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

CMP223-L

Compute innovation to enable any application in the cloud

Dave Brown (he/him)

Vice President, Amazon EC2
AWS

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

100 million

Instances launched everyday

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

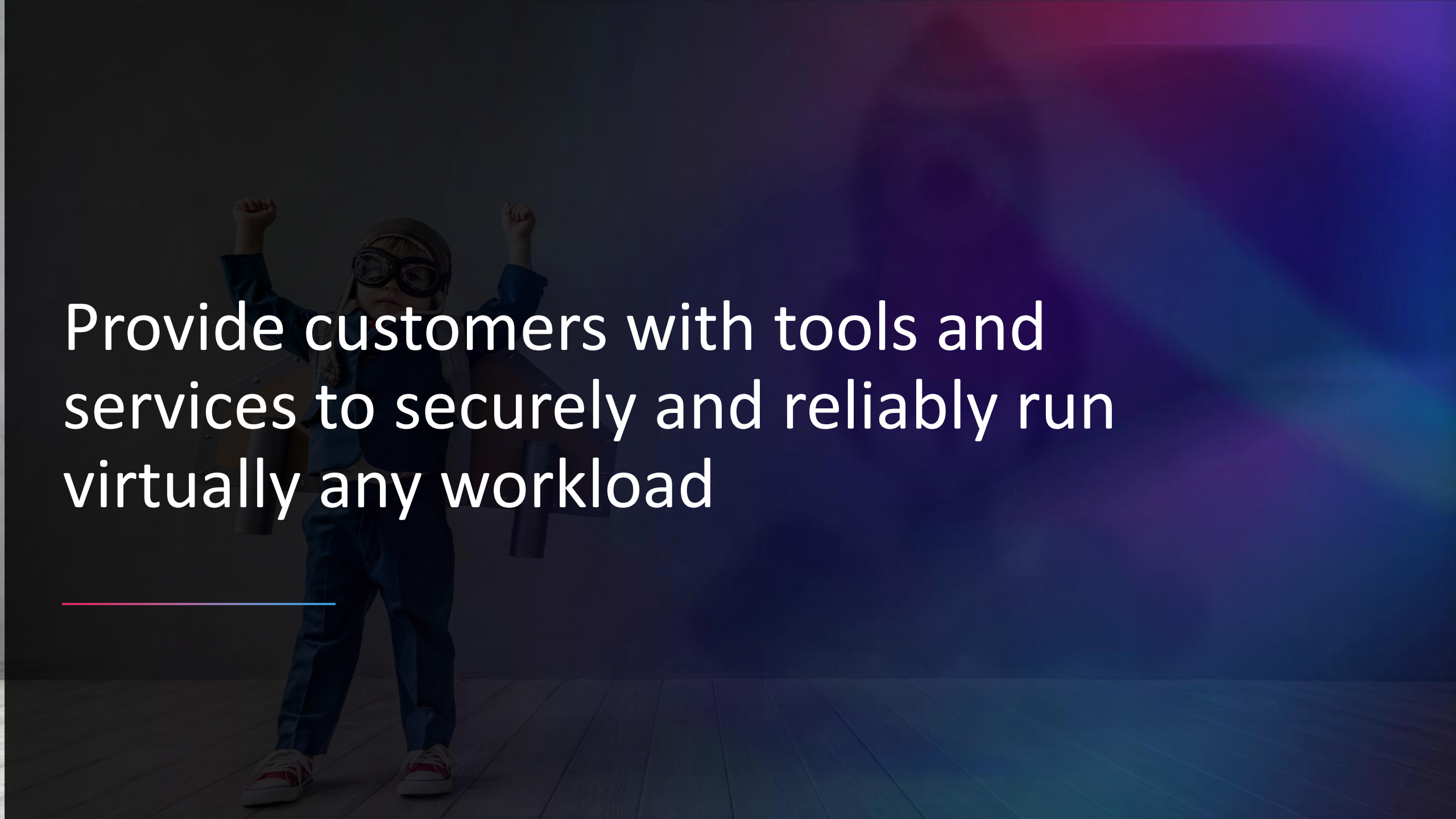
Over

30 billion

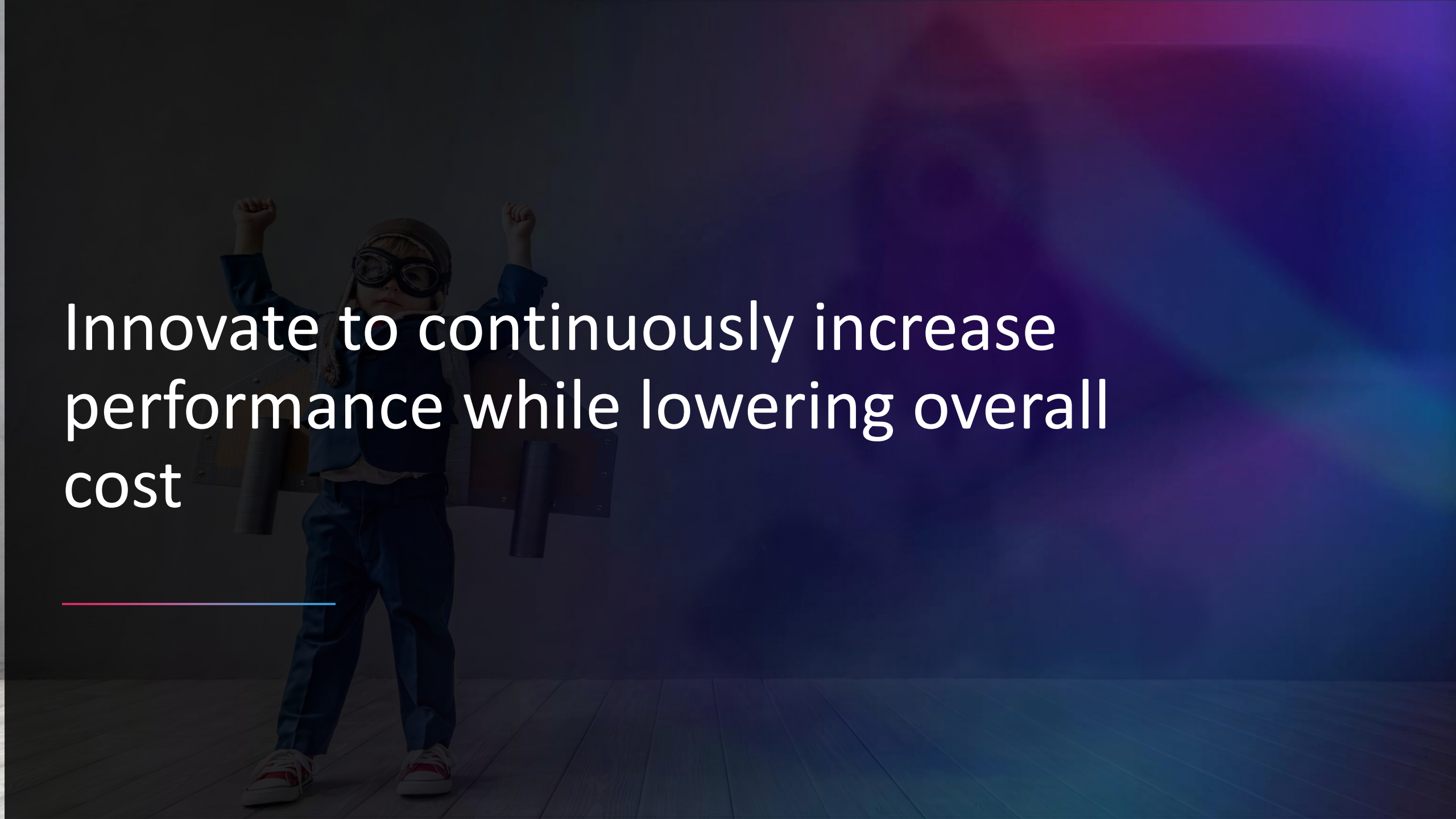
Instances launched since 2006

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





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



Innovate to continuously increase performance while lowering overall cost

“AWS really helps us scale up and down and helps us manage infrastructure in a way that keeps our teams from having to worry about infrastructure in a direct day to day way and to really focus on delivering experiences to our customers.”

Mark Imbriaco
Technical Director, Operations and Reliability,
Epic Games



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

COMPUTE INNOVATION FOR EVERY WORKLOAD



World class scale and performance



Operational excellence



Compute for every workload



Cost optimization and best practices



Compute where you need it

aws

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

COMPUTE INNOVATION FOR EVERY WORKLOAD



World class scale and performance



Operational excellence



Compute for every workload



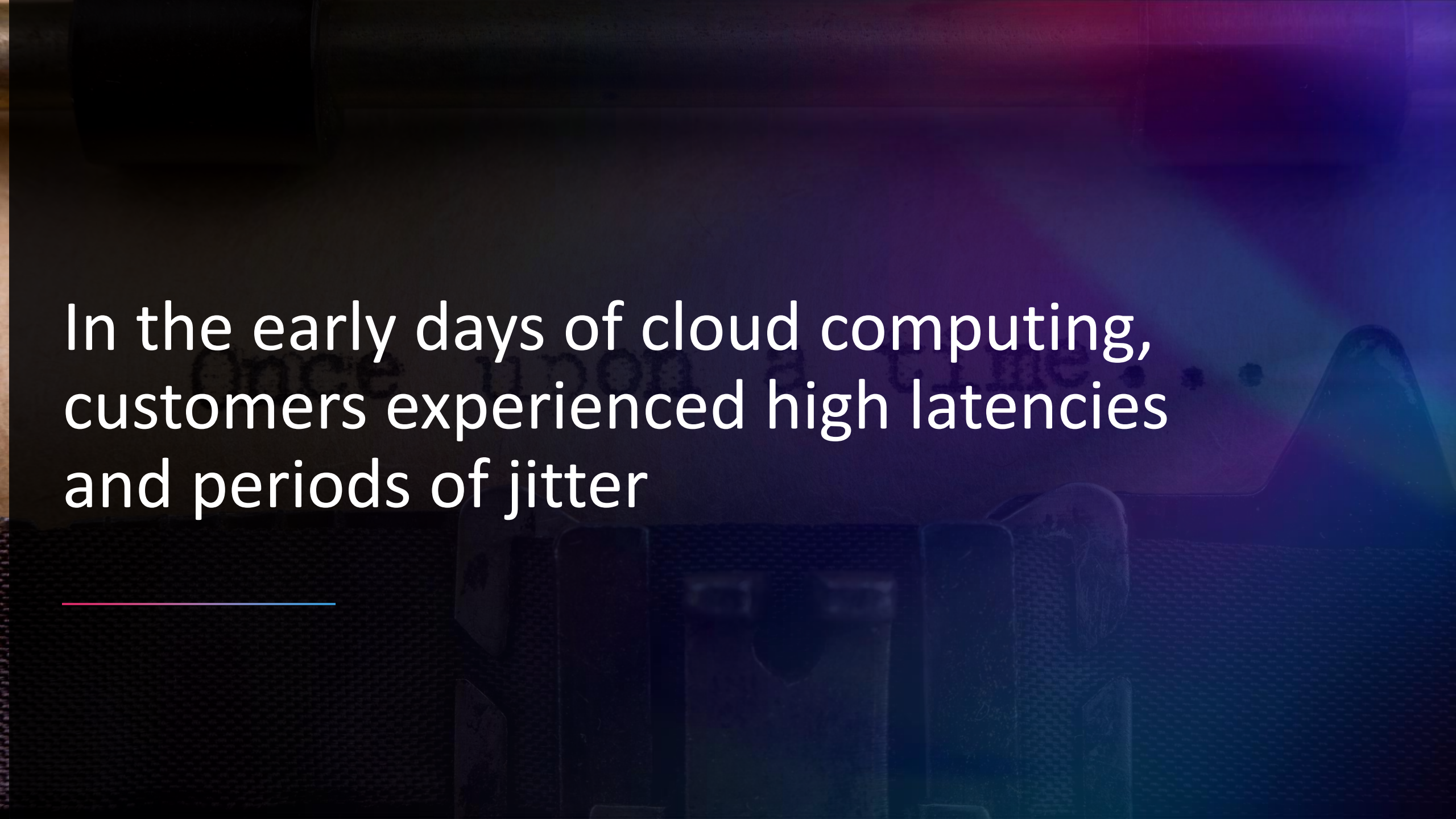
Cost optimization and best practices



Compute where you need it

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


Once upon a time...


In the early days of cloud computing, customers experienced high latencies and periods of jitter

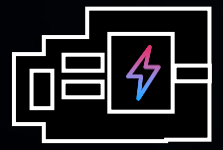

Virginia St 1900
HISTORICAL DISTRICT
Pike Place Market
ONE WAY
NO STOPS 6-9AM 3-6PM
Virginia
INN


powered by
annapurnalabs
an amazon company

Innovation enabled by AWS Nitro System

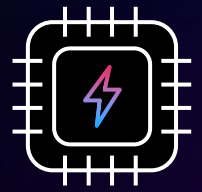
MODULAR BUILDING BLOCKS FOR RAPID DESIGN AND DELIVERY OF EC2 INSTANCES

NITRO CARD



Local NVMe storage,
Amazon EBS, networking,
monitoring, and security

NITRO SECURITY CHIP



Integrated into motherboard

Protects hardware resources

NITRO HYPERVISOR



Lightweight hypervisor

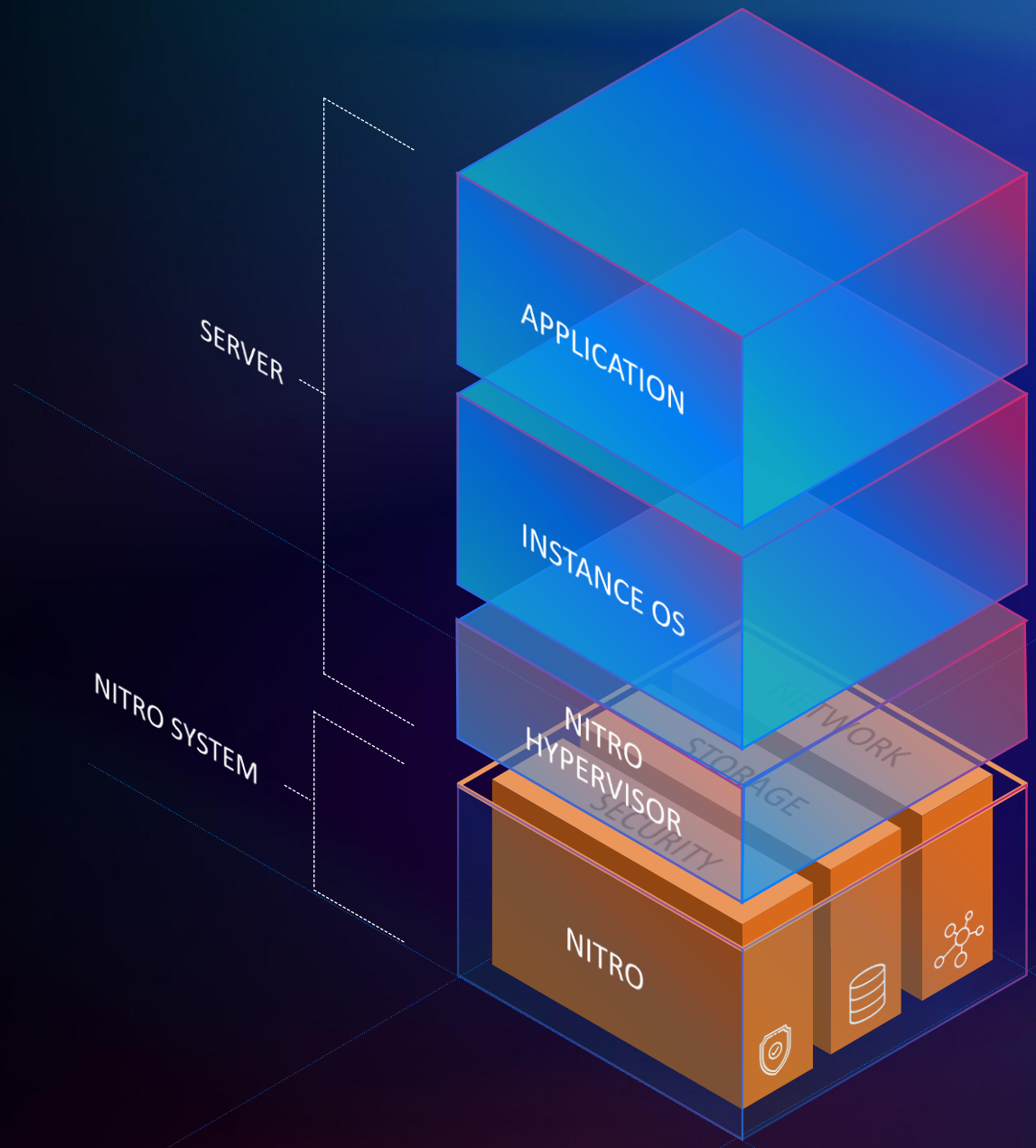
Memory and CPU allocation

Bare metal-like performance

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

The AWS Nitro System architecture

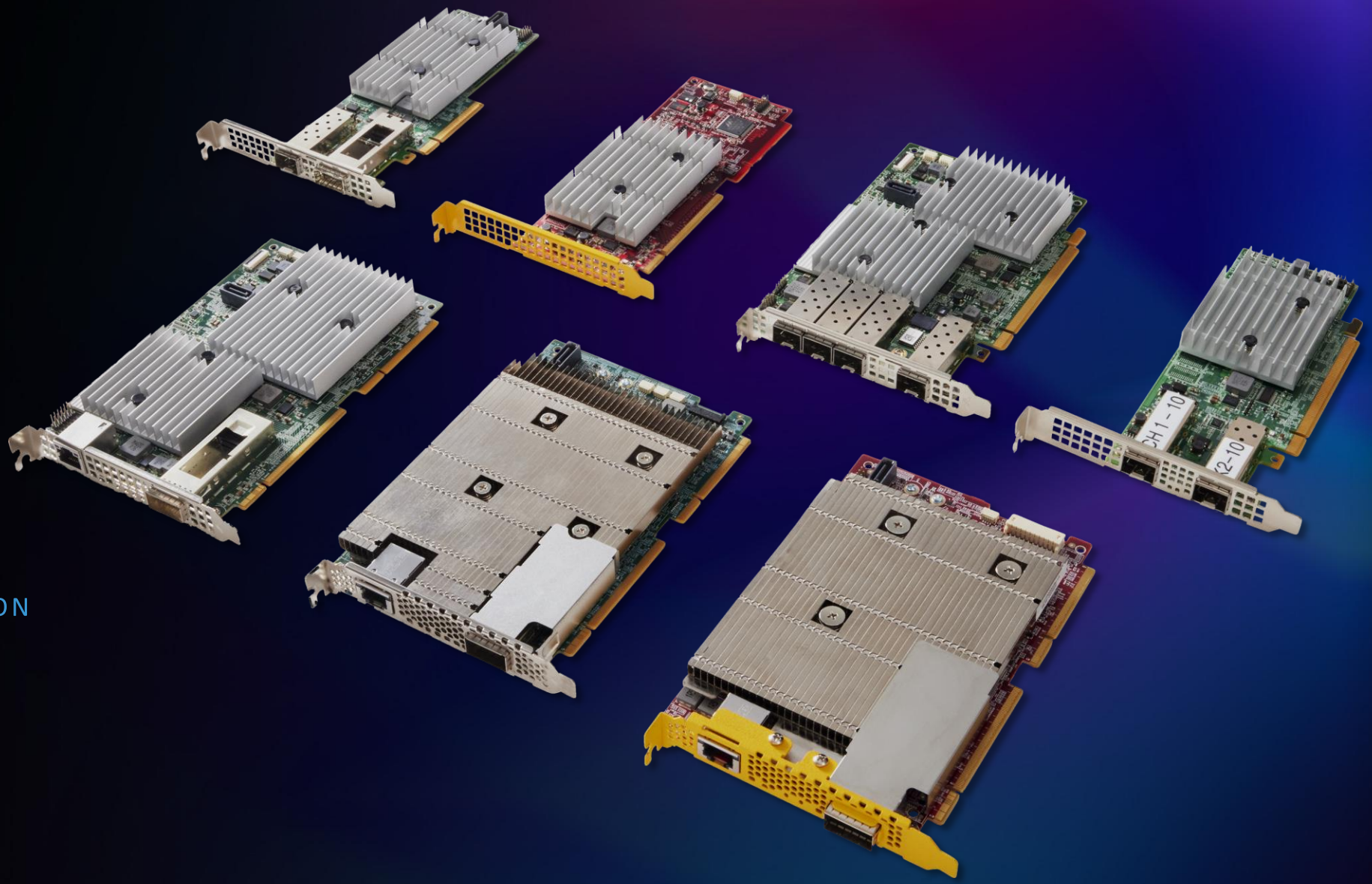
OFFERING THE BEST SECURITY, PERFORMANCE, AND INNOVATION IN THE CLOUD



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

The AWS Nitro System hardware

OFFERING THE BEST SECURITY, PERFORMANCE, AND INNOVATION IN THE CLOUD



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

20 million

Nitro Chips shipped

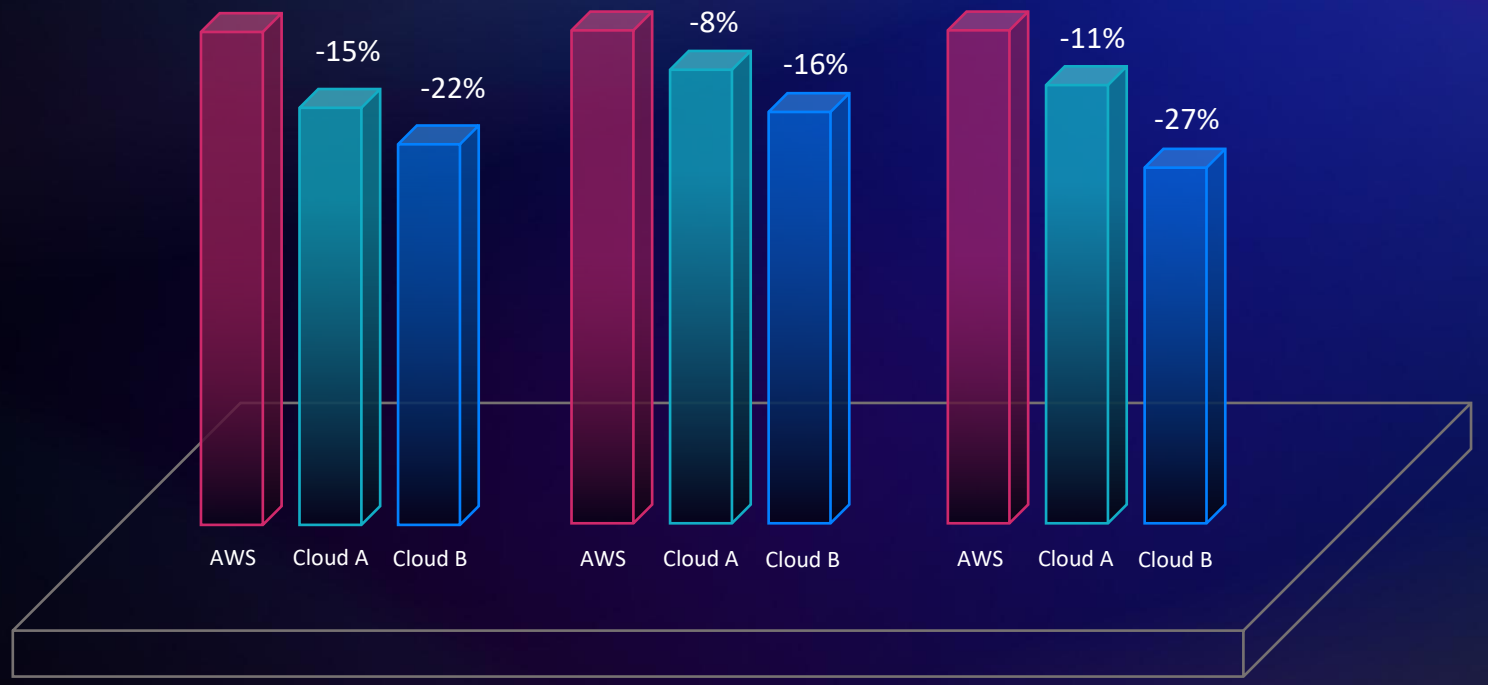
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Nitro performance for real-world workloads

AMAZON EC2 INSTANCES CAN DELIVER OVER 15% HIGHER THROUGHPUT PERFORMANCE



SPECINT2017



MEMCACHED

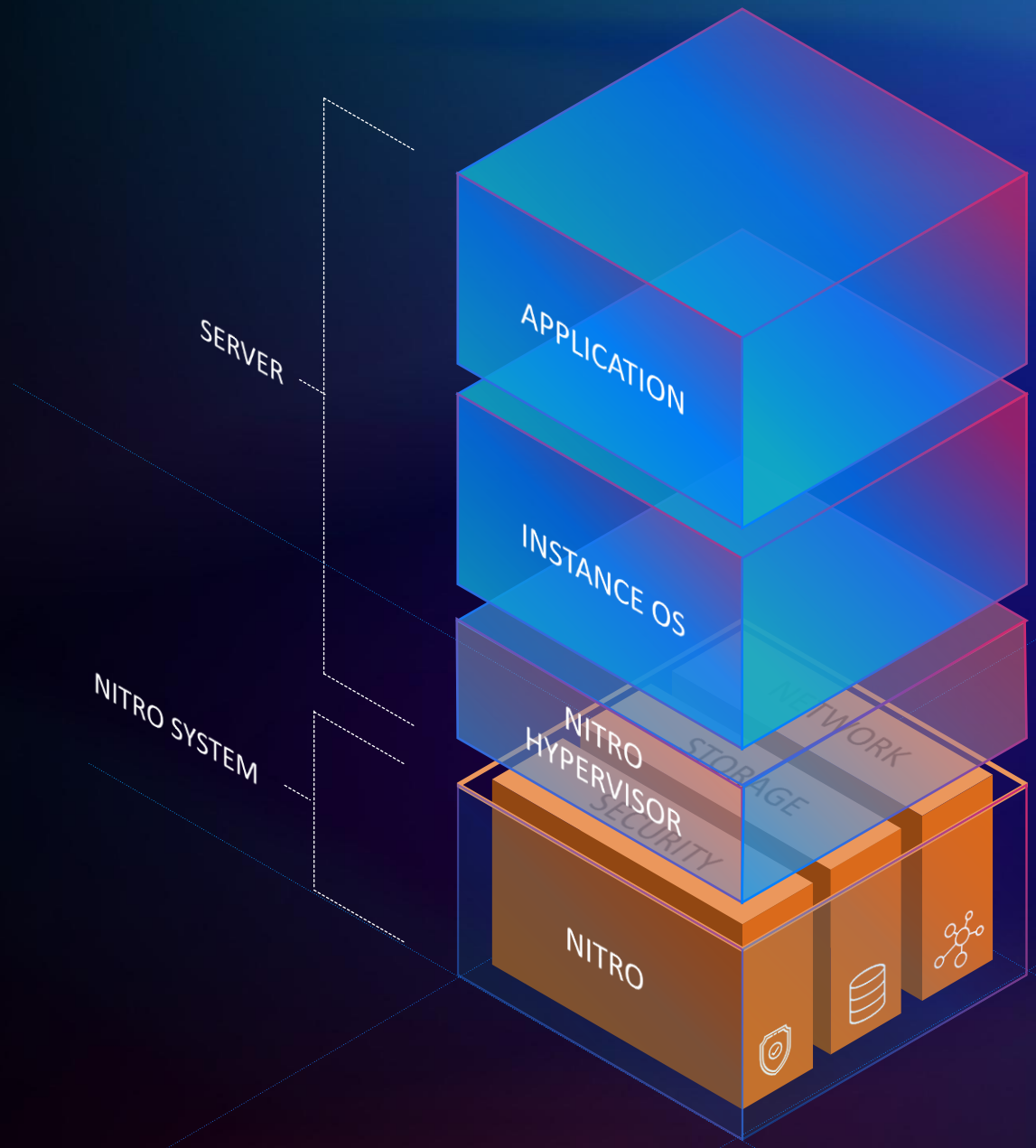
NGINX WEBSERVING

REDIS

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

The AWS Nitro System networking

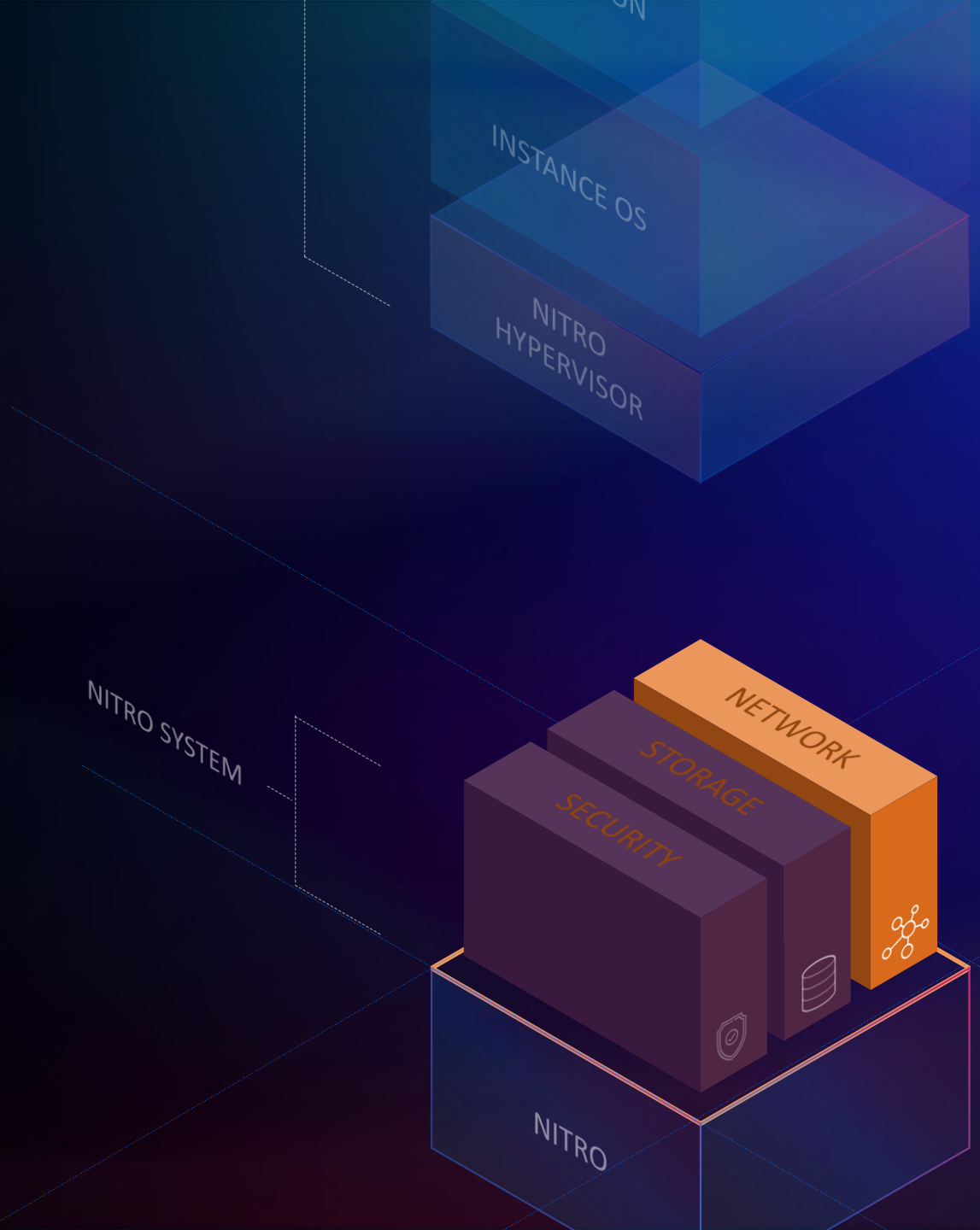
OFFERING THE BEST SECURITY, PERFORMANCE, AND INNOVATION IN THE CLOUD



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

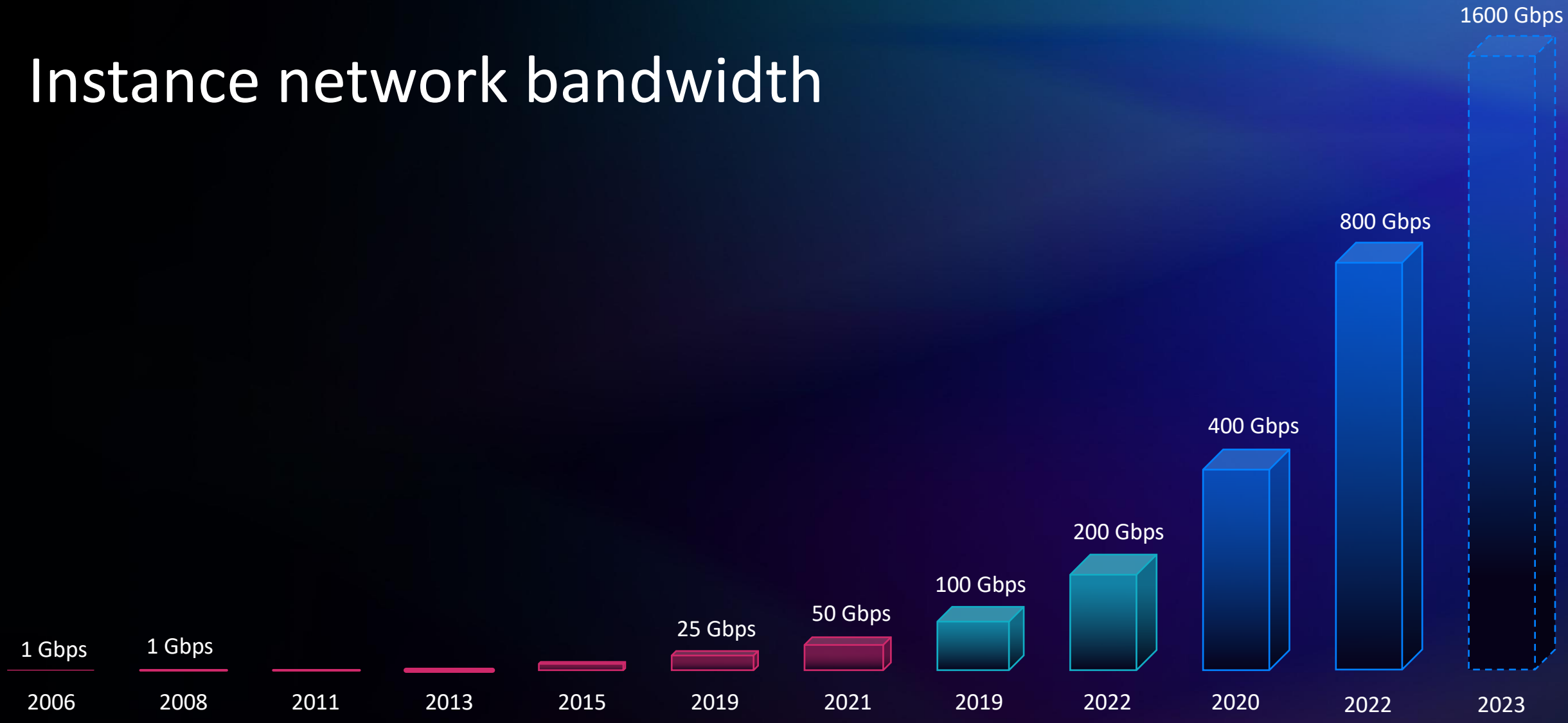
The AWS Nitro System networking

OFFERING THE BEST SECURITY, PERFORMANCE, AND INNOVATION IN THE CLOUD



© 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!

AVAILABLE NOW

6th Gen EC2 Network-optimized instances

BEST PERFORMANCE IN THE CLOUD FOR HIGH-NETWORKING AND EBS WORKLOADS



aws

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

NEW!

AVAILABLE NOW

6th Gen EC2 Network-optimized instances

BEST PERFORMANCE IN THE CLOUD FOR HIGH-NETWORKING AND EBS WORKLOADS



UP TO 200 GBPS

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



UP TO 2X PPS

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



EBS OPTIMIZED

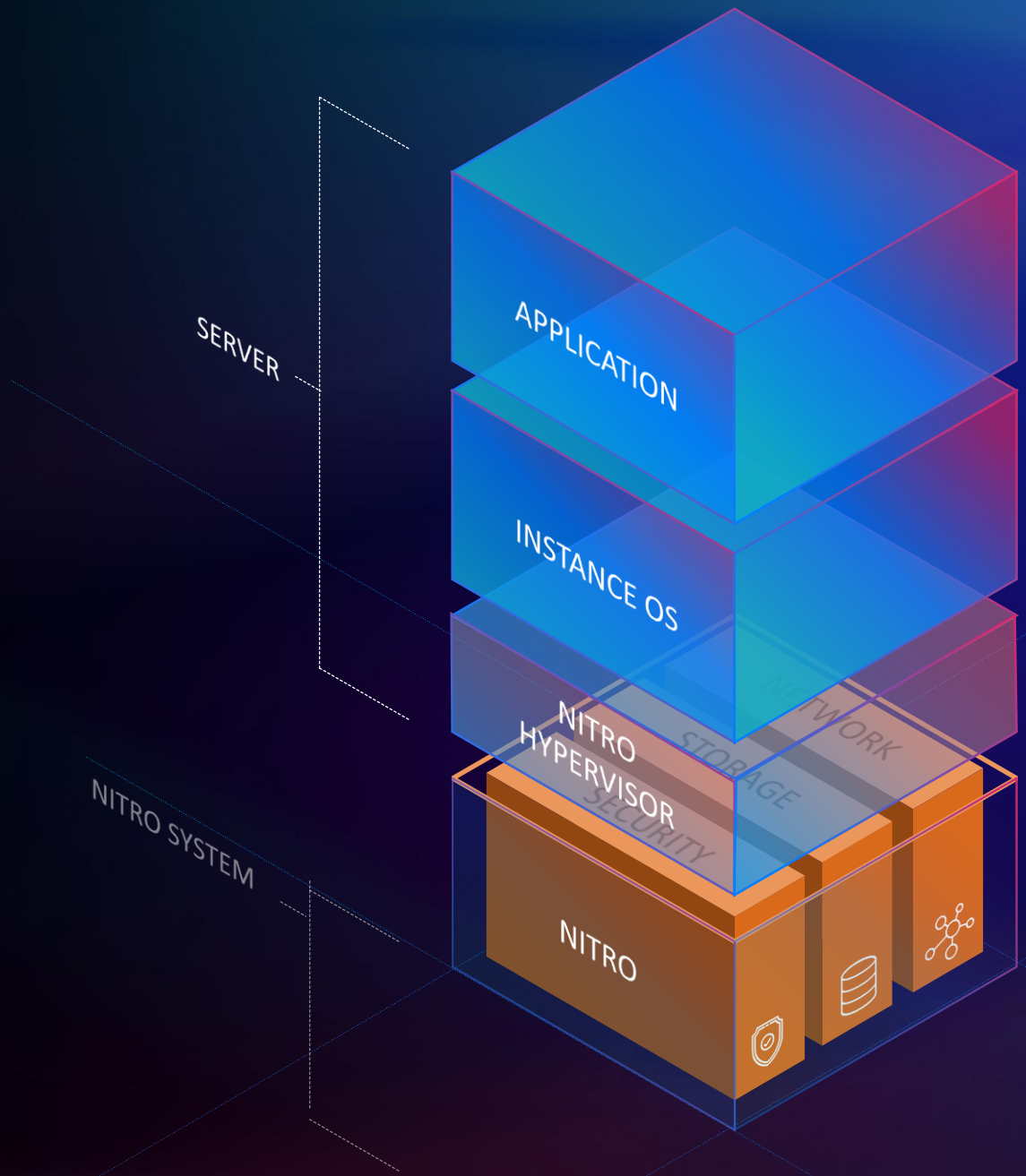
Supports up to 80 Gbps of Amazon EBS bandwidth to enable high performance file systems, databases, and big data analytics

aws

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

The AWS Nitro System storage

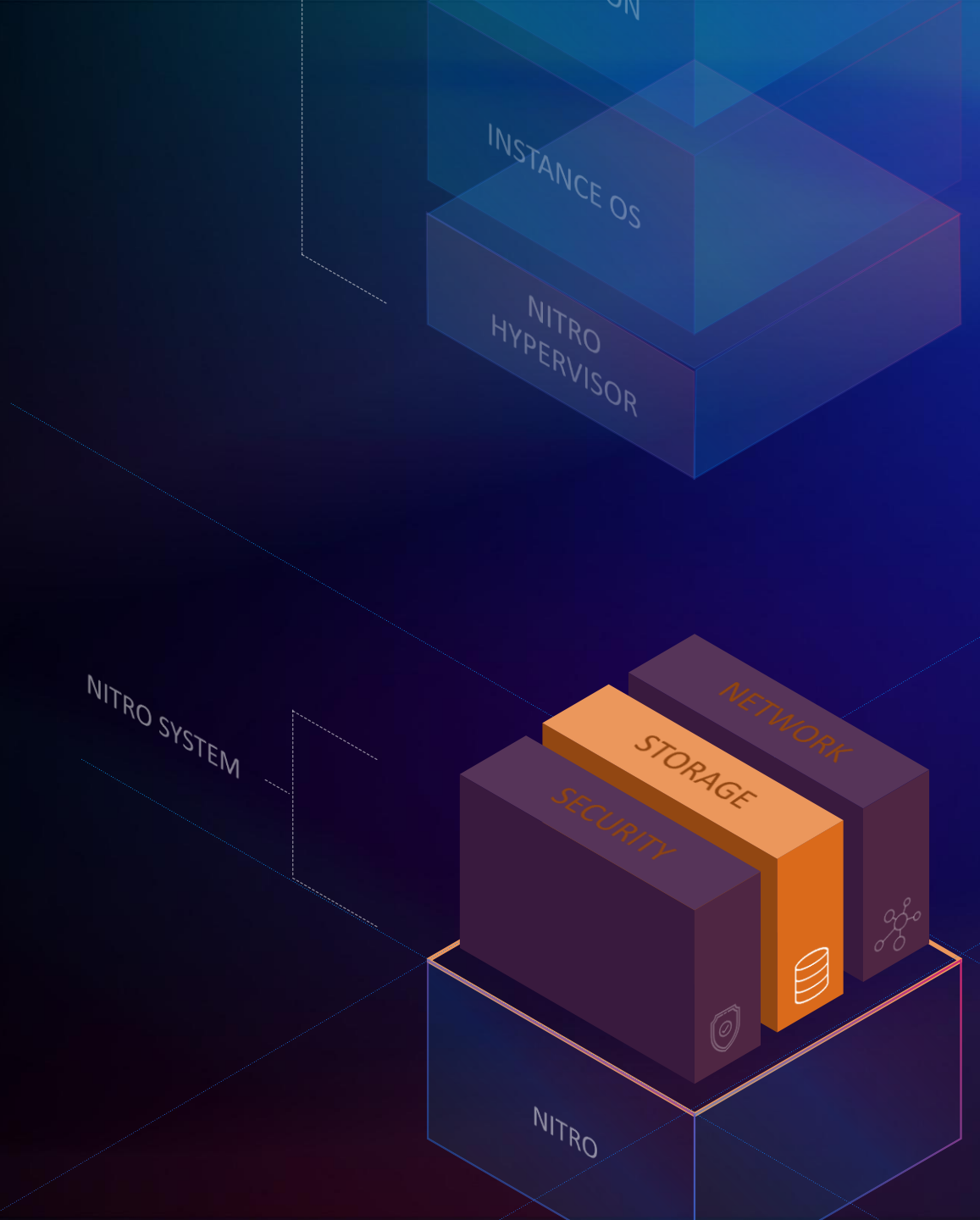
OFFERING THE BEST SECURITY, PERFORMANCE, AND INNOVATION IN THE CLOUD



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

The AWS Nitro System storage

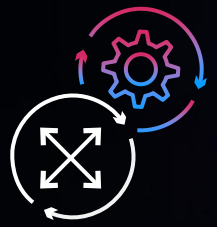
OFFERING THE BEST SECURITY, PERFORMANCE, AND INNOVATION IN THE CLOUD



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

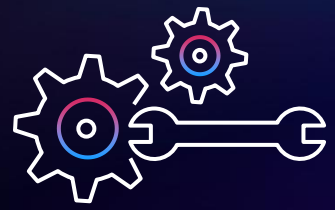
AWS Nitro SSD

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

aws

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

NEW!

AVAILABLE NOW

Announcing Torn Write Prevention

IMPROVE DATABASE PERFORMANCE WITH AMAZON EC2 I4I INSTANCES AND AMAZON EBS



aws

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

NEW!

AVAILABLE NOW

Announcing Torn Write Prevention

IMPROVE DATABASE PERFORMANCE WITH AMAZON EC2 I4I INSTANCES AND AMAZON EBS

AWS NITRO SSDs

Support for Amazon EC2 I4i, Im4gn, and Is4gen storage optimized instances and EBS volumes for all Nitro-based instances



DATABASE PERFORMANCE

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



REDUCE COSTS

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

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

“Aerospike running on EC2 I3en instances already delivers industry leading real-time performance for read/write workloads. With EC2 I4i instances, we have observed a 70% increase in performance on read workloads alone.”

Srini Srinivasan
CTO and Founder, Aerospike

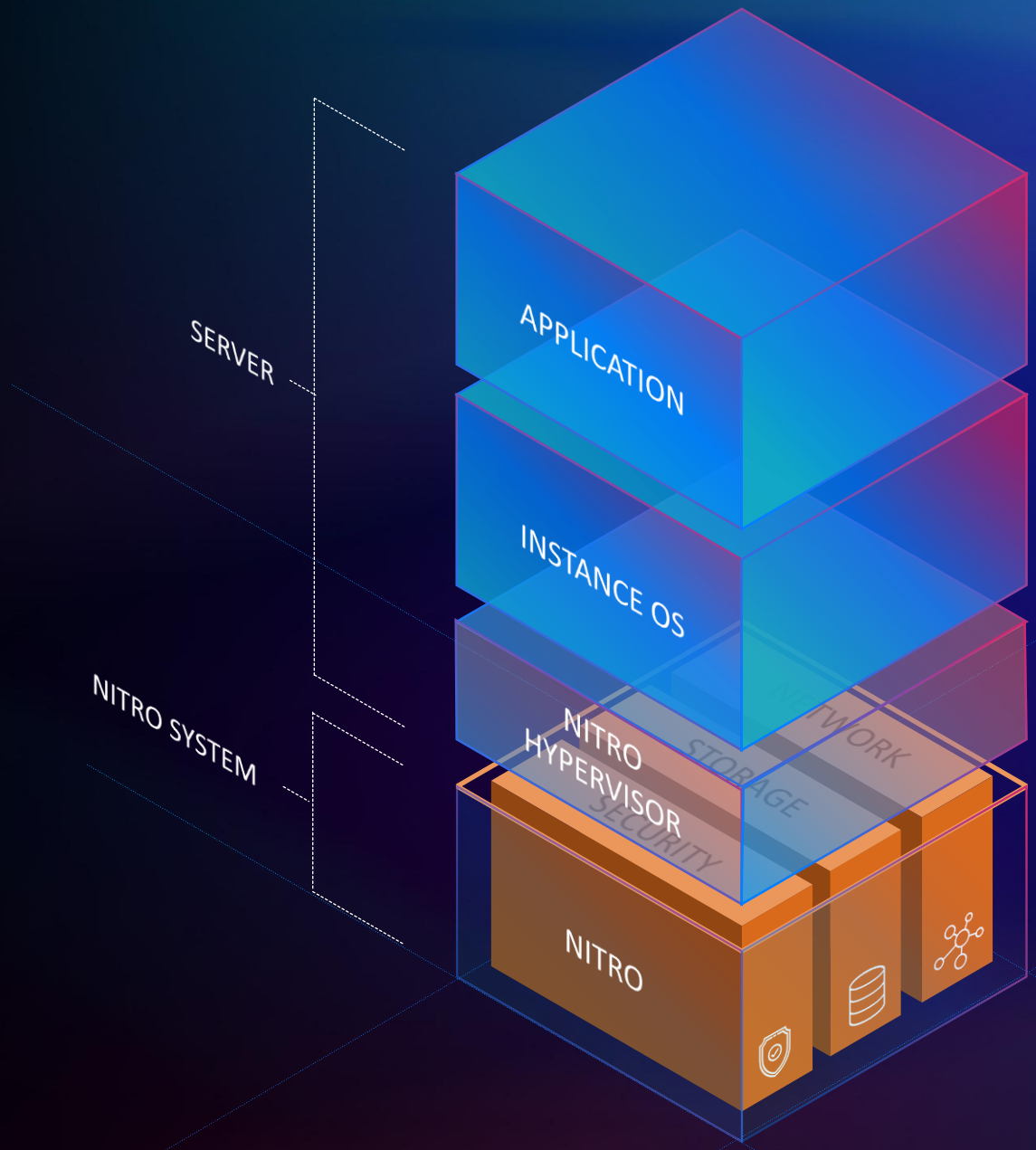


AEROSPIKE

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

The AWS Nitro System security

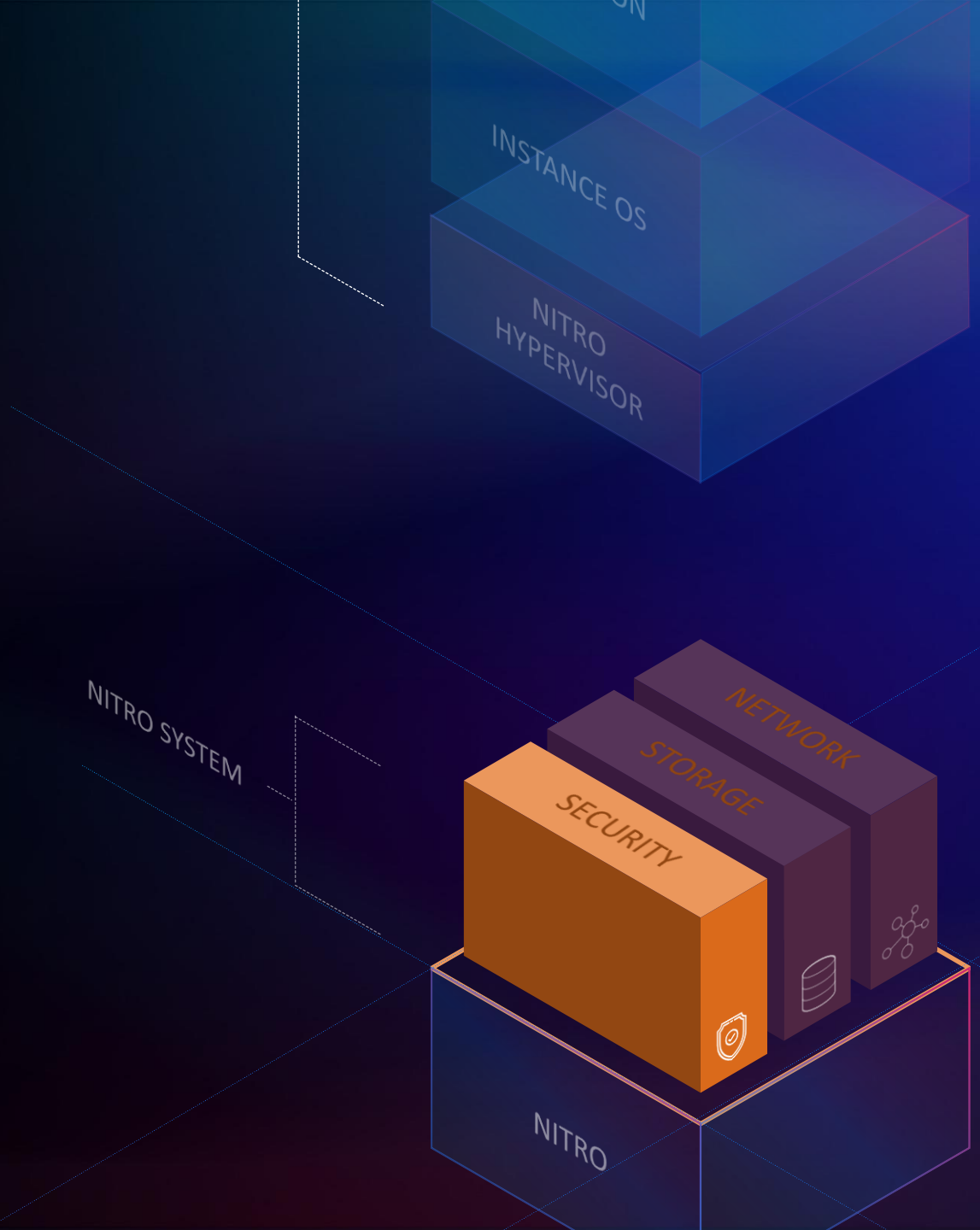
OFFERING THE BEST SECURITY, PERFORMANCE, AND INNOVATION IN THE CLOUD



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

The AWS Nitro System security

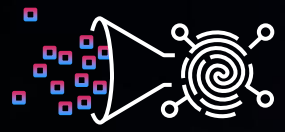
OFFERING THE BEST SECURITY, PERFORMANCE, AND INNOVATION IN THE CLOUD



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

Improved security with AWS Nitro

SECURITY IS ALWAYS OUR NUMBER ONE PRIORITY!



ENCRYPTION

All communication channels within the Nitro System are encrypted



SECURE BOOT

All hardware and software components are cryptographically validated on each boot



PATCHING

All Nitro components, including hypervisor, can be updated without any downtime



NO REMOTE ACCESS

No shell available on the Nitro System or underlying server

aws

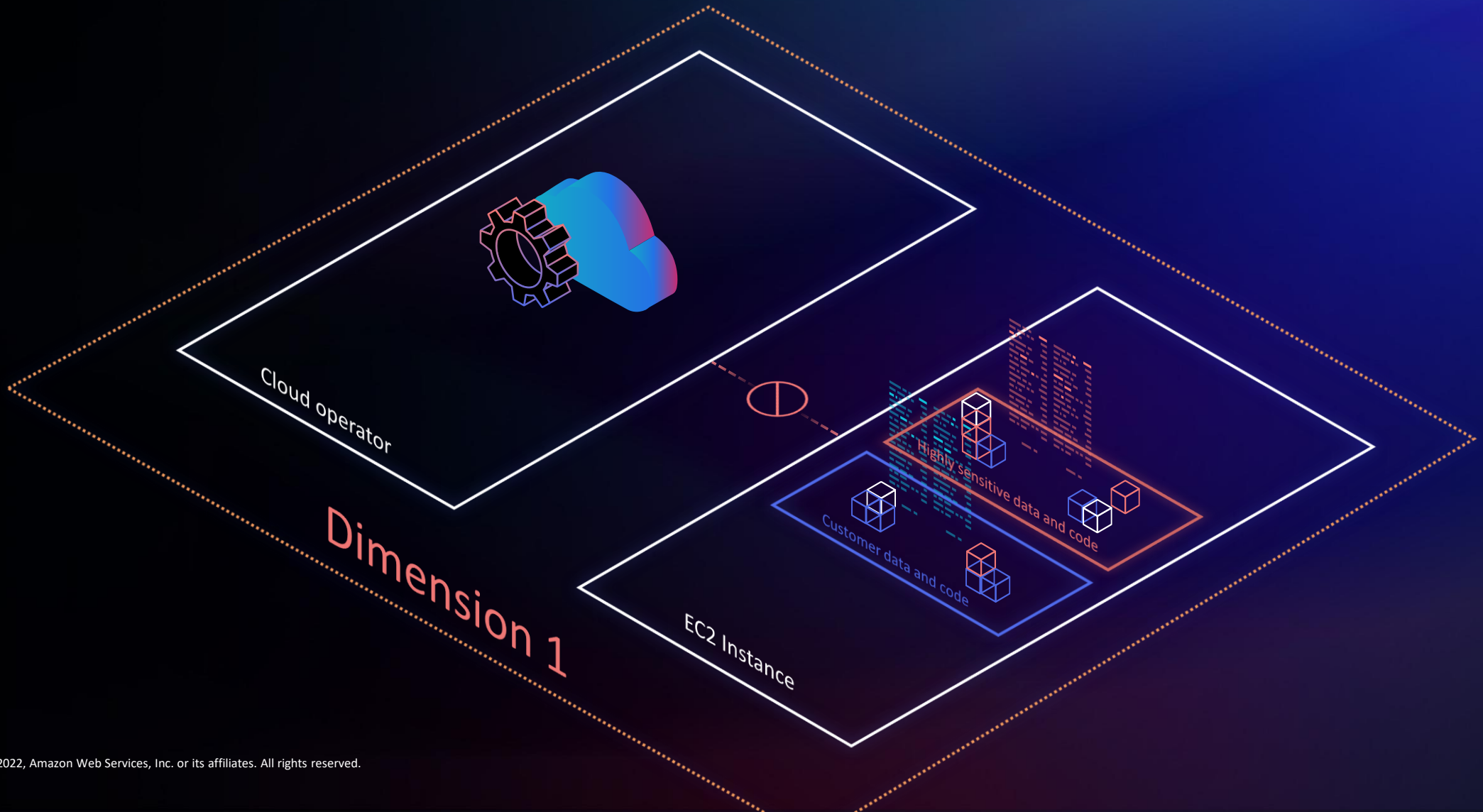
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Confidential Computing ensures that AWS protects customer code and data from any unauthorized access

What is confidential computing?

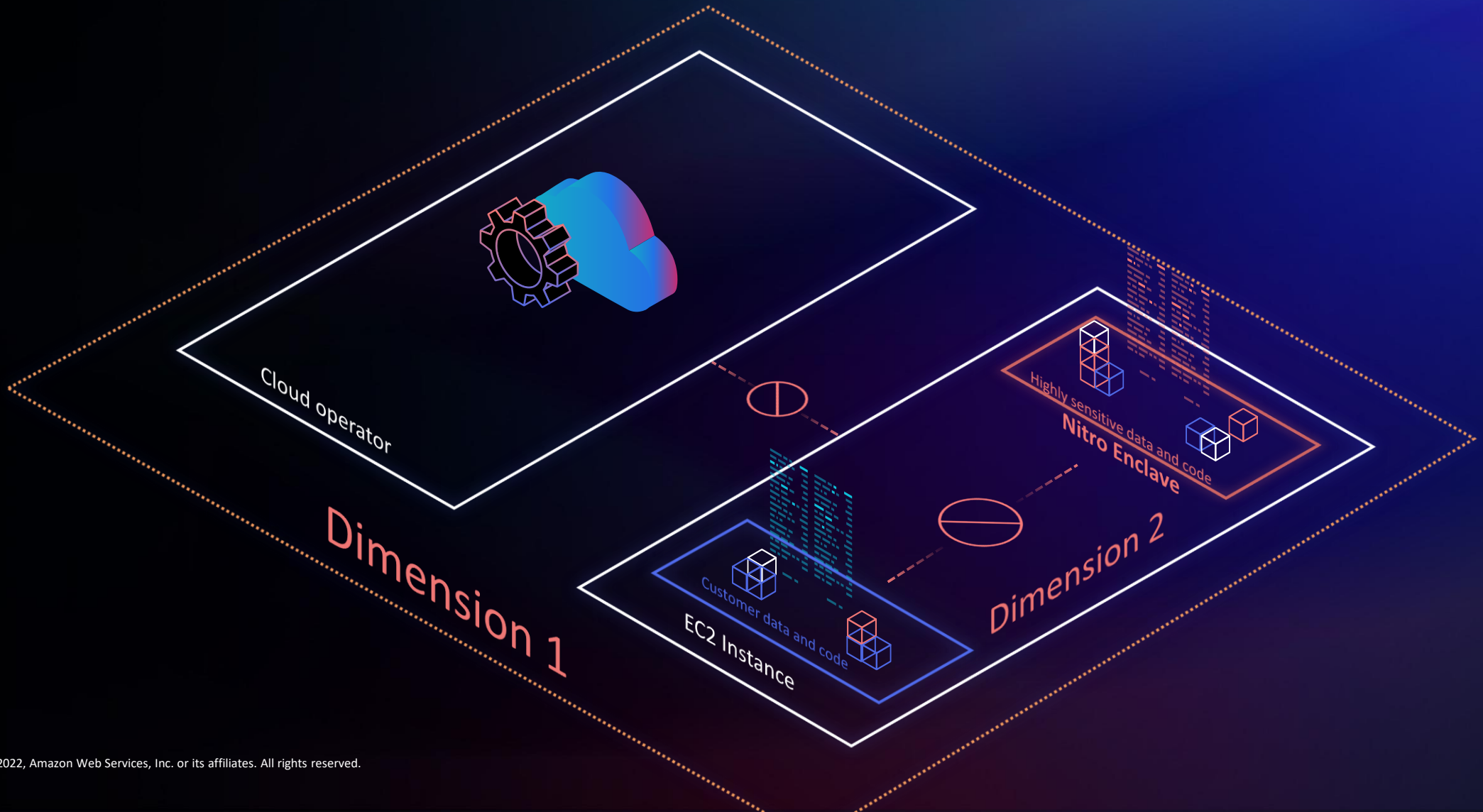
PROTECTING CUSTOMER CODE AND SENSITIVE DATA IN USE



© 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.

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 channels



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

aws

© 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

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

“Using Nitro Enclaves as a solution has helped ensure that customer’s privacy – and advertiser’s valuable first-party data – is protected and secured for more data driven media buying.”

Samantha Jacobson
Chief Strategy Officer, The Trade Desk



theTradeDesk®

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

COMPUTE INNOVATION FOR EVERY WORKLOAD



World class scale and performance



Operational excellence



Compute for every workload



Cost optimization and best practices



Compute where you need it

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





Jeremy Conley (he/him)

Principal Technical Program Manager, AWS

Instance launch times



LAUNCH TIMES

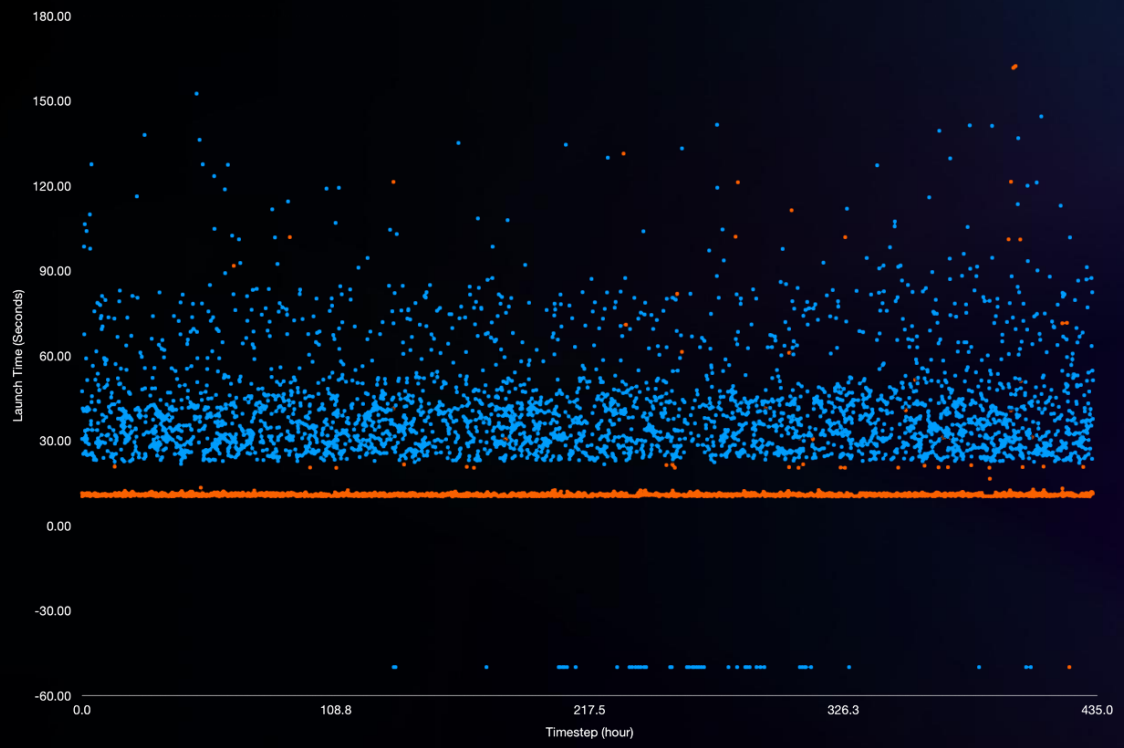


AVAILABILITY



RELIABILITY

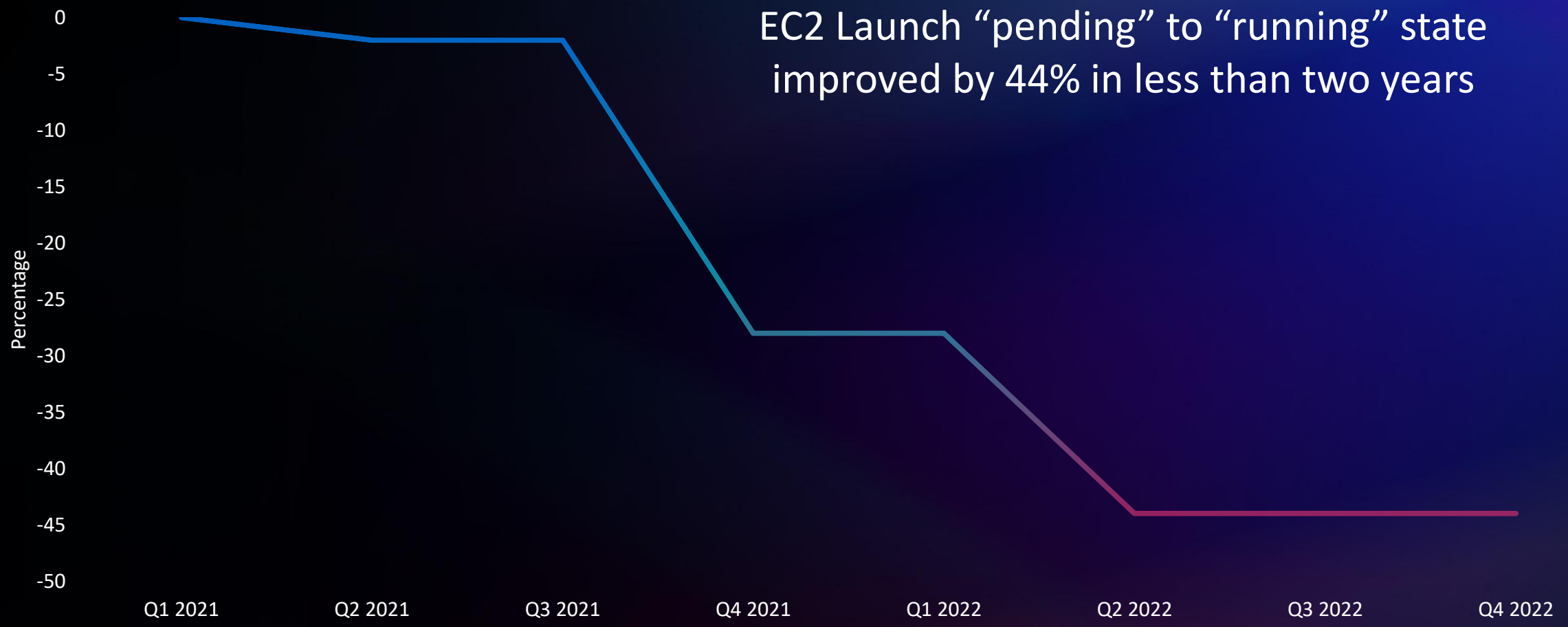
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.





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

Amazon EC2 Launch times



EC2 Launch “pending” to “running” state improved by 44% in less than two years

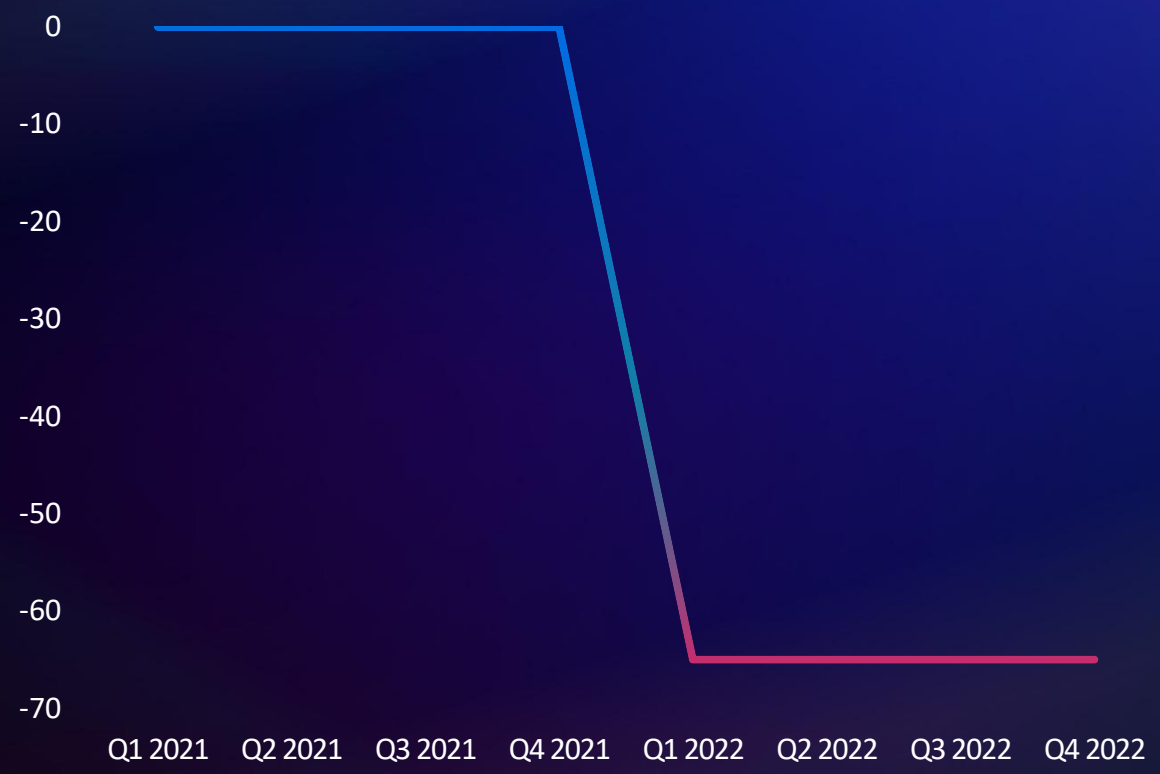
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon EC2 Launch times

Linux

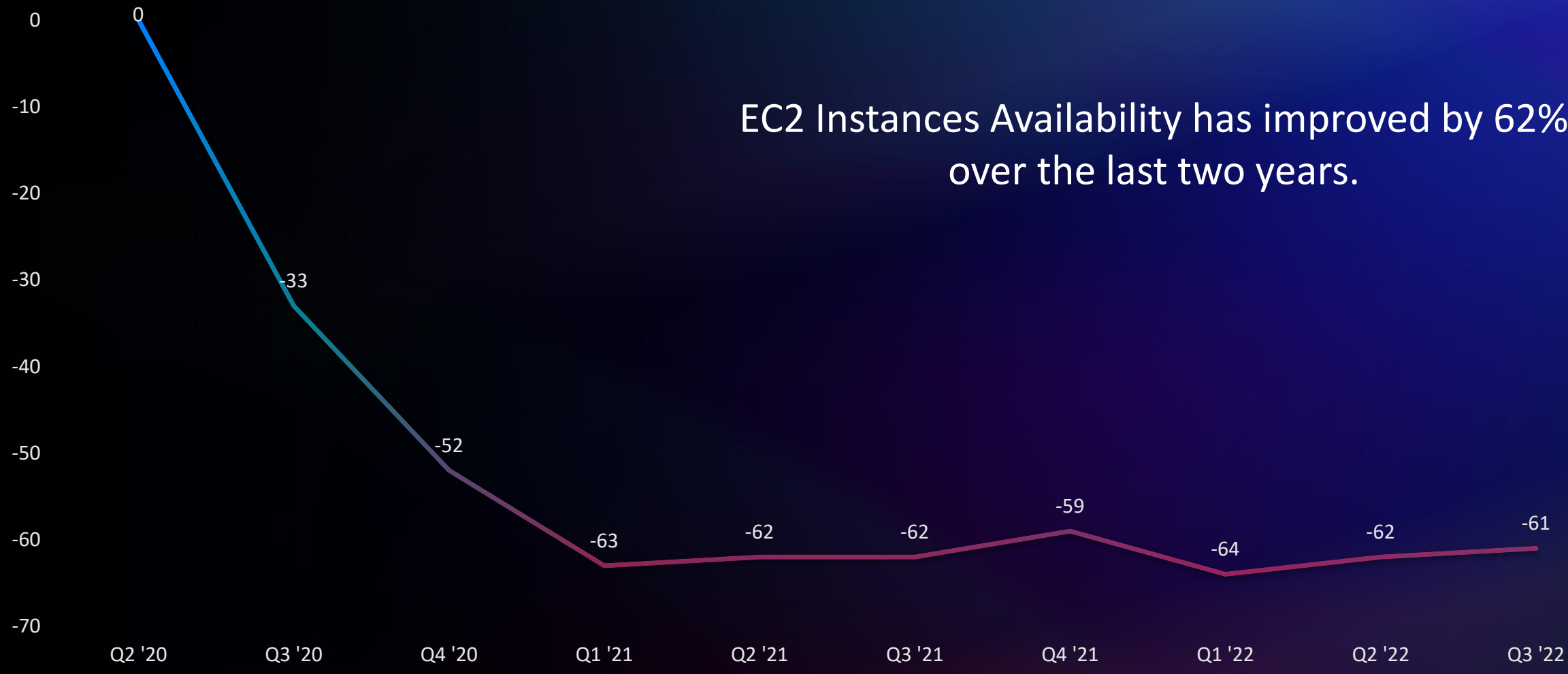


Windows



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

Amazon EC2 Instance Availability



EC2 Instances Availability has improved by 62% over the last two years.

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

Over

1 million

live migrations per week

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

COMPUTE INNOVATION FOR EVERY WORKLOAD



World class scale and performance



Operational excellence



Compute for every workload



Cost optimization and best practices

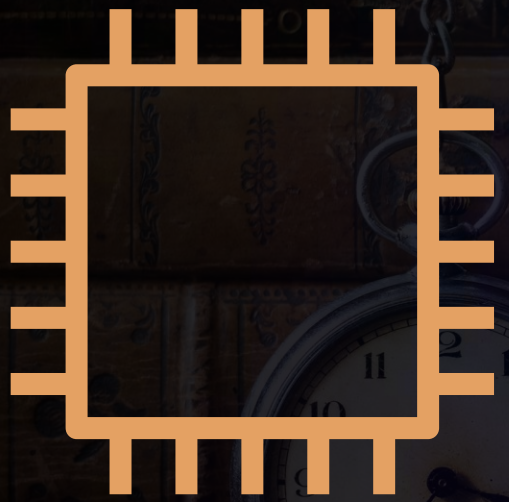


Compute where you need it

aws

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





1.7 GHz Xeon Processor

1.75 GB memory

160 GB HDD

250 Mbps bandwidth

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, FPGA)

Networking
(up to 1,600 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.

Broadest 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.

Broadest 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 the AWS portfolio



INTEGRATION

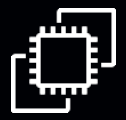
Over 350 Amazon EC2 instances are powered by Intel processors



FASTEST

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

RECENT INTEL-BASED INSTANCES



I4I
STORAGE-OPTIMIZED



M6I(D)N
NETWORK-OPTIMIZED



C6IN
NETWORK-OPTIMIZED



R6I(D)N
NETWORK-OPTIMIZED



M6ID
GENERAL PURPOSE



C6ID
COMPUTE-OPTIMIZED



R6ID
MEMORY-OPTIMIZED

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

AiSCOUT has modernized the expensive, time-consuming, and unsustainable process of scouting and identifying new talent

They have partnered with Intel to embed their 3DAT Cloud SDK running on Amazon EC2 DL1, C6i, and M6i instances



AiSCOUT

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

Broadest 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

Help you optimize both cost and performance for your workloads



BETTER ECONOMICS

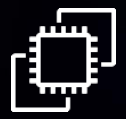
Deliver up to 10% lower cost versus comparable instances



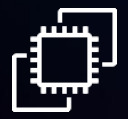
SEAMLESS MIGRATION

Easily migrate applications to the new AMD-based variants with little to no modification

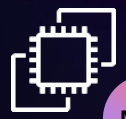
AMD-BASED INSTANCES



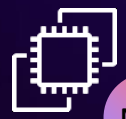
G4AD
GPU-OPTIMIZED



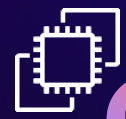
HPC6A
HPC-OPTIMIZED



M6A
GENERAL PURPOSE



C6A
COMPUTE-OPTIMIZED



R6A
MEMORY-OPTIMIZED

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

“We are always looking for ways to improve our cloud utilization and after using AWS’s EC2 M5a, R5a, and C5a, which are all based on AMD’s EPYC processors, we’ve secured substantial savings on our cloud costs.”

Mike Donnell
Senior DevOps Engineer, Sprinklr



sprinklr

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

Broadest choice of processors and architectures

Right compute for the workload



Intel Xeon
scalable processors



AMD EPYC
processors



AWS Graviton processors

aws

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

Innovating with AWS Graviton2

40% BETTER PRICE PERFORMANCE FOR A BROAD RANGE OF WORKLOADS



BEST PRICE PERFORMANCE

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



EXTENSIVE ECOSYSTEM

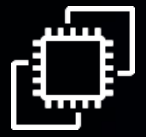
Supported by popular Linux OSes along with popular application and services from AWS and ISVs



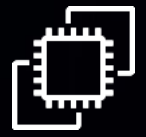
ENHANCED SECURITY

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

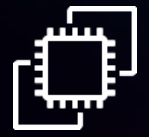
GRAVITON2-BASED INSTANCES



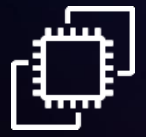
C6G(D)
COMPUTE-OPTIMIZED



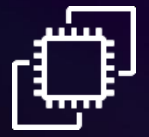
M6G(D)
GENERAL PURPOSE



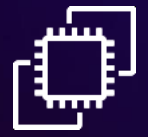
R6G(D)
MEMORY-OPTIMIZED



C6GN
NETWORK-OPTIMIZED



X2GD
ENHANCED MEMORY



T4G
BURSTABLE



IM4GN
STORAGE-OPTIMIZED



IS4GEN
STORAGE-OPTIMIZED

aws

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

Wide ecosystem and smooth adoption

OPERATING SYSTEMS

Amazon Linux 2

Ubuntu 16.04, 18.04 and newer

RedHat Enterprise Linux 7.6 and 8.0

SUSE Linux Enterprise Server for Arm 15

Docker Desktop Community and Docker Enterprise Engine

CONTAINERS

Most Docker official images support Graviton

Amazon ECS
AVAILABLE TODAY

Amazon EKS
PUBLIC PREVIEW

INTEGRATED SERVICES

AWS Marketplace

AWS Systems Manager

Amazon CloudWatch

AWS CodeCommit

AWS Cloud9

AWS CodePipeline

Amazon Inspector

AWS Batch

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


Jeff Barr (@) @jeffbarr
It's time to join the Graviton Challenge. Same exact workload, same or better performance, and up to 40% better price performance!
3
6
36
Oliver perks @OllyPerksHPC
I have been really impressed with the AWS Graviton2 chip for HPC workloads, this blog highlights some of the out-of-the-box experiences we have been having scaling OpenFOAM out to 2k cores. The price performance is even more impressive.
Arthur Petitpierre @ArthurPtP
Eager to try a #CFD application on #AWS #C6g #EC2 instances ? Have a look to @Dr_NeilA latest blog post about running @OpenFOAM on #Arm based #Graviton2 instances: aws.amazon.com/blogs/compute/...
Show this thread
24



Migration is now simple and fast, in some cases we have seen customers move in as little as 4 days!

“With the superior performance and lower per-instance cost of Graviton2-based instances compared to x86-based instances, we are estimating 25% lower operating costs for these microservices.”

Jonathan Tronson
VP Software Engineering, DIRECTV STREAM



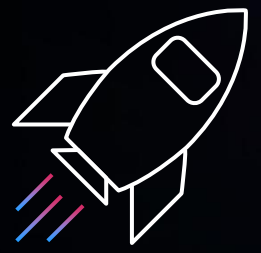
DIRECTV stream

aws

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

Innovating with AWS Graviton3

PROVIDES A 25% PERFORMANCE IMPROVEMENT OVER GRAVITON2



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

aws

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

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

Pat Symonds
CTO at Formula 1 Management



TM

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



Graviton3 uses up to 60% less energy to compute the same workload as comparable x86 processors

NEW!

PREVIEW

Amazon EC2 C7gn instances

HIGHEST NETWORK BANDWIDTH FOR EC2 NETWORK-OPTIMIZED INSTANCES



aws

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

NEW!

PREVIEW

Amazon EC2 C7gn instances

HIGHEST NETWORK BANDWIDTH FOR EC2 NETWORK-OPTIMIZED INSTANCES



HIGHEST NETWORK PERFORMANCE

Scale data intensive workloads workloads with 200 Gbps of bandwidth and 2x higher throughput per vCPU compared to current C6gn instances



HIGHEST PPS PERFORMANCE

The new Nitro v5 card enables the highest packet processing performance across network-optimized instances to scale network virtual appliances



LATEST GRAVITON3 PROCESSORS

Up to 25% better compute performance and up to 2x faster performance for cryptographic workloads compared to Graviton2 instances

aws

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



Electronic Design Automation (EDA): software tools for designing electronic systems such as integrated circuits and printed circuit boards

“Performing EDA workloads in the cloud will transform the way semiconductors are developed. By utilizing AWS’s EDA in the cloud services, Marvell will be able to optimize our chip development and accelerate our time to market.”

Raghib Hussain
President of Products and Technologies at Marvell



MARVELL™

aws

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

NEW!

PREVIEW

Amazon EC2 R7iz instances

THE HIGHEST COMPUTE PERFORMANCE IN EC2 FOR RUNNING X86-BASED WORKLOADS



aws

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

NEW!

PREVIEW

Amazon EC2 R7iz instances

THE HIGHEST COMPUTE PERFORMANCE IN EC2 FOR RUNNING X86-BASED WORKLOADS



COMPUTE PERFORMANCE

- Up to 15% better compute performance vs. previous generation instances
- First EC2 instance with Intel Sapphire Rapids supporting an all core turbo frequency up to 3.9 GHz



WORKLOAD PERFORMANCE

- Up to 40% workload performance improvements for databases, video/audio processing and load balancing



SCALE UP YOUR WORKLOADS

- More than 2.5x compute and memory resources versus previous generation to scale your EDA, gaming, and databases workloads

aws

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


E
SE
S
NE
SW
N
NW
W
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
320
340

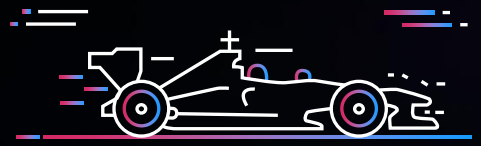


Sometimes supporting a workload in the cloud is a multi-year journey

High performance computing

HIGHEST NETWORK BANDWIDTH FOR EC2 NETWORK-OPTIMIZED INSTANCES

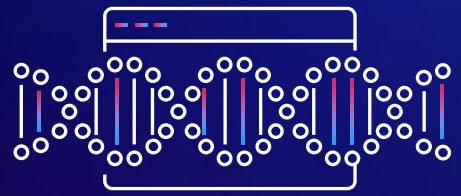
COMPUTATION FLUID DYNAMICS



LIFE SCIENCES



GENOMICS



RISK ANALYSIS



ENERGY



RESEARCH



aws

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

Amazon EC2 Hpc6a instances

THE BEST PRICE PERFORMANCE FOR COMPUTE-INTENSIVE HPC WORKLOADS IN AMAZON EC2



100G NETWORKING WITH EFA

With 100 Gbps Elastic Fabric Adapter (EFA) networking, you have access to a high-throughput, low latency, low jitter channel for tightly coupled workloads



PRICE PERFORMANCE BENEFITS

Hpc6a instances are purpose-built to cost effectively run HPC workloads on AWS, up to 70% less than similar compute-optimized, x86 based instances



DEPLOY WITH AWS PARALLELCLUSTER

Deploy and manage your Hpc6a instances with the open source cluster management tool by AWS

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.

NEW!

AVAILABLE NOW

Amazon EC2 Hpc6id instances

BEST PRICE PERFORMANCE FOR MEMORY AND DATA-INTENSIVE HPC WORKLOADS IN AMAZON EC2



aws

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

NEW!

AVAILABLE NOW

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 DATA INTENSIVE 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!

COMING SOON

Amazon EC2 Hpc7g instances

BEST PRICE PERFORMANCE FOR COMPUTE INTENSIVE HPC WORKLOADS IN AMAZON EC2



aws

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

NEW!

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 instances

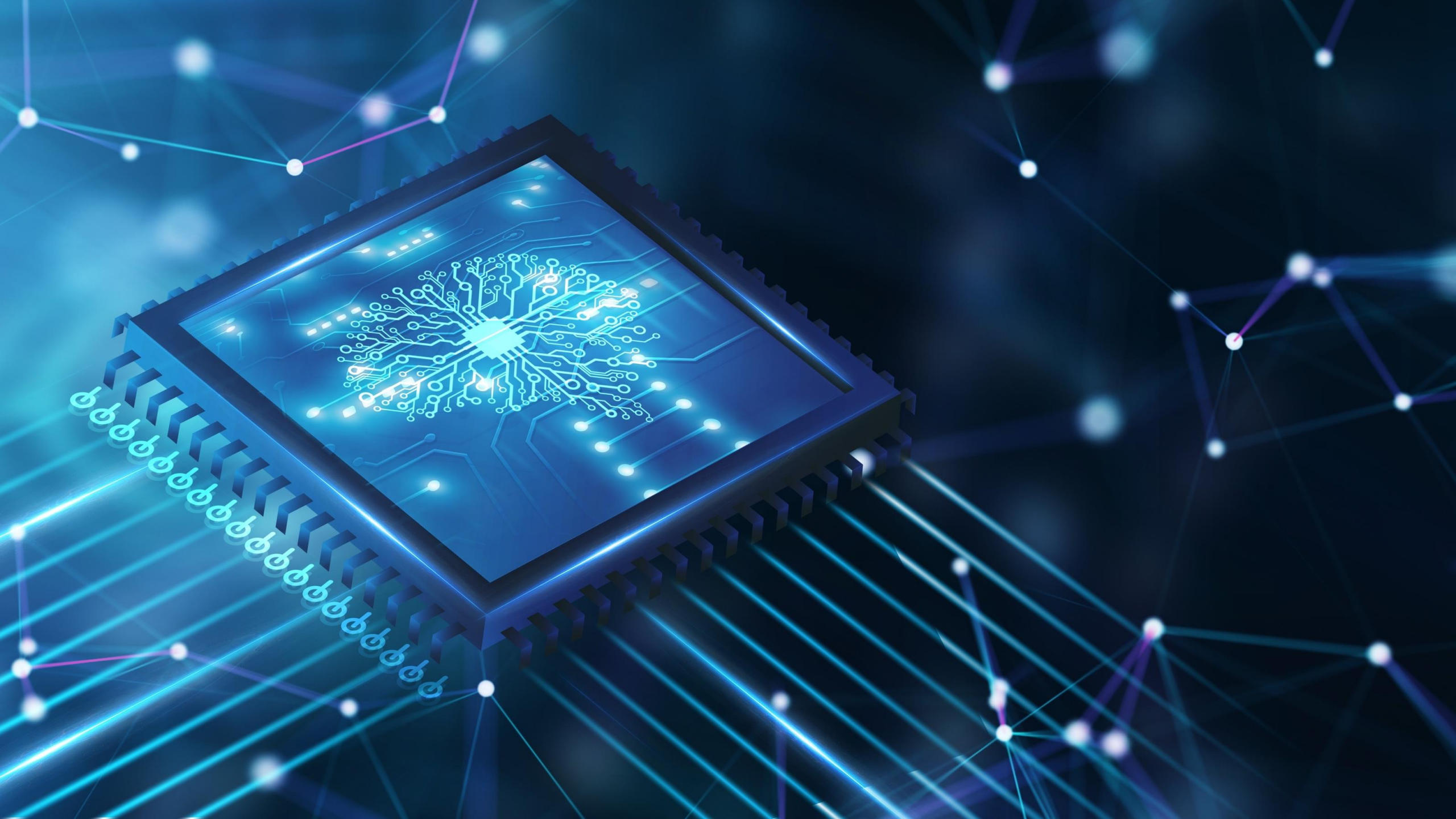


DEPLOY WITH AWS PARALLELCLUSTER

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

aws

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



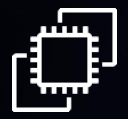
More and more companies are deploying machine learning to improve their customer experience

Supporting ML workloads

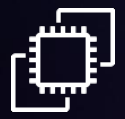
LARGEST SELECTION OF ML INSTANCES IN THE CLOUD

MACHINE LEARNING

TRAINING



P3DN
NVIDIA V100
TENSOR CORE

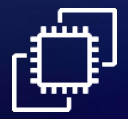


P4D
NVIDIA A100
TENSOR CORE

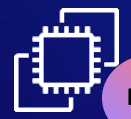


NEW

P4DE
NVIDIA A100
TENSOR CORE



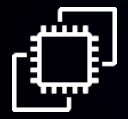
DL1
HABANA GAUDI
INTEL



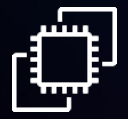
NEW

TRN1
AWS TRAINIUM

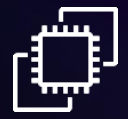
INFERENCE



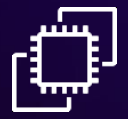
G4DN
NVIDIA T4 GPU



G5
NVIDIA A10G
TENSOR CORE



G5G
NVIDIA T4G
TENSOR CORE



INF1
AWS INFERENTIA

PLATFORM

TENSORFLOW

MXNET

PYTORCH

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

NEW!

PREVIEW

Amazon EC2 P4de instances

HIGHEST PERFORMANCE FOR ML TRAINING AND HPC APPLICATIONS IN THE CLOUD

NVIDIA

POWERED BY A100 TENSOR CORE GPUS

Each A100 GPU offers over 2.5x the compute performance compared to the V100 GPU and comes with 80 GB HBM2e of high-performance GPU memory



400G NETWORKING WITH EFA

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



DEPLOY 4,000+ GPUS IN AN ULTRACLUSTER

EC2 UltraClusters of P4d instances combine high performance compute, networking, and storage into one of the most powerful supercomputers in the world

aws

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


OPTIMIZATION

Given the growth in machine learning, there are few places where an improvement in price performance can have a larger impact

NEW!

AVAILABLE NOW

Amazon EC2 Trn1 instances

HIGHEST PERFORMANCE AND LOWEST COST FOR TRAINING DEEP LEARNING MODELS



POWERED BY AWS TRAINIUM

16 Trainium accelerators with 512 GiB of HBM2 high speed memory, 800 Gbps of networking throughput, and 8 TB of local NVMe storage

DEEP LEARNING PERFORMANCE

Lower time-to-train when compared to the latest comparable Amazon EC2 instances



DEPLOY IN AN ULTRACLUSTER

Deployable in Amazon EC2 UltraClusters, a supercomputer in the cloud with tens of thousands of Trainium accelerators with petabit scale networking bandwidth

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

NEW!

AVAILABLE NOW

Amazon EC2 Trn1 instances

HIGHEST PERFORMANCE AND LOWEST COST FOR TRAINING DEEP LEARNING MODELS

UP TO **50%**

COST TO TRAIN SAVINGS OVER

COMPARABLE GPU-BASED INSTANCES

AWS TRAINIUM

16 Trainium accelerators with 512 GiB of HBM2 high speed memory, 800 Gbps of networking throughput, and 8 TB of local NVMe storage

DEEP LEARNING PERFORMANCE

Lower time-to-train when compared to the latest comparable Amazon EC2 instances

DEPLOY IN AN ULTRACLUSTER

Deployable in Amazon EC2 UltraClusters, a supercomputer in the cloud with tens of thousands of Trainium accelerators with petabit scale networking bandwidth

aws

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

“At Helixon, we build next-generation AI solutions to protein-based therapeutics. We are excited to utilize Amazon EC2 Trn1 instances, featuring the highest networking bandwidth (800 Gbps) available in AWS to improve the performance of our distributed training jobs and reduce our model training times, while also reducing our training costs.”

Jian Peng
CEO, Helixon



Helixon

aws

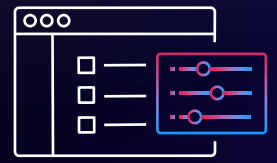
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon EC2 Inf1 instances

HIGH-PERFORMANCE, LOWEST-COST MACHINE LEARNING INFERENCE

POWERED BY AWS INFERENTIA

Featuring AWS Inferentia, the first ML chip designed by Amazon with up to 70% lower cost than comparable GPU instances



FRAMEWORK SUPPORT

Seamless software integration with ML frameworks like TensorFlow, PyTorch, and MXNet for quickly getting started and
with minimal code changes



INTEGRATED WITH AWS

Available through VMs, containers, Kubernetes, and Amazon SageMaker

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

“We launched a large-scale AI chatbot service on the Amazon EC2 Inf1 instances and reduced our inference latency by 97% over comparable GPU-based instances while also reducing costs. We expect Trn1 instances will provide additional value in improving end-to-end ML performance and cost.”

Takuya Nakade
CTO, Money Forward, Inc



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

NEW!

PREVIEW

Amazon EC2 Inf2 instances

BEST PERFORMANCE AT THE LOWEST COST IN THE CLOUD FOR DEEP LEARNING INFERENCE



aws

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

NEW!

PREVIEW

Amazon EC2 Inf2 instances

BEST PERFORMANCE AT THE LOWEST COST IN THE CLOUD FOR DEEP LEARNING INFERENCE



POWERED BY AWS INFERENTIA2

Up to 12 Inferentia2 accelerators and up to 384 GB of HBM2e high-speed accelerator memory



DEEP LEARNING PERFORMANCE

Up to 4x higher throughput and up to 10x lower latency than Inf1 instances



INFERENCE OPTIMIZED

Deploy 100B+ parameter models at scale

First inference platform with direct, 192 GB/s connectivity between multiple accelerators for distributed inference

aws

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

NEW!

PREVIEW

Amazon EC2 Inf2 instances

BEST PERFORMANCE AT THE LOWEST COST IN THE CLOUD FOR DEEP LEARNING INFERENCE

UP TO **45%**

BETTER PERFORMANCE PER WATT

THAN COMPARABLE GPU-BASED

INSTANCES

AWS INFERENTIA2

Up to 12 Inferentia2 accelerators and up to 384 GB of HBM2e high-speed accelerator memory

Up to 4x higher throughput and up to 10x lower latency than Inf1 instances

INFERENCE OPTIMIZED

Deploy 100B+ parameter models at scale

First inference platform with direct, 192 GB/s connectivity between multiple accelerators for distributed inference

aws

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





Supporting developer workloads improves productivity and quality by scaling out build, test, and deployment processes

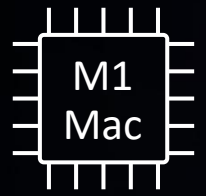


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

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 premises 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.

COMPUTE INNOVATION FOR EVERY WORKLOAD



World class scale and performance



Operational excellence



Compute for every workload



Cost optimization and best practices



Compute where you need it

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

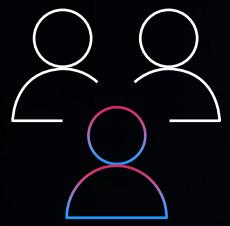

OFF



Many companies are looking for ways to lower costs while not affecting performance

Cost optimization best practices

HOW DOES AWS FOCUS ON COST REDUCTION



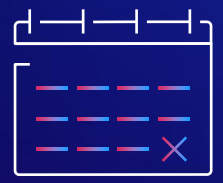
MAKE IT CULTURAL

At Amazon, our “frugality” leadership principal ensures that we consider the cost implications of every design or solution



TOOLS AND METRICS

Use the tools and metrics available to you to measure and control usage within the organization



ACCOUNTABILITY

Establish a mechanism to ensure teams identify cost savings opportunities and then track these closely over time

aws

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

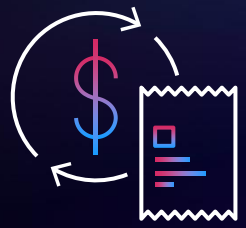
Cost optimization best practices

COST REDUCTION STRATEGIES FOR EC2



DIVERSIFY YOUR EC2 INSTANCE TYPES

AWS Graviton-based instances offer up to 40% better price performance and AMD-based instances deliver a 10%
savings versus comparable x86 based instances



CHOOSE THE RIGHT PURCHASE MODELS

Savings Plans offer a flexible pricing model with savings of up to 72% on your AWS compute usage compared to on-demand



MATCH CAPACITY WITH DEMAND

Available tools, AWS Compute Optimizer and AWS Cost Explorer, provide easy to implement rightsizing recommendations for your workloads

aws

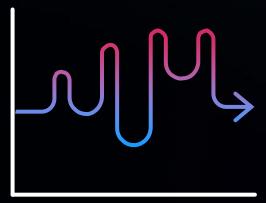
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon EC2 purchase models

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

ON-DEMAND

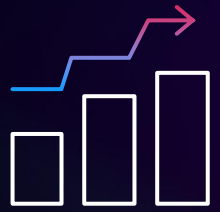
Pay for compute capacity by the second with no long-term commitments



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

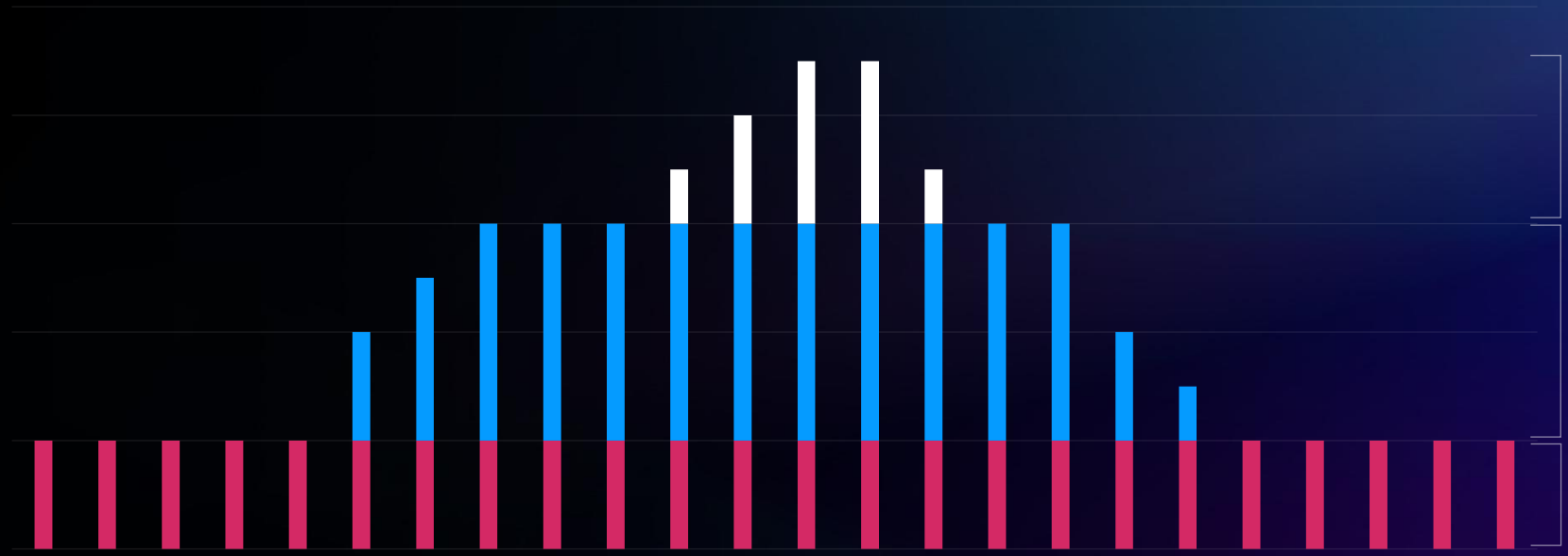


Fault-tolerant, flexible, stateless workloads

aws

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

Simplifying capacity and 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)
- Amazon EMR
- AWS CloudFormation
- AWS Batch

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

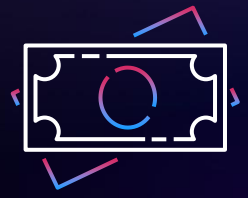
AWS Savings Plans

SAME GREAT SAVINGS AS RESERVED INSTANCES WITH MORE FLEXIBILITY



SIMPLE

Easy to use with recommendations in AWS Cost Explorer



COST SAVINGS

Significant savings of up to 72%



FLEXIBLE

Flexible across instance family, size, OS, tenancy, or AWS Region; also applies to AWS Fargate usage

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

Over

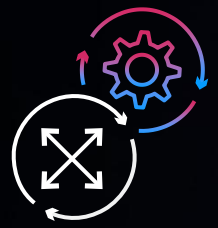
$15 billion

saved with Savings Plan

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

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

Right sized 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.

Over

10 billion

recommendations with Compute Optimizer

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

COMPUTE INNOVATION FOR EVERY WORKLOAD



World class scale and performance



Operational excellence



Compute for every workload



Cost optimization and best practices



Compute where you need it

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

AWS Global Infrastructure

AWS REGIONS, EDGE LOCATIONS, AND THE GLOBAL BACKBONE



AWS Global Infrastructure

AWS REGIONS, EDGE LOCATIONS, AND THE GLOBAL BACKBONE



REGIONAL EXPANSION

- Available today: 30 Regions
- Coming soon: 5 Regions

© 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

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

AWS Global Infrastructure

AWS REGIONS, EDGE LOCATIONS, AND THE GLOBAL BACKBONE

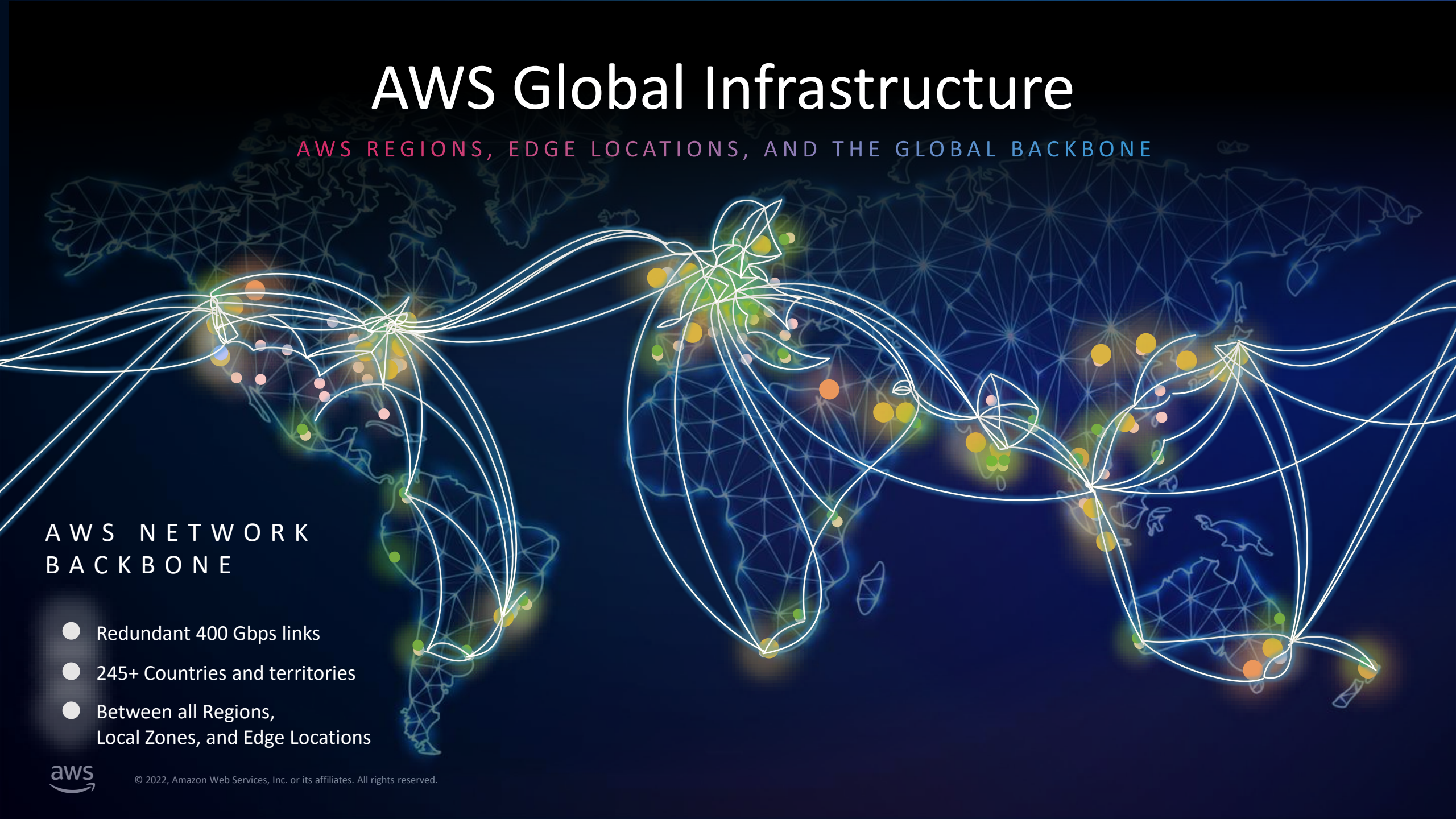


aws

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

AWS Global Infrastructure

AWS REGIONS, EDGE LOCATIONS, AND THE GLOBAL BACKBONE



AWS NETWORK BACKBONE

- Redundant 400 Gbps links
- 245+ Countries and territories
- Between all Regions, Local Zones, and Edge Locations

aws

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

AWS Outposts

AWS INFRASTRUCTURE AND SERVICES IN YOUR ON-PREMISES LOCATION



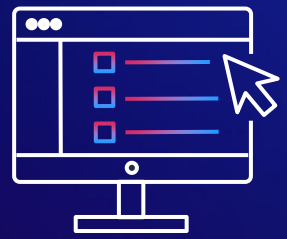
AWS DESIGNED

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



















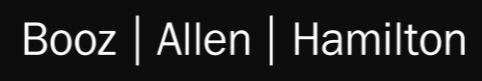








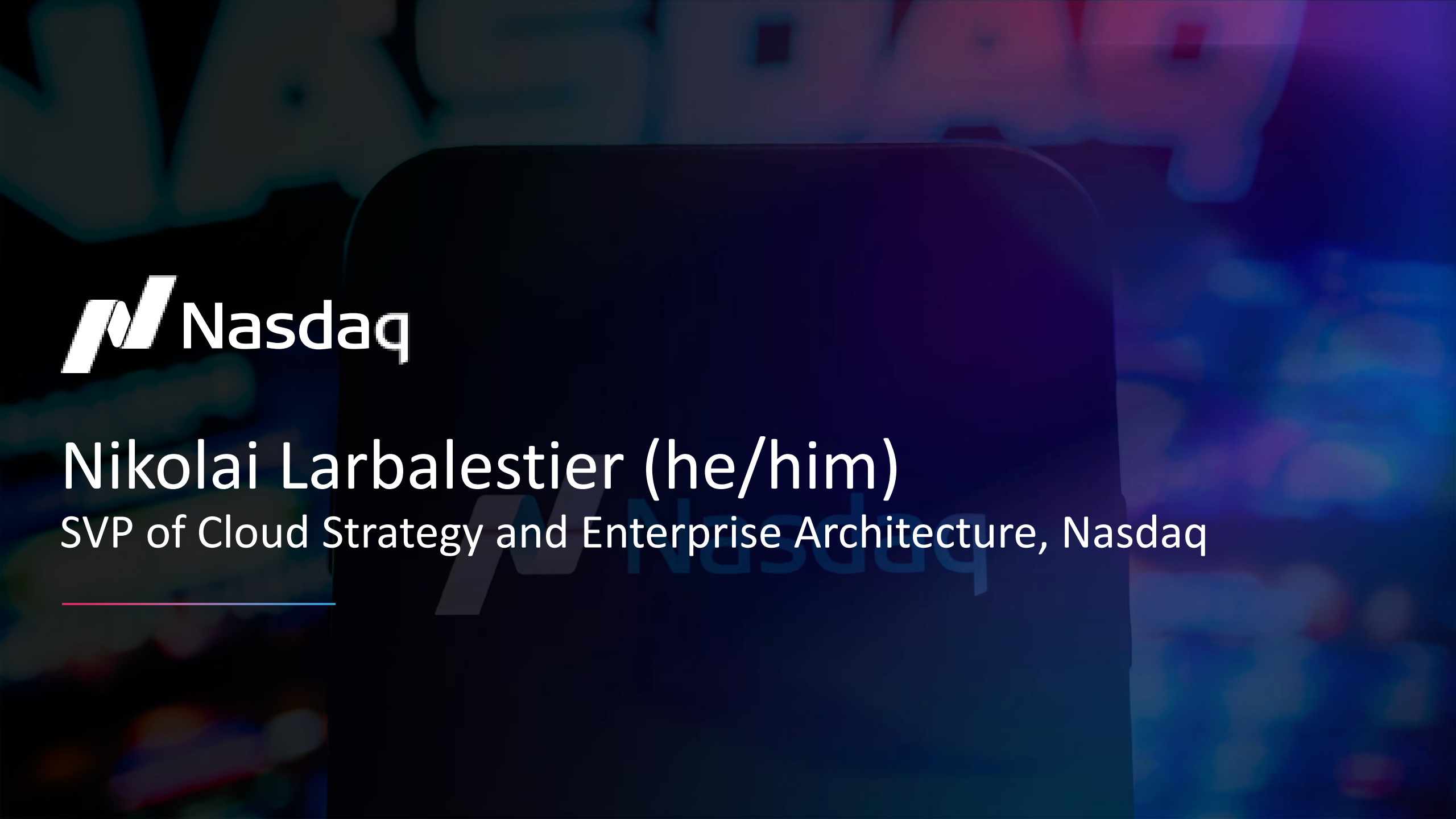




© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.



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



Nikolai Larbalestier (he/him)

SVP of Cloud Strategy and Enterprise Architecture, Nasdaq





aws

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

COMPUTE INNOVATION FOR EVERY WORKLOAD



World class scale and performance



Operational excellence



Compute for every workload



Cost optimization and best practices



Compute where you need it

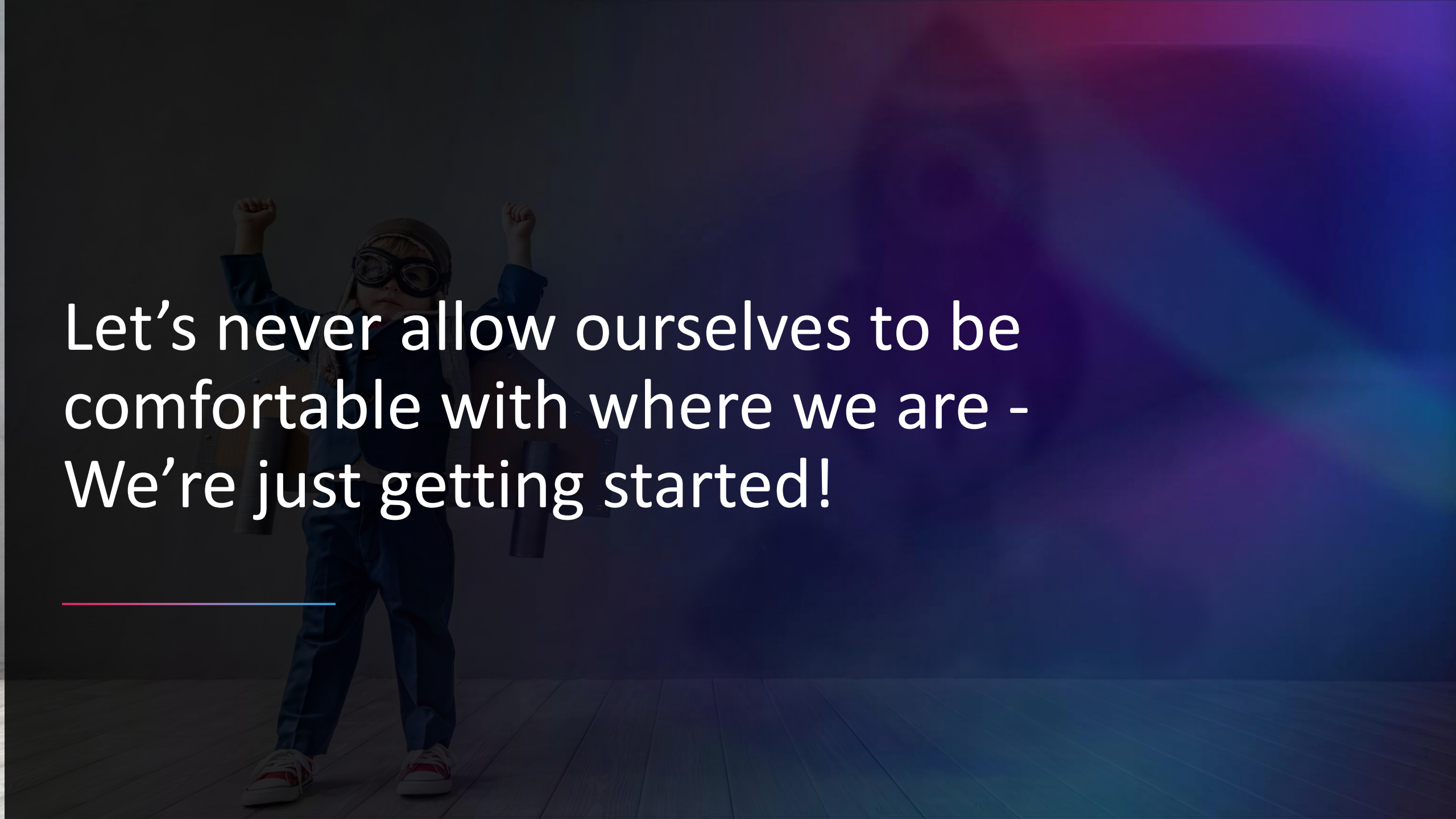
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Additional compute sessions

- CMP225 – What's New with Amazon EC2
- CMP221 – Modernize your iOS build pipelines with Amazon EC2 Mac instances
- CMP301 – Powering Amazon EC2: Deep dive on the AWS Nitro System
- CMP302 – Confidential Computing with AWS compute
- CMP201 – Silicon Innovation at AWS
- CMP410 – Framework for adopting AWS Graviton-based instances
- CMP205 – HPC on AWS: Solve complex problems with pay-as-you-go infrastructure
- CMP313 – Accelerate deep learning and innovate faster with AWS Trainium
- CMP319 – Optimize for cost and availability with capacity management

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




Let’s never allow ourselves to be comfortable with where we are - We’re just getting started!

Thank you!



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

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