How moving onto the AWS cloud reduces carbon emissions

Contents

Executive summary

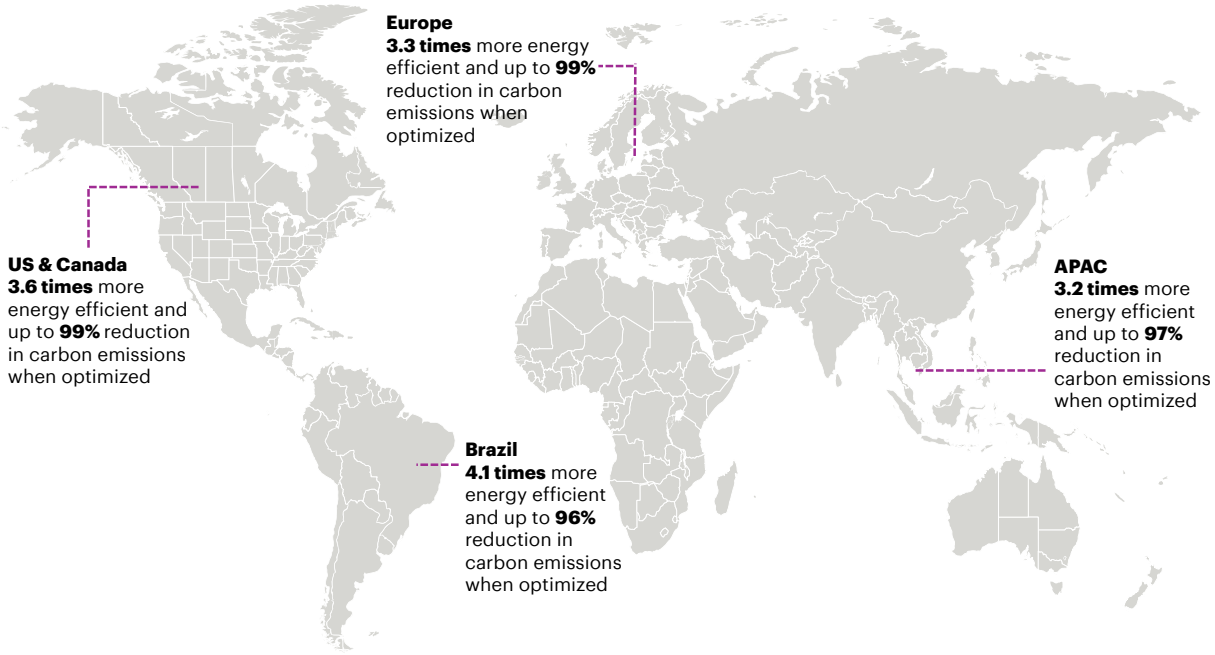
Advancements in digital transformation and adoption of advanced technologies such as artificial intelligence (AI) are driving up global demand for data center capacity. At the same time, there is growing interest to understand the environmental cost of this demand and how the impact of the IT sector overall can be reduced. In that vein, this research set out to quantify the energy efficiency and carbon reduction opportunity of moving customer workloads from on-premises to Amazon Web Services (AWS) and further optimizing the workloads on AWS.

AWS commissioned Accenture to study the sustainability metrics of on-premises versus AWS deployments. On-premises describes IT infrastructure hardware and software applications that exist on-site, for example locally within an organization's own physical office or space, in contrast to assets that are hosted off-site in the cloud.

Accenture based its study off the ISO standard for Software Carbon Intensity (SCI) and analyzed operational carbon emissions and embodied carbon emissions (hardware components) across simulated on-premises and AWS deployments for a representative storage-heavy workload and compute-heavy workload. The research went beyond the SCI methodology to include procurement of carbon-free energy to reduce workload carbon emissions from electricity consumption as well.

For the reference workloads, the results show that running workloads on AWS is up to 4.1 times more energy-efficient than on-premises and can reduce the associated carbon footprint by up to 99% when optimized.

Carbon emissions reduction and energy efficiency by moving to AWS (for compute-heavy workloads)



Similarly, customers running compute-heavy workloads can see carbon emissions reduce by up to 99%, with reduction of up to 94% by moving from on-premises to AWS and an additional reduction up to 81% using AWS's purpose-built silicon. For storage-heavy workloads, customers can reduce carbon emissions by up to 93%, with reduction of up to 88% by moving from on-premises to AWS and an additional up to 47% reduction by leveraging a modernized AWS architecture.

These results are made possible because AWS works to continually improve efficiencies across their data centers—from rack layouts to electrical distribution to cooling techniques – so they strive to move towards operating closer to peak energy efficiency. AWS optimizes resource utilization across their global footprint, minimizes idle capacity and continuously improves the efficiency of their hardware.

AWS's latest data center design seamlessly integrates optimized air-cooling and free-air cooling solutions alongside liquid cooling capabilities for the most powerful AI chipsets. Aligning with Amazon's commitment to achieving net-zero carbon emissions across all operations by 2040, AWS is rapidly working to match 100% renewable energy for the electricity powering its global operations. AWS also leverages purpose-built silicon like the AWS Trainium chip (for training) and AWS Inferentia chip (for inference) to achieve significantly higher throughput than comparable accelerated compute instances. Many of these factors are challenging at the smaller operational scale of a typical on-premises data center.

Introduction

According to the International Energy Agency (IEA), global data centers' electricity use is expected to double by 2026 from 460 TWh in 2022. Much of this projected increase comes from AI workloads and the amount of data feeding into the models.[1] Accenture's estimates suggest that organizations are achieving AI transformation 16 months faster than their digital transformation.[2] As AI workloads become more complex and data-intensive, they demand higher levels of performance from CPUs, GPUs, memory, storage and networking infrastructure, all of which can be energy and carbon intensive. On-premises data centers are struggling to keep pace due to their inherent limitations in driving scalability and energy efficiency.[3,4] This is where AWS can make a difference.

AWS commissioned Accenture to study the sustainability metrics of on-premises versus AWS deployments based on the ISO-certified (ISO/IEC 21031:202) Software Carbon Intensity (SCI) standard from Green Software Foundation, which defines a methodology for calculating the rate of carbon emissions for a software system (see the box on next page). Accenture analyzed energy consumption profiles across simulated on-premises and AWS deployments using a representative storage-heavy workload and compute-heavy workload. It went beyond the SCI methodology, which focuses on absolute carbon reduction to include corporate procurement of carbon-free energy, which reduces emissions from electricity consumption. Accenture incorporated data on carbon-free energy procured by on-premises data centers, based on secondary research and its experience working with organizations that host these centers. Additionally, data on carbon-free energy procured by AWS was included as reported.

The findings of the research show workloads are more energy efficient with a lower carbon intensity when run on AWS compared to on-premises. These findings are attributable to AWS's infrastructure designed for optimal energy usage such as purpose-built silicon, higher server utilization rates, economies of scale and using carbon-free energy sources. AWS also achieves lower embodied emissions per unit of computing power or storage by consolidating workloads and individual customers onto shared, highly utilized infrastructure. Thus, embodied emissions from construction materials, IT equipment and power infrastructure are shared across a larger customer base.

It is these efficiencies that will reduce associated operational and embodied emissions of workloads when they are migrated away from on-premises to run on AWS as is. But the potential benefits do not stop there. Optimization of workloads on AWS can further reduce the carbon and energy intensity of workloads.

Accenture used proprietary tools and models to estimate potential carbon emissions from on-premises versus cloud applications in selected use cases, focusing on two types of workloads: storage-heavy and compute-heavy.

ISO-Standard Software Carbon Intensity

The Software Carbon Intensity (SCI) specification from the Green Software Foundation (GSF) aims to articulate the methodology for calculating total carbon emissions to overcome the problems of opaque metrics and lack of standardization across industries. It is inspired by the Life-Cycle Assessment and GHG Protocol for assessing environmental impact.

The SCI is the rate or intensity of carbon emissions per functional unit. This functional unit is customizable and is defined based on how the workload scales, for example, per inference, per tebibytes or TiB storage or per user. The functional unit is what makes SCI useful in comparing two software systems or, in other words, workloads with the same use-case to enable carbon-saving decisions. SCI captures embodied emissions too,broadening its scope beyond operational emissions.

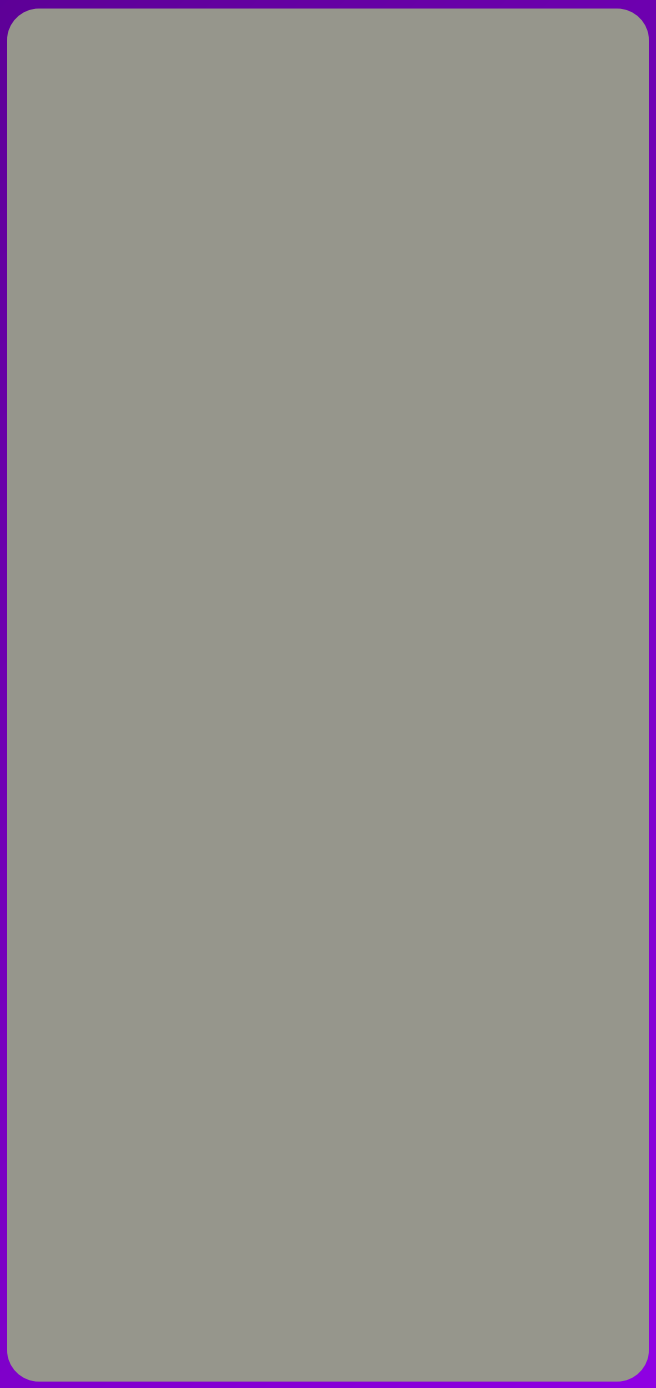
For each workload, Accenture designed two sub-scenarios. The first compared an on-premises deployment with the Lift-and-Shift Scenario—where the workload was migrated to AWS as-is, without any optimization. The second compared the Lift-and-Shift Scenario with the AWS-Optimized Scenario—where the workload was migrated to AWS and then optimized by leveraging a modernized AWS architecture and purpose-built silicon. Comparing the Lift-and-Shift Scenario against the AWS-Optimized Scenario enabled an assessment of the potential benefits of adopting AWS optimization techniques for each workload. These sub-scenarios were run across four regions—the United States and Canada, Europe, Asia Pacific, and Brazil—considering the grid energy mix in each.

To gain further insights into energy efficiency and carbon emissions of deployed workloads across regions, two variables were selected and analyzed—Power Usage Effectiveness (PUE) and idle capacity—along with two methods:

1. Location-Based Method (LBM) for scope 2 emissions—based on average energy generation emission factors where the energy consumption occurs, and

2. Market-Based Method (MBM) for scope 2 emissions—based on emissions by the generators from which the company contractually purchases electricity bundled with instruments, or unbundled instruments on their own. (For details on these variables, see the Appendix).

To estimate the embodied emissions from running storage-heavy and compute-heavy workloads on premises and on AWS, the study analyzed the full lifecycle emissions for each relevant piece of IT hardware (blade/rack servers, HDD, SSD, CPU, GPU, RAM) and its average lifespan. Using the average lifespan indicated by manufacturers is a conservative estimate, given that AWS typically uses hardware longer than that. Carbon emissions from component manufacturers were considered to represent embodied emissions, since those related to transportation and end-of-life were negligible.



The study did not account for embodied emissions from non-IT infrastructure (such as buildings, HVAC, lighting) as this is outside the bounds of SCI, even though AWS does work across its supply chain to build data centers with lower-carbon concrete and steel wherever possible.[5]

While specific situations may differ from these simulations and performance may vary, the findings remain indicative of the improved efficiencies and carbon reduction opportunities that can be achieved on AWS.

Storage-heavy workloads

With the world anticipated to produce 180ZB of data every year by 2025, the enterprise data landscape is expanding at an astonishing speed and scale.[6] Considering the data requirements for training and inference of AI models, organizations should factor in the potential carbon savings associated with their storage requirements.

The research considered a storage-heavy workload to study how organizations at any stage of their digital technology maturity can benefit from the carbon reduction opportunity when running workloads on AWS. The representative storage workload selected was Network Attached Storage (NAS), a file-dedicated storage device that makes data continuously available for effective collaboration over a network. The research focused on simulating the carbon emissions associated with hosting the file system. Specifically, it considered a set of requirements for throughput, data storage for active data sets and inactive data sets, and simulated workloads which met those requirements on premises, on AWS and optimized on AWS. The functional unit used as the basis for this analysis was provisioned storage volume in TiB. The workload consisted of two file systems, where each file system environment included two file servers and storage volumes for both active and inactive data.

The research considered the benefits of migrating a NAS to AWS using Amazon FSx for NetApp ONTAP, a storage service that allows customers to launch and run fully managed ONTAP file systems on AWS (Lift-and-Shift Scenario) and the additional efficiencies gained by leveraging the flexibility of AWS by resizing the relative active and inactive data to reflect actual usage (AWS-Optimized Scenario).

Since lift-and-shift cloud deployment minimizes workload architecture changes, Accenture modeled an on-premises storage solution equivalent to the AWS lift-and-shift architecture.

This ensured comparable storage deployment parameters, including active/inactive data size, throughput, SSD IOPS, replication, storage efficiency and offline backups. The on-premises storage solutions were carefully selected to match the performance of the workload on AWS.

The results showed that an AWS-managed storage service with carbon-free energy procurement offers up to 88% lower carbon emissions and is up to 2.5 times more energy efficient than comparable NAS systems on-premises (Table 1). AWS-optimized file systems then lowered carbon emissions by up to an additional 47% than comparable AWS file systems on AWS. This is due to optimized utilization of inactive storage volumes and reduced file system size (in terms of throughput) for the file system storing secondary data copies. Therefore, running optimized storage workloads on AWS can reduce the associated carbon footprint by up to 93% compared to on-premises.

About Amazon FSx for NetApp ONTAP

ONTAP is NetApp's file system technology that provides a widely adopted set of data access and data management capabilities. FSx for ONTAP provides the same familiar features, performance and APIs as on-premises NetApp file systems with the additional agility, scalability and simplicity of a fully managed AWS service. This helps businesses reduce costs, improve scalability and enhance performance. For example, Amazon Annapurna Labs which designs custom AWS chips and accelerators, migrated to the fully managed Amazon FSx for NetApp ONTAP service, helping improve storage scalability and availability and reducing costs by 35%.[1]

Geographical Region	% Reduction in carbon emissions intensity			Energy efficiency factor
	On-premises vs AWS Lift-and-Shift	AWS Lift-and-Shift vs AWS-Optimized	On-premises vs AWS-Optimized	On-premises vs AWS Lift-and-Shift
US & Canada	86%	42%	92%	1.8
Europe	88%	40%	93%	1.9
Asia Pacific	73%	47%	86%	2.3
Brazil	65%	45%	81%	2.5

Table 1: Comparison of operational and embodied emissions for a storage-heavy workload—% reduction in carbon emissions intensity (gCO2 per hour per TiB)

The functional unit used as the basis for this analysis was provisioned storage volume in TiB. The workload consisted of two file systems, where each file system environment included two file servers and storage volumes for both active and inactive data. Each file system environment included two file servers and storage volumes for both active and inactive data.

Regional variations

United States & Canada

The United States is home to leading cloud computing providers and large technology companies, who are responsible for the development of large-scale data centers. The United States and Canada, along with Europe, have some of the lowest PUE values. The region also offers dedicated programs and resources through the Center of Excellence for Energy Efficiency in Data Centers, established by the Federal Energy Management Program (FEMP), to help agencies reduce their PUE.[7]

The United States and Canada showed a significant carbon reduction opportunity in the Lift-and-Shift Scenario for storage workloads. This is attributable to the gap in PUE between on-premises and AWS data centers and higher carbon-free energy procurement by AWS.

Compared to running storage-heavy workloads on-premises, even with carbon-free energy, AWS helps lower carbon emissions by up to 86% through a combination of reduction and abatement strategies.

Efficient hardware with improved utilization	**18%**
Power and cooling efficiencies	**21%**
Carbon-free energy procurement	**47%**

Migrating workloads to AWS-optimized hardware with efficient replication can lead to an additional reduction in carbon emissions of up to 42%. Therefore, running optimized storage workloads on AWS can reduce the associated carbon footprint by up to 92% compared to on-premises.

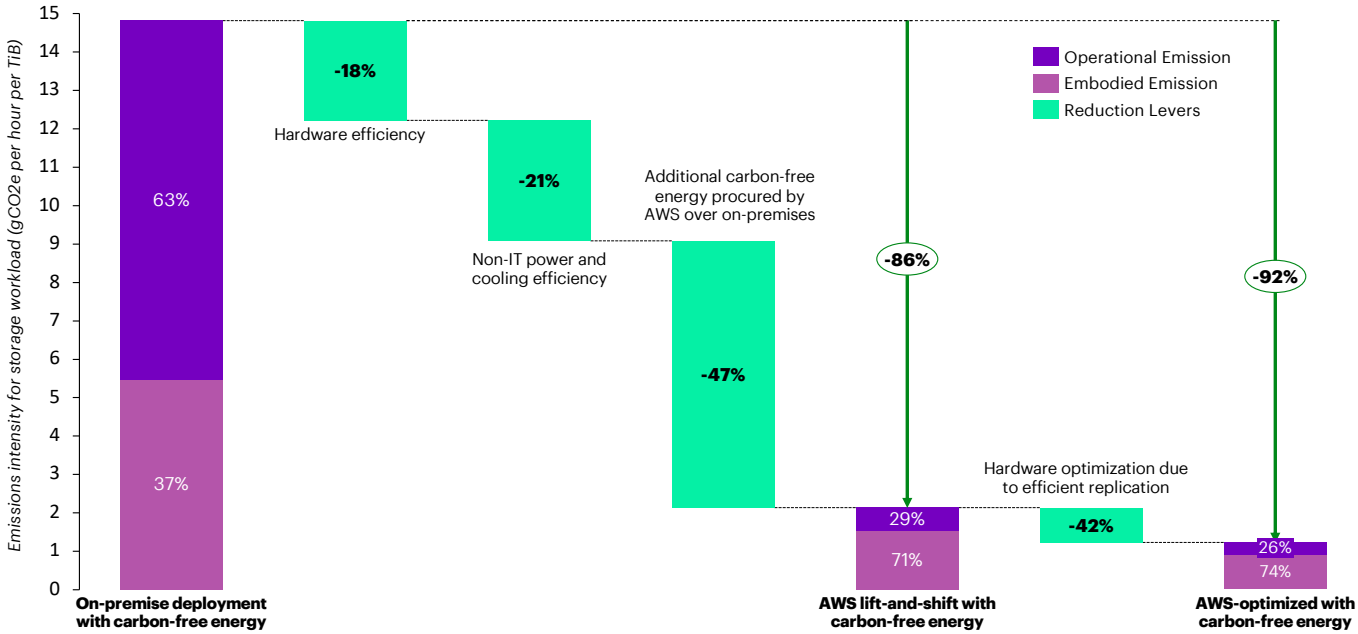


Figure 1: Reduction in carbon emissions for storage-heavy workloads in the United States & Canada

Europe

Europe is the second-largest region for data centers after the United States and Canada, accounting for around 1.5% of the region's total electricity consumption.[8] Upcoming laws such as the Energy Efficiency Directive, German Energy Efficiency Act and Corporate Sustainability Reporting Directive, influence data center operations.

Europe is also incentivizes reduction of PUE through the EU Code of Conduct for Data Centres and awards centers.[9] These regulations and incentives, combined with favorable climatic conditions, result in data centers in Europe having the lowest PUE value for both AWS and on-premises data centers among the regions studied. Even then, migrating storage-heavy workloads from on-premises to AWS can still reduce the associated carbon emissions by up to 88%. This is because of a combination of reduction and abatement strategies that help AWS reduce carbon emissions.

Efficient hardware with improved utilization	**28%**
Power and cooling efficiencies	**17%**
Carbon-free energy procurement	**43%**

Leveraging a modernized AWS architecture that resizes active and inactive data sets based on actual usage can lead to additional carbon emission reductions of up to 40% over the Lift-and-Shift Scenario. Therefore, running optimized storage workloads on AWS can reduce the associated carbon footprint by up to 93% compared to on-premises.

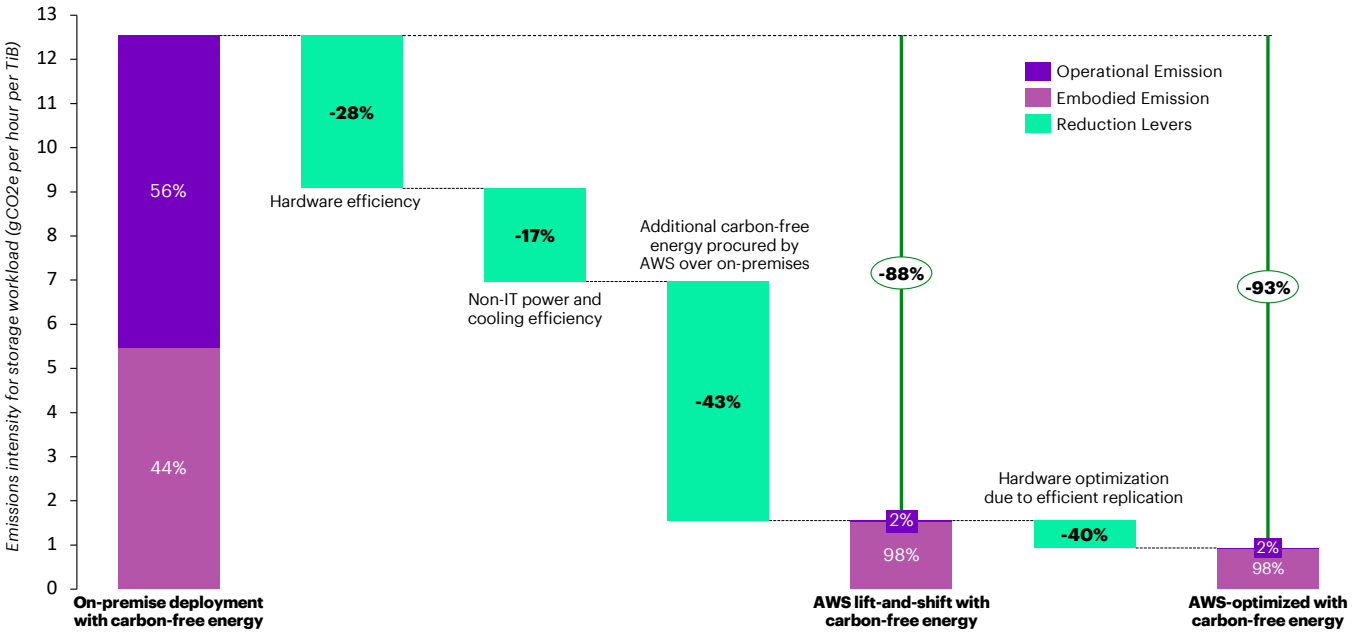


Figure 2: Reduction in carbon emissions for storage-heavy workloads in Europe

Asia Pacific (Singapore, Australia, India, South Korea and Japan)

Data centers in the Asia-Pacific (APAC) region are growing fast, with Japan, Australia, India and Singapore dominating operational capacity.[15] PUE values are higher in the region due to higher temperatures and humid climate and the lack of regulations or incentives around improving the sustainability of data centers, as in Europe. On-premises data centers in APAC often build redundant systems with multiple power and cooling paths to mitigate the risks associated with unreliable access to utilities and delays in repair services.[10]

Given the above circumstances, shifting storage-heavy workloads to AWS data centers in the APAC region presents an opportunity for organizations to reduce the associated carbon footprint. Even when on-premises environments use carbon-free energy, migrating to AWS can help reduce carbon emissions by up to 73%. Reduction and abatement strategies that help lower carbon emissions include:

Efficient hardware with improved utilization	**23%**
Power and cooling efficiencies	**32%**
Carbon-free energy procurement	**18%**

Leveraging a modernized AWS architecture can lead to additional carbon emission reductions of up to 47% over the Lift-and-Shift Scenario. Therefore, running optimized storage workloads on AWS can reduce the associated carbon footprint by up to 86% compared to on-premises.

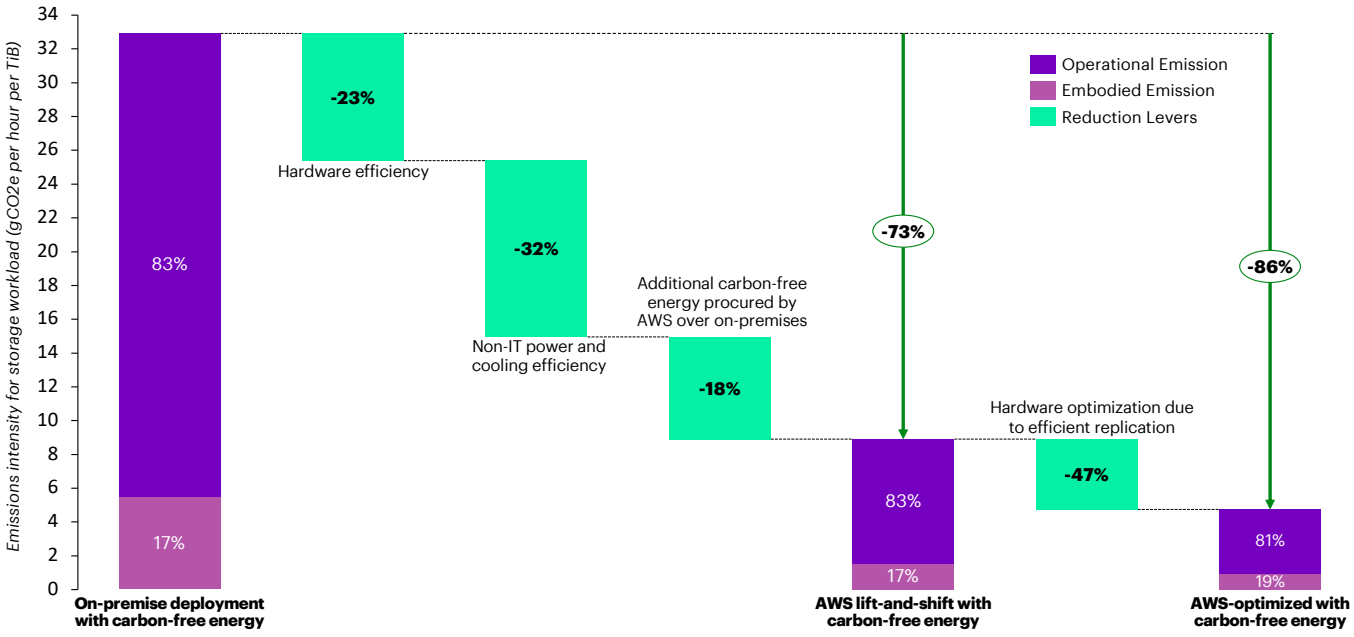


Figure 3: Reduction in carbon emissions for storage-heavy workloads in APAC

Brazil

Brazil's large population, internet connectivity and digital demand make it a data center hub for Latin America (67% inventory).[17] However, energy efficiency lags behind other regions due to climate and data center practices. Despite higher PUE values, Brazil's relatively "green" grid makes it an attractive option.

90% of electricity production in Brazil comes from a mix of carbon-free and low-carbon sources.[19] This limits the opportunity for further carbon reduction through corporate carbon-free energy procurement. For this reason, across regions, Brazil showed the lowest carbon reduction in Lift-and-Shift Scenario even though it has the largest difference between on-premises and AWS PUE. Even then, migrating workloads to AWS can reduce associated emissions by up to 65%. Reduction and abatement strategies that help AWS reduce carbon emissions include:

Efficient hardware with improved utilization	**37%**
Power and cooling efficiencies	**27%**
Carbon-free energy procurement	**1%**

Organizations can achieve additional reductions of up to 45% by leveraging a modernized AWS architecture. Therefore, running optimized storage workloads on AWS can reduce the associated carbon footprint by up to 81% compared to on-premises.

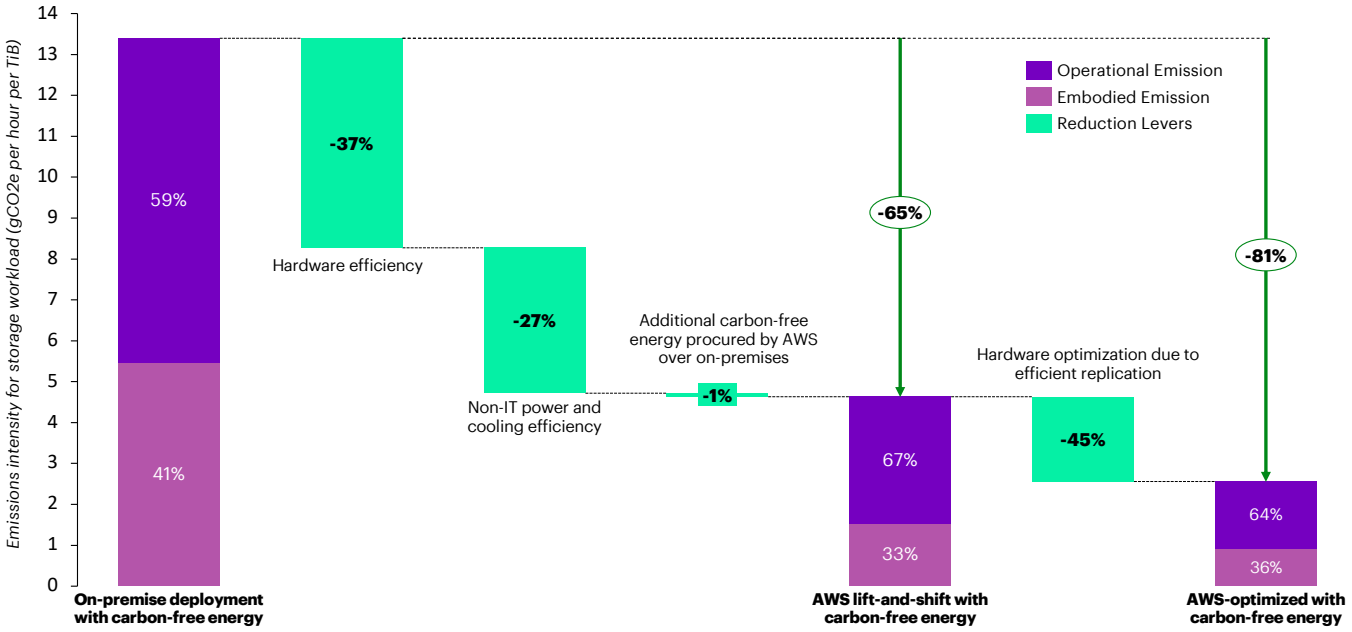


Figure 4: Reduction in carbon emissions for storage-heavy workloads in Brazil

Compute-heavy workloads

The use of AI and complex simulations, deep learning and other resource-intensive tasks across industries is increasing demand for compute power. By embracing innovative solutions, adopting more sustainable practices and carefully managing cloud resources, businesses can optimize their compute-heavy workloads on AWS to reduce the associated carbon footprint.

The research considered the extent to which customers can reduce their carbon footprint—while continuing to transform their organizations—by running compute-heavy workloads on AWS. The representative compute workload selected was LayoutLM for document processing. Accenture modeled a workload running on-premises that is equivalent to the LayoutLM model, with minimal changes to the workload architecture (Lift-and-Shift Scenario) and then considered the same workload deployed on AWS purpose-built silicon chips (AWS-Optimized Scenario).

About LayoutLM

LayoutLM helps automate real-world workflows for processing of documents, such as invoices, receipts and forms, with high levels of accuracy. For example, a leading American airline built an active learning framework that used the LayoutLM model to automate processing and information extraction from passenger documents like passports to verify passenger identities.[1] The framework intelligently samples unlabeled data for manual labeling to reduce costs while iteratively retraining LayoutLM to improve accuracy.

A lift-and-shift migration to the cloud generally involves minimal alterations to the workload architecture. Consequently, Accenture designed an on-premises workload architecture that closely matches the performance of the AWS-hosted LayoutLM model, aiming to meet the target SLA by closely replicating the compute capabilities of the AWS environment.

The on-premises architecture is meticulously modeled to mirror the AWS setup, featuring processors with similar architecture, core count and maximum frequency, as well as an equivalent number and type of GPU chipsets. It also includes comparable memory, storage and network infrastructure to help ensure equivalent performance without differences in capabilities or generation of components.

The results showed that LayoutLM on AWS with carbon-free energy procurement offers up to 94% lower carbon emissions and is up to 4.1 times more energy efficient than the equivalent workload running on-premises. Deploying Layout LM on AWS purpose-built silicon chips then lowered carbon emissions by up to an additional 81% than comparable AWS computing instances. This improvement highlights the efficiency AWS offers organizations who want to further optimize energy-intensive AI workloads. Therefore, running optimized AI workloads on AWS can reduce the associated carbon footprint by up to 99% compared to on-premises.

Geographical Region	% Reduction in carbon emissions intensity			Energy efficiency factor
	On-premises vs AWS Lift-and-Shift	AWS Lift-and-Shift vs AWS-Optimized	On-premises vs AWS-Optimized	On-premises vs AWS Lift-and-Shift
US & Canada	93%	81%	99%	3.6
Europe	94%	79%	99%	3.3
Asia Pacific	87%	80%	97%	3.2
Brazil	82%	79%	96%	4.1

Table 2: Comparison of change in total carbon emissions for LayoutLM (% carbon reductions in emissions intensity (gCO2 per million inferences)

The research leveraged the SCI standard, using the metric 'number of inferences' for comparison. Compute-accelerated hardware that provides substantial performance gains over traditional CPUs was used. The intent was to maintain equivalent hardware capabilities in the on-premises system so that it could meet the target Service Level Agreement (SLA) of 1 million inferences per hour, comparable to the Lift-and-Shift Scenario. This target is often considered real-time.

While modeling the operational emissions of AI workloads, the study considered resources related to:

- **Compute:** Running AI workloads requires significant computational power, including CPUs, GPUs and other specialized hardware accelerators. The energy consumption and associated emissions from these compute resources can be significant, especially for large-scale AI models and training processes. Reduction and abatement strategies like hardware efficiency, utilization rates and data center energy sources determine the emissions associated with these compute resources.

- **Storage:** The data required for inferencing uses instance storage. There are no additional storage systems associated with the workload. The energy consumption and emissions associated with this instance storage contributes to the overall impact of AI operations and should be accounted for.

- **Networking:** AI workloads frequently involve the transfer of large datasets and model parameters both within data centers and across the internet. Network bandwidth, data transfer volumes and the efficiency of networking hardware all contribute to the carbon emissions associated with networking resources.

United States & Canada

Organizations running compute-heavy (AI) workloads in AWS data centers in the United States and Canada benefit from a 93% reduction in carbon emissions compared to on-premises. This reduction is a result of several reduction and abatement strategies.

Efficient hardware with improved utilization	**32%**
Power and cooling efficiencies	**31%**
Carbon-free energy sources procured by AWS	**30%**

Moving workloads to AWS infrastructure and using purpose-built silicon leads to another 81% reduction in carbon emissions (see Figure 5). Therefore, running optimized AI workloads on AWS can reduce the associated carbon footprint by up to 99% compared to on-premises.

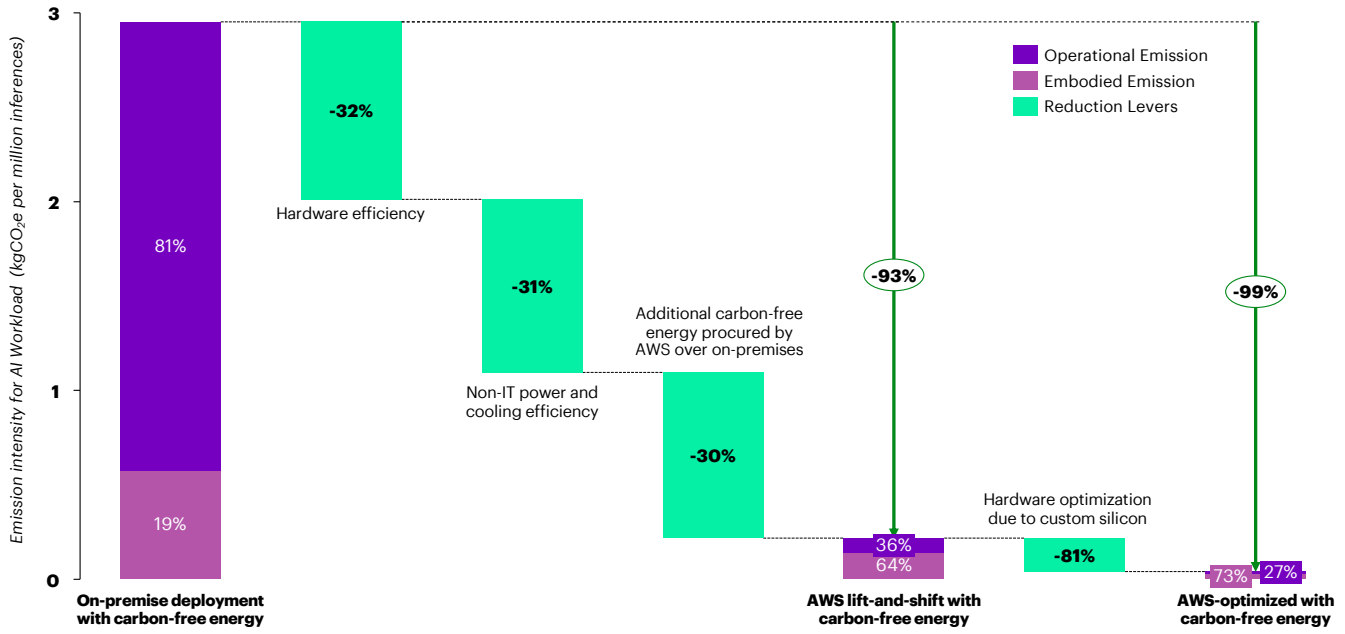


Figure 5: Reduction in carbon emissions for compute-heavy (AI) workloads in the United States and Canada

Europe

Organizations with compute-heavy (AI) workloads hosted in AWS data centers across Europe derive a sustainability advantage compared to on-premises deployments. Compared to running compute-heavy (AI) workloads on-premises, even with carbon-free energy, AWS helps lower carbon emissions by up to 94% through a combination of reduction and abatement strategies.

Efficient hardware with improved utilization	**36%**
Power and cooling efficiencies	**25%**
Additional carbon-free energy procurement	**33%**

By transitioning to AWS and using purpose-built silicon, companies can unlock an additional 79% carbon emissions reduction compared to the Lift-and-Shift Scenario, compounding the environmental benefits of their cloud migration. Therefore, running optimized AI workloads on AWS can reduce the associated carbon footprint by up to 99% compared to on-premises.

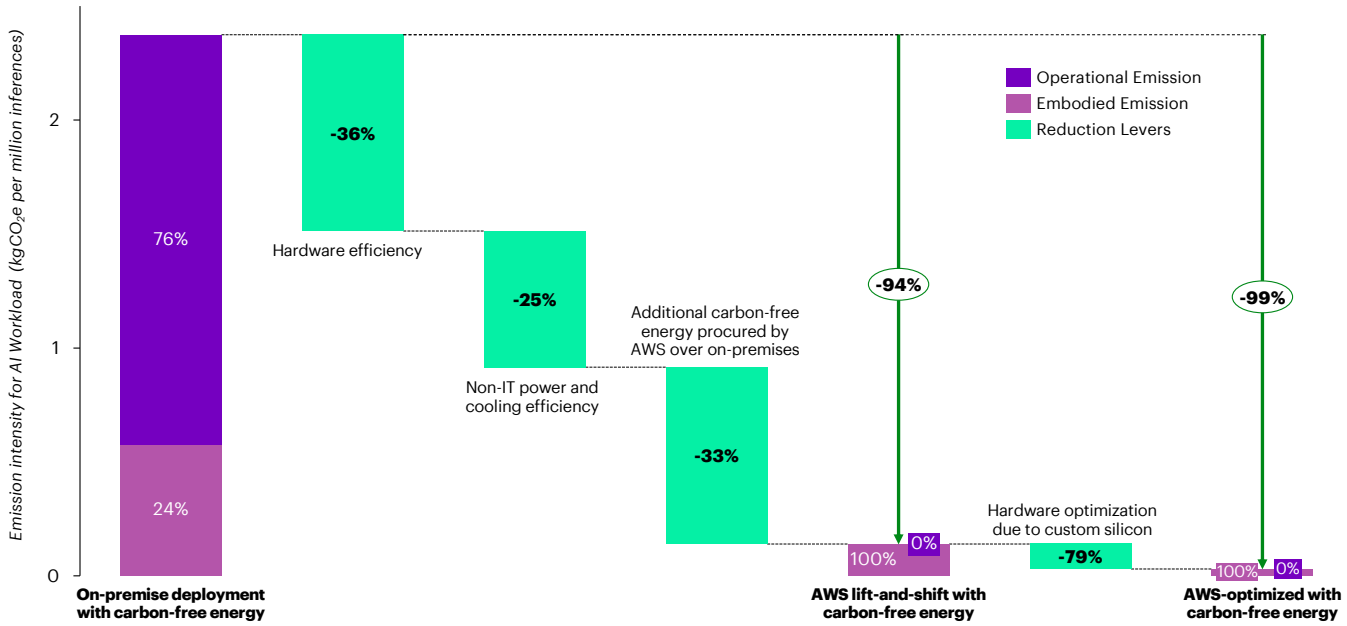


Figure 6: Reduction in carbon emissions for compute-heavy (AI) workloads in Europe

Asia Pacific (Singapore, Australia, India, South Korea and Japan)

Compared to running compute-heavy (AI) workloads on premises, even with corporate carbon-free energy procurement, AWS helps lower carbon emissions by up to 87% through a combination of reduction and abatement strategies.

Hardware efficiencies	**29%**
Non-IT power and cooling efficiencies	**38%**
Additional carbon-free energy procured by AWS compared to on-premises	**20%**

Transitioning to AWS and using purpose-built silicon enables an additional carbon emission reduction of 80% on top of the Lift-and-Shift Scenario. Therefore, running optimized AI workloads on AWS can reduce the associated carbon footprint by up to 97% compared to on-premises.

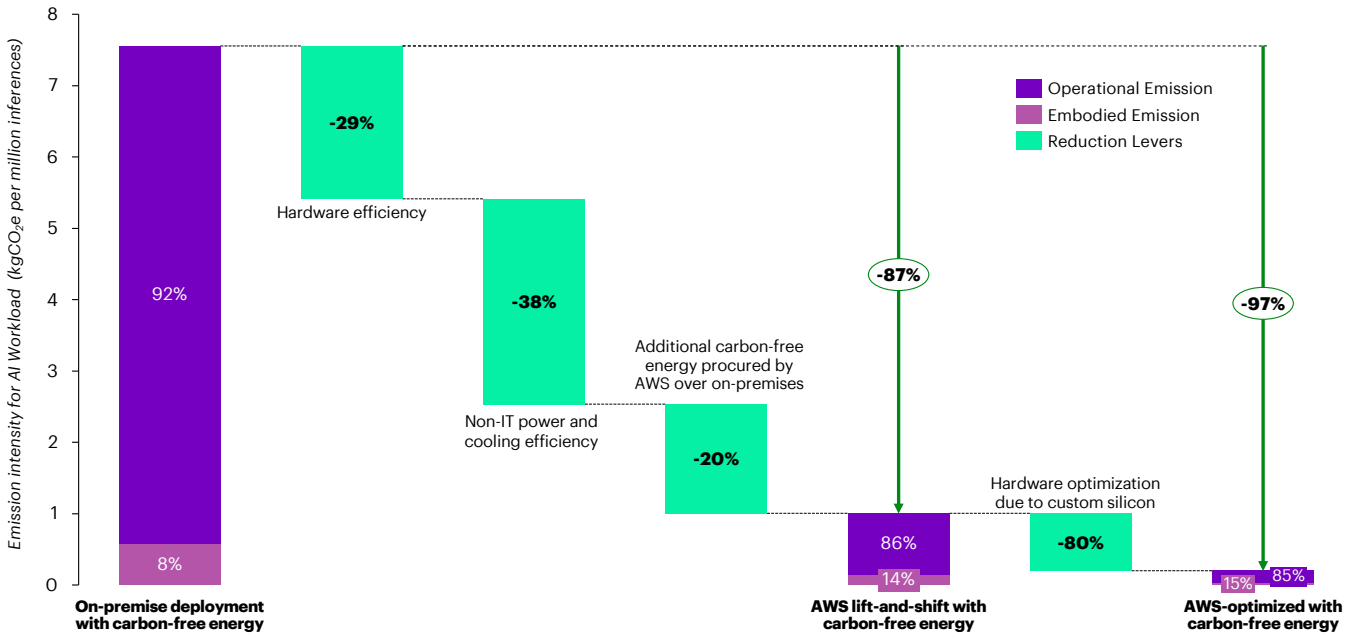


Figure 7: Reduction in carbon emissions for compute-heavy (AI) workloads in Asia Pacific

Brazil

Compared to running compute-heavy (AI) workloads on-premises even with carbon-free energy, AWS helps lower carbon emissions by up to 82% through a combination of reduction and abatement strategies.

Efficient hardware with improved utilization	**38%**
Power and cooling efficiencies	**37%**
Carbon-free energy procurement	**7%**

By transitioning to AWS and using purpose-built silicon, these companies can unlock an additional 79% carbon emissions reduction compared to the Lift-and-Shift Scenario, compounding the environmental benefits of their cloud migration. Therefore, running optimized AI workloads on AWS can reduce the associated carbon footprint by up to 96% compared to on-premises.

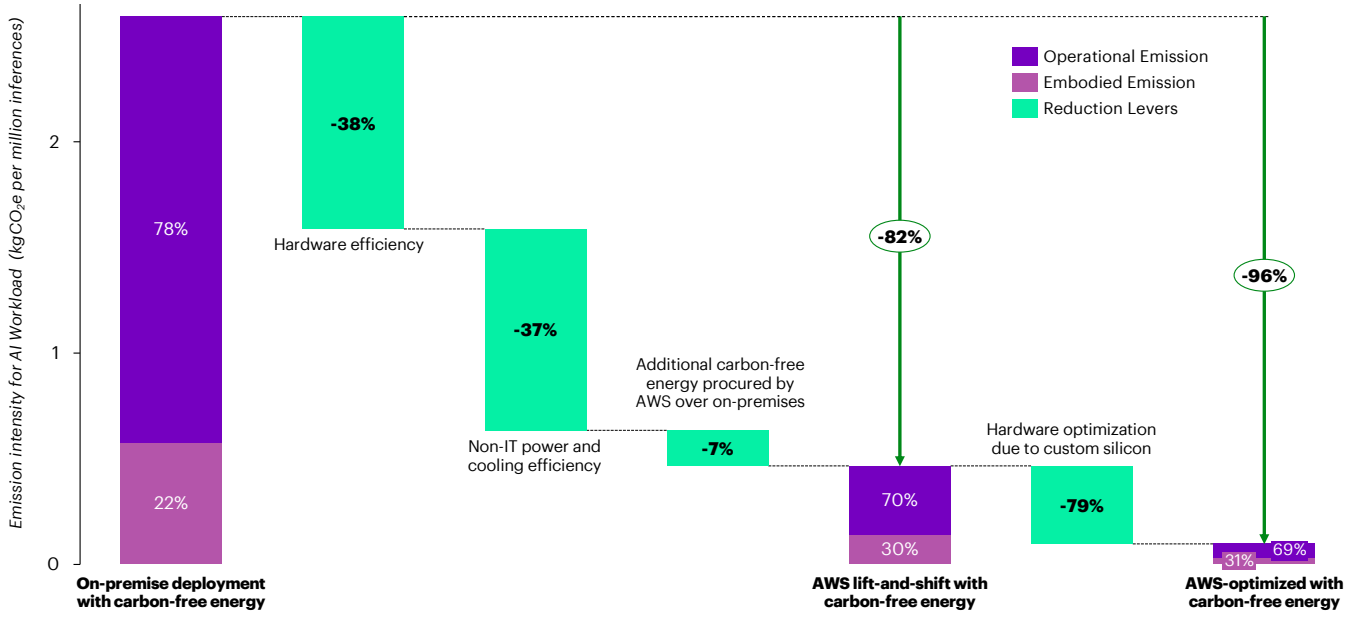


Figure 8: Reduction in carbon emissions for compute-heavy (AI) workloads in Brazil

Primary drivers for reduction of energy usage and carbon footprint

AWS has provided cloud services to customers for over 18 years, and they continue to innovate to increase efficiency with each generation of their data center designs, spanning across all aspects of their infrastructure, from the design of data centers and hardware to modeling operational performance for continuous enhanced efficiency. Energy efficiency enhancement strategies include transitioning to carbon-free energy, reducing embodied carbon, using water responsibly and transitioning to a circular economy.

- **Predicting performance:** AWS uses advanced modeling methods, such as computational fluid dynamics tools, to optimize data center design. This allows them to understand how a data center will perform before it is even built, enabling them to optimize their systems for higher reliability and energy efficiency. Once data centers are operational, AWS uses computer-based modeling and weather data from the Amazon Sustainability Data Initiative to predict and learn how to improve operations. These innovations, combined with economies of scale, enhance operational efficiency and improve AWS's PUE.

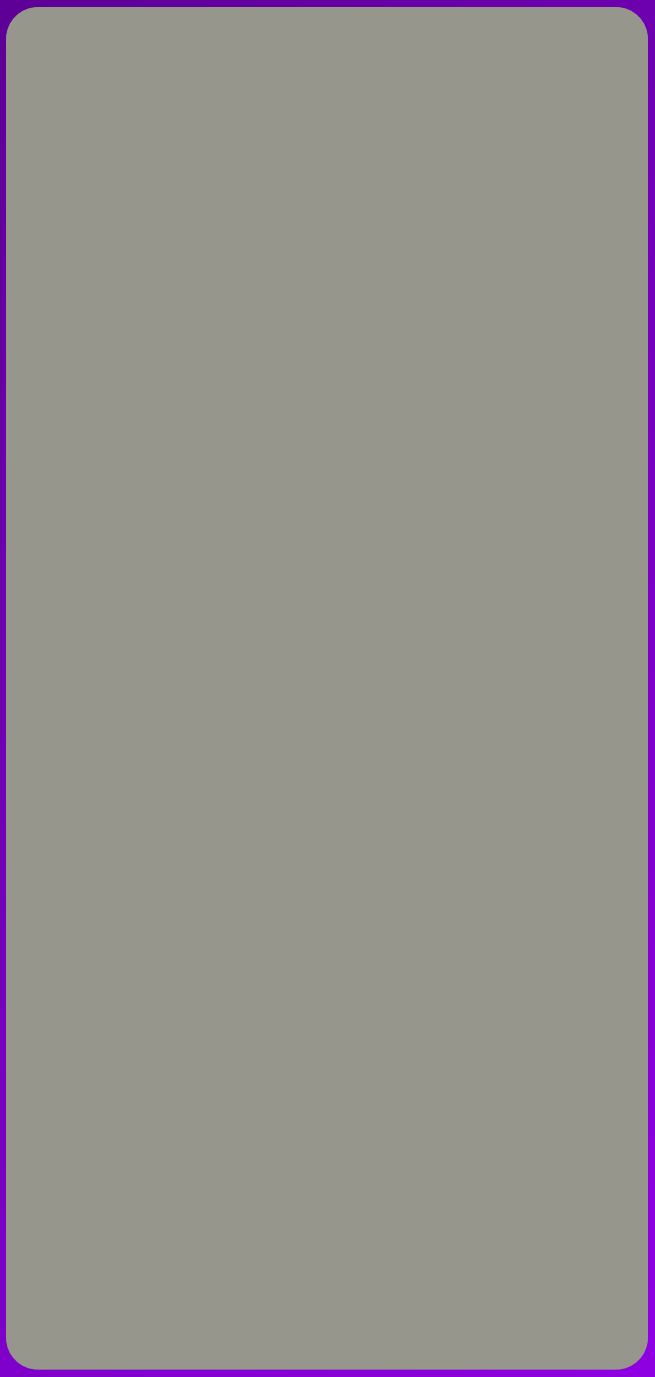
- **Cooling efficiency:** Cooling data centers consume substantial energy. AWS uses different cooling techniques, including free-air cooling depending on the location and time of year and adapts to changing weather conditions using real-time sensors. AWS is also working to optimize the longevity and airflow performance of the cooling equipment used in their data center cooling systems.[11] GPUs—chips that perform mathematical calculations at high speed—are critical for Machine Learning (ML) models. They generate much more heat than other types of chips, and higher densities in the future will require liquid cooling. AWS's latest data center design seamlessly integrates optimized air-cooling solutions alongside liquid cooling capabilities for the most powerful AI chipsets. This flexible, multi-modal cooling allows AWS to extract maximum performance and efficiency, whether running traditional workloads or AI/ML models. The AWS team works to continuously improve efficiencies across their data centers—from rack layouts to electrical distribution to cooling techniques—so they strive to move towards operating closer to peak energy efficiency, no matter the compute demands.

- **Hardware efficiency:** AWS's IT infrastructure, designed for high performance and energy efficiency, often contributes to lower operational and embodied emissions compared to on-premises data centers. By leveraging a shared-economy model, AWS optimizes resource utilization, reduces idle capacity, and enhances hardware efficiency. In contrast, on-premises data centers typically maintain excess capacity to accommodate unpredictable demand spikes and future growth due to limited capability to scale dynamically, leading to underutilized resources and a higher and more carbon intensive operations.

- **Hardware optimization from purpose-built silicon (AI workload only):** One of the most visible ways AWS improves power efficiency is by investing in AWS chips. AWS's latest Graviton processor, Graviton4, delivers the best price performance and energy efficiency for a broad range of workloads running on Amazon EC2. AWS Trainium is a high-performance ML chip designed to reduce the time and cost of training generative AI models—cutting training time for some models from months to hours. This, in turn, cuts down on the money and power required for building new models, with potential cost savings as well as energy-consumption reductions.[12] The AI inference chip, AWS Inferentia, is also built for improved sustainability. Inferentia2 AI delivers up to 50% better performance per watt against comparable instances.[13] AWS's purpose-built silicon helps them achieve higher throughput than comparable compute-accelerated instances, enabling efficient execution of complex AI models like LLMs at scale. This translates to a reduced infrastructure footprint for similar workloads, resulting in enhanced performance per watt of power consumption.[14]

- **Hardware optimization through efficient replication (storage workload only):** AWS enables workload owners to modernize their storage strategies by efficiently segregating active and inactive data sets using managed services. Additionally, workload replications can be optimized by reducing their size and throughput, leading to decreased energy consumption associated with storage.

- **Additional carbon-free energy procured by AWS compared to on-premises:** In 2023, Amazon was the world's largest corporate purchaser of renewable energy for the fourth year in a row, according to BloombergNEF and publicly available sources. It has enabled more than 500 renewable energy projects worldwide as of January 2024.[15] Amazon is on track to match electricity use with 100% renewable energy by 2025, and the electricity consumed in 19 AWS regions was attributable to 100% renewable energy in 2022.[16] Carbon-free procurement is usually more of a challenge for on-premises data centers given AWS can acquire in larger tranches resulting in a likely cost benefit compared to on-premises operators. This difference between carbon-free energy procurement by AWS and on-premises is a major reduction lever for certain regions.



AWS also has a range of tools and resources to help organizations achieve their sustainability goals. Professional services and solutions architecture teams are knowledgeable about sustainability and work with organizations to advise them on best practices. For example, the AWS Well-Architected pillar for sustainability provides guidance to organizations on how to modernize workloads.

Conclusion

As global data centers' electricity use continues to grow, including compute power required to train and run AI models, there is an increasing strain on traditional data centers. It is crucial that the associated energy demands, and associated carbon emissions are addressed.

This research demonstrates that migrating compute- and storage-heavy workloads to AWS can reduce emissions across different regions. For the reference workloads, customers running compute-heavy workloads can see carbon emissions reduce by up to 99%, with reduction of up to 94% by moving from on-premises to AWS and an additional reduction up to 81% using AWS's purpose-built silicon. Similarly for storage-heavy workloads, customers can reduce carbon emissions by up to 93%, with reduction of up to 88% by moving from on-premises to AWS and an additional reduction up to 47% by leveraging a modernized AWS architecture.

While the energy requirements of generative AI such as large language models are daunting—and this study did not cover its potential optimization on AWS—there are strategies to mitigate their associated carbon emissions as well. Techniques like model optimization, efficient hardware utilization and transitioning to carbon-free energy sources can reduce the carbon footprint of these workloads. Moreover, generative AI can contribute to grid decarbonization by enabling energy utilities to optimize energy production and distribution, enhance operational efficiency and improve safety and emissions transparency.

Ultimately, as customers mature in their digital and AI journey, it will be imperative to prioritize more sustainable practices—embracing cloud migration, optimizing workload efficiency and accelerating the transition to carbon-free energy. By leveraging AWS, the potential of AI can be realized while minimizing the associated carbon emissions.

Organizations should focus on three key actions as they consider the transition to the cloud with sustainability goals in mind:

1. **Migrate on-premises workloads to the cloud:** Even a simple lift-and-shift migration can significantly reduce carbon emissions because cloud providers like AWS can achieve higher hardware utilization rates, better power and cooling efficiencies and greater use of carbon-free energy sources compared to typical on-premises data centers. Additionally, cloud providers like AWS benefit from lower embodied emissions due to the large scale of their operations.

2. **Use optimized configurations for improved results:** AWS offers hardware optimization, including purpose-built silicon, which can substantially improve carbon efficiency for workloads running in the cloud. This is an additional benefit on top of the simple lift-and-shift approach.

3. **Consider regional variations in data center carbon emissions:** The geographic location of a cloud provider's data centers affects the total carbon emissions associated with running workloads in those facilities. To minimize emissions, organizations should select an optimal mix of regions while considering any relevant regulatory requirements.

Contributing team

Key experts

AWS

Prasad Kalyanaraman
VP of Infrastructure Services

Chris Walker
Director of AWS Sustainability

Jenna Leiner
Head of ESG and External Engagement

Margaret O'Toole
Tech Lead for Sustainability

Accenture

Sanjay Podder
Managing Director – Global Green Cloud, Software & AI Lead

Akshay Kasera
Strategy Lead – Green Cloud, Software & AI

Josh Whitney
Managing Director - Sustainability Measurement Analytics & Performance

Dr. Vibhu S. Sharma
Research Lead – Green Cloud, Software & AI

Acknowledgements

We'd like to express our thanks for the valuable insights and guidance of Charles Inglis, Ryan Bradley and Avinash Murthy from AWS and Navveen Balani, Nasim Shomali, Shalabh Kumar Singh, Matthew C. Robinson, Rohit Mehra, Bhushan Jagyasi, Sayali Karekar, Siddharth Gupta, Sanjeet Kumar De, Tulika Poddar, Giju Mathew and Ramani Moses from Accenture for their contributions to this study.

Appendix

Methology

Accenture leveraged the **Software Carbon Intensity** (SCI), an ISO-certified standard, to quantify potential carbon emissions from on-premises versus cloud applications in selected use cases.

SCI is defined as a rate or intensity of carbon emissions per functional unit. The equation used to calculate SCI score of a workload is calculated as:
SCI = C per R

Where:

- **C** is the total amount of carbon the workload emits
- **R** is the functional unit (e.g., carbon emissions per additional user, API-call or ML training run)

Estimating operational emissions for an on-premises equivalent:

Accenture's estimation of on-premises data center (DC) energy consumption involves a detailed analysis of individual workload components, which vary depending on the workload type (for example, AI or storage). This energy consumption is then multiplied by the grid emission factor to determine location-based emissions. The calculation is further refined by considering typical regional carbon-free energy procurement practices, leading to market-based emissions. Finally, regional factors like power usage effectiveness (PUE) and data center idle capacity are applied for a comprehensive assessment.

Specific components considered for AI workloads include compute (CPU, GPU, etc.), storage disks and data center network infrastructure. For storage workloads, the analysis encompasses compute (CPU, etc.), storage solutions for block and object storage and data center network infrastructure.

Estimating embodied emissions for an on-premises equivalent:

In determining embodied emissions, Accenture employed a methodology that assesses lifecycle emissions for each server component across both workload types. These emissions were then annualized by dividing them by the lifecycle data provided by server manufacturers. To attribute these emissions to the specific workload running on the server, a ratio was calculated based on the workload's resource consumption compared to the total server resources.

AWS operational and embodied emissions

AWS workload emissions for Lift-and-Shift and AWS-Optimized Scenarios across both workload types were estimated in accordance with the SCI standard as well as including corporate carbon-free energy procurement. While operational emissions were derived through AWS's simulation of the most likely scenario, Accenture modelled AWS's embodied emissions leveraging a methodology similar to an on-premises workload. Accenture has used the parameters required as inputs to estimate AWS's emissions intensity directly, as shared by AWS.

Four variables selected to analyze and compare data centers across regions

1. Power usage effectiveness (PUE): A measure of data center efficiency calculated as the ratio of total power consumption (including cooling and lighting) to the power consumed by IT equipment. A lower PUE value indicates greater data center efficiency, resulting in lower carbon emissions.

2. Idle capacity: The percentage of compute resources (such as servers and storage) in the data center that are online but not being used, kept idle to handle peaks and future growth.

3. Grid electricity—location-based method (LBM) emissions factor: The greenhouse gas emission intensity of the local electricity grid where the data center operates.

4. Carbon-free energy—market-based method (MBM) emissions factor: MBM emissions factor estimated by applying a carbon-free energy percentage per region on top of the grid carbon intensity. This method uses the residual grid mix, which accounts for allocation of carbon-free energy procurement through market instruments such as Renewable Energy Certificates (RECs) and Power Purchase Agreements (PPAs).

Generative AI workloads: Take a lifecycle approach

Organizations can make significant strides via optimization across data efficiency, algorithms, infrastructure utilization and hardware efficiency. Organizations deploying LLMs and generative AI should critically assess their technical infrastructure, architecture, operational model and governance. This helps to ensure they can handle the high computational demands while keeping costs and energy consumption in check.

Accenture recommends assessing the lifecycle of generative AI models as an important first step in evaluating organizational "green quotient" and creating transparency to facilitate development of green energy metrics and energy-efficient decisions. Starting with smaller models with shorter context windows is recommended. These models require less power and resources, and specialized smaller models can even achieve similar performance to larger ones for specific tasks. This approach allows for scaling up model size and complexity only if necessary.

Organizations can adopt readily available best practices using AWS at each stage of the lifecycle. For example, to train models from scratch, organizations can use efficient silicon, high-quality data and scalable data curation and adopt distributed training. AWS's purpose-built hardware offers high-performance deep learning model training with optimized energy efficiency. In 2022, training models on this infrastructure helped reduce energy consumption by up to 29% compared to comparable instances.[17] Additionally, the use of chips designed for faster performance—with less precise computations—is likely to stem the inflation of compute costs.

Similarly, for model inference and deployment, AWS helps organizations use deep learning containers for large model inference, set appropriate inference model parameters, adopt efficient inference infrastructure and align inference Service Level Agreement (SLA) with sustainability goals.

Choose wisely

It's crucial to evaluate the cost-benefit analysis of generative AI compared to alternative AI or analytical approaches that might be better suited for specific tasks, often at a fraction of the computational cost. A useful categorization can be to bucket AI techniques into diagnostic, predictive and generative approaches.

Diagnostic AI techniques help explain why something happened. They can provide insights into known problems, create scenarios, rapidly and accurately segment datasets and even establish likely solutions based on which part of the system is failing.

Predictive AI techniques can interpret underlying patterns and forecast what might happen in the future, thus helping simulate the most likely future scenarios, optimize operations and recommend solutions based on inference. It can be highly relevant in shaping business strategy, managing evolving customer expectations and predictive and preventive maintenance of assets and infrastructure, among other things.

Generative AI techniques are particularly effective during the execution phase, where the aim is to produce new content, automate processes or provide recommendations. These methods not only excel in creating innovative machine learning algorithms but also in enhancing data sets to boost the performance of these systems. Additionally, they are useful for automating mundane tasks and improving more complex activities, such as responding to customer questions and promoting products and services.

For many applications, a simpler diagnostic AI technique will be more appropriate, and more sustainable, than a generative AI approach.

Focus on net zero

Not all aspects of the technology are related to its hardware and software; the source of the energy it uses is also critically important. Generative AI can help businesses in their transition to carbon-free energy. It can accelerate the energy transition by optimizing carbon-free energy generation, storage and distribution, reducing greenhouse gases and boosting energy efficiency. It can predict energy demand and pinpoint optimal locations for solar panels and wind turbine designs based on weather patterns.

Beyond the transition to carbon-free energy, generative AI is emerging as a powerful tool for organizations aiming to enhance their sustainability initiatives. By leveraging generative AI capabilities in data analysis, pattern recognition and automation, organizations can optimize various aspects of their value chain. However, success hinges on using high-quality data and selecting the right tools for specific needs.[18]

It is still early days, but it is already clear that the sustainability promise of generative AI is unparalleled. In the next few years, the evolution of generative AI and its uses is likely to see explosive growth. If properly managed, this creative side of technology can empower businesses to build a more secure and resilient future.

References

1 https://iea.blob.core.windows.net/assets/6b2fd954-2017-408e-bf08-952fdd62118a/Electricity2024-Analysisandforecastto2026.pdf

2 https://www.accenture.com/content/dam/system-files/acom/custom-code/ai-maturity/Accenture-Art-of-AI-Maturity-Report-Global-Revised.pdf

3 https://techhq.com/2024/01/how-the-demands-of-ai-are-impacting-data-centers-and-what-operators-can-do/

4 https://techhq.com/2024/01/how-the-demands-of-ai-are-impacting-data-centers-and-what-operators-can-do/

5 https://www.aboutamazon.com/news/sustainability/aws-decarbonizing-construction-data-centers

6 https://www.accenture.com/content/dam/accenture/final/capabilities/technology/cloud/document/Accenture-Cloud-Data-Value-A-New-Dawn-for-Dormant-Data-vF.pdf

7 https://www.energy.gov/femp/energy-efficiency-data-centers

8 https://joint-research-centre.ec.europa.eu/jrc-news-and-updates/eu-code-conduct-data-centres-towards-more-innovative-sustainable-and-secure-data-centre-facilities-2023-09-05_en#:~:text=Data%20centres%20consume%20a%20significant%20amount%20of%20energy%2C,accounting%20for%201.4-1.6%25%20of%20total%20EU%20electricity%20consumption.

9 https://e3p.jrc.ec.europa.eu/communities/data-centres-code-conduct

10 https://www.mordorintelligence.com/industry-reports/asia-pacific-data-center-market

11 https://sustainability.aboutamazon.com/products-services/the-cloud?energyType=true

12 https://aws.amazon.com/machine-learning/trainium/

13 https://aws.amazon.com/machine-learning/inferentia/

14 https://sustainability.aboutamazon.com/products-services/the-cloud?energyType=true

15 https://sustainability.aboutamazon.com/climate-solutions/carbon-free-energy?energyType=true

16 https://sustainability.aboutamazon.com/products-services/the-cloud?energyType=true

17 https://aws.amazon.com/blogs/machine-learning/optimize-generative-ai-workloads-for-environmental-sustainability/

18 https://aws.amazon.com/blogs/machine-learning/the-executives-guide-to-generative-ai-for-sustainability/