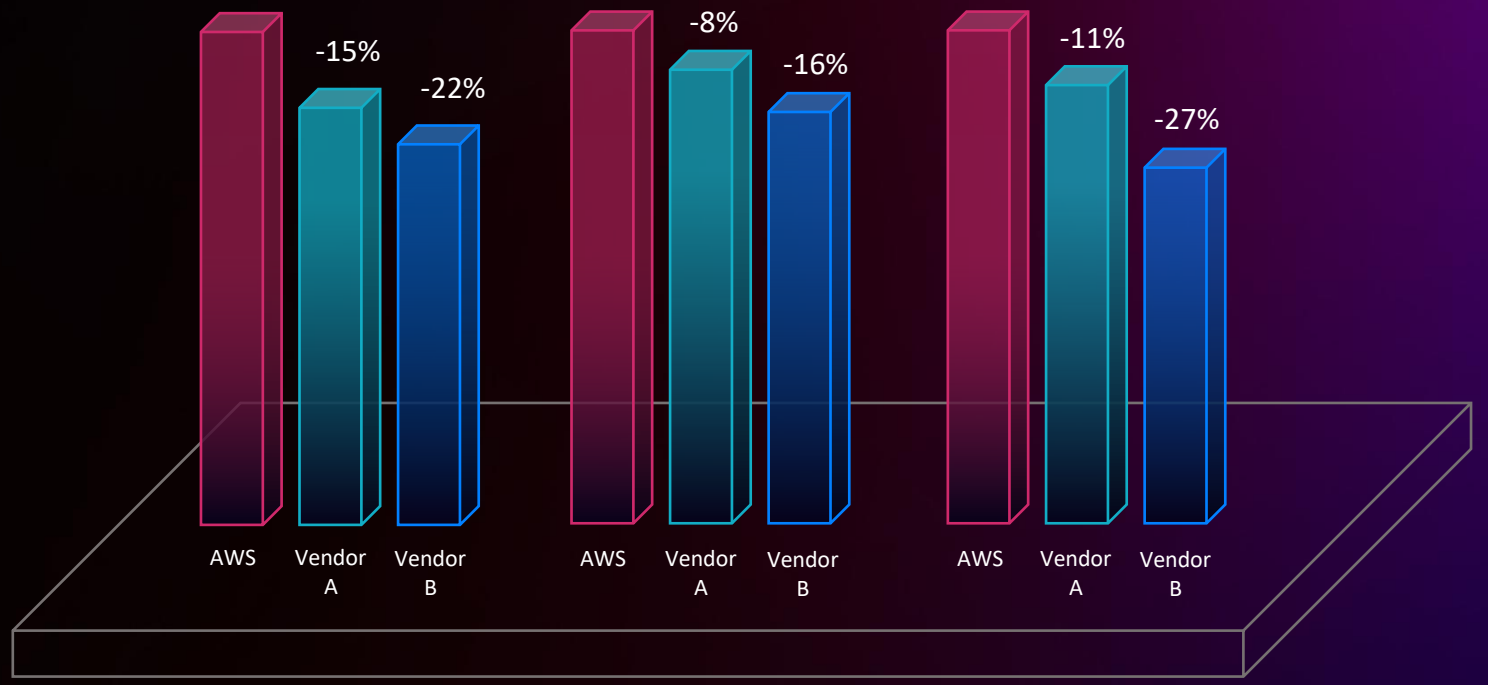
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Nitro performance for real-world workloads

Amazon EC2 instances can deliver over 15% higher throughput performance



SPECINT2017



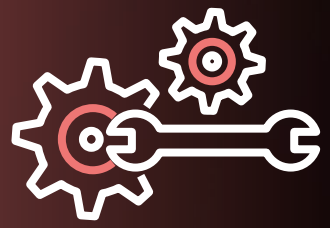
MEMCACHED

NGINX WEBSERVING

REDIS

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

Why build our own chips?



Specialization



Speed



Innovation



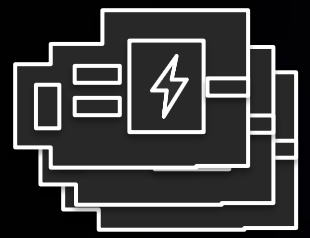
Security

aws

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

The Nitro System

Nitro Cards



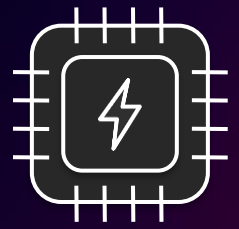
VPC Networking

Amazon Elastic Block Store (Amazon EBS)

Instance Storage

System Controller

Nitro Security Chip



Integrated into motherboard

Protects hardware resources

Hardware Root of Trust

Nitro Hypervisor



Lightweight hypervisor

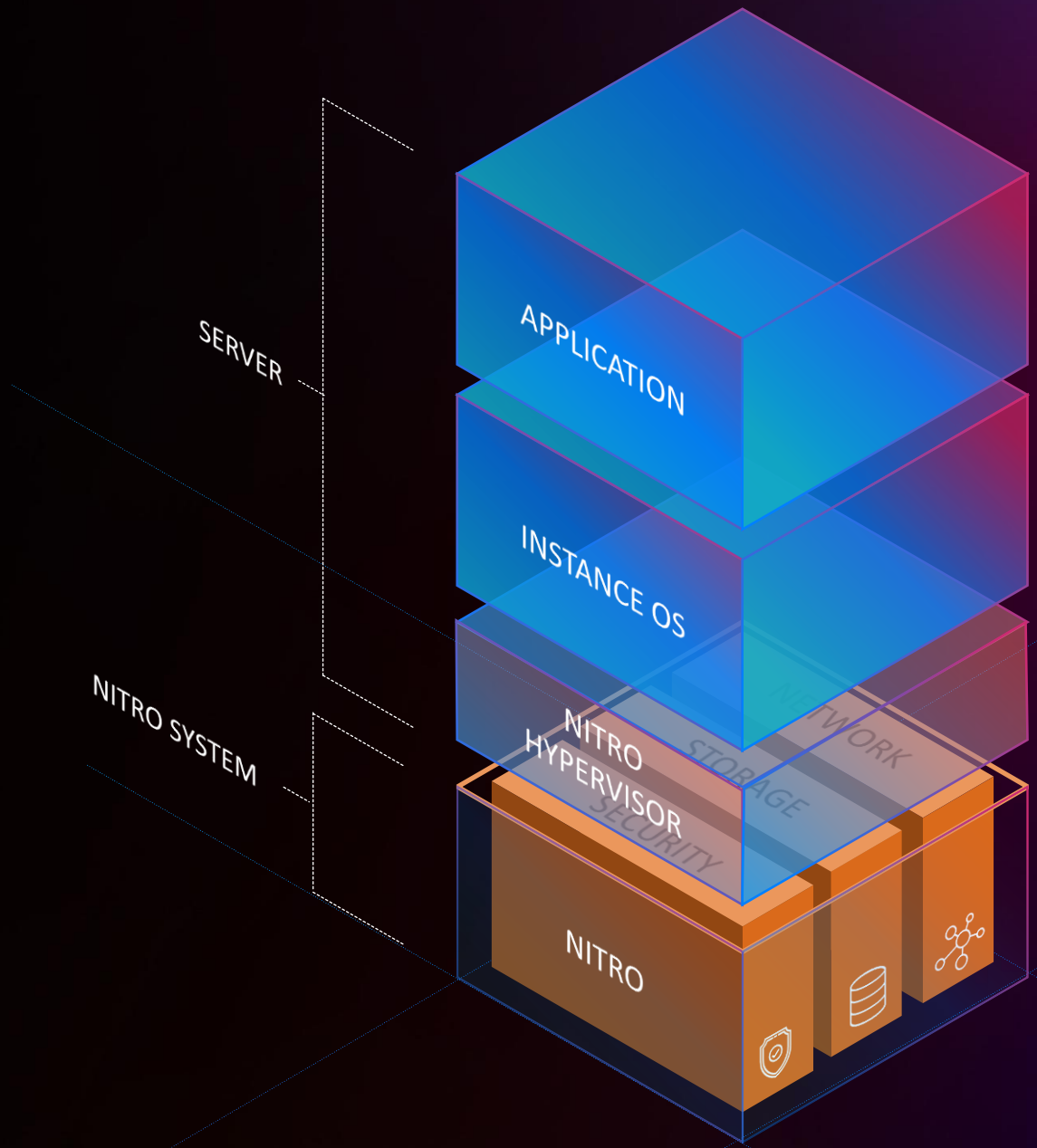
Memory and CPU allocation

Bare Metal-like performance

aws

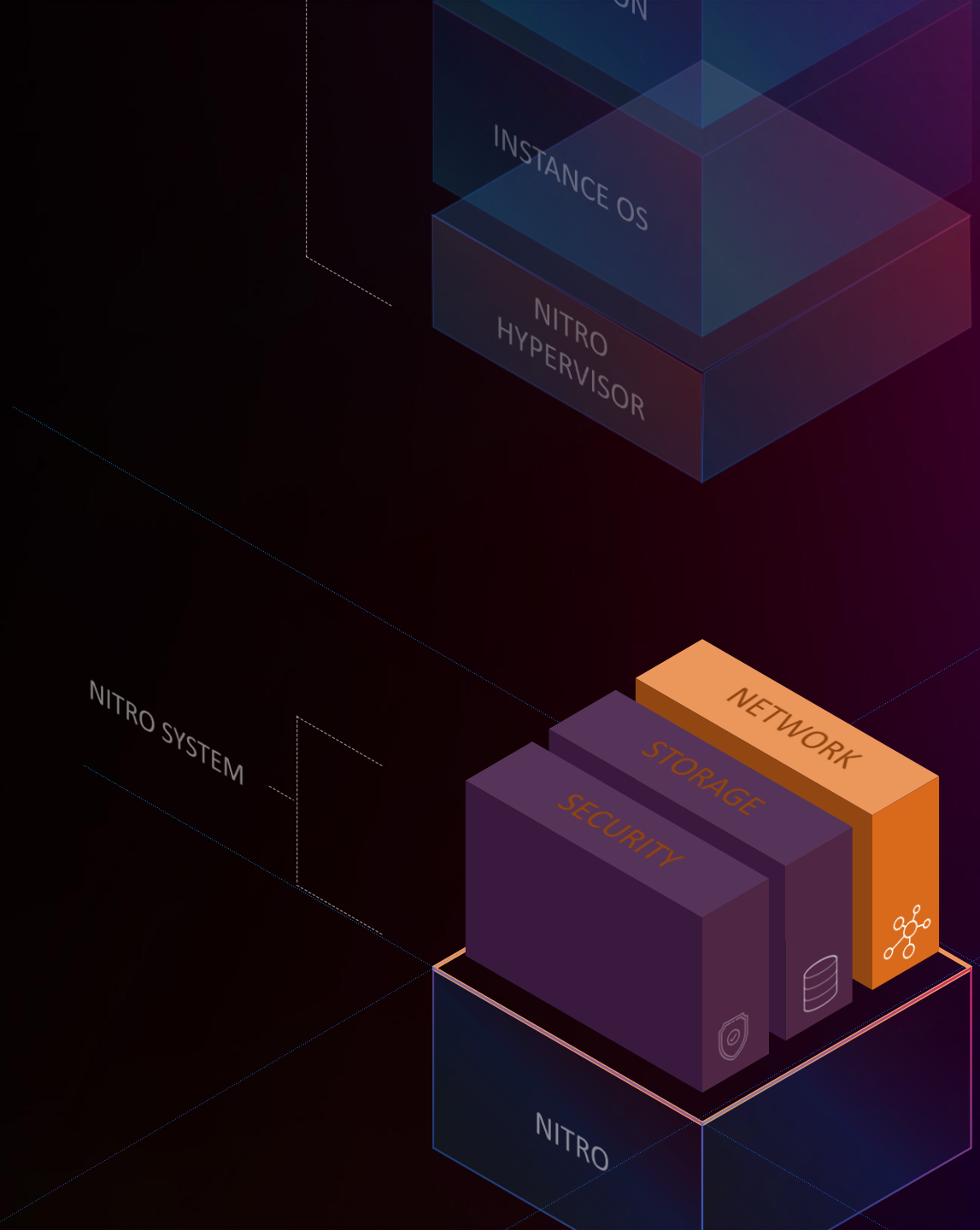
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Networking



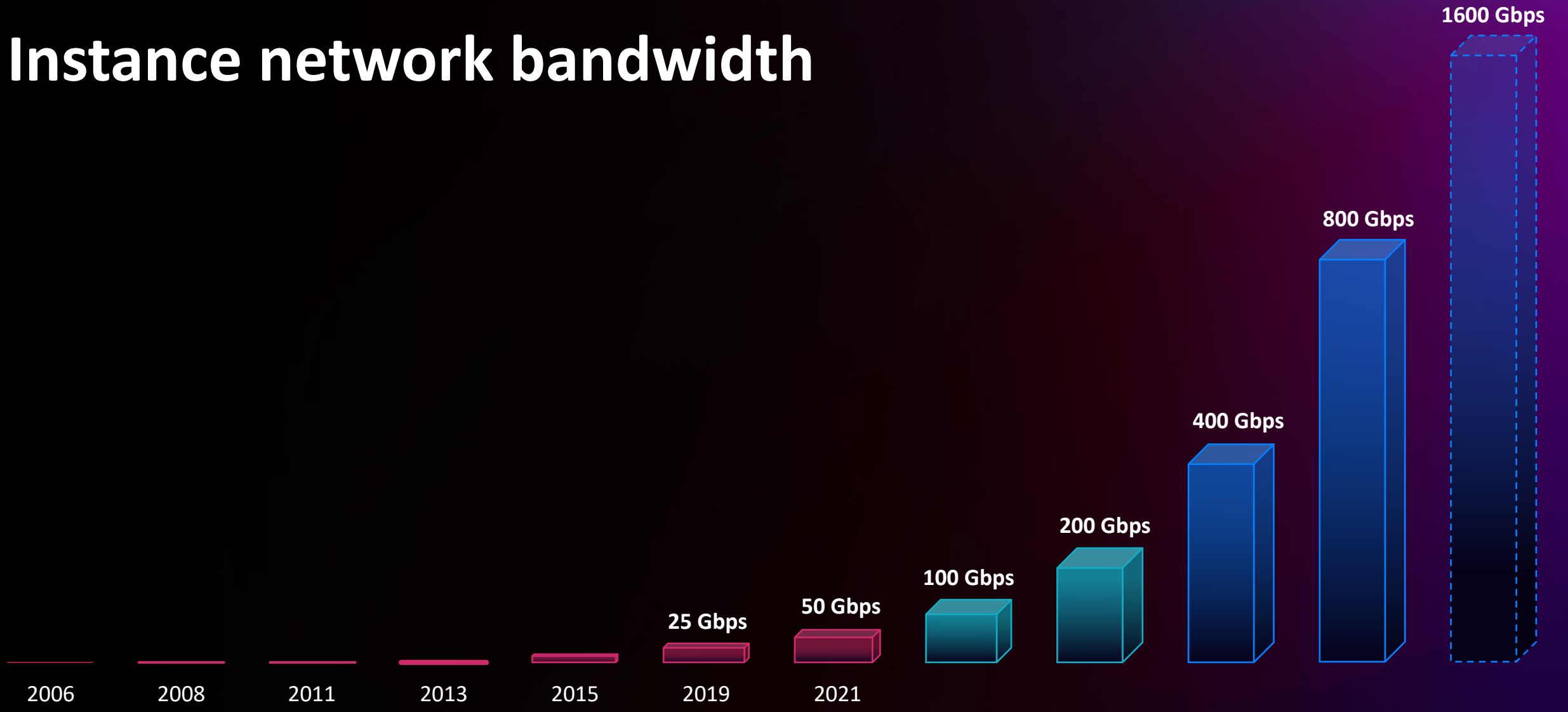
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Networking



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

Instance network bandwidth



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

NEW

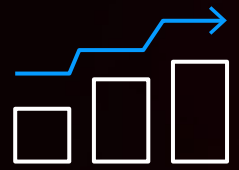
6th Gen EC2 network-optimized instances

Best performance in the cloud for high-networking and EBS workloads



200 Gbps network bandwidth

First network-optimized instance in the cloud to support 200 Gbps to accelerate data intensive analytics workloads



Up to 2x Packets-Per-Second vs previous generation

Scale performance of network virtual appliances, Telco, and in-memory database workloads



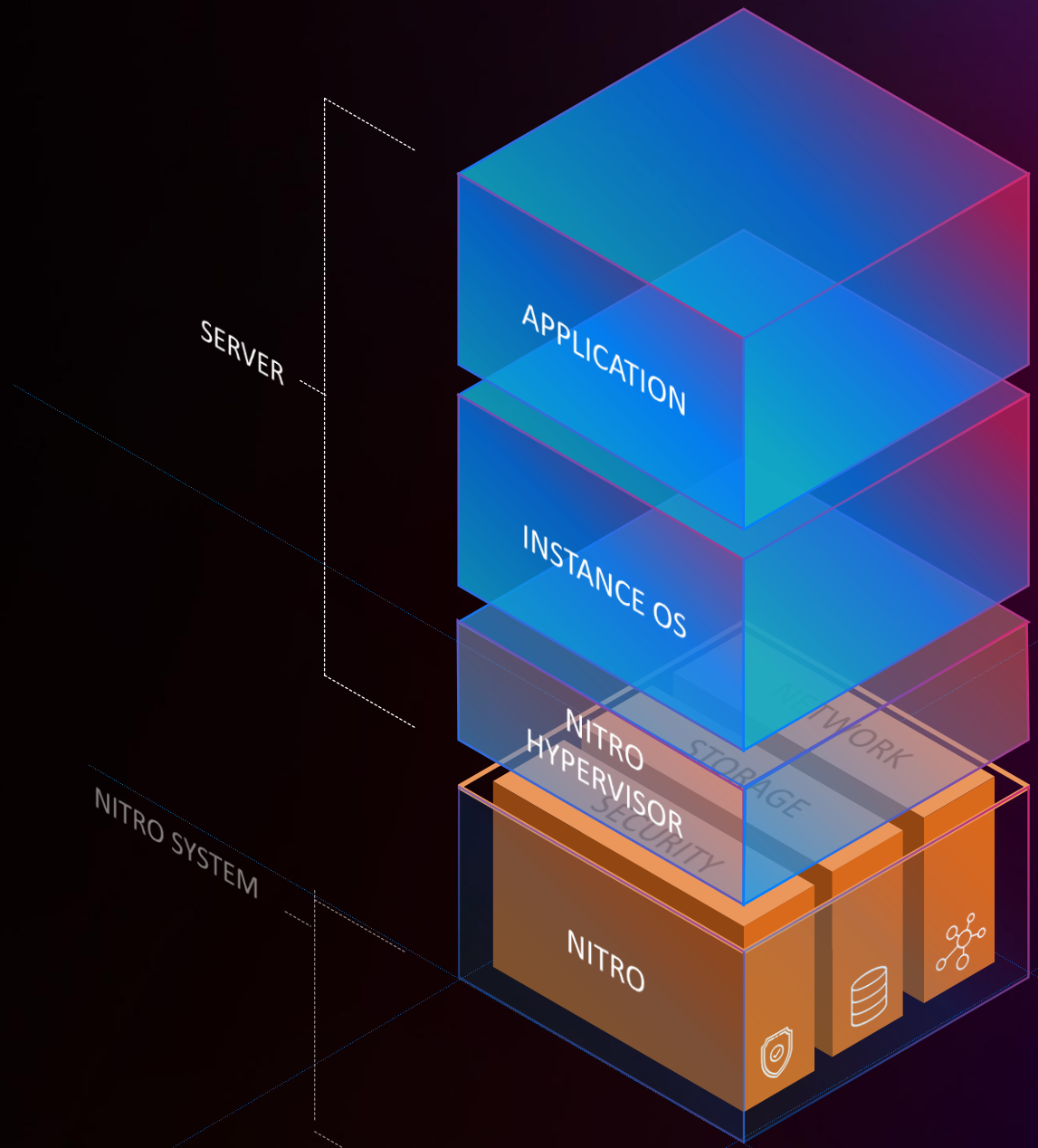
Fastest EBS performance in EC2

Supports up to 80 Gbps of EBS bandwidth to enable high-performance file systems, databases, and big data analytics

aws

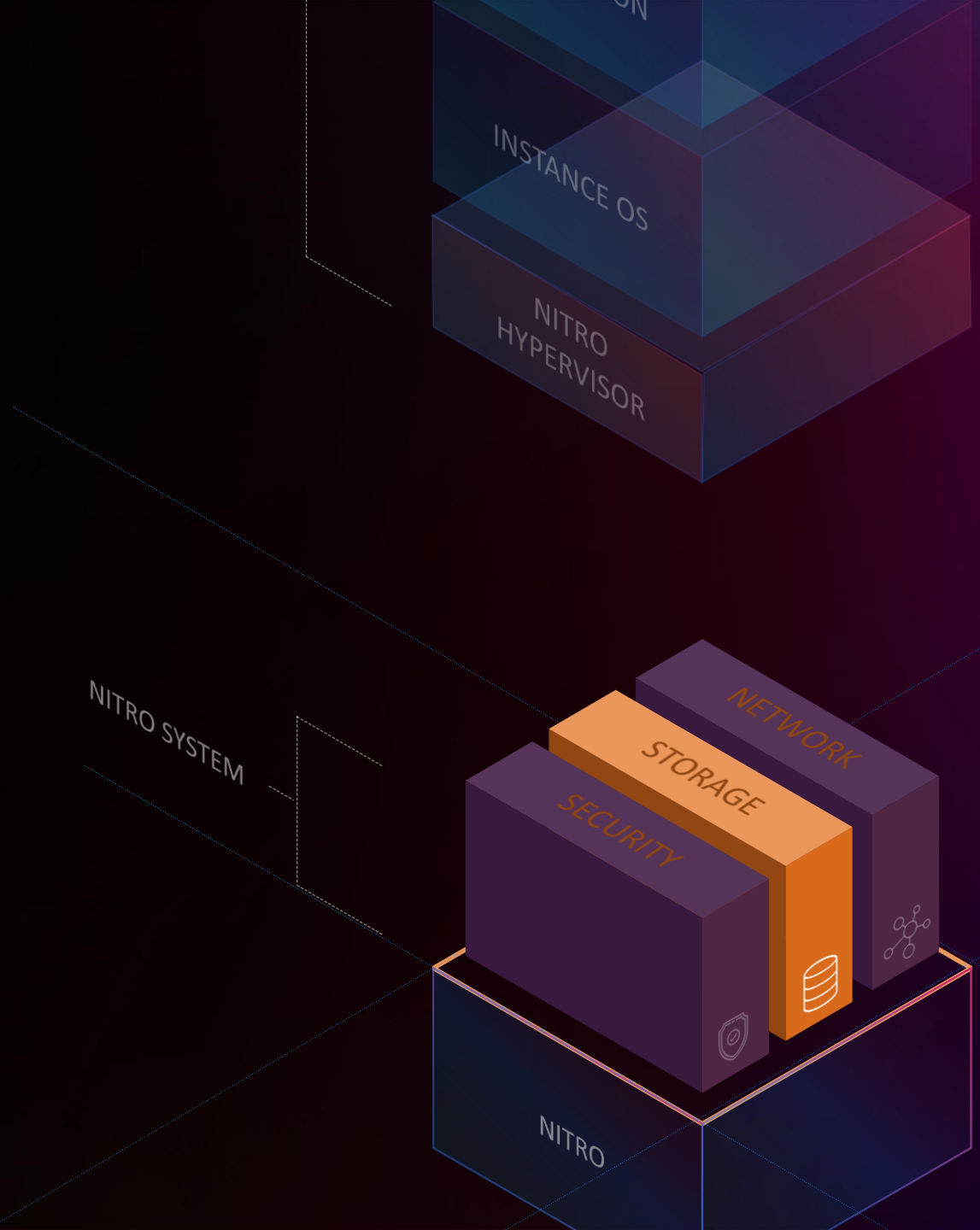
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Storage



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

Storage



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

Introducing AWS Nitro SSD

AVAILABLE NOW

High-performance, low-latency SSD custom designed by AWS



Lower latencies

Tightly integrated with the AWS Nitro System to provide 60% lower I/O latency



Improved reliability

Faster firmware updates to improve reliability without any downtime to the instance



Nitro security

All data stored on the disks is encrypted at rest with AES-256 ephemeral keys

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

Intel-based storage-optimized instances

AVAILABLE NOW



Highest X86-based local storage performance within Amazon EC2

- Powered by third-generation Intel Xeon Scalable processors (Ice Lake)
- Up to 128 vCPUs, 1 TiB memory, 75 Gbps networking, and 30 TiB of high-performance AWS Nitro SSD NVMe storage
- Up to 30% better compute price performance and 60% lower storage I/O latency than I3 instances
- Ideal for Microsoft SQL Server, Oracle databases, NoSQL databases, and data analytics workloads

aws

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

AWS Graviton2-based storage-optimized instances

AVAILABLE NOW

The best price performance for storage intensive workloads in Amazon EC2

- Up to 64 CPU cores, 256 GiB memory, and 100 Gbps networking
- Up to 40% better price performance and up to 44% lower cost per TB of storage compared to I3 instances
- Ideal for file systems, NoSQL databases, data indexing and search, and data analytics

---

- Up to 32 CPU cores, 192 GiB memory, and 50 Gbps networking
- Provide up to 45% better price performance and up to 15% lower cost per TB of storage compared to I3en instances
- Ideal for filesystems, stream processing and real-time databases, and log analytics





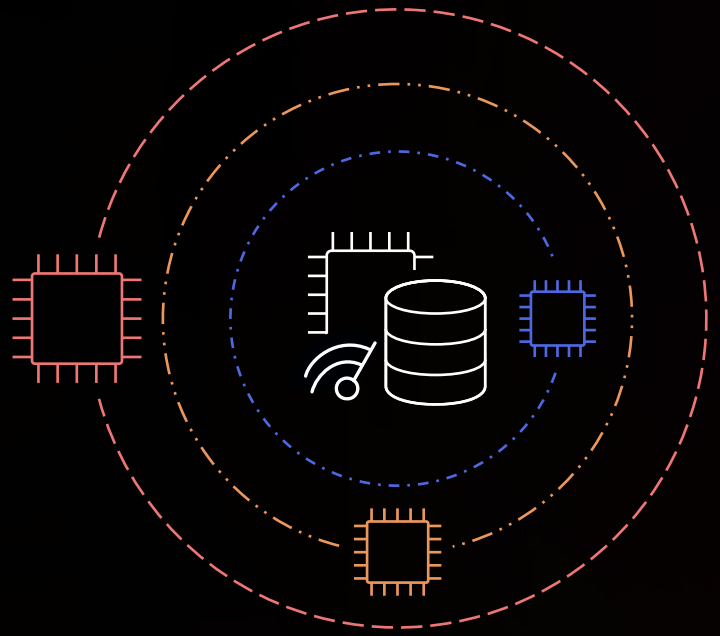
aws

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

NEW

Introducing Torn Write Prevention

Improve database performance with Amazon EC2 I4i instances and Amazon EBS



Enabled on AWS Nitro SSDs

Available on EC2 I4i storage optimized instances and EBS volumes for all Nitro-based instances

Database Performance

Up to 30% higher database Transactions Per Second (TPS)

Reduce Cost

Improved database performance eliminates the need to overprovision or scale up the cluster to meet business growth, reducing TCO

aws

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

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."

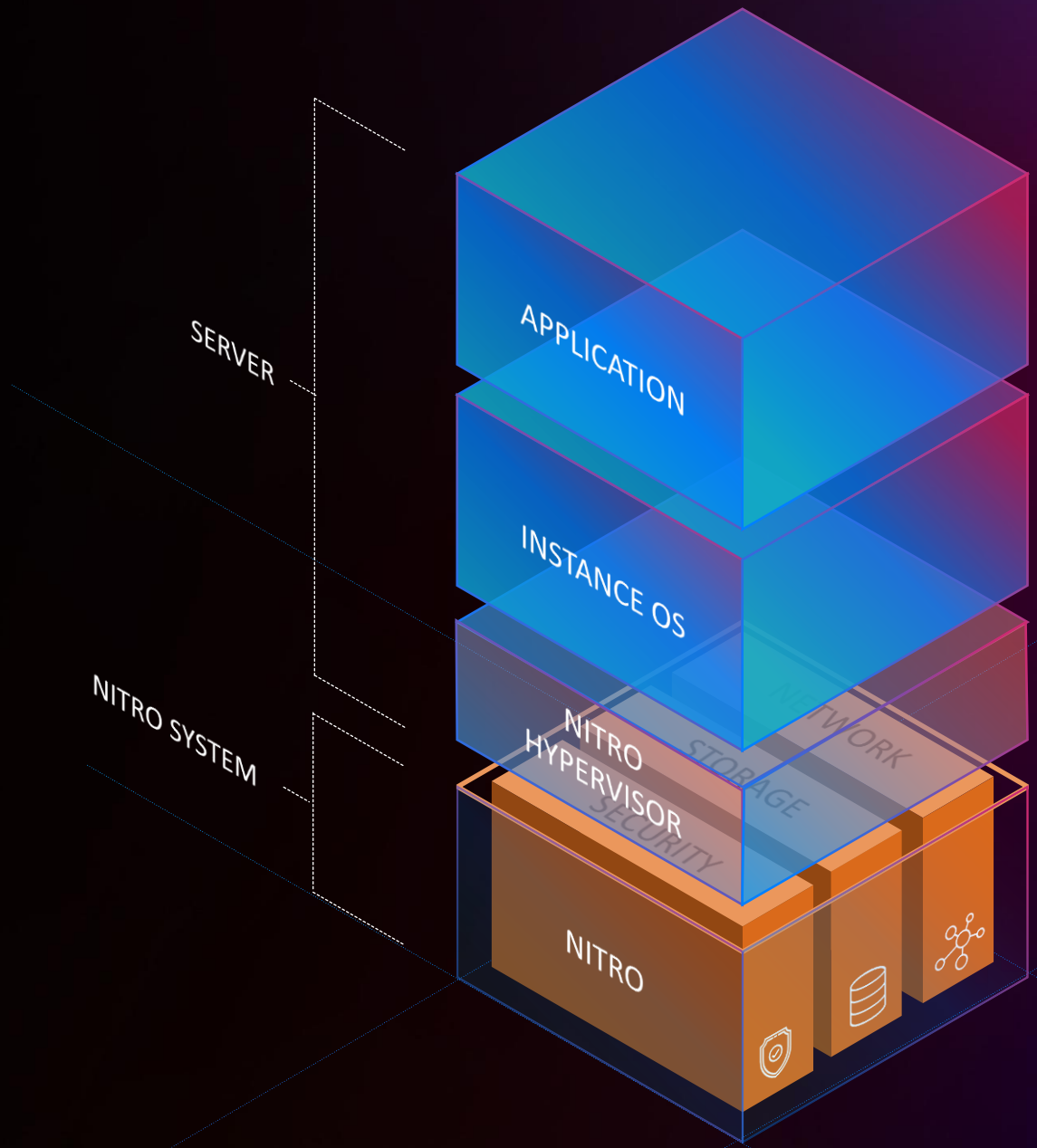
—Brad Murphy, Vice President
Cloud Operations & Infrastructure, Splunk



aws

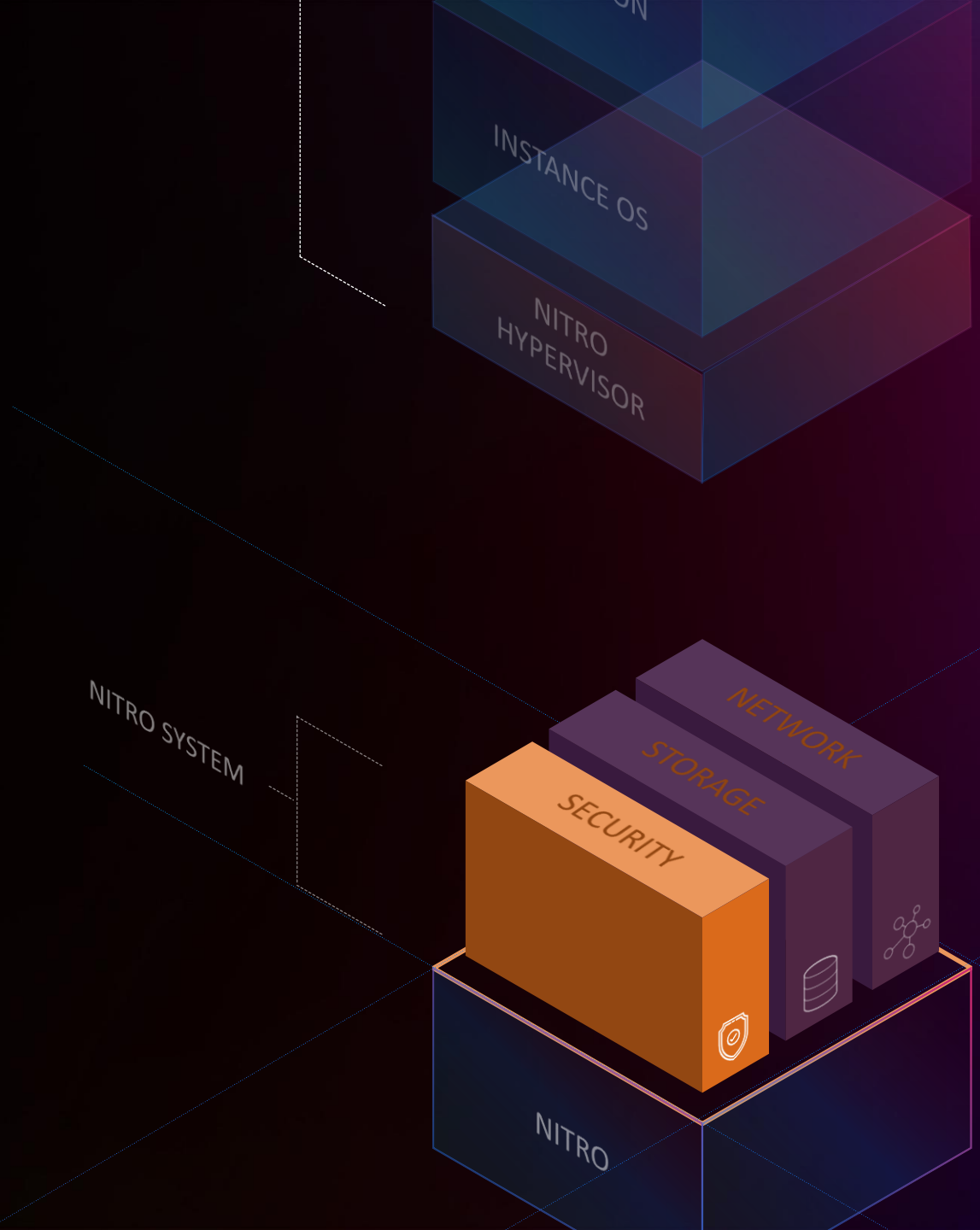
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Security



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

Security



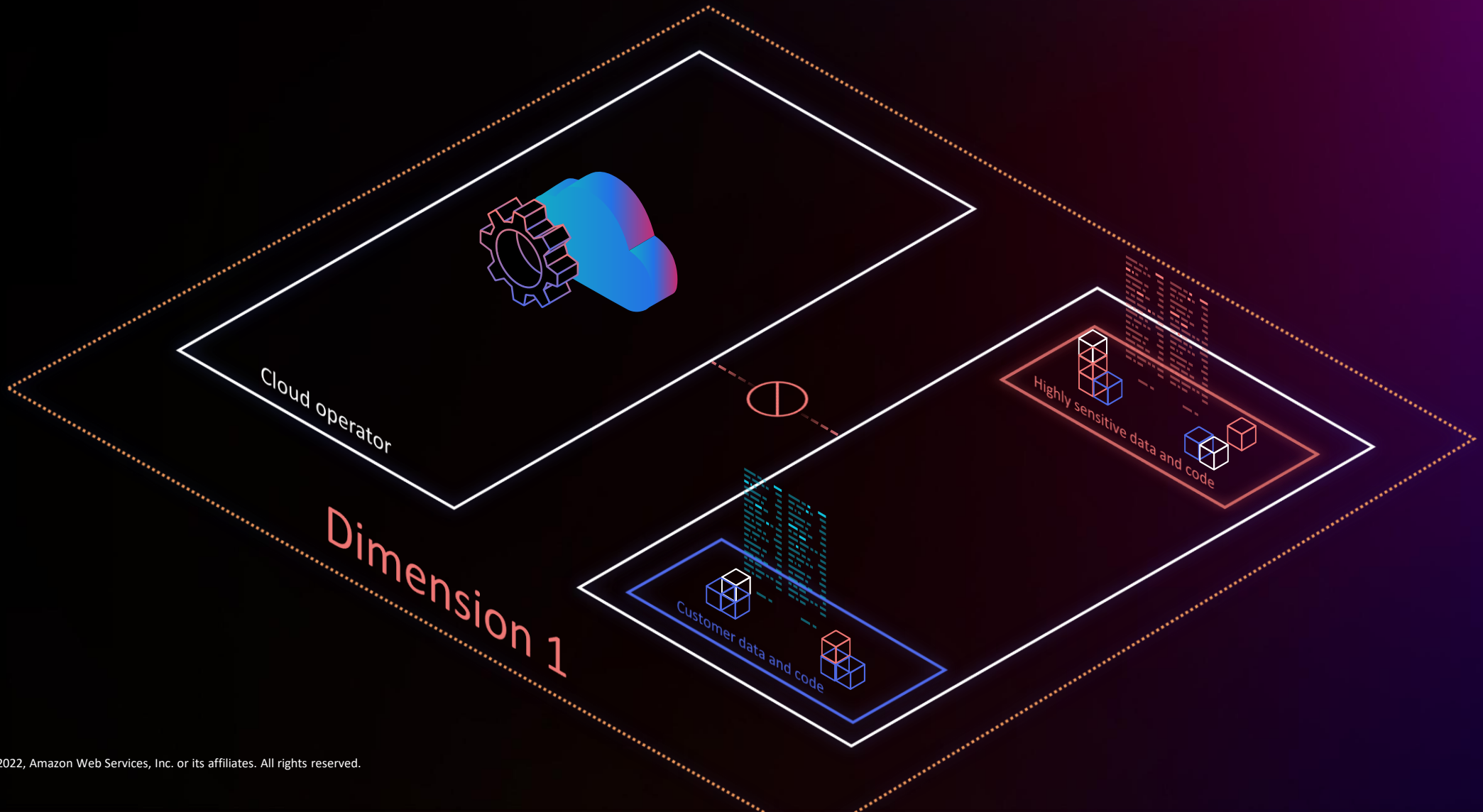
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Confidential computing

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

What is confidential computing?

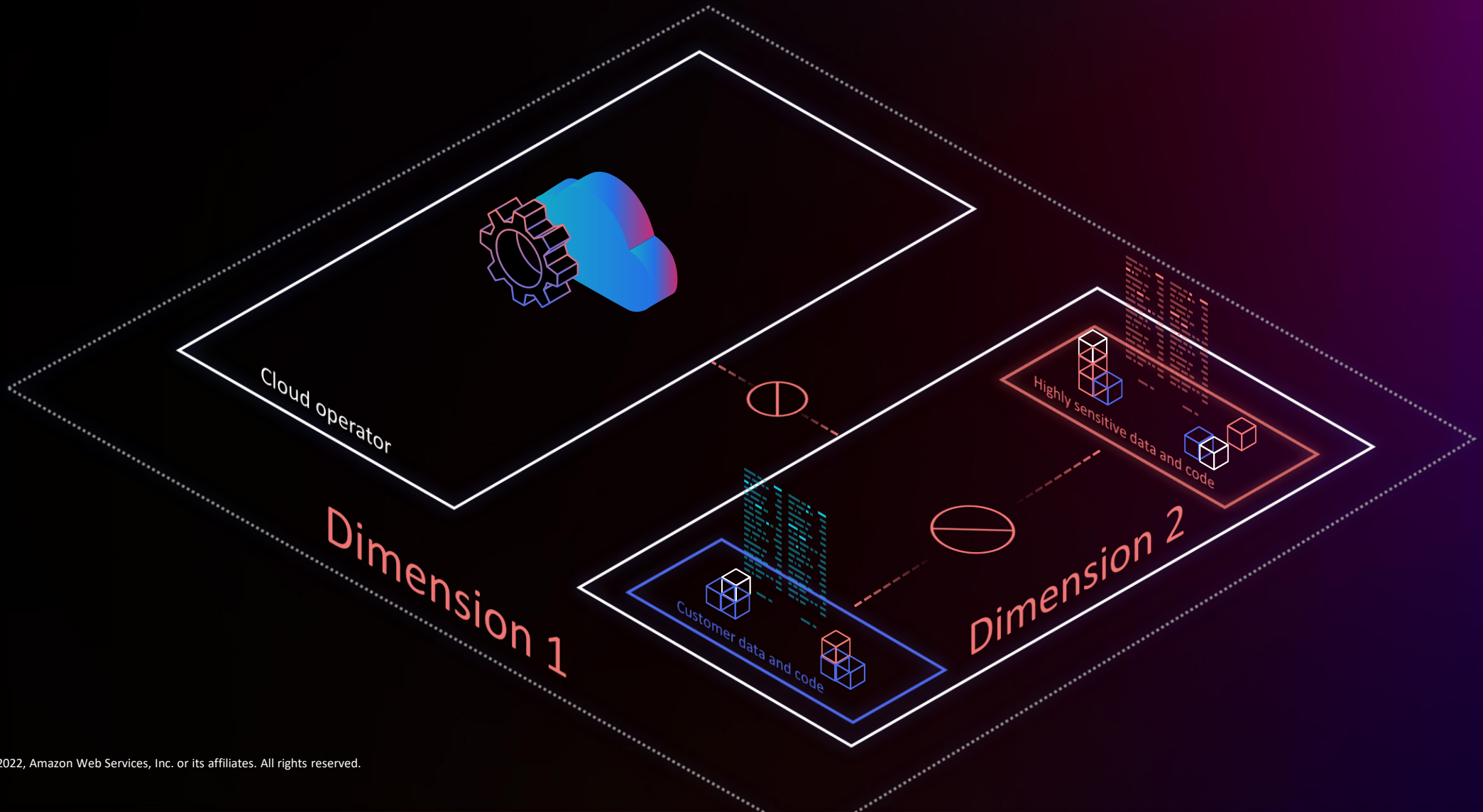
Protecting customer code and sensitive data in use



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

What is confidential computing?

Protecting customer code and sensitive data in use



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

Introducing AWS Nitro TPM

AVAILABLE NOW

Additional protection of sensitive information in Amazon EC2 Instances

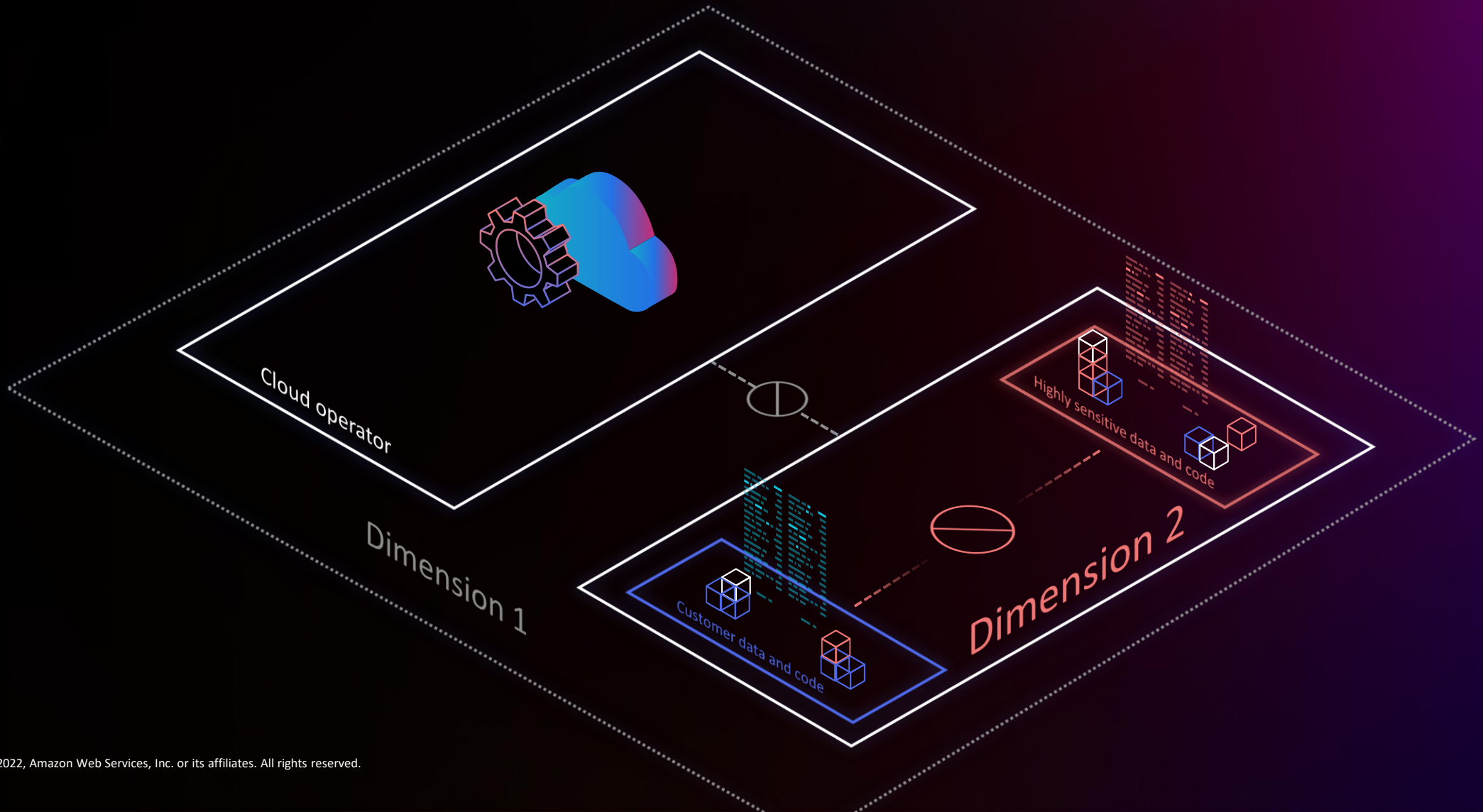
- Allows attestation of the health of Amazon EC2 instances and provides a source of trust that is rooted in silicon
- Easy to migrate on-premises applications that depend on a Trusted Platform Module (TPM)
- Conforms to the Trusted Platform Module 2.0 spec



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

What is confidential computing?

Protecting customer code and sensitive data in use



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

AWS Nitro Enclaves

CREATE ISOLATED COMPUTE ENVIRONMENTS TO FURTHER PROTECT AND PROCESS HIGHLY SENSITIVE DATA



FULLY ISOLATED

Enclaves are hardened and constrained environments that can only be accessed through secure local channel



SECURE

Attestation is integrated with AWS Key Management Service to ensure sensitive data is only accessed by the Nitro Enclave



USE CASES

Further protect and process highly sensitive data such as personally identifiable information (PII), healthcare, and financial data

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

NEW!

AVAILABLE NOW

AWS Nitro Enclaves Enhancements

CREATE ISOLATED COMPUTE ENVIRONMENTS TO FURTHER PROTECT AND PROCESS HIGHLY SENSITIVE DATA



GRAVITON

Nitro Enclaves is now supported across the majority of Graviton, Intel, and AMD-based Amazon EC2 instances



KUBERNETES

Use Kubernetes tools to orchestrate, scale, and deploy enclaves from a Kubernetes pod

aws

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

Broad choice of processors and architectures

RIGHT COMPUTE FOR THE WORKLOAD



Intel Xeon
scalable processors



AMD EPYC
processors



AWS Graviton processors

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

Broad choice of processors and architectures

RIGHT COMPUTE FOR THE WORKLOAD



Intel Xeon
scalable processors



AMD EPYC
processors



AWS Graviton processors

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

Innovating with Intel

16 YEARS OF COLLABORATION AND INNOVATION WITH AWS



Collaboration

Deep engineering collaboration across AWS portfolio



Extensive integration

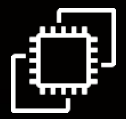
Over 350 Amazon EC2 instances are powered by Intel processors



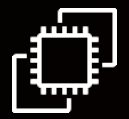
Fastest

Fastest processor in the cloud and widest selection of Ice Lake instances

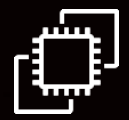
Recent intel-based instances



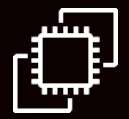
I4I
STORAGE-OPTIMIZED



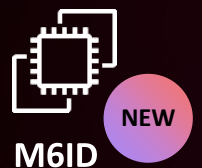
X2IDN
MEMORY-OPTIMIZED



X2IEDN
MEMORY-OPTIMIZED



X2IEZN
HIGH-FREQUENCY



M6ID
GENERAL PURPOSE



C6ID
COMPUTE-OPTIMIZED



R6ID
MEMORY-OPTIMIZED

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

Amazon EC2 C6id, M6id, and R6id instances

AVAILABLE NOW

NEW!

EC2 instances powered by 3rd gen Intel Xeon Scalable processor and NVMe attached storage

- Equipped with up to 7.6 TB of local NVMe-based SSD block-level storage for workloads that needs access to high-speed, low-latency storage
- Deliver up to 15% better price performance compared to previous gen C5d, M5d, and R5d instances
- Up to 2x faster networking and 20% higher memory bandwidth
- Support for Total Memory Encryption (TME)
- Ideal for core computing workloads that need access to high-speed, low latency storage.



aws

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

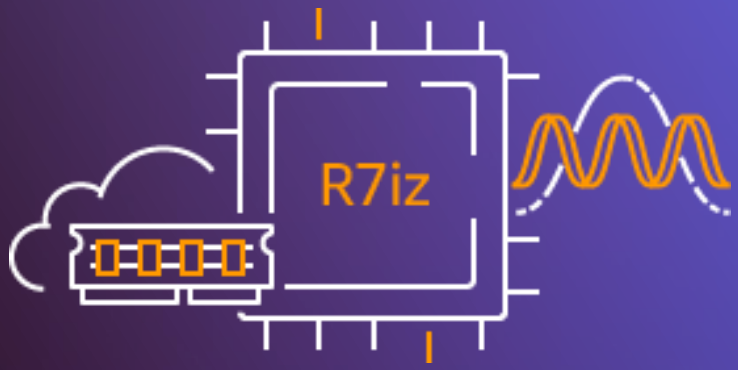
Amazon EC7 R7iz instances

PREVIEW

NEW!

High-frequency memory-optimized instances powered by 4th generation Intel Xeon Scalable processor

- Up to 128 vCPU, up to 1 TiB of memory to provide up to 2.6x more vCPU and memory compared to comparable high frequency instances
- Up to 20% higher performance when compared to comparable high frequency instances
- First x86-based EC2 instance to use DDR5 memory and deliver up to 2.4x higher mem bandwidth over comparable high frequency instances
- Designed for workloads such as front-end Electronic Design Automation (EDA), relational database workloads with high per-core licensing fees, and financial, actuarial, and data analytics simulation workloads



aws

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

Broad choice of processors and architectures

RIGHT COMPUTE FOR THE WORKLOAD



Intel Xeon
scalable processors



AMD EPYC
processors



AWS Graviton processors

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

Innovating with AMD

10% LOWER COST VERSUS COMPARABLE X86 INSTANCES



Flexibility and choice

Help you optimize both cost and performance for your workloads



Better economics

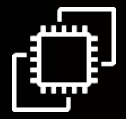
Deliver up to 10% lower cost versus comparable instances



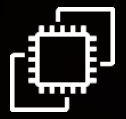
Seamless migration

Easily migrate applications currently running on existing Amazon EC2 instances to the new AMD-based variants with little to no modification

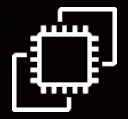
AMD-based instances



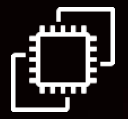
M5A
GENERAL PURPOSE



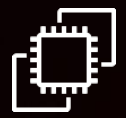
R5A
MEMORY-OPTIMIZED



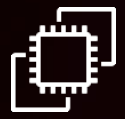
C5A
COMPUTE-OPTIMIZED



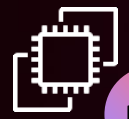
T3A
BURST-OPTIMIZED



G4AD
GPU-OPTIMIZED

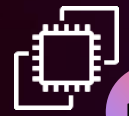


HPC6A
HPC-OPTIMIZED



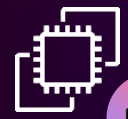
NEW

M6A
GENERAL PURPOSE



NEW

C6A
COMPUTE-OPTIMIZED



NEW

R6A
MEMORY-OPTIMIZED

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

Amazon EC2 C6a, M6a, and R6a instances

AVAILABLE NOW

NEW!

EC2 instances powered by 3rd gen AMD EPYC processor

- Up to 192 vCPUs, 1.5 TiB memory, 50 Gbps networking, and 40 Gbps EBS bandwidth
- Up to 35% better price performance compared to M5a/R5a instances and 15% better price performance compared to C5a instances.
- Up to 2.5x faster networking and up to 2x faster EBS bandwidth
- 10% lower cost than comparable x86-based Amazon EC2 instances
- Support for AMD Transparent Single Key Memory Encryption (TSME)
- Ideal for a broad set of core computing workloads



aws

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

Broad choice of processors and architectures

RIGHT COMPUTE FOR THE WORKLOAD



Intel Xeon
scalable processors



AMD EPYC
processors



AWS Graviton processors

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

AWS Graviton2

40% BETTER PRICE PERFORMANCE FOR BROAD RANGE OF WORKLOADS



Best price performance

Delivers up to 40% better price performance over comparable x86-based instances



Extensive partner community

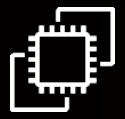
Supported by popular Linux operating systems along with many popular application and services from AWS and ISVs



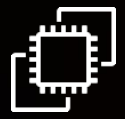
Enhanced security

Provide key capabilities for application security, including always on 256-bit DRAM encryption

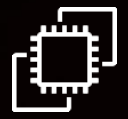
Graviton2 instances



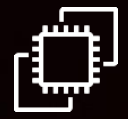
C6G(D)
COMPUTE-OPTIMIZED



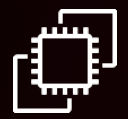
M6G(D)
GENERAL PURPOSE



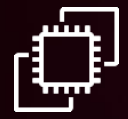
R6G(D)
MEMORY-OPTIMIZED



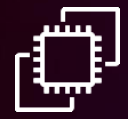
C6GN
ENHANCED NETWORKING



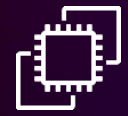
X2GD
ENHANCED MEMORY



T4G
BURSTABLE



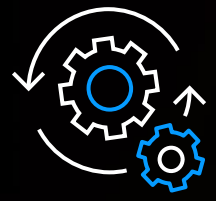
IM4GN
STORAGE-OPTIMIZED



IS4GEN
STORAGE-OPTIMIZED

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

AWS Graviton3



Improved Performance

Up to 25% higher compute performance and 2x higher floating-point to accelerate compute-intensive workloads



Faster Memory

Supports DDR5 memory to provide 50% more memory bandwidth over DDR4 memory



Energy Efficient

Consumes 60% less power for the same performance compared to other CPUs

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

Amazon EC2 C7g instances

AVAILABLE NOW

NEW!

Best price performance for compute-intensive workloads on Amazon EC2

- Compute-optimized instances powered by AWS Graviton3 CPUs
- Up to 64 vCPUs, 128 GiB of memory, 30 Gbps of network performance, and 25 Gbps of EBS performance
- Up to 25% better compute and 2x higher floating-point performance over Graviton2 to accelerate compute workloads
- First in the cloud to feature the cutting edge DDR5 memory, that provides 50% more bandwidth compared to DDR4 memory



aws

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

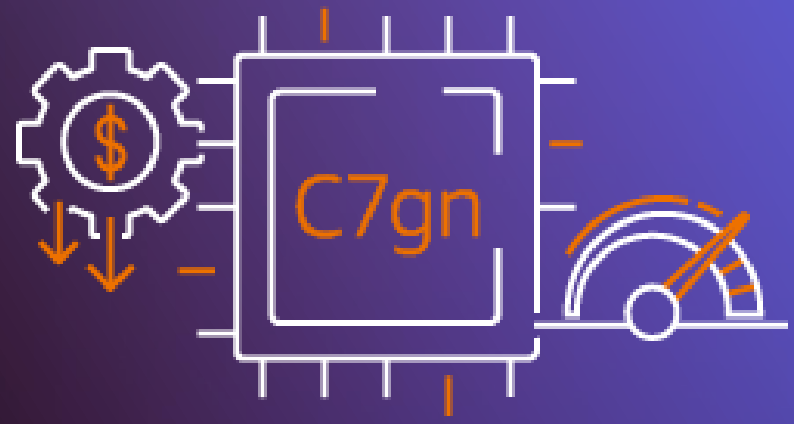
Amazon EC2 C7gn instances

PREVIEW

NEW!

Network optimized for highest network bandwidth for Graviton-based EC2 instances

- New Nitro cards enable up to 200 Gbps of network bandwidth and up to 50% higher pack-processing performance
- Up to 25% better compute and 2x higher floating-point performance over Graviton2 to accelerate compute workloads
- Ideal for network intensive workloads such as network virtual appliances, data analytics, and CPU-based AI/ML inference



aws

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

Customers adopting AWS Graviton









































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

Graviton3 customer feedback



"Our testing has shown they are suitable for even the most demanding latency-sensitive workloads while providing significant price performance benefits and expanding what is possible within Fortnite and any Unreal Engine-created experience."

—Mark Imbriaco, Senior Director of Engineering at Epic Games



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

—Pat Symonds, CTO at Formula 1 Management



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

—Liz Fong-Jones, Principal Developer Advocate, honeycomb.io

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

aws

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

AWS Graviton Ready Program

CERTIFIED PARTNER SOLUTIONS FOR GRAVITON CUSTOMERS

















































aws

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

AWS managed services supporting Graviton

EXTENDING 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 and containers



AWS Lambda



AWS Fargate



AWS Elastic Beanstalk



Amazon Elastic Kubernetes Service (EKS)



Amazon Elastic Container Service (ECS)

aws

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

Broad choice of processors and architectures

RIGHT COMPUTE FOR THE WORKLOAD



Intel Xeon
scalable processors



AMD EPYC
processors



AWS Graviton processors

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

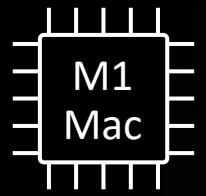


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

NEW!

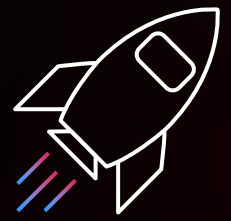
Amazon EC2 Mac instances

ON-DEMAND APPLE SILICON MACOS ENVIRONMENTS FOR THE FIRST TIME ON AWS



POWERED BY APPLE SILICON

Apple M1 chip integrates the CPU, GPU, Neural Engine, I/O, and so much more onto a single tiny chip



IMPROVED PERFORMANCE

Up to 4x better build performance compared to on-premise and up to 60% better price performance compared to x86 Mac instances



HARNESS THE CLOUD

Provision macOS environments within minutes and only pay for what you use. Offload the heavy lifting that comes with managing infrastructure to AWS

aws

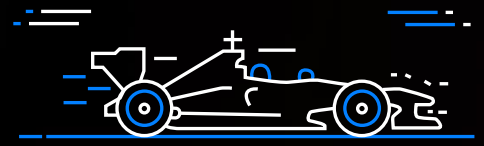
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

High-performance computing is a fast-growing space and one that demands continuous innovation

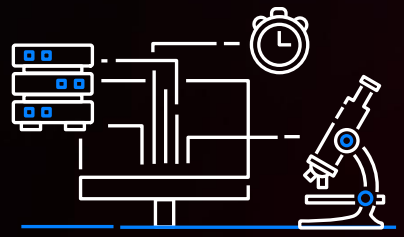
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

High-performance Computing

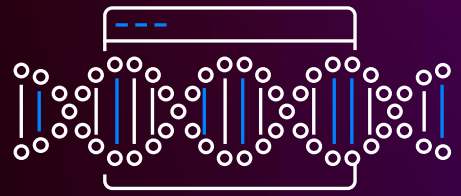
Computation Fluid Dynamics



Life Sciences



Genomics



Risk Analysis



Energy



Research



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

High-performance computing

OPTIMIZED FOR VIRTUALLY ANY HPC WORKLOAD WITH THE ELASTICITY TO SCALE



AWS Batch



AWS ParallelCluster



Elastic Fabric Adapter
+100 Gbps networking

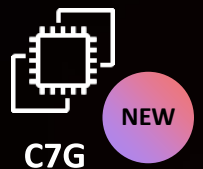


Amazon FSx
for Lustre

HPC-optimized instances



HPC6A
COMPUTE-OPTIMIZED



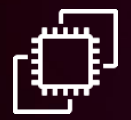
C7G
COMPUTE-OPTIMIZED



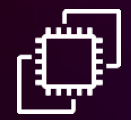
C6I
COMPUTE-OPTIMIZED



C6A
COMPUTE-OPTIMIZED



C6GN
COMPUTE-OPTIMIZED



C5N
COMPUTE-OPTIMIZED

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

High-performance computing

OPTIMIZED FOR VIRTUALLY ANY HPC WORKLOAD WITH THE ELASTICITY TO SCALE



AWS Batch



AWS ParallelCluster



Elastic Fabric Adapter
+100 Gbps networking



Amazon FSx
for Lustre

HPC-optimized instances



HPC6A
COMPUTE-OPTIMIZED



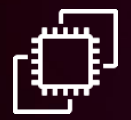
C7G
COMPUTE-OPTIMIZED



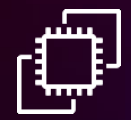
C6I
COMPUTE-OPTIMIZED



C6A
COMPUTE-OPTIMIZED



C6GN
COMPUTE-OPTIMIZED



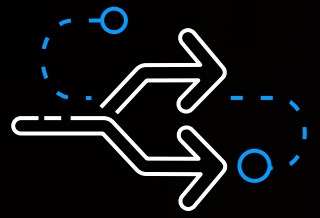
C5N
COMPUTE-OPTIMIZED

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

NEW

Amazon EC2 Hpc6id instances

Best price performance for memory and data-intensive HPC workloads in Amazon EC2



200G networking with EFA

2x higher Elastic Fabric Adaptor performance over current generation HPC instances for increased application performance



Price performance benefits

Up to 2.2x better price-performance for data-intensive HPC workloads such as Finite Element Analysis (FEA) over comparable x86-based instances



Optimized for for data intensive HPC workloads

1TB of instance memory and 15.2 TB of NVMe storage to accelerate seismic, energy, and FEA workloads

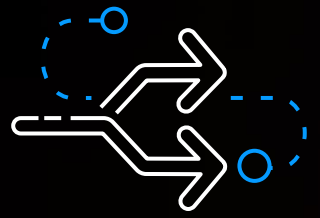
aws

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

NEW

Amazon EC2 Hpc6id instances

Best price performance for memory and data-intensive HPC workloads in Amazon EC2



200G networking with EFA

2x higher Elastic Fabric Adaptor performance over current generation HPC instances for increased application performance



Price performance benefits

Up to 2.2x better price-performance for data-intensive HPC workloads such as Finite Element Analysis (FEA) over comparable x86-based instances



Optimized for for data intensive HPC workloads

1TB of instance memory and 15.2 TB of NVMe storage to accelerate seismic, energy, and FEA workloads

aws

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

COMING SOON

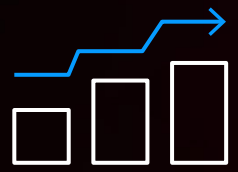
Amazon EC2 Hpc7g instances

Best price performance for compute intensive HPC workloads in Amazon EC2



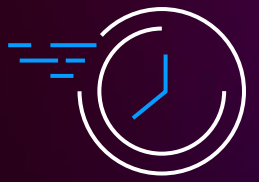
200G networking with EFA

The new Nitro v5 card enables 2x higher network performance and up to 25% lower EFA latency to efficiently scale your HPC cluster



Best price performance for HPC workloads

New Graviton3E processor enables up to 60% better performance for compute-intensive workloads such as weather and CFD over comparable x86-based instances



Deploy with AWS ParallelCluster

Deploy and manage your Hpc7g instances with AWS' open source cluster management tool

aws

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

AWS Best HPC Cloud Platform 5 years running









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

Broadest choice of accelerators

RIGHT ACCELERATOR FOR THE WORKLOAD

AMD/XILINX FPGAs, video transcoding, GPUs

NVIDIA A100 and T4 Tensor Core GPUs

Habana Gaudi accelerators

AWS Inferentia and AWS Trainium

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

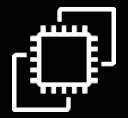
Supporting ML workloads

LARGEST SELECTION OF ML INSTANCES IN THE CLOUD

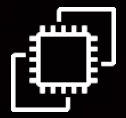
Training



TRN1
AWS TRAINIUM



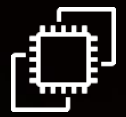
P3DN
NVIDIA V100
TENSOR CORE



P4D
NVIDIA A100
TENSOR CORE

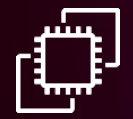


P4DE
NVIDIA A100
TENSOR CORE

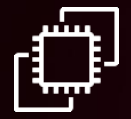


DL1
HABANA GAUDI

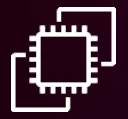
Inference



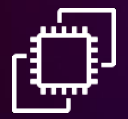
INF1
AWS INFERENTIA



G4
NVIDIA T4 GPU



G5
NVIDIA A10G
TENSOR CORE



G5G
NVIDIA T4G
TENSOR CORE

aws

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

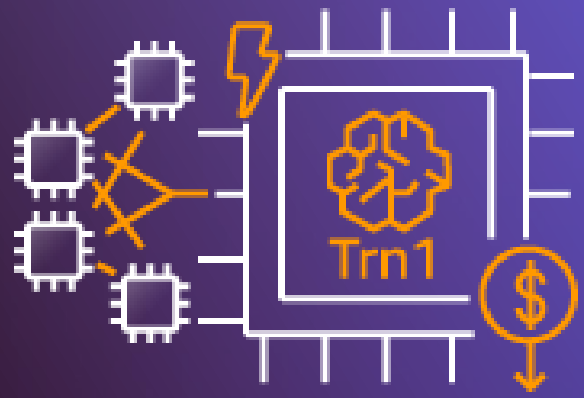
Amazon EC2 Trn1 instances

AVAILABLE NOW

NEW!

Highest performance and lowest cost for training deep learning models

- 16 Trainium accelerators with 512 GiB of HBM2 high speed memory, 800 Gbps of networking throughput, and 8 TB of local NVMe storage
- Lower time-to-train when compared to the latest comparable Amazon EC2 instances
- Deployable in Amazon EC2 UltraClusters, a supercomputer in the cloud with tens of thousands of Trainium accelerators with petabit scale networking bandwidth
- Amazon EC2 Trn1n instance coming soon!



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

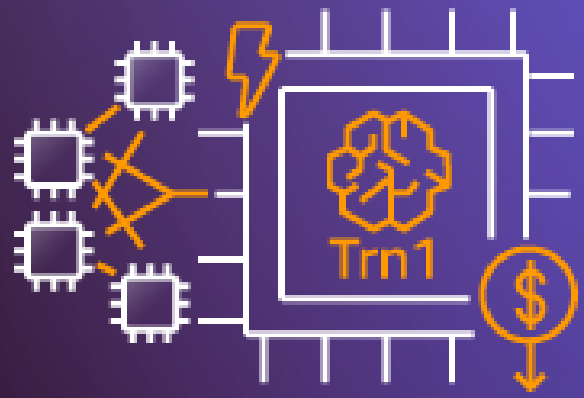
Amazon EC2 Trn1 instances

AVAILABLE NOW

NEW!

Highest performance and lowest cost for training deep learning models

- 16 Trainium accelerators with 512 GiB of HBM2 high speed memory, 800 Gbps of networking throughput, and 8 TB of local NVMe storage
- Lower time-to-train when compared to the latest comparable Amazon EC2 instances
- Deployable in Amazon EC2 UltraClusters, a supercomputer in the cloud with tens of thousands of Trainium accelerators with petabit scale networking bandwidth
- Amazon EC2 Trn1n instance coming soon!



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

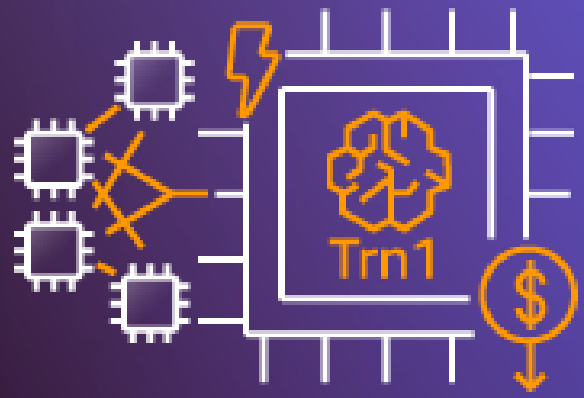
Amazon EC2 Trn1 instances

AVAILABLE NOW

NEW!

Highest performance and lowest cost for training deep learning models

- 16 Trainium accelerators with 512 GiB of HBM2 high speed memory, 800 Gbps of networking throughput, and 8 TB of local NVMe storage
- Lower time-to-train when compared to the latest comparable Amazon EC2 instances
- Deployable in Amazon EC2 UltraClusters, a supercomputer in the cloud with tens of thousands of Trainium accelerators with petabit scale networking bandwidth
- Amazon EC2 Trn1n instance coming soon!

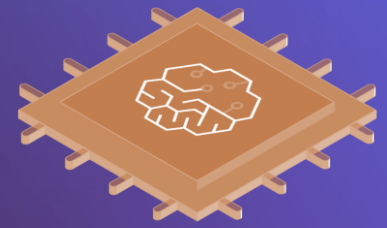


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon EC2 Inf1 instances

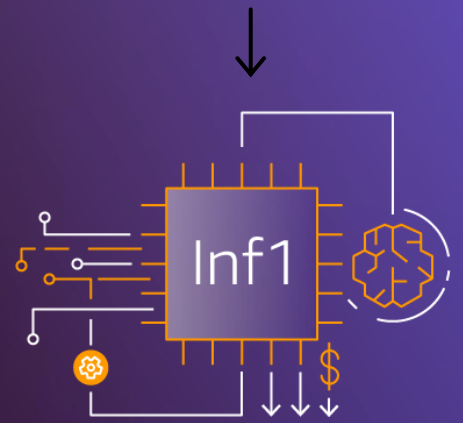
High-performance, lowest-cost machine learning inference

- Featuring **AWS Inferentia**, the first ML chip designed by Amazon
- **Lowest in the cloud** for running deep learning models
- Up to 70% lower cost than comparable GPU instances
- Seamless software **integration with ML frameworks** like TensorFlow, PyTorch, and MXNet for quickly getting started and 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

High performance and lowest cost inference in the cloud

aws

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

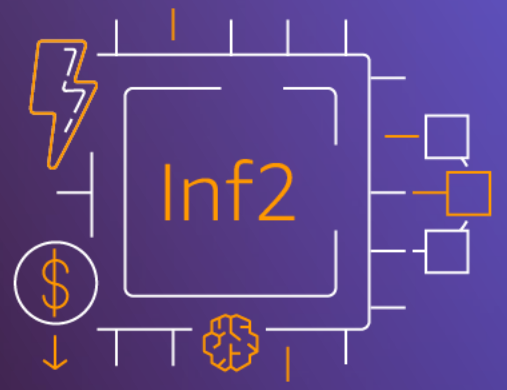
Amazon EC2 Inf2 instances

PREVIEW

NEW!

Best performance at the lowest cost in the cloud for DL inference

- Optimized to deploy 100B+ parameter models at scale
- Up to 4x higher throughput and up to 10x lower latency than Inf1 instances
- Up to 12 Inferentia2 accelerators and up to 384 GB of HBM2e high speed accelerator memory
- First inference platform with direct, 192 GB/s connectivity between multiple accelerators for distributed inference
- Native integration with PyTorch and TensorFlow
- 45% better performance/watt than GPU-based instances



aws

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

AWS Inferentia Momentum

airbnb

AUTODESK

Finch COMPUTING

sprinklr

NTTPC

amazonrobotics

SCREENING EAGLE

amazon alexa

amazon ads

The Asahi Shimbun

CONDÉ NAST

Talroo™

Anthem

DISCO

SKYWATCH

Money Forward

prime video

DMP DIGITAL MEDIA PROFESSIONALS

Amazon Rekognition

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

ML Silicon customer momentum

AUTODESK

"Autodesk is advancing the cognitive technology of our AI-powered virtual assistant, Autodesk Virtual Agent (AVA) by using Inferentia. Piloting Inferentia, we are able to obtain a 4.9x higher throughput over G4dn for our NLU models."

—Binghui Ouyang, Sr. Data Scientist, Autodesk

Anthem

"We seamlessly deployed our deep learning inferencing workload onto Amazon EC2 Inf1 instances powered by the AWS Inferentia processor. The new Inf1 instances provide 2x higher throughput to GPU-based instances and allowed us to streamline our inference workloads."

—Numan Laanait PhD, Principal AI/Data Scientist, Anthem
—Miro Mihaylov PhD, Principal AI/Data Scientist, Anthem

---

airbnb

"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

aws

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

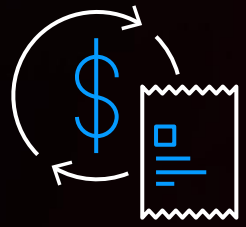
Amazon Cost Optimization

Best performance at the lowest cost in the cloud for DL inference



Instances

Graviton instances offer up to 40% better price performance while AMD instances deliver a 10% savings versus other x86 based instances



Purchase Models

Savings Plans offer up to 72% savings while Spot instances deliver up to 90% discount versus on-demand



Tools

Tools such as Compute Optimizer provide easy recommendations to reduce cost with no performance impact

aws

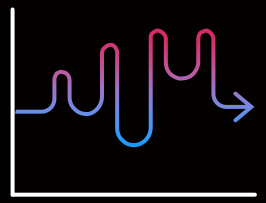
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

EC2 Purchase Models

EC2 OFFERS SIGNIFICANT COST REDUCTION OPOTUNITIES THROUGH ON-DEMAND, SAVINGS PLANS AND SPOT INSTANCES

ON-DEMAND

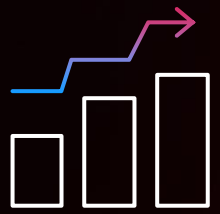
Pay for compute capacity by the second with no long-term commitments



Spiky workloads, to define needs

SAVINGS PLANS

Make a 1- or 3-year commitment and receive a significant discount off On-Demand prices



Committed and steady-state usage

SPOT INSTANCES

Spare EC2 capacity at savings of up to 90% off On-Demand prices



Fault-tolerant, flexible, stateless workloads

aws

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

Over

$15 billion

saved with Savings Plan

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

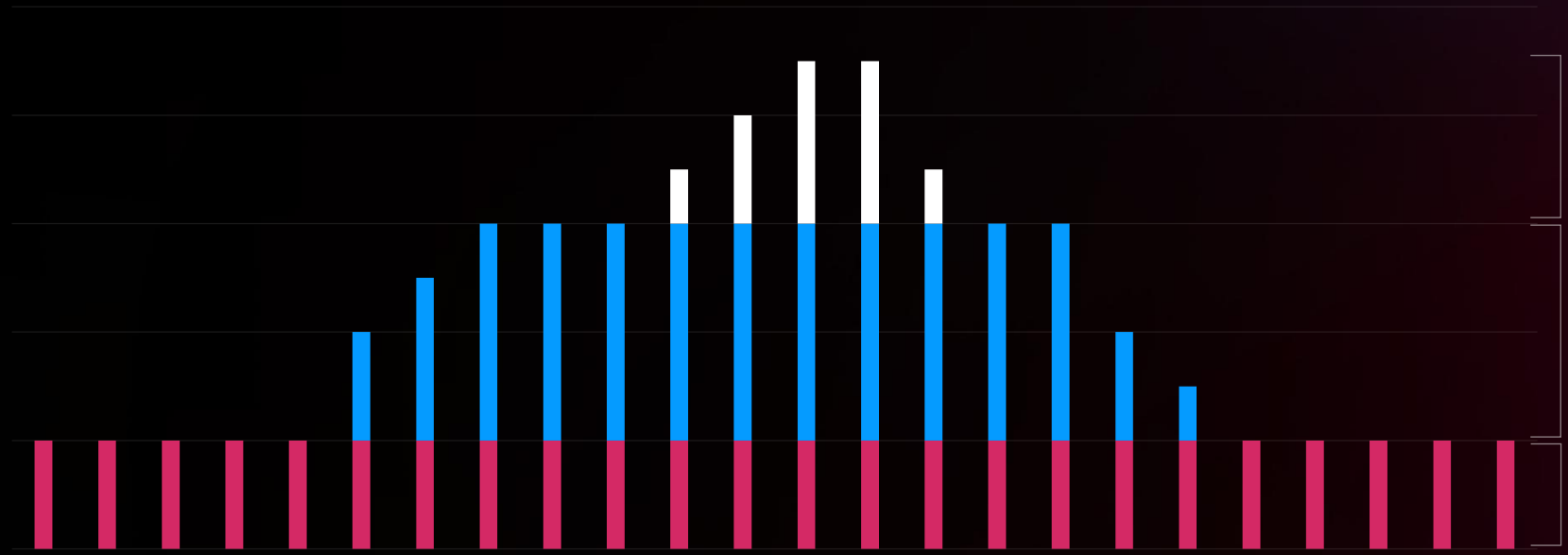
Over

$10 billion

saved with
EC2 Spot instances

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

Simplifying capacity & cost optimization



SCALE USING **SPOT** FOR FLEXIBLE, FAULT-TOLERANT WORKLOADS

SCALE USING **ON-DEMAND** FOR NEW OR STATEFUL SPIKY WORKLOADS

USE **SAVINGS PLANS** FOR KNOWN/STEADY-STATE WORKLOADS

AWS SERVICES MAKE THIS EASY AND EFFICIENT

- Amazon EC2 Auto Scaling
- EC2 Fleet
- Amazon Elastic Container Service (Amazon ECS)
- Amazon Elastic Kubernetes Service (Amazon EKS)
- AWS Thinkbox
- Amazon EMR
- AWS CloudFormation
- AWS Batch

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

Over

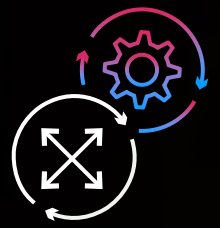
10 billion

recommendations with Compute Optimizer

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

AWS Compute Optimizer

REDUCE COST AND INCREASE PERFORMANCE BY OPTIMIZING YOUR AWS RESOURCES



EC2 INSTANCE TYPE OPTIMIZATION

Take advantage of EC2's broad portfolio of instances to optimize workload performance and eliminate under provisioned resources



DATA-BACKED RECOMMENDATIONS

Rightsize workloads with machine learning-based recommendations to reduce costs by up to 25%



SUPPORT FOR KEY AWS RESOURCES

Amazon Elastic Compute Cloud (EC2) instance types, Amazon Elastic Block Store (EBS) volumes, and AWS Lambda functions

aws

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

Thank you!



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

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