AWS Announces Three Amazon EC2 Instances Powered by New AWS-Designed Chips

Hpc7g instances featuring new AWS Graviton3E chips deliver the best price performance for HPC workloads on Amazon EC2

C7gn instances featuring new AWS Nitro Cards with enhanced networking offer the highest network bandwidth and packet rate performance across Amazon EC2 network-optimized instances

Inf2 instances powered by new AWS Inferentia2 chips deliver the lowest latency at the lowest cost on Amazon EC2 for running the largest deep learning models at scale

LAS VEGAS—Nov. 29, 2022—At AWS re:Invent, Amazon Web Services, Inc. (AWS), an Amazon.com, Inc. company (NASDAQ: AMZN), today announced three new Amazon Elastic Compute Cloud (Amazon EC2) instances powered by three new AWS-designed chips that offer customers even greater compute performance at a lower cost for a broad range of workloads. Hpc7g instances, powered by new AWS Graviton3E chips, offer up to 2x better floating-point performance compared to current generation C6gn instances and up to 20% higher performance compared to current generation Hpc6a instances, delivering the best price performance for high performance computing (HPC) workloads on AWS. C7gn instances, featuring new AWS Nitro Cards, offer up to 2x the network bandwidth and up to 50% higher packet-processing-per-second performance compared to current generation networking-optimized instances, delivering the highest network bandwidth, the highest packet rate performance, and the best price performance for network-intensive workloads. Inf2 instances, powered by new AWS Inferentia2 chips, are purpose built to run the largest deep learning models with up to 175 billion parameters and offer up to 4x the throughput and up to 10x lower latency compared to current-generation Inf1 instances, delivering the lowest latency at the lowest cost for machine learning (ML) inference on Amazon EC2.

AWS has a decade of experience designing chips developed for performance and scalability in the cloud at a lower cost. In that time, AWS has introduced specialized chip designs, which make it possible for customers to run even more demanding workloads with varying characteristics that require faster processing, higher memory capacity, faster storage input/output (I/O) and increased networking bandwidth. Since the introduction of the AWS Nitro System in 2013, AWS has developed multiple AWS-designed silicon innovations, including five generations of the Nitro System, three generations of Graviton chips optimized for performance and cost for a wide range of workloads, two generations of Inferentia chips for ML inference, and Trainium chips for ML training. AWS uses cloud-based electronic design automation as part of an agile development cycle for the design and verification of AWS-designed silicon, enabling teams to innovate faster and make chips available to customers more quickly. AWS has demonstrated that it can deliver a new chip based on a more modern, power-efficient silicon process at a predictable and rapid pace. With each successive chip, AWS delivers a step function improvement in performance, cost, and efficiency to the Amazon EC2 instances hosting them, giving customers even more choice of chip and instance combinations optimized for their unique workload requirements.

“Each generation of AWS-designed silicon—from Graviton to Trainium and Inferentia chips to Nitro Cards—offers increasing levels of performance, lower cost, and power efficiency for a diverse range of customer workloads,” said David Brown, vice president of Amazon EC2 at AWS. “That consistent delivery, combined with our customers’ abilities to achieve superior price performance using AWS silicon, drives our continued innovation. The Amazon EC2 instances we’re introducing today offer significant improvements for HPC, network-intensive, and ML inference workloads, giving customers even more instances to choose from to meet their specific needs."

Hpc7g instances are purpose built to offer the best price performance for running HPC workloads at scale on Amazon EC2

Organizations across numerous sectors rely on HPC to solve their most complex academic, scientific, and business problems. Today, customers like AstraZeneca, Formula 1, and Maxar Technologies run conventional HPC workloads like genomics processing, computational fluid dynamics (CFD), and weather forecasting simulations on AWS to take advantage of the superior security, scalability, and elasticity it offers. Engineers, researchers, and scientists run their HPC workloads on Amazon EC2 network-optimized instances (e.g., C5n, R5n, M5n, and C6gn) that deliver virtually unlimited compute capacity and high levels of network bandwidth between servers that process and exchange data across thousands of cores. While the performance of these instances is sufficient for most HPC use cases today, emerging applications such as artificial intelligence (AI) and autonomous vehicles require HPC-optimized instances that can further scale to solve increasingly difficult problems and reduce the cost of HPC workloads, which can scale to tens of thousands of cores or more.

Hpc7g instances powered by new AWS Graviton3E processors offer the best price performance for customers' HPC workloads (e.g., CFD, weather simulations, genomics, and molecular dynamics) on Amazon EC2. Hpc7g instances provide up to 2x better floating-point performance compared to current generation C6gn instances powered by Graviton2 processors and up to 20% higher performance compared to current generation Hpc6a instances, enabling customers to carry out complex calculations across HPC clusters up to tens of thousands of cores. Hpc7g instances also provide high-memory bandwidth and 200 Gbps of Elastic Fabric Adapter (EFA) network bandwidth to achieve faster time to results for HPC applications. Customers can use Hpc7g instances with AWS ParallelCluster, an open-source cluster management tool, to provision Hpc7g instances alongside other instance types, giving customers the flexibility to run different workload types within the same HPC cluster. For more information on Hpc7g instances, visit aws.amazon.com/ec2/instance-types/hpc7g.

C7gn instances offer the best performance for network-intensive workloads with higher networking bandwidth, greater packet rate performance, and lower latency

Customers use Amazon EC2 network-optimized instances to run their most demanding network-intensive workloads like network virtual appliances (e.g., firewalls, virtual routers, and load balancers) and data encryption. Customers need to scale the performance of these workloads to handle increasing network traffic in response to spikes in activity, or to decrease processing time to deliver a better experience to their end users. Today, customers use larger instance sizes to get more network throughput, deploying more compute resources than required, which increases costs. These customers need increased packet-per-second performance, higher network bandwidth, and faster cryptographic performance to reduce data processing times.

C7gn instances, featuring new AWS Nitro Cards powered by new, fifth generation Nitro chips with network acceleration, offer the highest network bandwidth and packet-processing performance across Amazon EC2 network-optimized instances, while using less power. Nitro Cards offload and accelerate I/Ofor functions from the host CPU to specialized hardware to deliver practically all of an Amazon EC2 instance's resources to customer workloads for more consistent performance with lower CPU utilization. New AWS Nitro Cards enable C7gn instances to offer up to 2x the network bandwidth and up to 50%

higher packet-processing-per-second performance, and reduced Elastic Fabric Adapter (EFA) network latency compared to current generation networking-optimized Amazon EC2 instances. C7gn instances deliver up to 25% better compute performance and up to 2x faster performance for cryptographic workloads compared to C6gn instances. Fifth generation Nitro Cards also offer 40% better performance per watt compared to fourth generation Nitro Cards, lowering power consumption for customer workloads. C7gn instances let customers scale for both performance and throughput and reduced network latency to optimize the cost of their most demanding, network-intensive workloads on Amazon EC2. C7gn instances are available today in preview. To learn more about C7gn instances, visit aws.amazon.com/ec2/instance-types/c7gn.

Inf2 instances are purpose-built to serve today's most demanding deep learning model deployments, with support for distributed inference and stochastic rounding

In response to demand for better applications and even more tailored personalized experiences, data scientists and ML engineers are building larger, more complex deep learning models. For example, large language models (LLMs) with more than 100 billion parameters are increasingly prevalent, but they train on enormous amounts of data, driving unprecedented growth in compute requirements. While training receives a lot of attention, inference accounts for the majority of complexity and cost (i.e., for every $1 spent on training, up to $9 is spent on inference) of running machine learning in production, which can limit its use and stall customer innovation. Customers want to use state-of-the-art deep learning models in their applications at scale, but they are constrained by high compute costs. When AWS launched Inf1 instances in 2019, deep learning models were millions of parameters. Since then, the size and complexity of deep learning models have grown exponentially with some deep learning models exceeding hundreds of billions of parameters—a 500x increase. Customers working on next-generation applications using the latest advancements in deep learning want cost-effective, energy-efficient hardware that supports low latency, high throughput inference, with flexible software that enables engineering teams to quickly deploy their latest innovations at scale.

Inf2 instances, powered by new Inferentia2 chips, support large deep learning models (e.g., LLMs, image generation, and automated speech detection) with up to 175 billion parameters, while delivering the lowest cost per inference on Amazon EC2. Inf2 is the first inference-optimized Amazon EC2 instance that supports distributed inference, a technique that spreads large models across several chips to deliver the best performance for deep learning models with more than 100 billion parameters. Inf2 instances support stochastic rounding, a way of rounding probabilistically that enables high performance and higher accuracy as compared to legacy rounding modes. Inf2 instances support a wide range of data types including CFP8, which improves throughput and reduces power per inference, and FP32, which boosts performance of modules that have not yet taken advantage of lower precision data types. Customers can get started with Inf2 instances using AWS Neuron, the unified software development kit (SDK) for ML inference. AWS Neuron is integrated with popular ML frameworks like PyTorch and TensorFlow to help customers deploy their existing models to Inf2 instances with minimal code changes. Since splitting large models across several chips requires fast inter-chip communication, Inf2 instances support AWS's high-speed, intra-instance interconnect, NeuronLink, offering 192 GB/s of ring connectivity. Inf2 instances offer up to 4x the throughput and up to 10x lower latency compared to current-generation Inf1 instances, and they also offer up to 45% better performance per watt compared to GPU-based instances. Inf2 instances are available today in preview. To learn more about Inf2 instances, visit aws.amazon.com/ec2/instance-types/inf2.

The Water Institute is an independent, non-profit applied research organization that works across disciplines to advance science and develop integrated methods used to solve complex environmental and societal challenges. “The ability to make accurate, near-real-time numerical weather predictions to aid decision making is important to our clients. We’re excited to see Amazon EC2’s high performance computing offerings continue to evolve with the launch of Amazon EC2 Hpc7g instances,” said Zach Cobell, research engineer at The Water Institute. “With increased floating-point performance, higher efficiency using AWS Graviton3E processors, based on Arm architecture, and decreased inter-node latency using Elastic Fabric Adapter, we expect to continue to be able to deliver innovative and sustainable solutions across our computational portfolio.”

Arup is a global collective of designers, engineering and sustainability consultants, advisors and experts dedicated to sustainable development and to using imagination, technology and rigour to shape a better world. “We use AWS to run highly complex simulations to help our customers to build the next generation of high-rise buildings, stadiums, data-centres, and crucial infrastructure, along with assessing and providing insight into urban microclimates, global warming, and climate change that impacts the lives of so many people around the world,” said Dr. Sina Hassanli, senior engineer at Arup. “Our customers are constantly demanding faster, more accurate simulations at a lower cost to inform their designs at the early stages of development, and we are already anticipating how the introduction of Amazon EC2 Hpc7g instances with higher performance will help our customers innovate faster and more efficiently.”

HAProxy Technologies is the company behind HAProxy, the world’s fastest and most widely-used software load balancer. "HAProxy powers modern application delivery at any scale and in any environment, providing the utmost performance, observability, and security for some of the most popular websites in the world,” said Willy Tarreau, lead developer at HAProxy. “When HAProxy tested Amazon EC2 C6gn instances, we found unprecedented performance for a software load balancer. We are excited about the new C7gn instances with Graviton3E and fifth generation AWS Nitro Cards and the networking performance improvements they will bring to our customers.”

Aerospike Inc.'s real-time data platform is designed for organizations to build applications that fight fraud, enable global digital payments, deliver hyper-personalized user experiences to tens of millions of customers, and more. “The Aerospike Real-time Data Platform is a shared-nothing, multithreaded, multimodal data platform designed to operate efficiently on a cluster of server nodes, exploiting modern hardware and network technologies to drive reliably fast performance at sub-millisecond speeds across petabytes of data,” said Lenley Hensarling, chief product officer at Aerospike. “In our recent real-time database read tests, we were pleased to see a significant improvement in transactions per second on Amazon EC2 C7gn instances featuring new AWS Nitro Cards compared to C6gn instances. We look forward to taking advantage of C7gn instances and future AWS infrastructure improvements as they become available.”

Qualtrics designs and develops experience management software. “At Qualtrics, our focus is building technology that closes experience gaps for customers, employees, brands, and products. To achieve that, we are developing complex multi-task, multi-modal deep learning models to launch new features, such as text classification, sequence tagging, discourse analysis, key-phrase extraction, topic extraction, clustering, and end-to-end conversation understanding,” said Aaron Colak, head of Core Machine Learning at Qualtrics. “As we utilize these more complex models in more applications, the volume of unstructured data grows, and we need more performant inference-optimized solutions that can meet these demands, such as Inf2 instances, to deliver the best experiences to our customers. We are excited about the new Inf2 instances, because it will not only allow us to achieve higher throughputs, while dramatically cutting latency, but also introduces features like distributed inference and enhanced dynamic input shape support, which will help us scale to meet the deployment needs as we push towards larger, more complex large models."

Finch Computing is a natural language technology company providing artificial intelligence applications for government, financial services, and data integrator clients. "To meet our customers' needs for real-time natural language processing, we develop state-of-the-art deep learning models that scale to large production workloads. We have to provide low-latency transactions and achieve high throughputs to process global data feeds. We already migrated many production workloads to Inf1 instances and achieved an 80% reduction in cost over GPUs," said Franz Weckesser, chief architect at Finch Computing. "Now, we are developing larger, more complex models that enable deeper, more insightful meaning from written text. A lot of our customers need access to these insights in real-time and the performance on Inf2 instances will help us deliver lower latency and higher throughput over Inf1. With the Inf2 performance improvements and new Inf2 features, such as support for dynamic input sizes, we are improving our cost-efficiency, elevating the real-time customer experience, and helping our customers glean new insights from their data."

About Amazon Web Services
For over 15 years, Amazon Web Services has been the world's most comprehensive and broadly adopted cloud offering. AWS has been continually expanding its services to support virtually any cloud workload, and it now has more than 200 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 96 Availability Zones within 30 geographic regions, with announced plans for 15 more Availability Zones and five more AWS Regions in Australia, Canada, Israel, 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