



Differentiation with AWS Compute

Art Baudo (he/him)

Principal Product Marketing Manager
AWS

Martin Yip (he/him)

Senior Manager, Product Marketing
AWS

August 26, 2006

“Innovation never takes a break, and neither do I. From the steaming hot beaches of Cabo San Lucas I would like to tell you about the Amazon Elastic Compute Cloud, or Amazon EC2, now open for limited beta testing, with more beta slots to open soon.”

Jeff Barr

Vice President, AWS Evangelism



permalink | Share

break, and neither do I. From the steaming hot beaches of Cabo San Lucas I would like to tell you about the Amazon Elastic Compute Cloud, or Amazon EC2, now open for limited beta testing, with more beta slots to open soon.

access to a virtual computing environment. Your instance has a “virtual CPU”, the equivalent of a 1.7 GHz Xeon processor, 3 GB of RAM, 160 GB of local disk and 250 Mb/second of network bandwidth. You pay just 10 cents per clock hour (billed to your AWS account), and you can get as many virtual CPUs as you need. Learn more on the [EC2 Detail Page](#). We built Amazon EC2 on the machine monitor by the name of Xen.



instances in terms of AMIs, or Amazon Machine Images. Each AMI is a pre-configured boot disk — just a virtual operating system stored as an [Amazon S3](#) object. There are web service calls to create images, and to assign instances to run your application. If your application consists of the usual web server, business logic, and database, you can build distinct AMIs for each tier, and then spawn one or more instances of each type based on the

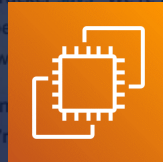
post, [Sometimes You Need Just a Little...](#), I alluded to the new world of scalable, on-demand web services. I also talked about the fact that sometimes a little bit of storage is all you need.

Some of you need a lot of processing power, and sometimes you need just a little. Sometimes you need a lot, but you need it for a limited amount of time. Perhaps you are doing some number crunching, some in-depth text processing, scientific research, or your end-of-month accounting. Or perhaps you want to experiment with some radical new processing algorithm for a week or two. In any of these situations, acquiring sufficient hardware to accommodate your needs and your water mark of your usage would definitely not be economical. There are already some interesting examples of

in the [Amazon EC2 Discussion Forums](#). For example:

Daniel Drucker says “We’re planning on using it for functional MRI analysis. We have large datasets which, when they’re being processed, require a cluster of 15-20 machines... but we only need those machines for a couple hours every few days.”

In the same thread, another user says “I have a small application that runs on a dedicated server, but that you’re only using from 7am to 7pm. You can cut your server costs by 50%. Take this to its logical conclusion — only start up an instance when you actually need your inventory app...”



Amazon EC2

In another way, time is another interesting axis of scalability.

Over
30 billion / instances launched
since 2006

100 million

instances launched
every day



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

Innovate to continuously
increase performance while
lowering overall cost



Differentiating with AWS Compute



World-class scale and performance



Compute for every workload



Cost optimization and best practices



Compute where you need it

Differentiating WITH AWS Compute



World-class scale and performance



Compute for every workload



Cost optimization and best practices



Compute where you need it

Innovation enabled by AWS Nitro System

MODULAR BUILDING BLOCKS FOR RAPID DESIGN AND DELIVERY OF AMAZON EC2 INSTANCES

NITRO CARD



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

NITRO SECURITY CHIP



**Integrated into motherboard
Protects hardware resources**

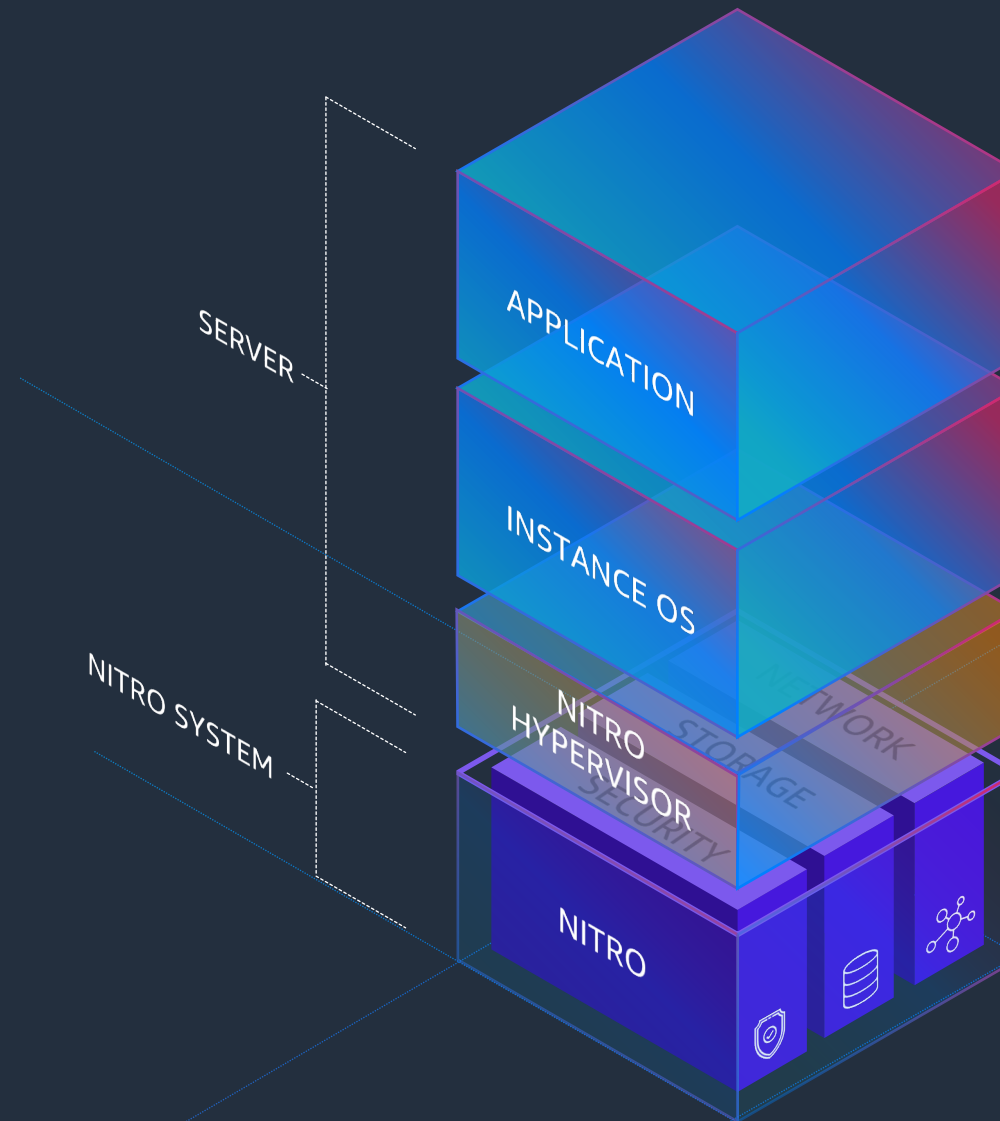
NITRO HYPERVISOR



**Lightweight hypervisor
Memory and CPU allocation
Bare metal-like performance**

The AWS Nitro System architecture

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

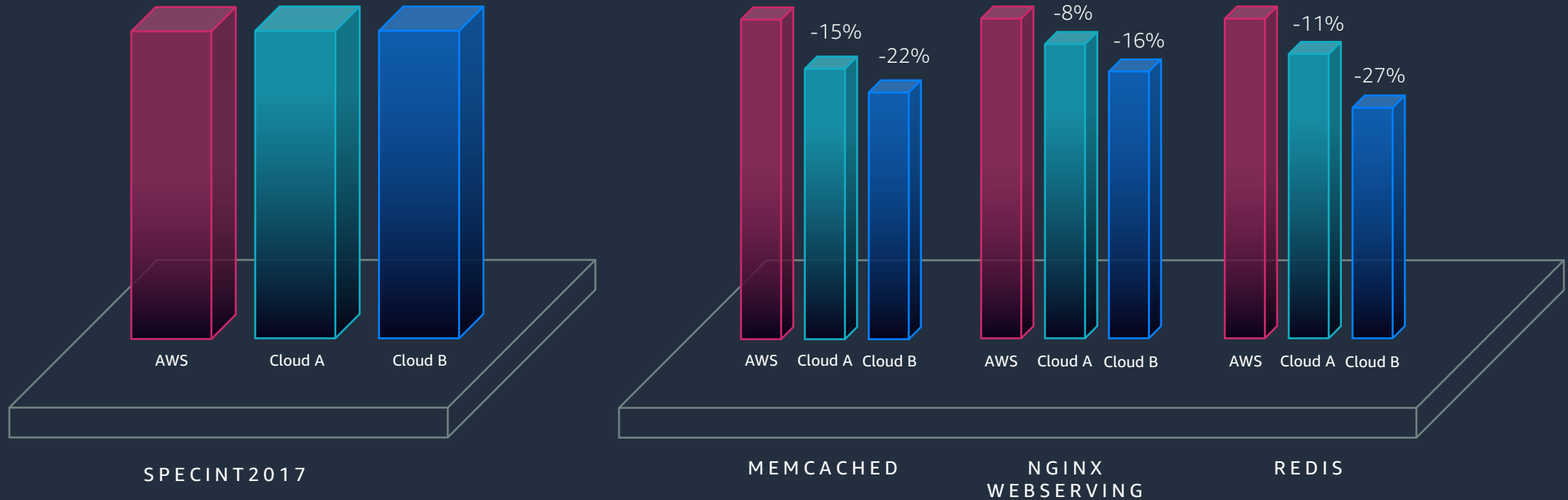


20 million

Nitro chips
shipped

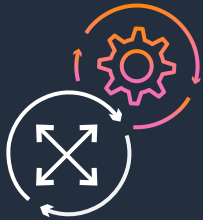
Nitro performance for real-world workloads

AMAZON EC2 INSTANCES CAN DELIVER OVER 15% HIGHER THROUGHPUT PERFORMANCE



AWS Nitro SSD

High-Performance, Low-Latency SSD Custom Designed By AWS



LOWER LATENCIES

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



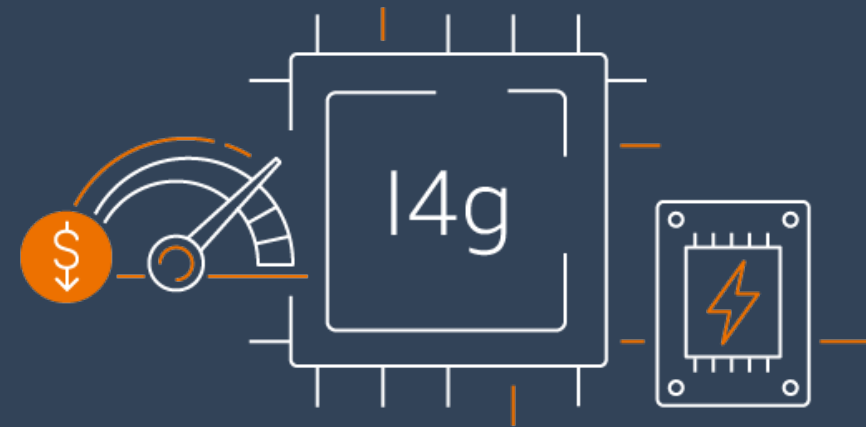
NITRO SECURITY

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

AWS Graviton2 storage-optimized instances

BEST STORAGE PERFORMANCE PER TB AND COMPUTE PRICE PERFORMANCE PER TB FOR GRAVITON-BASED, STORAGE-OPTIMIZED INSTANCES

- Powered by AWS Graviton2 processors
- Up to 64 vCPUs, 1 TiB memory, 25 Gbps networking, and 512 GiB of high-performance AWS Nitro SSD NVMe storage
- Deliver up to 15% better compute performance compared to similar storage-optimized instances
- Ideal for workloads that perform a high mix of random read/write operations and require low I/O latency, such as transaction databases (Amazon DynamoDB, MySQL, and PostgreSQL), Amazon OpenSearch Service, and real-time analytics such as Apache Spark



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 Nitro Hypervisor, can be updated without any downtime



NO REMOTE ACCESS

No shell available on the Nitro System or underlying server

Differentiating WITH AWS Compute



World-class scale and performance



Compute for every workload



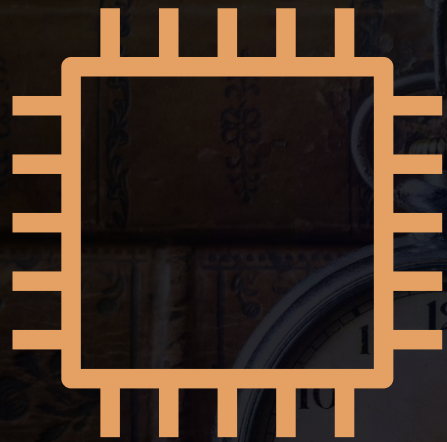
Cost optimization and best practices



Compute where you need it



© 2023, 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 choice of processors and architectures

RIGHT COMPUTE FOR THE WORKLOAD

The Intel logo, featuring the word "intel" in a lowercase, sans-serif font with a small blue square above the letter "i".

Intel Xeon
Scalable processors

The AMD logo, consisting of the word "AMD" in a bold, uppercase, sans-serif font followed by a square icon containing a stylized "A" shape.

AMD EPYC
processors

The AWS logo, featuring the lowercase letters "aws" in a white, sans-serif font with a curved orange arrow underneath that points from the "a" to the "s".

AWS Graviton
processors



Apple M1
processors

Broadest choice of processors and architectures

RIGHT COMPUTE FOR THE WORKLOAD

The Intel logo, consisting of the word "intel" in a lowercase, sans-serif font. The letter "i" has a small blue square above it. A registered trademark symbol (®) is located to the right of the word.

Intel Xeon
Scalable processors



Innovating with Intel

16 YEARS OF COLLABORATION AND INNOVATION WITH AWS



COLLABORATION

Deep engineering collaboration across the AWS portfolio



INTEGRATION

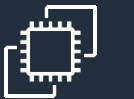
Over 350 Amazon EC2 instances are powered by Intel processors



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

Broadest choice of processors and architectures

RIGHT COMPUTE FOR THE WORKLOAD

AMD 

AMD EPYC
processors

Innovating with AMD

10% LOWER COST VERSUS COMPARABLE X86 INSTANCES



FLEXIBILITY

Help you optimize both cost and performance for your workloads



BETTER ECONOMICS

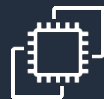
Deliver up to 10% lower cost versus comparable instances



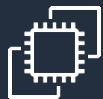
SEAMLESS MIGRATION

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

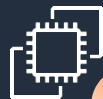
AMD-BASED INSTANCES



G4AD
GPU-
OPTIMIZED



HPC6A
HPC-
OPTIMIZED



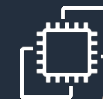
M6A
GENERAL
PURPOSE

NEW



C6A
COMPUTE-
OPTIMIZED

NEW



R6A
MEMORY-
OPTIMIZED

NEW

Broadest choice of processors and architectures

RIGHT COMPUTE FOR THE WORKLOAD



AWS Graviton
processors



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

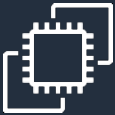
Supported by popular Linux OSes along with popular application and services from AWS and ISVs



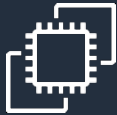
ENHANCED SECURITY

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

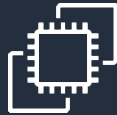
GRAVITON2-BASED INSTANCES



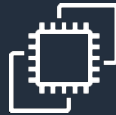
C6G(D)
COMPUTE-
OPTIMIZED



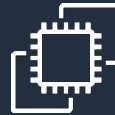
M6G(D)
GENERAL
PURPOSE



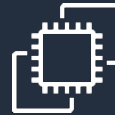
R6G(D)
MEMORY-
OPTIMIZED



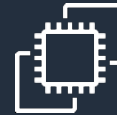
C6GN
NETWORK-
OPTIMIZED



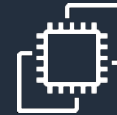
X2GD
ENHANCED
MEMORY



T4G
BURSTABLE



IM4GN
STORAGE-
OPTIMIZED



IS4GEN
STORAGE-
OPTIMIZED

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

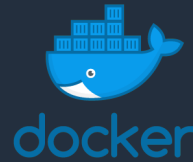


SUSE Linux Enterprise Server
for Arm 15



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**

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

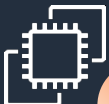
Supports DDR5 memory to provide 50% more memory bandwidth over DDR4 memory



ENERGY EFFICIENT

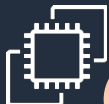
Consumes 60% less power for the same performance compared to other CPUs

GRAVITON3-BASED INSTANCES



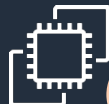
NEW

C7G
COMPUTE-
OPTIMIZED



NEW

M7G
GENERAL
PURPOSE



NEW

R7G
MEMORY-
OPTIMIZED

“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



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

Broadest choice of processors and architectures

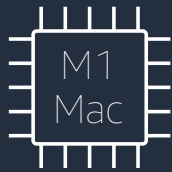
RIGHT COMPUTE FOR THE WORKLOAD



Apple M1
processors

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 onto AWS

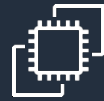
More and more companies are
deploying **machine learning** to
improve their customer experience

Supporting ML workloads

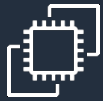
LARGEST SELECTION OF ML INSTANCES IN THE CLOUD

MACHINE LEARNING

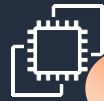
TRAINING



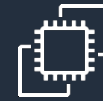
P3DN
NVIDIA V100
TENSOR CORE



P4D
NVIDIA A100
TENSOR CORE



P4DE
NVIDIA A100
TENSOR CORE



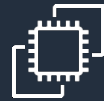
DL1
HABANA GAUDI
INTEL



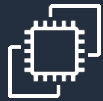
TRN1
AWS TRAINIUM



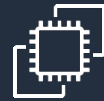
INFERENCE



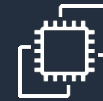
G4DN
NVIDIA T4 GPU



G5
NVIDIA A10G
TENSOR CORE



G5G
NVIDIA T4G
TENSOR CORE



INF1
AWS INFERENCE

PLATFORM

TENSORFLOW

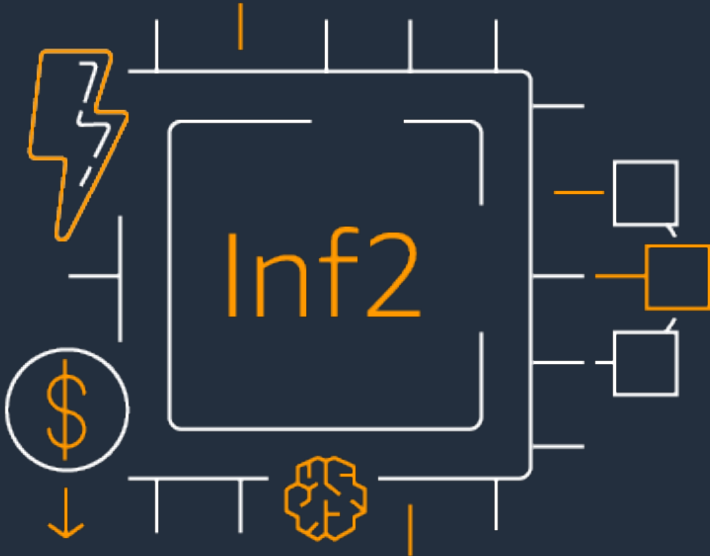
MXNET

PYTORCH

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

AWS Inferentia2-based Inf2 instances

HIGH PERFORMANCE, ENERGY EFFICIENT, AND LOWEST COST INFERENCE



Optimized to deploy 100B+ parameter models at scale

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

First inference platform with direct, ultra-high-speed connectivity between accelerators for distributed inference

70% better price performance and 50% better performance/watt than comparable Amazon EC2 instances

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

AWS Inferentia2: High performance, less power, lower cost

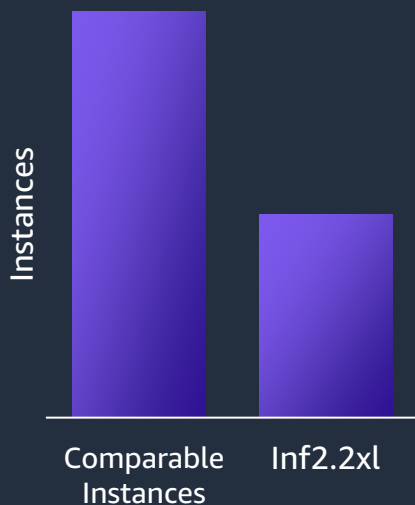
NEW!

REAL-TIME DEPLOYMENT BERT-LARGE WITH AWS INFERENTIA2

50%

Fewer instances

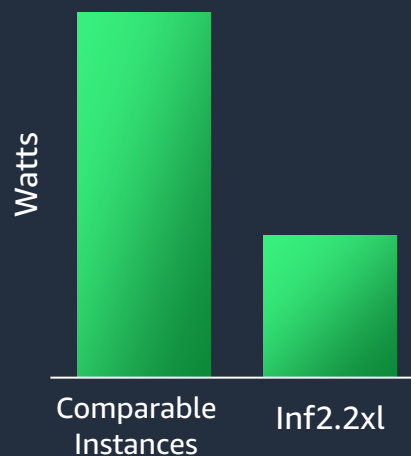
Number of instances



50%

Less energy

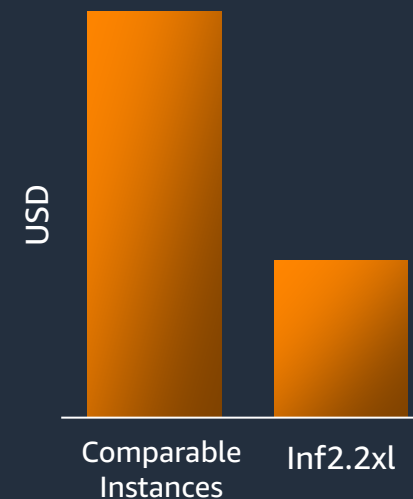
Power



65%

Lower cost

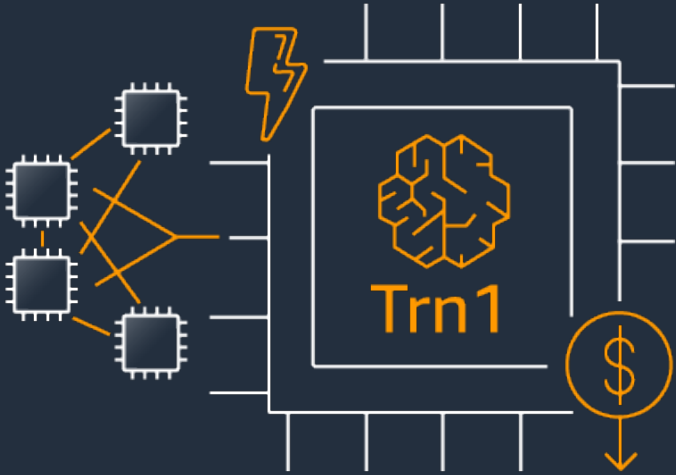
Inference cost



NEW!

AWS Trainium-based Trn1/Trn1n instances

HIGH PERFORMANCE, ENERGY EFFICIENT, AND COST-EFFECTIVE TRAINING



Highest performance for training deep learning models such as NLP models on Amazon EC2

Save up to 50% on training costs over comparable GPU-based instances in Amazon EC2

Up to 16 Trainium accelerators, 512 GB HBM2e memory, 800/1600 Gbps of networking, & 8 TB of local NVMe storage

Deployable in Amazon EC2 UltraClusters—tens of thousands of accelerators connected with petabit scale network

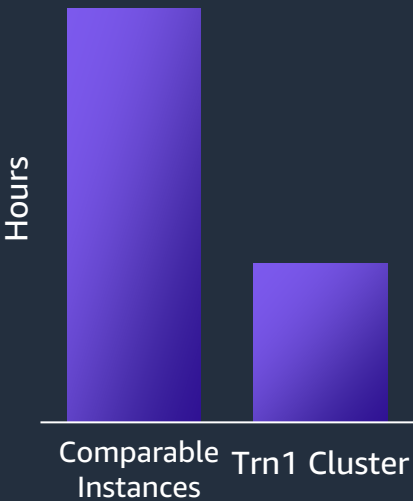
AWS Trainium: High performance, less power, lower cost

TRAINING BERT LARGE WITH AWS TRAINIUM

2.3x

Faster training

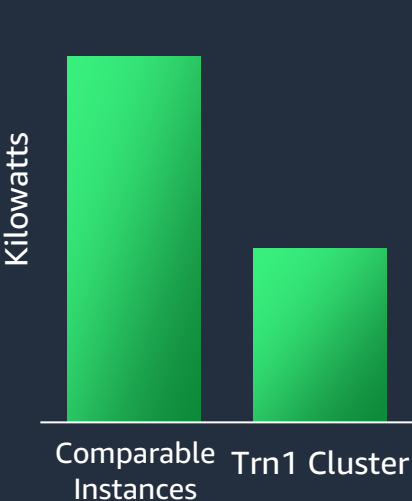
Time to train



47%

Less energy

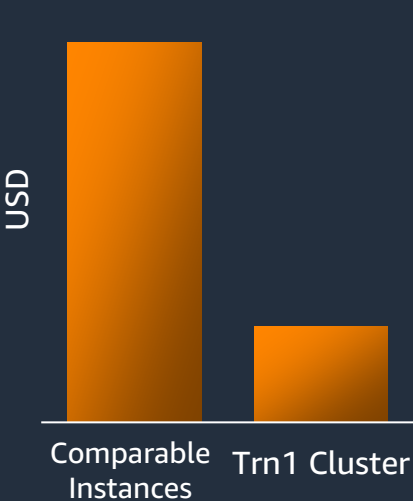
Power



72%

Lower cost

Cost to train



Differentiating WITH AWS Compute



World-class scale and performance



Compute for every workload



Cost optimization and best practices



Compute where you need it





Many companies are looking for ways to **lower costs** without affecting performance

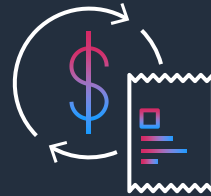
Cost-optimization best practices

COST REDUCTION STRATEGIES FOR AMAZON EC2



DIVERSIFY YOUR AMAZON 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 like AWS Compute Optimizer and AWS Cost Explorer provide easy-to-implement right-sizing recommendations for your workloads

Differentiating WITH AWS Compute



World-class scale and performance



Compute for every workload



Cost optimization and best practices



Compute where you need it

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: 31 Regions
- Coming soon: 5 Regions



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 them to support low-latency applications



FULLY MANAGED

Fully owned, managed, and supported by AWS

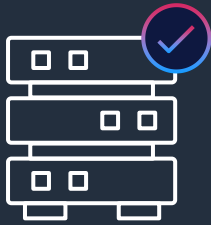


CITIES

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

AWS 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



Nasdaq



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Nasdaq REWRITE TOMORROW



Nasdaq REWRITE TOMORROW



Let's never allow ourselves to be comfortable with where we are – we're just getting started!



Thank you!



Thank you!