Driving energy efficiency for HPC workloads with accelerated computing from AWS and NVIDIA

The momentum towards a clean energy economy is accelerating. Increasing global demands for energy and mandates to lower carbon emissions are making organizations identify the most energy efficient ways to run high-performance workloads.

Over the last two years, global demand for electricity increased by six percent in 2021 and 2.4 percent in 2022[1], reaching the highest increase on record. To meet growing demands, countries had to generate power with coal, which prompted coal-fired power generation to continue its record-breaking streak for a second year in a row to around 10,400 TWh.[2]

At the same time, artificial intelligence (AI) and machine learning (ML) models are advancing and becoming more complex. Generally, more complex models tend to have a higher number of parameters sometimes reaching into the billions or even trillions. In the last ten years, the number of parameters has shown an increase of 10,000 times.[3] While these large models can achieve remarkable performance results, they come with a significant computational demand.

As high performance computing (HPC) and AI workloads continue to grow in complexity and size, they require increasingly large computational resources that use accelerated computing.

As a result, organizations are looking for an energy-efficient approach to run HPC workloads at scale with less total energy consumption.

Key challenges faced in running AI and HPC workloads

Rising energy costs

Organizations operating data centers on-premises are faced with high energy costs for cooling, electricity, and other infrastructure. The energy cost to power a single server rack in a data center in the US can be as high as almost $30,000 a year.[4] Data centers and data transmission networks can consume an estimated 2 percent of worldwide electricity[5], a number projected to double by 2030.[6] Organizations are looking to reduce electricity used in data centers while running HPC simulations and building AI/ML models at scale.

Limited power allocation

Increasing global demands for energy and mandates to lower carbon emissions could reduce power allocated to data centers. Subsequently, power constraints can cause interruptions in data center operations and lead to delays. Therefore, IT teams need to find a way to ensure sufficient power to support key HPC workloads and applications.

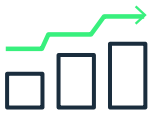
Growing complexity of AI and HPC workloads

Training massive AI models and running large-scale HPC simulations consume large amounts of energy. The increase in complexity of AI models and HPC workloads, along with the rising number of government regulations regarding carbon emissions is driving organizations to rethink their approach to energy sourcing and costs. According to the International Energy Agency (IEA), data center workloads account for almost two percent of global energy.[7] One of the major factors driving the cost of training ML models is the need for more computing power. Thus, there is a need for smarter workflow approaches able to use HPC in a more energy-efficient way.

Realizing energy efficiency with AWS and NVIDIA

Cloud efficiency

Amazon Web Services (AWS) is focused on improving efficiency in every aspect of its infrastructure, from innovative server designs to cooling solutions. In fact, 90 percent of the electricity consumed by Amazon in 2022 was attributable to renewable energy sources, and the organization remains on a path to 100 percent by 2025.[8] Transitioning to renewable energy is one of the highest-impact ways to immediately lower emissions. With the broadest energy partner ecosystem, AWS empowers organizations to improve performance, accelerate innovation, and minimize their carbon footprint.



Shortening time to insight

Organizations can reduce energy costs by improving the efficiency of physical resources used for compute, networking, and storage through virtualized servers and secured infrastructure. Amazon Elastic Compute Cloud (Amazon EC2) instances, powered by NVIDIA GPUs, enable accelerated computing for faster time to results. On top of highly-performant GPUs, NVIDIA also offers a portfolio of software libraries dedicated to AI training, inferencing, visualization, and data processing to help reduce the time needed to complete a job and reduce overall energy consumption for comparable HPC workloads.

Scaling workloads

Organizations can overcome limited on-premises power capacity by scaling out HPC and AI/ML workloads to run on thousands of NVIDIA GPUs on AWS. Accelerated computing in the cloud, with low-latency and high-bandwidth networking, is lowering the amount of time taken to compute. Therefore, HPC teams can perform more computation for typically the same energy unit and in the same timeframe which saves organizations from consuming their power allocation and frees resources for additional compute needs.

The advantages of AWS and NVIDIA

Cloud computing from AWS, powered by NVIDIA GPUs, along with a full-stack AI computing platform, purpose-built HPC tools and services allow scientists and engineers to access faster compute, storage, and networking solutions to increase the speed of their workloads and perform more in typically the same amount of time.

Build on energy-efficient infrastructure

Amazon is the largest corporate purchaser of renewable energy in the world and AWS infrastructure is 3.6 times more energy efficient than the median of the surveyed US organizations data centers and up to five times more energy efficient than the average data center surveyed in Europe.[9]

Lower and measure carbon footprint

By moving on-premises computing workloads to the cloud, AWS can help lower workload carbon footprints by nearly 80 percent compared to surveyed organization data centers.[10] Customers can access the Customer Carbon Footprint Tool on AWS to measure the estimated carbon emissions of using AWS services.

Increase scalability with the cloud

By moving all or some of their HPC and AI/ML workloads to the cloud, organizations can take advantage of advanced technologies, such as the latest NVIDIA GPUs and NVIDIA AI Enterprise, to accelerate workloads using AWS' global reach and pay-as-you-go infrastructure to scale up or down as demand requires.

Optimize AI/ML workload efficiency

Engineers can use model training services from AWS and NVIDIA, and automatically launch training instances, shutting them down as soon as the training job is complete. This minimizes idle compute resources and can improve energy efficiency.

Align with cloud best practices

The AWS Well-Architected Framework helps organizations understand the benefits of building systems in the cloud. The six pillars of the framework allow HPC users to learn architectural best practices for building secure, high-performing, resilient, and efficient infrastructure for a wide range of applications and workloads.

Key solutions

Amazon EC2 P5 instances, powered by NVIDIA H100 Tensor Core GPUs, are purpose-built to accommodate growing HPC and AI demands while lowering the cost to run them. The high performing H100 powered P5 instances with 80 GB memory per GPU and high speed Elastic Fabric Adapter (EFA) deliver fast networking bandwidth enabling organizations to train large AI models faster, for typically the same amount of energy.

AWS Batch allows developers, scientists, and engineers to efficiently run hundreds of thousands ML computing jobs and optimize compute resources.

AWS ParallelCluster helps quickly build HPC compute environments, ensuring simpler deployment and management of HPC clusters.

NVIDIA AI Enterprise on AWS Marketplace provides a production-ready, containerized stack, for organizations to build, fine-tune, train, and deploy AI models faster. NVIDIA AI Enterprise includes NVIDIA NeMo—an end-to-end cloud framework, and NVIDIA AI Workbench—a toolkit that allows developers to quickly create, test, and customize pretrained generative AI and large language models (LLMs).

NVIDIA RAPIDS Accelerator for Apache Spark speeds up data processing time.

PyTorch, TensorFlow and the NVIDIA TAO Toolkit help shorten model training times.

NVIDIA TensorRT helps accelerate application performance up to 40 times.[11]

NVIDIA Triton Inference Server and NVIDIA Triton Management Service simplifies and optimizes the deployment of AI models at scale, aiding lower power consumption.

NVIDIA HPC SDK is a comprehensive suite of compilers, libraries, and tools that helps maximize developer productivity and the performance and portability of HPC applications.

How key industries are benefiting from AWS and NVIDIA

Automotive

ML is one of the biggest workloads in the development of autonomous mobility (AM). Automotive companies develop a large variety of AI models to solve problems in data curation, mapping, perception, prediction, and planning, and are moving to larger models to support new use cases, thereby increasing their compute and storage consumption. AM development also requires simulating billions of miles of driving across a multitude of scenarios to be deemed safe. In addition, high-fidelity computer aided engineering physics simulations are used in product development. High performing GPUs, fast storage, high speed networking, and efficient batch processing reduce the time associated with these complex computations.

Financial Services

Financial institutions are scaling up HPC and AI/ ML on AWS to address the rise in number of risk assessments, financial modeling exercises, and portfolio optimizations due to ever increasing regulatory and business demands. In addition, real-time access to data analytics and AI/ ML enables financial firms to improve customer experience by reducing time in loan application and customer services responses; reduce operational expenses with fraud detection and streamlined document processing; as well as expand customer revenues with tailored product recommendations. Using EC2 instances, powered by NVIDIA GPUs, organizations can optimize data processing in real-time, feed into sophisticated ML models, and gain insights faster.

Energy

Scientists and engineers working with high-fidelity, 3D geophysics visualizations for reservoir simulation and seismic processing can use EC2 instances, powered by NVIDIA GPUs. This scalable HPC infrastructure and AI/ML tools can fine-tune models faster.

Public Sector

With easy-to-manage hardware and software tools from AWS and NVIDIA, government institutions can build secure infrastructure to run large-scale HPC and AI workloads.

Healthcare and Life Sciences

Scientists and researchers can run large-scale HPC simulations and train large models for drug discovery more efficiently. Hospitals, pharmaceutical companies, biotech firms, and healthcare providers are faced with 10 year drug development cycles[12] to get a new drug from phase 1 to regulatory approval and $1 billion in R&D costs.[13] NVIDIA GPU-accelerated tools and services running on AWS reduce data processing times and accelerate genomic sequencing, resulting in shorter drug development time.

Industrial Manufacturing

Engineers can run computational fluid dynamics (CFD) simulations needed to optimize product design with high throughput and low latency. With HPC and AI/ML tools and services from AWS and NVIDIA, they can run high-fidelity simulations faster, reducing total compute demands.

Final thoughts

Organizations around the world are using AWS and NVIDIA to run their HPC and AI/ML workloads. Powered by the latest NVIDIA GPUs and accelerated by various tools and services, organizations can gain insights faster while improving energy efficiency.

Learn more about running HPC workloads on AWS ›

[1] "Electricity Market Report", iea.org, 2022
[2] "Coal", iea.org, 2022
[3] "Parameters in notable artificial intelligence systems," Our World in Data, 2023
[4] "How Much Does It Cost to Power One Rack in a Data Center?" Sunbird, 2020
[5] "Data Centres and Data Transmission Networks", iea.org, 2022
[6] "Usage impact on data center electricity needs: A system dynamic forecasting mod-el", Science Direct, 2021
[7] "Data Centres and Data Transmission Networks", iea.org, 2022
[8] "Renewable Energy" Report, Amazon Web Services, 2022
[9] "Delivering Progress Every Day, Amazon's 2021 Sustainability Report," Amazon Web Services, 2021
[10] "Delivering Progress Every Day, Amazon's 2021 Sustainability Report," Amazon Web Services, 2021
[11] "Host ML models on Amazon SageMaker using Triton: TensorRT models," Amazon Web Services, 2023
[12] Highlights from the AWS Life Sciences Executive Symposium 2023: Accelerating Pharma Drug Discovery with ML and Generative AI, Amazon Web Services, 2023
[13] Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018, Report, National Library of Medicine, 2020