uptime INTELLIGENCE

INTELLIGENCE UPDATE

How AWS's own silicon and software deliver cloud scalability


Dr. Owen Rogers

20 Jan 2025

Amazon Web Services (AWS) was the world's first hyperscale cloud provider, and it remains the largest today. It represents around one-third of the global market, offering more than 200 infrastructure, platform and software services across 34 regions. To efficiently deliver so many services at such a scale, AWS designs and builds much of its own hardware.

The core AWS service is Amazon EC2 (Elastic Cloud Compute), which delivers virtual machines as a service. Not only is Amazon EC2 a service for customers, but it is also the underlying, hidden foundation for AWS's platform and software services. The technology deployed in AWS data centers is often used by its parent company, Amazon, to deliver e-commerce, streaming and other consumer capabilities.

A hyperscale cloud provider does more than just manage "someone else's computer," as the joke goes. At the annual AWS re:Invent conference in November 2024, one speaker stated that AWS EC2 users create around 130 million new instances daily, which is well beyond anything colocation or enterprise data centers can achieve. Managing the IT infrastructure to meet such demand requires servers and silicon specifically designed for the task. Since 2017, a core capability in AWS infrastructure has been the Nitro system, which enables such scale by offloading virtualization, networking and storage management from the server processor and onto a custom chip.

Nitro architecture

Virtualization software divides a physical server into many virtual machines. It is a vital component of the public cloud because it enables the provider to create, sell and destroy computing units purchased on demand by users.

The AWS Nitro system consists of a custom network interface card containing a system-on-chip (SoC) and a lightweight hypervisor (virtualization software layer) installed on each server. Designed by Annapurna Labs, which Amazon acquired in 2015, the hardware and firmware are developed and maintained by AWS engineering teams.

The system offloads many of the functions of software virtualization onto dedicated hardware. This offloading reduces CPU overhead, freeing up resources previously consumed by virtualization software for running customer workloads. It also offloads some security and networking functionality.

A full breakdown of Nitro's capability is provided in **Table 1**.

Table 1 Features of Nitro card

Feature	Description	Benefits
Virtualization Offloading	Moves hypervisor functions to dedicated hardware.	Frees CPU resources for customer workloads; reduces virtualization overhead.
Hardware-Based Isolation	Dedicated Nitro hardware for storage, network and compute management.	Enhances security; prevents data leakage in multi-tenant environments.
Hardware Root of Trust	Each Nitro chip has a unique cryptographic fingerprint for attestation.	Verifies system integrity from manufacturing to runtime.
Measured Boot Process	Sequential cryptographic validation of each boot stage.	Prevents execution of unauthorized code; ensures secure boot process.
Bare-Metal Support	Provides direct access to hardware for applications that require low-level control.	Bypasses virtualization for maximum performance in HPC and specialized workloads.
macOS Virtualization	Supports running macOS instances on AWS infrastructure.	Expands AWS offerings to Apple developers and workloads that require macOS compatibility.
Dynamic Resource Management	Allocates compute, storage, and networking resources on demand.	Optimizes infrastructure for variable workloads; supports elastic scaling for large events.

UPTIME INSTITUTE 2025 uptime INTELLIGENCE

AWS has millions of servers that are connected and ready to use. Nitro enables users (or applications) to provision resources and start them up securely within seconds without requiring human interaction. It also provides AWS with the ability to control and optimize its estate.

Through Nitro, AWS can manage all its servers regardless of the underlying hardware, operating system, or the AWS service provisioned upon them. Nitro allows x86 and ARM servers to be managed using the same technology, and it can also support accelerators such as Nvidia GPUs and AWS's own Inferentia and Trainium application-specific integrated circuits for AI workloads.

Although AWS uses servers from original equipment manufacturers, such as Dell and HPE, it also designs its own, manufacturing them via original design manufacturers (ODMs), usually based in Asia. These servers are stripped of nonessential components to reduce cost overheads and optimize performance for AWS's specific requirements, such as running its ARM-based CPU, Graviton. In addition, AWS designs its own networking equipment, which is also manufactured by ODMs, reportedly including Wiwynn and Quanta.

The Graviton CPU

Graviton is ASW's family of ARM-based chips, designed and manufactured by Annapurna Labs. Just like Nitro, Graviton is becoming an increasingly important enabler for AWS, and the two capabilities are becoming more entwined.

The use of Graviton is growing, according to speakers at the re:Invent conference. In the past two years, 50% of AWS's new CPU capacity has been based on Graviton. Customers can consume Graviton directly through a range of EC2 virtual machines, but AWS also utilizes Graviton to power platforms and services where the customer has no visibility to (or interest in) the underlying technology — for example, 150,000 Graviton chips power the AWS DynamoDB database service.

Graviton is also employed by the parent company: Amazon used 150,000 Graviton chips during its annual Prime Day sale to meet its e-commerce demand.

The growth in Graviton processor adoption is driven primarily by economics. Compared with instances using x86 designs by Intel and AMD, AWS prices Graviton instances lower at comparable configurations (vCPUs, memory, bandwidth) as it tries to steer customers towards its own platform.

For AWS, selling access to its own chips captures revenue that would otherwise have gone to its partners Intel and AMD. It also gives AWS a differentiator in the market and a degree of lock-in; AWS's competitors are now offering ARM services, but Graviton is more mature and widely adopted in the cloud market.

The downside for cloud customers is that chips based on ARM instruction sets cannot run the vast library of x86 codes and have a less mature software toolchain. This makes it harder for developers to implement some features or extract optimal performance, making them unsuitable for many commercial business applications.

Nitro enhances AWS's latest Graviton chip (version 4) by providing a secure foundation through hardware-based attestation and isolation. Graviton4 processors and Nitro chips verify each other's identity cryptographically and establish encrypted communication channels, which helps protect workloads running on AWS from unauthorized access with minimal performance impact.

Scalable storage

Nitro also enables storage to be disaggregated from compute, making it independently scalable.

Compute and storage do not necessarily scale with each other. One application might need a lot of compute and little disk, while another might need the complete opposite. This presents a problem in a static server with a fixed capacity of compute and storage.

In a traditional storage array, a head node is a server that manages the interactions between

storage users and the actual disks. A storage array is provisioned with a head node and many disks connected directly to it.

The problem with this setup is that the maximum number of disks that the array can support is decided at setup. If an array is full, a new array has to be purchased.

As the size of the array design grows, practical challenges arise. AWS scaled a single storage array to 288 drives, with the hardware holding nearly six petabytes and weighing two tons. The sheer size of the appliance meant:

- Data center floors had to be reinforced.
- Specialized equipment was required to move and install arrays.
- Vibrations from all drives moving in unison created performance issues.
- A single failure of a head node would render 288 drives inaccessible.

To allow storage to scale independently and reliably from compute without such deployment challenges, AWS designed its own storage system, effectively utilizing Nitro as a lightweight head node.

In AWS's method, each disk enclosure contains its own Nitro card. The Nitro card acts as a basic head node, managing the disks contained within the enclosure, and interacting with virtual machines hosted on servers elsewhere.

The primary benefits for AWS are easier maintenance and increased reliability. If a Nitro card fails, only a few drives lose connectivity, as opposed to an entire array of disks. Any failed drive can be removed from the service and a replacement added without causing downtime of the other disks or compute server. If a virtual machine goes down due to a failure of a compute server, it can be restarted elsewhere and the disks reconnected automatically, without loss of data.

The Uptime Intelligence View

Enterprises and colocation providers should focus on what the hyperscalers cannot do — supporting a wide range of hardware configured for each customer (internal or external), ensuring that hardware is secure (physically and virtually) and accessible only by that customer, and offering hands-on support tweaked to customer needs. They should also accept that customers will use the cloud for some applications simply because the hyperscalers can squeeze efficiency and provide scalability to a degree that is impossible for most organizations. Colocations and private facilities should enable the use of both on-premises and cloud infrastructure for their applications.

ABOUT THE AUTHOR

Dr. Owen Rogers

Dr. Owen Rogers is Uptime Institute's Sr. Research Director of Cloud Computing. Dr. Rogers has been analyzing the economics of cloud for over a decade as a product manager, a PhD candidate and an industry analyst. Rogers covers all areas of cloud, including economics, sustainability, hybrid infrastructure, quantum computing and edge.

orogers@uptimeinstitute.com

About Uptime Institute

Uptime Institute is the Global Digital Infrastructure Authority. Its Tier Standard is the IT industry's most trusted and adopted global standard for the proper design, construction, and operation of data centers – the backbone of the digital economy. For over 25 years, the company has served as the standard for data center reliability, sustainability, and efficiency, providing customers assurance that their digital infrastructure can perform at a level that is consistent with their business needs across a wide array of operating conditions.

With its data center Tier Standard & Certifications, Management & Operations reviews, broad range of related risk and performance assessments, and accredited educational curriculum completed by over 10,000 data center professionals, Uptime Institute has helped thousands of companies, in over 100 countries to optimize critical IT assets while managing costs, resources, and efficiency.