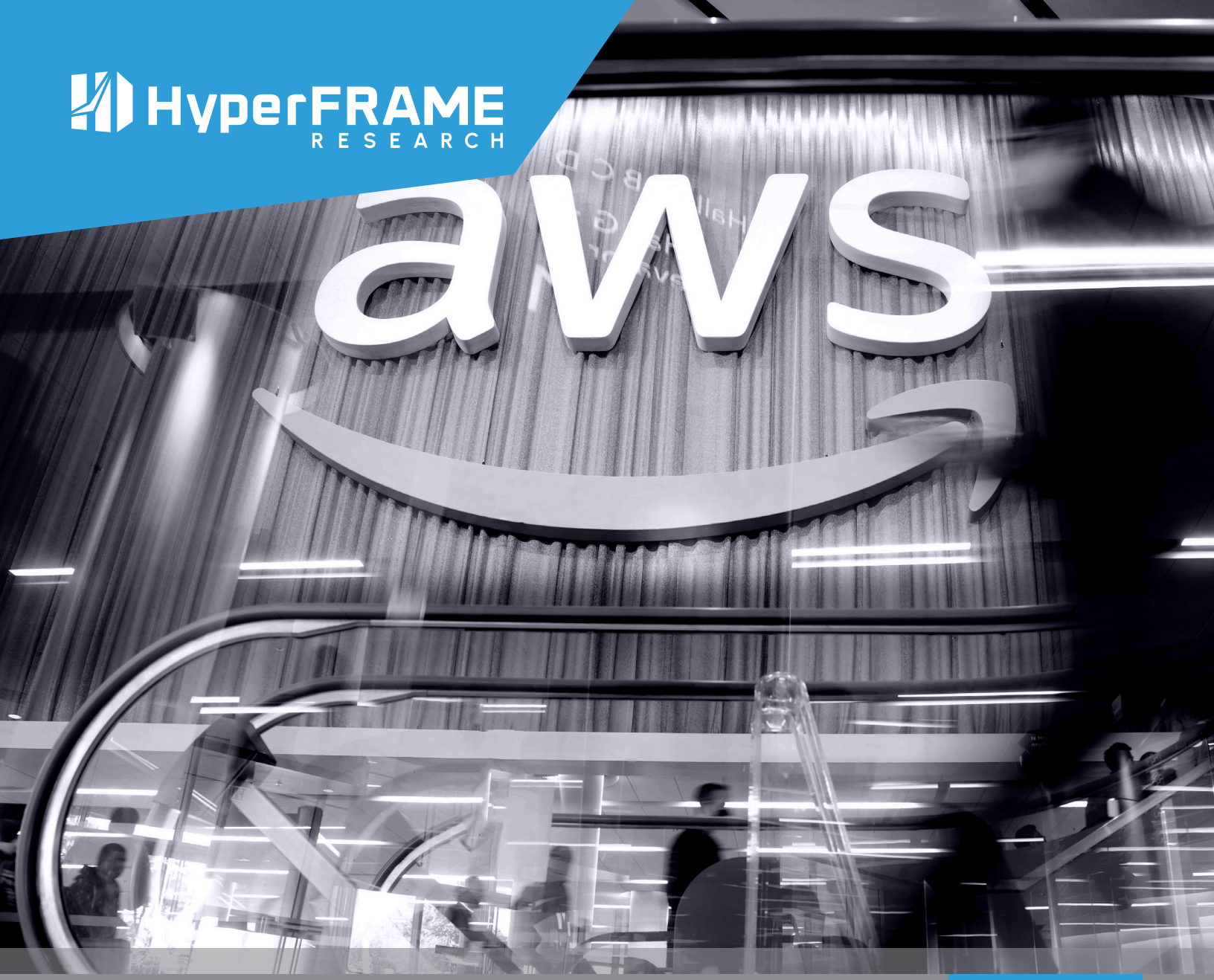
HyperFRAME RESEARCH



WHITEPAPER

AWS re:Invent 2024 Re-Cap

Generative AI, Silicon Innovation, and Automation dominate AWS re:Invent 2024 Keynote Themes

Author:
Steven Dickens
CEO and Principal Analyst

DECEMBER 2024

Image source : AWS Press

Key Highlights

- AWS introduced Trainium 3, advancing its silicon leadership and AI model efficiency.
- Generative AI capabilities, including agentic workflows, aim to transform enterprise productivity.
- Amazon Q expands with tools for automating business workflows and developer efficiencies.
- Resilience and security in AI applications were emphasized as “secure by design” and became a focal point.
- Partnerships with NVIDIA, advancements in mainframe, and VMware transformations highlight AWS’s multi-modal strategy.
- AWS highlights details of a new Supercomputer to power its collaboration with Anthropic for AI workloads
- AWS Launched a multi-modal foundation AI model, Nova, which aims to compete head on with the likes of OpenA.

The News

AWS re:Invent 2024 showcased major innovations, including Trainium 2 (general availability) and Trainium 3 (announced) for AI training, which, according to the company, delivers 30–40% better price performance than competing GPUs. Generative AI and automation tools like QuickSight Q and Q Business aim to streamline workflows and democratize AI adoption. Additionally, AWS launched Q Developer for automating developer tasks and tools for transforming legacy workloads to cloud-native systems. Find out more about the key announcements here.

Analyst Take

Keeping up with the sheer volume of announcements at re:Invent was no small feat, as the company’s culture and service breadth leads to a hectic announcement driven re:Invent every year. At AWS re:Invent 2024, Matt Garman’s keynote (with a guest appearance by Amazon CEO Andy Jassy) attempted to position AWS as a leader in the race to define AI’s role in enterprise computing. These announcements emphasized AWS’s ability to balance innovation in its silicon and AI portfolio with the flexibility to support multi-vendor solutions. Here is HyperFRAME Research’s take on a curated list of the numerous announcements from the event.

Image source : AWS Press

Building Blocks - Computer, Storage and Security

It was refreshing for a keynote by a big tech vendor in 2024 to focus on the core product before diving into the hot topic of AI. In his keynote, Garman spent a surprisingly disproportionate amount of time focused on compute, storage and security before the keynote turned its attention to AI. One key takeaway from the focus on the building block section stated that in 2019 the whole of AWS was ~$35bn, and now AWS is able to attribute that same number to its own in house silicon based offerings with Nitro, Traininum and Inferentia.

Compute

AWS's recent announcements underscore its dual strategy of advancing in-house silicon while maintaining partnerships with the industry leader NVIDIA, an approach that aligns with what we are seeing among hyperscalers. The general availability of AWS Trainium2 powered Trn2 instances and Trn2 UltraServers marks a step in AWS's effort to continue to lead in the AI/ML infrastructure market. According to the company's own benchmarks, Trn2 instances deliver 30-40% better price performance than GPU-based EC2 instances, with a focus on training and deploying large language models (LLMs) and foundation models (FMs). The introduction of Trn2 UltraServers, which features 64 interconnected Trainium2 chips, demonstrates AWS's commitment to scaling AI workloads for trillion-parameter models and beyond. The number of clients that will need this scale is questionable, but those that do will be paying a huge bill for the privilege.

These announcements come alongside the unveiling of Trainium3 chips, expected in 2025, which, based on AWS provided benchmark data, will promise 4x the performance of Trn2 UltraServers. With Trainium2 and Neuron SDK, AWS is optimizing AI performance while integrating seamlessly with popular frameworks like PyTorch and JAX. Partnerships mentioned with the likes of Anthropic, Databricks, Hugging Face, and others further highlight AWS's hybrid approach of leveraging its silicon innovations while collaborating on ecosystem development.

We will need to see independent benchmarks from the likes of testing shops like Signal 65 to be able to stress test the benchmarks and provide comparative, but given that vendors are usually easier to believe with generation-to-generation comparisons rather than relative benchmarks, I am prone to take the AWS numbers on face value. As such, a 4x performance gain will be welcomed by clients, although it is unclear what the price premium will be for an increase in performance over the Tm2 offerings.

This mirrors a broader industry trend where hyperscalers like Google (TPU) and Microsoft increasingly develop proprietary silicon for AI while continuing to partner with NVIDIA for GPU-based solutions, balancing innovation, cost-efficiency, and flexibility to meet growing AI demands.

Storage

Storage, particularly EC, was one of the OG offerings for AWS back in the early days 18+ years ago. Garman took us down memory lane to share some of the growth of this core AWS offering; suffice to say the growth has been 'exponential' over that time. Amazon EC2 provides a range of instance types optimized for diverse workloads, including I/O-intensive storage tasks, machine learning training, and capacity reservation for critical events. Key recent announcements include storage-optimized instances, machine learning instances, and capacity reservations.

Amazon EC2 I7ie instances deliver up to 120 TB of low-latency NVMe storage and leverage the 5th generation Intel Xeon Scalable Processors and 3rd generation AWS Nitro SSDs. Compared to previous generations and based on in-house benchmark data provided, I7ie instances offer up to 65% better real-time storage performance per TB, 50% lower I/O latency, and 40% better compute performance. Available in nine sizes with up to 192 vCPUs and 1.5 TiB memory, these instances are ideal for NoSQL databases, analytics, and distributed file systems.

Amazon EC2 I8g instances introduce 22.5 TB of local NVMe storage with third-generation AWS Nitro SSDs and Graviton4 processors, delivering 65% better storage performance per TB, 60% lower latency variability, and 60% better compute performance than I4g instances. These instances target transactional databases, real-time analytics, and similar I/O-intensive workloads.

Amazon EC2 now supports future-dated Capacity Reservations (CRs), allowing AWS customers to schedule resources up to 120 days in advance. These CRs accommodate workloads requiring guaranteed capacity for critical events such as product launches and migrations. While the whole premise of cloud storage is its elastic nature, this reserve ahead ability will be welcomed by many customers.

These innovations demonstrate AWS's commitment to optimizing performance, cost, and flexibility for diverse workloads, and while I need to see third party benchmarks, the focus on continuing to drive innovation in what many would consider a less 'sexy' offering is to be applauded.

Security By Design

Garman went to great lengths to stress that everything AWS does starts and stops with Security. AWS's commitment to "secure by design" AI applications is timely, addressing the vulnerabilities that have emerged in the rapid adoption of generative AI. Initiatives like Bedrock's Automated Reasoning Check aim to mitigate risks like hallucination and ensure AI outputs align with customer expectations. To ram home the point about security being inherent in everything AWS does bringing JPMC to stage was a particular

Image source : AWS Press

Image source : AWS Press

AI Supercomputer

Amazon's Project Rainer marks a significant collaboration with Anthropic, an AI research lab and OpenAI competitor, that AWS has invested in. Announced at re:Invent conference, Project Rainer is set to feature hundreds of thousands of Amazon's Trainium 2 chips, designed specifically for AI training tasks and aims to build one of the most powerful AI supercomputers in the world. Once completed, the supercomputer will be five times larger than the infrastructure used to train Anthropic's most advanced models to date. Amazon claims this project will produce the largest reported AI machine globally, reflecting the scale and ambition of its push into generative AI. Are we in an arms race for this claim? You bet.

This collaboration aligns with Amazon's continued investment strategy in Anthropic, the company's commitment to advancing AI research and applications and, perhaps more crucially, being on the forefront of the AI hardware battleground. Anthropic, known for its work on safety-focused generative AI, stands to benefit from Amazon's infrastructure and expertise, while Amazon gains deeper integration into cutting-edge AI development, enhancing its competitiveness against rivals with deep-pockets such as; OpenAI, Microsoft, and Google.

The partnership is emblematic of a broader strategy to democratize AI capabilities, with Amazon providing Anthropic and other customers with the tools and infrastructure to train increasingly complex models. Beyond hardware, Project Rainer reflects a convergence of resources and expertise aimed at addressing the computational challenges posed by frontier AI development.

Garman's keynote emphasized the cost-efficiency and scale of Trainium 2-powered systems compared to traditional GPU-based clusters. This affordability, paired with the claimed high-performance capabilities of Trainium 2 and the forthcoming Trainium 3 chips, is expected to position AWS as a compelling alternative for enterprises developing large-scale AI models. Project Rainer not only highlights Amazon's technical innovation but also its role as a partner in advancing safe and scalable AI development. By collaborating with Anthropic, Amazon is accelerating progress in AI research while establishing itself as a foundational player in the generative AI ecosystem. It is too early to predict how this partnership based approach will play out against Google's TPU and Gemini andOpenAI and Microsoft, but AWS is certainly on the right track.

Apple and JPMC On The Mainstage

re:Invent 2024 saw Benoit Dupin, Apple's Senior Director of Machine Learning and AI, discuss Apple's extensive use of AWS services across products like iPad, Apple Music, Apple TV, News, App Store, and Siri. We have long suspected that Apple was a large AWS customer, but this is the first time it has been confirmed. Dupin highlighted that employing AWS's Graviton and Inferentia chips has led to over 40% efficiency gains in Apple's machine learning inference workloads compared to x86 instances. Additionally, Dupin noted that Apple is evaluating AWS's Trainium 2 chips and was expecting to see up to a 50% improvement in efficiency for pre-training its models.

Image source : AWS Press

To continue Garman's keynote at re:Invent 2024, Lori Beer, Global CIO of JPMorgan Chase, discussed the company's extensive use of AWS to modernize business platforms and drive innovation. She emphasized the development of an internal generative AI assistant, enhancing efficiency for 200,000 employees. Beer placed special focus on the fact that the firm's cloud architecture has been pivotal in unlocking generative AI use cases, enabling continuous innovation within the organization.

These tier one enterprises taking that stage, prove that AWS has long grown beyond its start up roots and is now firmly solidified as a big supplier to Fortune 500 companies, and is now the largest and most regulated within that cohort.

Generative AI

Unsurprisingly, generative AI remained a dominant theme. Agentic workflows, an area where AWS is looking to be an innovator, promise to redefine enterprise productivity by leveraging autonomous AI agents to execute tasks ranging from supply chain optimization to customer service. These workflows, paired with QuickSight Q, empower business users to interact with AI through natural language queries, lowering the technical barrier to entry.

Q Business and Amazon Q Developer represent AWS's broader push into enterprise automation. By mechanising business workflows and streamlining developer tasks (e.g., unit tests and code reviews), these offerings align with a growing market demand for efficiency tools that reduce costs and improve time-to-market. Additionally, the Amazon Q transformation initiatives for VMware, .NET to Linux, and mainframe workloads showcase AWS's ambition to capture legacy workloads while maintaining its commitment to cloud-native architectures. Mixed with announcements of Windows to Linux and VMware Migrations was the news that AWS is bringing Amazon Q to help with mainframe migrations.

AWS has introduced new generative AI–powered capabilities for its Amazon Q Developer tool, which is now available in public preview. According to AWS, these enhancements aim to accelerate the assessment and modernization of mainframe applications, providing enterprises with a streamlined web experience designed for large-scale transformation projects.

AWS highlights that Amazon Q Developer supports collaborative workflows with federated identity and natural language interaction. The tool employs generative AI to classify application assets, primarily COBOL, generate documentation, and create modernization plans tailored to specific business objectives. AWS asserts that these plans can be reviewed, adjusted, and approved iteratively by users, ensuring alignment with organizational goals.

Once approved, AWS claims that Amazon Q Developer autonomously refactors COBOL code into cloud-optimized Java while maintaining the integrity of business logic. The tool reportedly enables teams to scale modernization projects more effectively by automating tedious tasks, such as code analysis and documentation, through its generative AI agents.

AWS places emphasis on the importance of governance and compliance, stating that the tool provides a transparent trail of transformation decisions, fostering accountability and trust. By delegating routine tasks to AI agents, AWS hints that organizations can achieve faster project timelines, improved performance, and enhanced transformation quality.

I have written a whitepaper on the burgeoning proliferation of AI in the code discovery and explanation space of 2024, and AWS's move into this space to augment its M2 offering is going to be welcomed by GSI such as Kyndryl, Accenture and Deloitte while providing heartburn for the likes of AveriSource, who have competing offerings.

AWS gets into the Multimodal foundation models game with NOVA

Andy Jassey, now the Amazon overall CEO, took to the stage amidst obvious excitement from the crowd to drop the wider AI announcements. Amazon Nova introduces a new generation of foundation models on Amazon Bedrock, offering advanced capabilities for generative AI tasks with a focus on enterprise needs. Nova models are categorized into two groups: understanding models and creative content generation models.

Understanding models, including Nova Micro, Lite, Pro, and Premier, handle multimodal inputs like text, image, and video for tasks such as document analysis, visual question answering, and coding. These models support retrieval-augmented generation (RAG) and custom fine-tuning to meet specific business requirements. Creative content models like Nova Canvas and Reel enable high-quality image and video generation with detailed control over output style and features.

Key features of Amazon Nova include processing up to 300,000 tokens, multimodal fine-tuning, and robust safety measures like watermarking and content moderation. Integration with Amazon Bedrock allows seamless customization, deployment, and scaling while supporting over 200 languages and a wide range of business applications.

Compared to ChatGPT by OpenAI and Llama by Meta, Amazon Nova stands out in its multimodal capabilities, enterprise-focused features, and integration with AWS services. Nova offers scalability and higher token limits, surpassing ChatGPT's 32,000-token capacity, while falling short of Gemini, especially with the 2.0 release. ChatGPT is more accessible and conversational, and Llama focuses on research and open fine-tuning, but Nova targets enterprise workflows with tools tailored for customization and agentic applications. Lines are being drawn for AI across the industry, and AWS is now indisputably in the mix, if not leading in some areas

Looking Ahead

As I look to the future of a rapidly developing topic, AWS's announcements at re:Invent 2024 reflect its dual strategy of innovation in AI infrastructure, developer assistance through Amazon Q, and a further increase in the capability of its enterprise compute, and storage capabilities. The introduction of Trainium 2-powered Trn2 instances and Trn2 UltraServers stood out to me and showcases AWS's focus on enhancing AI training efficiency for large-scale models, while the unveiling of Trainium 3 promises a significant leap in performance by 2025 which will be sorely needed if the current trends continue. AWS's commitment to advancing proprietary silicon, alongside its ongoing partnership with NVIDIA, underscores its balanced approach to addressing diverse customer needs in AI/ML workloads.

On the enterprise automation front, Amazon Q Developer represents a pivotal step in addressing legacy modernization challenges. AWS is betting on generative AI-powered tools to simplify mainframe application transformation, offering enterprises a collaborative platform for analyzing, documenting, and refactoring code. By integrating governance and compliance capabilities, AWS positions Q Developer as a solution for streamlining complex modernization projects while maintaining trust and accountability. While HyperFRAME remains sceptical on the mainframe modernization and migration to the cloud strategy in general, AWS entering this space will certainly heighten the focus and spur a renewed focus by many enterprises.

AWS's broader AI initiatives, such as Nova foundation models on Bedrock, highlight its ambition to lead in multimodal AI capabilities tailored for enterprise workflows and cement the company's position on the battlefield for AI workloads. These developments, combined with partnerships like Project Rainer with Anthropic, signal AWS's intent to solidify its leadership in both AI innovation and enterprise adoption. Going forward, AWS's ability to deliver tangible value from these innovations will shape its competitive edge in the evolving AI landscape and will dictate the bottom line impact for Amazon overall.

ABOUT HYPERFRAME RESEARCH:

HyperFRAME Research delivers indepth research and insights across the global technology landscape, spanning everything from hyperscale public cloud to the mainframe and everything in between. We offer strategic advisory services, custom research reports, tailored consulting engagements, digital events, go to market planning, message testing, and lead generation programs.

Our industry analysts specialize in rigorous qualitative and quantitative assessments of technology solutions, business challenges, market forces, and end user demands across industry sectors. HyperFRAME Research collaborates closely with your Analyst Relations, Product, and Marketing teams to build and amplify your thought leadership, positioning your expertise to enhance brand and product recognition. Through content that engages readers, viewers, and listeners alike, we ensure your voice resonates across channels.

CONTACT HYPERFRAME RESEARCH:

Steven Dickens
CEO & Principal Analyst | HyperFRAME Research

Email Address:

steven.dickens@hyperframeresearch.com

Telephone Number:

+1 845 505 1678

X: - @StevenDickens3

LinkedIn: Steven Dickens

BlueSky: Steven Dickens

CONTRIBUTORS

Steven Dickens CEO & Principal Analyst
HyperFRAME Research

INQUIRIES

Contact us if you would like to discuss this report and HyperFRAME Research will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "HyperFRAME Research." Non-press and non-analysts must receive prior written permission by HyperFRAME Research for any citations

LICENSING

DISCLOSURES

HyperFRAME Research provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

COVER IMAGE SOURCE: AWS Press


HyperFRAME
RESEARCH