

AWS announces Trainium2 AI training chip, with Anthropic among early users

Source copy kept by wheresthe.ai for the story "AWS announces Trainium2 AI training chip, with Anthropic among early users" (<https://wheresthe.ai/d/aws-trainium2-graviton4-2023-11/>).

Source: <https://press.aboutamazon.com/2023/11/aws-unveils-next-generation-aws-designed-chips>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: 54fcce7bc1718004387942fa902e5d01266126eef3d13d0e248de11f3a67e277

AWS Graviton4 is the most powerful and energy-efficient AWS processor to date for a broad range of cloud workloads AWS Trainium2 will power the highest performance compute on AWS for training foundation models faster and at a lower cost, while using less energy Anthropic, Databricks, Datadog, Epic, Honeycomb , and SAP among customers using new AWS-designed chips

LAS VEGAS--(BUSINESS WIRE)-- At AWS re:Invent, Amazon Web Services, Inc. (AWS), an Amazon.com, Inc. company (NASDAQ: AMZN), today announced the next generation of two AWS-designed chip families—AWS Graviton4 and AWS Trainium2—delivering advancements in price performance and energy efficiency for a broad range of customer workloads, including machine learning (ML) training and generative artificial intelligence (AI) applications. Graviton4 and Trainium2 mark the latest innovations in chip design from AWS. With each successive generation of chip, AWS delivers better price performance and energy efficiency, giving customers even more options—in addition to chip/instance combinations featuring the latest chips from third parties like AMD, Intel, and NVIDIA—to run virtually any application or workload on Amazon Elastic Compute Cloud (Amazon EC2).

AWS Graviton4 and AWS Trainium2 (prototype) (Photo: Business Wire)

Graviton4 provides up to 30% better compute performance, 50% more cores, and 75% more memory bandwidth than current generation Graviton3 processors, delivering the best price performance and energy efficiency for a broad range of workloads running on Amazon EC2.

Trainium2 is designed to deliver up to 4x faster training than first generation Trainium chips and will be able to be deployed in EC2 UltraClusters of up to 100,000 chips, making it possible to train foundation models (FMs) and large language models (LLMs) in a fraction of the time, while improving energy efficiency up to 2x. "Silicon underpins every customer workload, making it a critical area of innovation for AWS," said David Brown, vice president of Compute and Networking at AWS. "By focusing our chip designs on real workloads that matter to customers, we're able to deliver the most advanced cloud infrastructure to them. Graviton4 marks the fourth generation we've delivered in just five years, and is the most powerful and energy efficient chip we have ever built for a broad range of workloads. And with the surge of interest in generative AI, Trainium2 will help customers train their ML models faster, at a lower cost, and with better energy efficiency." Graviton4 raises the bar on price performance and energy efficiency for a broad range of workloads Today, AWS offers more than 150 different Graviton-powered Amazon EC2 instance types globally at scale, has built more than 2 million Graviton processors, and has more than 50,000 customers—including the top 100 EC2 customers—using Graviton-based instances to achieve the best price performance for their applications. Customers including Datadog, DirecTV, Discovery, Formula 1 (F1), NextRoll, Nielsen, Pinterest, SAP, Snowflake, Sprinklr, Stripe, and Zendesk use Graviton-based instances to run a broad range of workloads, such as databases, analytics, web servers, batch processing, ad serving, application servers, and microservices. As customers bring larger in-memory databases and analytics workloads to the cloud, their compute, memory, storage, and networking requirements increase. As a result, they need even higher performance and larger instance sizes to run these demanding workloads, while managing costs. Furthermore, customers want more energy-efficient compute options for their workloads to reduce their impact on the environment. Graviton is supported by many AWS managed services, including Amazon Aurora, Amazon ElastiCache, Amazon EMR, Amazon MemoryDB, Amazon OpenSearch, Amazon Relational Database Service (Amazon RDS), AWS Fargate, and AWS Lambda, bringing Graviton's price performance benefits to users of those services. Graviton4 processors deliver up to 30% better compute performance, 50% more cores, and 75% more memory bandwidth than Graviton3. Graviton4 also raises the bar on security by fully encrypting all high-speed physical hardware interfaces. Graviton4 will be available in memory-optimized Amazon EC2 R8g instances, enabling customers to improve the execution of their high-performance databases, in-memory caches, and big data analytics workloads. R8g instances offer larger instance sizes with up

to 3x more vCPUs and 3x more memory than current generation R7g instances. This allows customers to process larger amounts of data, scale their workloads, improve time-to-results, and lower their total cost of ownership. Graviton4-powered R8g instances are available today in preview, with general availability planned in the coming months. To learn more about Graviton4-based R8g instances, visit aws.amazon.com/ec2/instance-types/r8g. EC2 UltraClusters of Trainium2 are designed to deliver the highest performance, most energy efficient AI model training infrastructure in the cloud. The FMs and LLMs behind today's emerging generative AI applications are trained on massive datasets. These models make it possible for customers to completely reimagine user experiences through the creation of a variety of new content, including text, audio, images, video, and even software code. The most advanced FMs and LLMs today range from hundreds of billions to trillions of parameters, requiring reliable high-performance compute capacity capable of scaling across tens of thousands of ML chips. AWS already provides the broadest and deepest choice of Amazon EC2 instances featuring ML chips, including the latest NVIDIA GPUs, Trainium, and Inferentia2. Today, customers including Databricks, Helixon, Money Forward, and the Amazon Search team use Trainium to train large-scale deep learning models, taking advantage of Trainium's high performance, scale, reliability, and low cost. But even with the fastest accelerated instances available today, customers want more performance and scale to train these increasingly sophisticated models faster, at a lower cost, while simultaneously reducing the amount of energy they use. Trainium2 chips are purpose-built for high performance training of FMs and LLMs with up to trillions of parameters. Trainium2 is designed to deliver up to 4x faster training performance and 3x more memory capacity compared to first generation Trainium chips, while improving energy efficiency (performance/watt) up to 2x. Trainium2 will be available in Amazon EC2 Trn2 instances, containing 16 Trainium chips in a single instance. Trn2 instances are intended to enable customers to scale up to 100,000 Trainium2 chips in next generation EC2 UltraClusters, interconnected with AWS Elastic Fabric Adapter (EFA) petabit-scale networking, delivering up to 65 exaflops of compute and giving customers on-demand access to supercomputer-class performance. With this level of scale, customers can train a 300-billion parameter LLM in weeks versus months. By delivering the highest scale-out ML training performance at significantly lower costs, Trn2 instances can help customers unlock and accelerate the next wave of advances in generative AI. To learn more about Trainium, visit aws.amazon.com/machine-learning/trainium/.

A leading advocate for the responsible deployment of generative AI, Anthropic is an AI safety and research company that creates reliable, interpretable, and steerable AI systems. An AWS customer since 2021, Anthropic recently launched Claude—an AI assistant focused on being helpful, harmless, and honest. "Since launching on Amazon Bedrock, Claude has seen rapid adoption from AWS customers," said Tom Brown, co-founder of Anthropic. "We are working closely with AWS to develop our future foundation models using Trainium chips. Trainium2 will help us build and train models at a very large scale, and we expect it to be at least 4x faster than first generation Trainium chips for some of our key workloads. Our collaboration with AWS will help organizations of all sizes unlock new possibilities, as they use Anthropic's state-of-the-art AI systems together with AWS's secure, reliable cloud technology." More than 10,000 organizations worldwide—including Comcast, Condé Nast, and over 50% of the Fortune 500—rely on Databricks to unify their data, analytics, and AI. "Thousands of customers have implemented Databricks on AWS, giving them the ability to use MosaicML to pre-train, finetune, and serve FMs for a variety of use cases," said Naveen Rao, vice president of Generative AI at Databricks. "AWS Trainium gives us the scale and high performance needed to train our Mosaic MPT models, and at a low cost. As we train our next generation Mosaic MPT models, Trainium2 will make it possible to build models even faster, allowing us to provide our customers unprecedented scale and performance so they can bring their own generative AI applications to market more rapidly." Datadog is an observability and security platform that provides full visibility across organizations. "At Datadog, we run tens of thousands of nodes, so balancing performance and cost effectiveness is extremely important. That's why we already run half of our Amazon EC2 fleet on Graviton," said Laurent Bernaille, principal engineer at Datadog. "Integrating Graviton4-based instances into our environment was seamless, and gave us an immediate performance boost out of the box, and we're looking forward to using Graviton4 when it becomes generally available." Epic is a leading interactive entertainment company and provider of 3D engine technology. Epic operates Fortnite, one of the world's largest games with over 350 million accounts and 2.5 billion friend connections. "AWS Graviton4 instances are the fastest EC2 instances we've ever tested, and they are delivering outstanding performance across our most competitive and latency sensitive workloads," said Roman Visintine, lead cloud engineer at Epic. "We look forward to using Graviton4 to improve player experience and expand what is possible within Fortnite." Honeycomb is the observability platform that enables engineering teams to find and solve problems they couldn't before. "We are thrilled to have evaluated AWS Graviton4-based R8g instances," said Liz Fong-Jones, Field CTO at Honeycomb. "In recent tests, our Go-based OpenTelemetry data ingestion workload required 25% fewer replicas on the Graviton4-based R8g instances compared to Graviton3-based C7g/M7g/R7g instances—and additionally achieved a 20% improvement in median latency and

10% improvement in 99th percentile latency. We look forward to leveraging Graviton4-based instances once they become generally available." SAP HANA Cloud, SAP's cloud-native in-memory database, is the data management foundation of SAP Business Technology Platform (SAP BTP). "Customers rely on SAP HANA Cloud to run their mission-critical business processes and next-generation intelligent data applications in the cloud," said Juergen Mueller, CTO and member of the Executive Board of SAP SE. "As part of the migration process of SAP HANA Cloud to AWS Graviton-based Amazon EC2 instances, we have already seen up to 35% better price performance for analytical workloads. In the coming months, we look forward to validating Graviton4, and the benefits it can bring to our joint customers." About Amazon Web Services Since 2006, Amazon Web Services has been the world's most comprehensive and broadly adopted cloud. AWS has been continually expanding its services to support virtually any workload, and it now has more than 240 fully featured services for compute, storage, databases, networking, analytics, machine learning and artificial intelligence (AI), Internet of Things (IoT), mobile, security, hybrid, virtual and augmented reality (VR and AR), media, and application development, deployment, and management from 102 Availability Zones within 32 geographic regions, with announced plans for 15 more Availability Zones and five more AWS Regions in Canada, Germany, Malaysia, New Zealand, and Thailand. Millions of customers—including the fastest-growing startups, largest enterprises, and leading government agencies—trust AWS to power their infrastructure, become more agile, and lower costs. To learn more about AWS, visit aws.amazon.com . About Amazon Amazon is guided by four principles: customer obsession rather than competitor focus, passion for invention, commitment to operational excellence, and long-term thinking. Amazon strives to be Earth's Most Customer-Centric Company, Earth's Best Employer, and Earth's Safest Place to Work. Customer reviews, 1-Click shopping, personalized recommendations, Prime, Fulfillment by Amazon, AWS, Kindle Direct Publishing, Kindle, Career Choice, Fire tablets, Fire TV, Amazon Echo, Alexa, Just Walk Out technology, Amazon Studios, and The Climate Pledge are some of the things pioneered by Amazon. For more information, visit amazon.com/about and follow @AmazonNews .

Amazon.com, Inc. Media Hotline Amazon-pr@amazon.com www.amazon.com/pr

Source: Amazon.com, Inc.