White Paper

Energy Efficiency is Driving Many HPC Users to the Cloud

Sponsored by: AWS and NVIDIA

Jaclyn Ludema and Mark Nossokoff
March 2024

HYPERION RESEARCH OPINION

Recent years have witnessed a convergence of global events and evolving trends in the high performance computing (HPC) market, propelling energy efficiency to the forefront of data center priorities for many sites. Pivotal factors driving this shift to energy efficiency prioritization include a growing awareness of climate change and its severe repercussions, the rising cost of energy across the globe, and the increasing demand for HPC resources for scientific research, simulations, artificial intelligence (AI), and large language models (LLMs). As the impact of these factors on data center energy usage becomes increasingly pronounced in budgets and consumption requirements, both organizations and regulators are advocating for more energy-efficient computing solutions.

HPC use cases are growing in number and computational intensity, with new industrial users looking to use HPC to address their computational challenges in areas like AI. The goal of providing these new and existing end users with energy-efficient HPC resources has challenged data center managers, with some opting to look outside their on-premises infrastructure for a solution. A growing number of data center operators are turning to cloud service providers (CSPs) to address their need for energy-efficient computing to run their most energy-intensive workloads. With the proper expertise and guidance on building, configuring, and operating cloud resources, organizations can leverage the energy-efficient compute offerings of CSPs to supplement or even replace on-premises data center capacity, and see substantial reductions in the energy used for certain workloads.

Cloud energy efficiency is a collective effort between CSPs and their customers. The onus on CSPs involves the design and construction of energy-efficient cloud infrastructure. This entails the establishment of robust and streamlined architectures, offering a diverse array of options to cater to the needs of every customer, for a wide variety of workloads, and providing access to the latest technology. It also includes locating resources in areas with lower carbon footprint energy sources. Additionally, many CSPs provide tools that empower users to manage their cloud resources more efficiently. On the customer's end, the responsibility for promoting energy efficiency lies in the strategic development of workloads that minimize overall resource requirements and adopting practices that optimize resource utilization.

Moreover, the forefront of GPU innovation is often witnessed in cloud environments, where CSPs deploy cutting-edge GPUs before they become widely available in traditional computing setups. Cloud users have early access to the most energy-efficient GPU solutions, empowering them to drive energy-efficient computing practices. This transition not only accelerates the execution of complex workloads but also contributes to reduced energy consumption per computation, as GPUs excel at parallelizing many tasks, enabling faster completion times.

OVERVIEW OF THE CURRENT ENERGY DEMANDS ON HPC DATA CENTERS

Data centers are encountering significant challenges as they strive to scale their energy infrastructure to accommodate the escalating demand for HPC. The demand is growing exponentially, as HPC projects generate more data than ever before, and more HPC resources are then required to extract valuable information from the data. For instance, the use of HPC for genome research has exploded since the completion of the Human Genome Project in 2003. At the time, sequencing the human genome took $1 billion and 13 years to complete. Now, using Next-generation Sequencing (NGS), which uses HPC parallel file processing, the human genome can be sequenced for about $1,000 in less than 2 days. This advancement has resulted in a staggering increase of 15,000 times more data generated per day over the last two decades. By utilizing HPC and these massive genome datasets, genomic data scientists are working to decipher the information they contain to better understand human health and disease. The National Human Genome Research Institute estimates that genomic research will generate between 2 and 40 exabytes of data by 2025.

The demand for new, higher-resolution modeling and simulation further amplifies the energy needs of HPC data centers. As simulations become more intricate and detailed, the computational requirements escalate, necessitating continuous advancements in hardware and infrastructure.

Furthermore, the training of AI and LLMs involves processing massive datasets, leading to prolonged compute times and heightened energy usage. The adoption of AI at HPC sites has become ubiquitous. In 2022, 95 percent of HPC sites surveyed in the Hyperion Research multi-client study indicated utilizing at least one AI methodology in their data center, reflecting the proliferation of AI and high performance data analytics (HPDA) into the HPC ecosystem. Training AI and LLMs has become more data intensive as well. Over the last decade, there has been a remarkable 10,000 times increase in the size of LLMs, shifting from AlexNet with 60 million parameters to PaLM with 540 billion parameters in 2022. This surge in parameters drives higher computational demands, and greater memory needs, resulting in even higher energy consumption.

HPC sites are actively exploring strategies such as hardware optimizations, algorithmic efficiency improvements, and the adoption of energy-efficient technologies to address the growing demands of these new and expanding data-intensive workloads, all without sacrificing accuracy of results.

CONCERNS FOR POWER-CONSTRAINED DATA CENTERS

The concerns for power-constrained data centers encompass a broad spectrum, from financial challenges due to energy price increase and fluctuations to competition for reliable energy sources, heat management complexities, and evolving regulatory landscapes. Addressing these concerns requires a holistic and adaptive approach to ensure the sustainability and efficiency of HPC data center operations.

Fluctuations in global energy prices pose a significant challenge for HPC data centers, as spikes can exert strain on budgets and affect the affordability of running energy-intensive operations. The rapid growth in HPC and compute-intensive workloads handled by large data centers has led to a substantial increase in energy use, growing by an annual rate of 20-40%, as reported in the IEA Data Centers and Data Transmission Networks Report.

The escalating global demand for energy resources introduces potential competition for access to reliable energy sources for HPC data centers. According to the International Energy Agency,

"*Estimated global data center electricity consumption in 2022 was 240-340 TWh, or around 1-1.3% of global final electricity demand*." Global data center energy use is expected to grow moderately throughout the next few years, however, for certain countries and regions looking to increase their data center markets rapidly, energy use and availability are becoming increasing concerns. For example, Denmark data center energy use is projected to rise by six times by 2030 and account for 15% of the country's electricity use.

The impact of energy availability on expanding data center markets is already a concern for some regions. A real-world example of this is a US-based utility, which in 2022 postponed the construction of a billion-dollar data center in Ashburn, Virginia, due to an inability to guarantee the amount of electricity needed for operations. This issue didn't stem from a lack of power generation capabilities but rather from challenges in distributing electricity over high-voltage power lines to Ashburn.

As HPC systems become more powerful and introduce new energy-hungry computing elements, systems are becoming more challenging to cool at a rack level due to the increased energy density, necessitating advanced cooling solutions. Ensuring effective heat management without compromising system performance is a critical concern for the future. While liquid cooling is widely considered to be an energy-efficient method of heat management, it adds a new layer of resource usage concerns due to the amount of freshwater required. The dynamics of energy and water usage in data centers become a balancing act, wherein the usage of water is inescapable. Power plant cooling contributes significantly to freshwater withdrawals, accounting for 43 percent of total freshwater withdrawals in Europe and nearly 50 percent in the USA, according to UN-Water. This underscores the environmental impact of HPC energy usage as well as HPC cooling technology and emphasizes the need for energy efficient computing solutions.

Another concern for businesses is that evolving energy regulations and environmental policies may introduce new compliance requirements for HPC data centers. Remaining in compliance with these changing regulations is a concern for the future, not only to meet legal obligations but also to navigate potential operational challenges associated with evolving standards.

TRENDS IN CLOUD MIGRATION FOR HPC ENERGY EFFICIENCY

The HPC market is witnessing a notable shift towards cloud adoption, partially driven by the benefits of enhanced energy efficiency. Organizations are increasingly recognizing that migrating HPC workloads to the cloud has the potential to lower operating costs for certain workload types. For some organizations and workloads, cloud platforms offer the opportunity to optimize energy consumption, leverage shared resources, and benefit from the economies of scale inherent in cloud infrastructure more easily than on-premises. The potential reduction in energy usage can both provide cost savings and alignment with the growing emphasis on sustainable and environmentally conscious computing practices.

The potential benefits of moving to the cloud are enticing to many different workload types. In a recent Hyperion Research study, respondents using cloud resources were most likely to distribute AI (ML, DL, etc.) jobs to the cloud with a plurality of 40 percent. Cloud platforms have characteristics that lend themselves strongly to data-intensive workloads, which, when combined with the expertise of CSPs in the AI space, provide a strong platform for HPC users to deploy AI workloads in an energy conscious manner.

TABLE 1

Current Cloud Workload Distribution

Q: Distribute your current cloud computing cycles among the following categories:

	Overall Average
% Traditional Mod/Sim workloads:	28.9%
% AI (ML, DL, etc.):	40.0%
% Big Data/HPDA (excluding AI/ML/DL):	22.3%
% Quantum	3.4%
% Other	5.4%

n = 98

Source: Hyperion Research, 2023

In addition to potential cost savings, moving HPC workloads to the cloud can offer organizations the ability to scale computing capacity dynamically. CSPs can provide elastic resources, allowing HPC users to scale up or down based on their computational needs. This flexibility can provide organizations with the ability to efficiently match their computing resources with workload demands, avoiding underutilization or overprovisioning. Scalability is particularly crucial for HPC applications that require significant computational power for intermittent periods.

The move towards energy-efficient cloud solutions is not just a potential cost-saving strategy for users, but also a proactive measure to address the complexities associated with managing evolving HPC workloads. The leading CSPs offer managed environments where organizations can offload the intricacies of infrastructure management, data storage, and software maintenance. This abstraction of complexity allows HPC users to focus more on their core computational tasks, streamlining operations and boosting overall efficiency. As HPC workloads become increasingly sophisticated, the cloud provides a centralized and streamlined approach to managing computational complexities, helping organizations derive more value from their HPC endeavors.

THE AWS AND NVIDIA APPROACH TO ENERGY-EFFICIENT HPC

A prime example of a CSP and technology partnership providing industry leading energy-efficient computing solutions to HPC customers is the collaboration between Amazon Web Services (AWS) and NVIDIA. Both technology leaders contribute to the objective of HPC users reaching energy efficiency goals without sacrificing performance.

AWS offers a number of comprehensive energy-efficient solutions to customers by optimizing its data centers. The scale of their operations provides a number of advantages in computational choices and energy efficiency. This includes strategic decisions related to energy purchases, incorporating

renewable energy sources, enhancing data center energy efficiency, and optimizing the energy consumption of their infrastructure. This comprehensive approach allows organizations to scale up or down GPU consumption in alignment with the expansion or contraction of their HPC workloads. The ability to dynamically adjust resource allocation contributes to both optimized resource allocation and energy efficiency. AWS further contributes to energy cost reduction by improving the efficiency of physical resources used for computing, networking, and storage through virtualized servers and secured infrastructure. Users can easily spin-up resources when needed and spin-down resources when not in use, conserving energy during idle periods. This on-demand utilization of resources helps organizations optimize their HPC workloads by ensuring that resources are active only when needed.

The AWS and NVIDIA collaboration addresses the critical concern of potential power shortages by allowing organizations to scale computing capacity. Accelerated computing with low-latency, high-bandwidth networking provided by AWS can potentially save organizations from consuming their allocated power capacity and free up resources for additional compute needs. This not only reduces the risk of not having enough power to meet business needs but also ensures that organizations can efficiently scale their computational resources based on workload demands.

Moving to the cloud with energy-efficient solutions from AWS and NVIDIA addresses many of the complexities inherent in HPC workloads. AWS, through its Amazon Elastic Compute Cloud (Amazon EC2) HPC-optimized instances powered by NVIDIA GPUs, facilitates accelerated computing for faster time to results. As workloads increase in complexity, accelerated computing in the cloud can speed up simulations resulting in less computing time and less energy consumed.

AWS & NVIDIA ARE IMPROVING ENERGY EFFICIENCY ACROSS GLOBAL INDUSTRIES

Across the HPC market, data centers are contending with the challenges of operating more energy efficiently, with some having to compromise performance as a result. The collaboration between AWS and NVIDIA is an excellent example of a CSP and technology provider delivering HPC solutions to users that satisfy both energy efficiency and performance goals.

Healthcare and life sciences

Cryo-Electron Microscopy (Cryo-EM) enables the creation of detailed 3D structures of vital drug targets, including the SARS-CoV-2 spike protein. Utilizing flash-freezing techniques, Cryo-EM Transmission Electron Microscopes capture near-native states of microscopic structures in vitreous ice, generating terabytes of data per sample. This data undergoes processing via HPC resources, accelerated by GPUs, with some interactive steps requiring scientist involvement and 3D visualization for result analysis. Cryo-EM analysis demands flexible computing access due to unpredictable processing speeds at various stages. Many new Cryo-EM workloads are leveraging Amazon EC2 resources on AWS, facilitated by AWS ParallelCluster deployments. To address increasing data sizes and optimize workflows, Cryo-EM applications have recently been adapted for higher-end NVIDIA GPUs. Various groups using Cryo-EM have seen efficiency advancements as a result of moving to AWS and the use of NVIDIA GPUs. For example, a recently published customer success story saw reduced processing time from 4 weeks on-premises to 4 days on AWS, at a fraction of the cost, while benefiting from enhanced resolution using the latest hardware available.

In the area of genomic sequencing, AWS and NVIDIA GPUs have played a pivotal role in revolutionizing bioinformatics tasks. The parallel processing capabilities of GPUs have been harnessed to accelerate the analysis of vast genetics datasets, drug development algorithms, and

medical imagery, leading to faster results and, notably, lower energy consumption compared to traditional CPU-based approaches.

Climate and weather

Similarly, in the weather industry, the combination of NVIDIA GPUs with AWS has ushered in a new era of energy-efficient computing for forecasting and modeling. The parallel processing capabilities of GPUs optimize computational workloads, contributing to overall energy efficiency. Many weather organizations leveraging this technology synergy benefit from more accurate predictions while operating with increased efficiency.

A weather technology group was recently awarded the AWS Public Sector Partner Award for its innovative HPC weather forecasting solution. The solution utilizes on-demand capacity reservations (ODCRs) to efficiently manage bursty workloads, ensuring sufficient EC2 instances are available without the interruptions of EC2 Spot or long-term commitments. This approach allows for quick pivoting to other HPC-optimized EC2 instance types, ensuring dependability and timely delivery of critical weather information to customers.

By incorporating ODCR workflows, the solution gains flexibility at scale, leveraging AWS' reliability, scalability, and agility to meet the rigorous demands of weather workflows. This makes the solution more attractive to users, offering benefits in terms of speed, ease of access, and dependability, crucial for supporting weather industry customer decision-making.

Engineering and manufacturing

A market leading engineering company using HPC in the cloud has witnessed a transformative impact on energy efficiency by harnessing the collaboration between AWS and NVIDIA GPUs. Leveraging Amazon EC2 P4d instances equipped with NVIDIA A100 Tensor Core GPUs, the engineering company has experienced a remarkable 3.5 times acceleration in simulations compared to the previous GPU generation. This significant speed-up translates to a substantial reduction in the time required to solve complex engineering problems, resulting in a lower energy consumption level along with a faster workload completion time. The AWS and NVIDIA GPU combination not only expedites time-to-market for customers but also enables higher fidelity simulations, contributing to the development of more efficient and environmentally friendly products. The collaboration between AWS and NVIDIA GPUs exemplifies aligning computational capabilities with energy efficiency goals in the engineering sector.

In the realm of industrial manufacturing, Computational Fluid Dynamics (CFD) code using industries have used NVIDIA GPUs to enhance energy-efficient computing in simulations. By offloading intensive calculations onto GPUs, CFD applications leverage parallel processing capabilities, accelerating simulations and reduce the time required to obtain results. This not only boosts productivity but also minimizes energy consumption, showcasing the transformative impact of AWS and NVIDIA GPUs in optimizing computational workflows within the industrial manufacturing sector.

A 2022 CFD customer success story shared by AWS highlighted significant productivity improvements, including a 98% reduction in wait time, and 26% faster runtimes for resource-intensive HPC jobs. The customer attributed these massive productivity improvements to the use of EC2 instances that are resizable for every type of workload, having access to the latest generation NVIDIA GPUs as soon as possible, and the dedicated 100 gigabits/second throughput offered by Elastic Fabric

Adapter (EFA). These productivity gains translate into reduced energy consumption and reduced time to results for this customer.

FUTURE OUTLOOK

Looking ahead, the outlook for HPC data centers involves grappling with the escalating energy demands posed by data-intensive workloads. From genomic sequencing to weather forecasting to CFD simulations to AI model training and inference, the computational requirements continue to rise, necessitating innovative strategies for energy-efficient computing. Cloud adoption emerges as a promising solution for many users, offering organizations the flexibility to scale GPU-accelerated compute capacity dynamically and access energy-efficient solutions provided by leading CSPs.

The collaboration between AWS and NVIDIA exemplifies how cloud services can contribute to energy-efficient computing, delivering leadership-class solutions to diverse industries. Together they stand out as an example of energy-efficient cloud services, offering flexible computing solutions to industries like engineering, genomics, weather forecasting, and industrial manufacturing. By utilizing AWS and NVIDIA, organizations have the opportunity to accelerate simulations, reduce time to results, and drive advancements in energy-efficient practices. The future of HPC data centers lies in navigating the complexities of running much larger workloads, reducing energy demands, embracing cloud solutions for appropriate workloads, and fostering collaborations that propel the industry toward energy efficiency in the face of evolving challenges.

About Hyperion Research, LLC

Hyperion Research provides data-driven research, analysis and recommendations for technologies, applications, and markets in high performance computing and emerging technology areas to help organizations worldwide make effective decisions and seize growth opportunities. Research includes market sizing and forecasting, share tracking, segmentation, technology, and related trend analysis, and both user & vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). Hyperion Research provides thought leadership and practical guidance for users, vendors, and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.HyperionResearch.com and www.hpcuserforum.com