Delivering price performance advantages with AWS silicon innovation

How AWS custom silicon provides
better application performance
with lower costs at scale

INTRODUCTION

Finding the right price performance with AWS custom silicon

As end users demand more from their applications and workloads become more complex, organizations need a way to solve for three critical needs: performance expectations for virtually any application, security of their data and their customers' data, and staying within their infrastructure budgets. It's a challenge that calls for a "secret sauce"—a solution that will make achieving the right balance of price and performance not only possible but also easy.

Innovations with Amazon Web Services (AWS) custom silicon is the solution to this price performance equation. AWS custom silicon have made it possible for organizations—from the smallest startups to the largest enterprises—to innovate for their customers in previously unimaginable ways.

This eBook highlights four common challenges facing modern organizations and illustrates how AWS—designed to be the most secure, reliable, and scalable cloud infrastructure—can help solve those challenges, enabling organizations to innovate with new capabilities and meet market demands, now and in the future.

INTRO TO AWS SILICON INNOVATION

Purpose-built silicon for today’s workloads

AWS is the leader in silicon designed and optimized exclusively for the cloud and developed for performance and scalability. As a result of our innovations, which include processors, machine learning (ML) chips, and high-performance storage products, we can deliver the best price performance at scale for a wide range of applications and workloads using AWS services.

In this eBook, you’ll learn about innovations with AWS custom silicon and how they can deliver value for your business:

- **AWS Nitro System:** Boosts performance, reduces costs, enables choice, and provides enhanced security
- **AWS Graviton:** Delivers the best price performance for a broad range of workloads
- **AWS Nitro SSD:** Enables low latency and latency variability for storage-intensive applications
- **AWS Inferentia:** Enables high performance and low cost for deep learning inference in the cloud
- **AWS Trainium:** Provides the best price performance in the cloud for training deep learning models

Together, AWS silicon innovations help you stay ahead of today’s ever-demanding workloads, innovate faster, save costs, and accelerate your business growth while delivering the experiences your customers expect.

Challenge 1: Securely running applications at scale

The need to innovate faster—to build, migrate, and run more types of workloads securely at scale and to do so with crunched compute budgets—is top of mind for organizations today.

The **AWS Nitro System**, the underlying platform for the majority of our modern Amazon Elastic Compute Cloud (Amazon EC2) instances, comprises three different components: Nitro Cards, which offload and accelerate I/O for functions, ultimately increasing overall system performance; the Nitro Security Chip, which features a minimized attack surface with virtualization and security functions that are offloaded to dedicated hardware and software; and the Nitro Hypervisor, a lightweight hypervisor that manages memory and CPU allocation and delivers performance that is virtually indistinguishable from bare metal.

On a broader level, the AWS Nitro System enables faster innovation, delivering benefits that include:

Improved performance: With the AWS Nitro System, Amazon EC2 instances can deliver more than 15 percent higher throughput performance on some workloads as compared to other major cloud providers running the same CPU. Dedicated Nitro Cards enable high-speed networking, high-speed Amazon Elastic Block Store (Amazon EBS), and I/O acceleration. Not having to hold back resources for management software means more savings and performance passed on to end users.

Enhanced security: The Nitro Security Chip provides the most secure cloud platform with a minimized attack surface. Virtualization and security functions are offloaded to dedicated hardware and software. Additionally, a locked-down security model prohibits all administrative access, including the access of Amazon employees, eliminating the possibility of human error and tampering.

Faster innovation: The AWS Nitro System delivers choice and the ability to bring more workloads to the cloud, enabling you to increase the pace of innovation. With the AWS Nitro System, functions are modularized, breaking the architecture of EC2 into smaller blocks by offloading the virtualization functions onto dedicated hardware. These blocks can be assembled in many ways, which delivers the flexibility to design and rapidly deliver EC2 instances with an ever-broadening number of compute, storage, memory, and networking options.

evervault

"Protecting and processing highly sensitive information such as financial, healthcare, identity, and proprietary data is one of the main use cases for Evervault's encryption infrastructure. At the core of Evervault is our Evervault Encryption Engine (E3), which performs all cryptographic operations and handles encryption keys for our customers. E3 is built on AWS Nitro Enclaves which provides an isolated, hardened, and highly constrained compute environment for processing sensitive data. Building E3 on Nitro Enclaves means that we can provide both security through cryptographic attestation, and a robust foundation for all other Evervault products and services. At no additional cost, Nitro Enclaves enable us to provide a highly secure, cost effective, and scalable service to our customers; a service that is capable of handling thousands of cryptographic operations per second."

Shane Curran, Founder & CEO at Evervault

anjuna crypto.com

evervault M10

Footprint

Explore more customer stories ›

Challenge 2: Achieving optimal performance for a wide range of applications

Organizations run a wide variety of applications to power their businesses, ranging from general-purpose application servers and microservices to open-source-based databases and caches to even compute-intensive workloads, such as media encoding, gaming, and ML. They need high performance for these workloads while keeping their costs optimal. Saving costs frees development teams to innovate and build new apps or improve existing ones, which enables them to keep the wheel of innovation turning, attract more customers, and stay competitive.

And with digital transformation evolving into sustainable transformation, energy efficiency is another priority organizations must take into account.

AWS Graviton processors are designed by AWS to deliver on all of these priorities:

Price performance: AWS Graviton–based instances deliver the best price performance for a broad range of workloads running on Amazon EC2: up to 40 percent better price performance over comparable x86-based instances for a wide variety of workloads, such as application servers, microservices, video encoding, high performance computing (HPC), electronic design automation, and more.

Energy efficiency: AWS Graviton is our most energy-efficient processor, using up to 60 percent less energy for the same performance than comparable EC2 instances.

Enhanced security: AWS Graviton processors feature key capabilities that enable you to run cloud-native applications securely and at scale. AWS Graviton3 processors feature always-on memory encryption, dedicated caches for every vCPU, and support for pointer authentication.

> "We benchmarked our Java-based search workloads on Amazon EC2 Im4gn/Is4gen instances powered by AWS Graviton2. Smaller Is4gen instances offer similar performance compared to larger I3en instances, presenting an opportunity to meaningfully reduce the TCO. We also saw a significant 50% reduction in latency for queries when moving our workloads from I3 to Im4gn instances, indicating a significant 40% price performance benefit. Moving to Graviton2-based instances was easy, taking 2 weeks to complete benchmarking. We are very happy with our experience and look forward to running these workloads in production on Im4gn and Is4gen instances."
>
> Abhay Bansal, VP of Engineering at Sprinklr

Read more about Sprinklr and AWS Graviton processors ›

Explore more customer stories ›

AWS NITRO SSD

Challenge 3: Increasing storage performance for I/O-intensive workloads

For decades, traditional hard drives (HDDs) were the primary devices for block storage. Today, HDDs still have their place, but most high-performance storage is based on modern solid state drives (SSDs). AWS set a goal to enable I/O-intensive workloads (relational databases, NoSQL databases, data warehouses, search engines, and analytics engines) to run faster and with more predictable performance. That goal was achieved with the launch of AWS Nitro SSD, a high-performance, low-latency SSD custom-built for I/O-intensive workloads.

AWS Nitro SSDs deliver 60 percent lower I/O latency compared to commercial SSDs. They provide faster firmware updates to improve reliability without instance downtime. And all stored data is encrypted at rest with AES-256 ephemeral keys.

With AWS Nitro SSDs, I/O-intensive applications can get faster, more predictable performance. Next, we'll see how AWS solutions are simplifying ML to make it accessible and affordable for organizations of every type and size.

AEROSPIKE

> "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. This combined with the proven ability of Aerospike to deliver high throughput of writes means Aerospike running on I4i instances will deliver the best price/performance in the industry for real-time access to petabytes of data running on EC2."
>
> Srini Srinivasan, CTO & Founder at Aerospike

Read more about Aerospike and AWS Nitro SSD ›

AEROSPIKE

honeycomb.io

redislabs HOME OF REDIS

SCYLLA.

splunk>

Explore more customer stories ›

Challenge 4: Scaling up machine learning while mitigating infrastructure costs

From enhancing customer experiences to boosting productivity and cutting costs, ML is being adopted by more organizations. While ML is proving to be an invaluable tool for solving business problems and optimizing processes, building, deploying, and optimizing ML models can be complex and costly.

In response to demand for better applications and tailored, personalized experiences, data scientists and ML engineers are building larger, more complex deep learning models. Large language models (LLMs) containing more than 100 billion parameters are increasingly prevalent. While these generative artificial intelligence (AI) models unlock several new use cases, such as text summarization, image generation, code generation, and more, training these models while meeting performance and accuracy goals and making them run efficiently in production is not only technically challenging but also extremely expensive and energy-intensive.

The good news is that many of these barriers can be overcome through cloud innovations. AWS offers two solutions that eliminate the cost and performance barriers for ML adoption: AWS Trainium and AWS Inferentia.

Reducing cost of inference in the cloud

AWS Inferentia accelerators are designed by AWS to deliver high performance at the lowest cost in Amazon EC2 for ML inference applications.

The first-generation AWS Inferentia accelerator powers **Amazon EC2 Inf1 Instances**, which deliver up to 2.3 times higher throughput and up to 70 percent lower cost per inference than comparable GPU-based Amazon EC2 instances.

AWS Inferentia2 accelerator is purpose-built to deploy ML models with more than 100 billion parameters. It delivers a major leap in performance and capabilities, delivering up to 4 times higher throughput and up to 10 times lower latency compared to first generation AWS Inferentia. AWS Inferentia2–based **Amazon EC2 Inf2 Instances** are designed to deliver high performance at the lowest cost in Amazon EC2 for your generative AI applications. Inf2 instances deliver 3 times higher throughput and 8 times lower latency than comparable GPU-based EC2 instances. They deliver up to 70 percent better price performance than comparable GPU-based instances in Amazon EC2. They are optimized to deploy increasingly complex models, such as LLMs and vision transformers, at scale.

AWS Neuron SDK helps developers train models on AWS Trainium and deploy models on AWS Inferentia accelerators. It integrates natively with frameworks, such as PyTorch and TensorFlow, so you can continue using your existing workflows and run on Inf1 or Inf2 instances. **Amazon SageMaker** offers a fully managed end-to-end workflow, that makes it easy to take advantage of these instances while reducing development time and costs.

airbnb

"Airbnb's Community Support Platform enables intelligent, scalable, and exceptional service experiences to our community of millions of guests and hosts around the world. We are constantly looking for ways to improve the performance of our Natural Language Processing models that our support chatbot applications use. 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 at Airbnb

Read more about how Airbnb uses AWS services to support its growth ›

airbnb Anthem. AUTODESK CONDÉ NAST Money Forward sprinklr

Explore more customer stories ›

Making training for machine learning models cost-efficient

Many development teams are limited by fixed ML training budgets. This puts a cap on the scope and frequency of training needed to improve their models and applications. AWS Trainium, the purpose-built ML accelerator optimized for high-performance deep learning training, answers this challenge by providing the most cost-efficient ML training in the cloud.

Amazon EC2 Trn1 Instances, powered by **AWS Trainium** accelerators, are purpose-built for high-performance (DL) training while offering up to 50 percent cost-to-train savings over comparable GPU-based instances. EC2 Trn1 instances deliver the highest performance on deep learning training of popular natural language processing (NLP) models on AWS. Trn1/Trn1n instances are the first EC2 instances with up to 1600 Gbps of Elastic Fabric Adapter (EFA) network bandwidth. They are deployed in EC2 UltraClusters that enable scaling up to 30,000 AWS Trainium accelerators, which are interconnected with a non-blocking petabit-scale network to provide up to 6.3 exaflops of compute. You can use EC2 Trn1 instances to train LLMs, diffusion models, and recommender models across a broad set of applications, such as text summarization, speech recognition, image generation, recommendation, and fraud detection.

With AWS Neuron SDK, you can get started on EC2 Trn1 instances by using your existing workflows and code in popular ML frameworks, such as PyTorch and TensorFlow. Again, the fully managed, end-to-end workflow of **Amazon SageMaker** allows you to easily and cost-effectively leverage these powerful ML instances.

Money Forward, Inc. serves businesses and individuals with an open and fair financial platform.

> “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. As we keep fine-tuning tailored NLP models periodically, reducing model training times and costs is also important. Based on our experience from successful migration of inference workload on Inf1 instances and our initial work on AWS Trainium-based EC2 Trn1 instances, we expect Trn1 instances will provide additional value in improving end-to-end ML performance and cost.”

Takuya Nakade, CTO at Money Forward, Inc.

Read about Amazon EC2 Trn1 instances powered by AWS Trainium ›

CONCLUSION

Bringing it all together

AWS custom silicon: The “secret sauce” of innovation

Over the last decade, the cloud has transformed the way organizations innovate. Now, AWS continues to transform the silicon industry to help you conquer the challenges of today’s workload demands and deliver on the ever-evolving expectations of your customers.

With custom silicon built for performance and scalability—AWS Nitro System, AWS Graviton, AWS Inferentia, AWS Trainium, and AWS Nitro SSD—AWS is helping drive next-generation innovation by enabling customers to run even more applications in the cloud while realizing value from unmatched price performance, better security, and energy efficiency.

No other cloud has the expertise and experience building custom silicon like AWS. Learn how your organization can lead the future of innovation and how AWS can make it possible.

Learn more about AWS silicon innovation ›