RETII, v. 3, n. 1, 2026
ISSN: 2966-2508

Revista de Engenharia, TI e Inovação - RETII

ENERGY CONSUMPTION IN DATA CENTERS AND EFFICIENCY STRATEGIES: A LITERATURE REVIEW

Leandro Aureliano da Silva[1]; Eduardo Silva Vasconcelos[2]; Luiz Fernando Ribeiro de Paiva[1]; Cleiton Silvano Goulart[1]; Welington Mrad Joaquim[1]; Marcelo Eustáquio Pereira Elias[1]; Edilberto Pereira Teixeira[1]; Maria Heliodora do Vale Romeiro Collaço[1]

[1]Universidade de Uberaba – UNIUBE, Brasil
[2]Instituto Federal Goiano - IF-GO, Brasil

Autor Correspondente: leandro.silva@uniube.br

RESUMO

A digitalização do século XXI, impulsionada pela computação em nuvem, big data e inteligência artificial generativa, tem intensificado o crescimento dos data centers como infraestrutura crítica da economia digital. Seu consumo energético saltou de 80 TWh em 2000 para 620 TWh em 2024, representando atualmente 0,43% da energia primária global, com projeções de atingir entre 1.000 e 1.600 TWh até 2030. Este artigo apresenta uma revisão de literatura sobre o consumo energético de data centers e as principais estratégias de eficiência disponíveis, abrangendo técnicas de resfriamento, otimização de hardware, orquestração de cargas de trabalho e integração com fontes renováveis. Os resultados evidenciam avanços significativos, mas também lacunas regulatórias e disparidades geográficas que demandam políticas públicas específicas.

Palavras-chave: data centers; consumo energético; eficiência energética; energias renováveis; inteligência artificial.

ABSTRACT

The twenty-first-century digitalization wave, propelled by cloud computing, big data, and generative artificial intelligence, has established data centers as critical infrastructure underpinning the global digital economy. Their energy consumption has surged from 80 TWh in 2000 to 620 TWh in 2024, currently accounting for 0.43% of global primary energy, with projections of reaching between 1,000 and 1,600 TWh by 2030. This article presents a literature review of data center energy consumption and the principal efficiency strategies available, covering cooling

techniques, hardware optimization, workload orchestration, and renewable energy integration. The findings highlight significant technical advances alongside persistent regulatory gaps and geographic disparities that demand targeted public policy responses.

Keywords: data centers; energy consumption; energy efficiency; renewable energy; artificial intelligence.

1 INTRODUCTION

The twenty-first-century digitalization wave, propelled by cloud computing, big data, and generative AI, has profoundly reconfigured economic, scientific, and social systems. Against this backdrop, data centers have consolidated their role as mission-critical infrastructure for the processing and storage of the information that sustains digital platforms, public services, financial systems, and large-scale machine-learning models (Silva *et al*., 2025).

Despite their centrality to the digital economy, the energy footprint of data centers is routinely underestimated. The majority of published studies consider only direct electricity consumption, overlooking the full conversion chain from primary sources such as coal, natural gas, petroleum, hydropower, or nuclear down to final end-use. Bodies including the International Energy Agency (IEA, 2023) and the Intergovernmental Panel on Climate Change (IPCC, 2022) stress the importance of adopting primary energy as the unit of analysis, enabling a more systemic understanding of the pressure that digital infrastructure places on global energy systems.

In 2022, data centers consumed approximately 460 TWh of electricity, roughly 2% of total world electricity demand, with projections pointing to a doubling by 2030 (IEA, 2023; Uptime Institute, 2024). The mainstreaming of generative AI from 2022 onward intensified that growth trajectory, posing new challenges to energy sustainability (Deloitte, 2024). Viewed through the lens of global primary energy, estimated at 141,500 TWh in 2022, data center consumption corresponds to just 0.33%, underscoring the need for more comprehensive metrics to guide both public policy and corporate strategy (SILVA *et al*., 2025).

Against this background, the present article conducts a literature review of data center energy consumption and the principal techniques employed to reduce it, encompassing energy efficiency strategies, renewable energy integration, and advances in hardware and physical infrastructure. The research is justified by the growing relevance of the topic in the context of global climate commitments and by the need to inform technical, regulatory, and strategic decision-making across the information technology sector.

2 THEORETICAL FRAMEWORK

2.1 Energy Infrastructure of Data Centers

Data centers have emerged as pillars of the digital economy, storing, processing, and distributing massive volumes of data that support cloud services, social networks, e-commerce, digital banking, and AI applications. Despite their socioeconomic benefits, concern is mounting over their energy impact, particularly in the face of the global climate crisis (IEA, 2023; Deloitte, 2024).

The sector's electricity consumption doubled between 2000 and 2005, then stabilized at around 1% to 1.3% of global electricity use through 2018, reflecting the efficiency gains achieved during that period (Koomey, 2008; Koomey, 2011; IEA, 2023). From 2022 onward, however, the widespread adoption of generative AI drove renewed acceleration, with consumption jumping from 205 TWh in 2020 to 620 TWh in 2024, a 675% increase relative to the year 2000 (Silva *et al*., 2025).

The literature distinguishes two main approaches to measuring this impact: direct electricity consumption, which is straightforward to quantify, and primary energy analysis, which accounts for the entire energy chain from extraction through distribution (Silva *et al*., 2025). The latter approach, recommended by the IPCC (2022) and the IEA (2023), provides a more systemic and epistemologically robust picture of the effects of digitalization on global energy systems.

2.2 Efficiency Metrics: PUE, DCiE, and CUE

Energy efficiency in data centers is assessed through a set of dedicated metrics. The most widely used is Power Usage Effectiveness (PUE), defined as the ratio of total facility energy consumption to the energy consumed exclusively by IT computing equipment. PUE values approaching 1.0 denote high efficiency, while values above 2.0 are considered wasteful (The Green Grid, 2008).

Complementary metrics include the Data Center Infrastructure Efficiency (DCiE), which is the reciprocal of PUE expressed as a percentage, and the Carbon Usage Effectiveness (CUE), which incorporates the carbon emissions associated with energy consumption (Uptime Institute, 2024). Using these metrics in combination enables a more complete assessment of a data center's environmental performance, capturing dimensions that electricity-only analyses fail to reflect (Silva *et al*., 2025).

3 METHODOLOGY

The present work is characterized as a narrative literature review, qualitative in nature and descriptive-exploratory in scope. Following the methodological guidelines of Gil (2019), the research systematizes the state of the art on data center energy consumption and available mitigation techniques, with an emphasis on publications from the past ten years.

The bibliographic search was conducted across the IEEE Xplore, ACM Digital Library, Scopus, and Web of Science databases, supplemented by technical reports from international bodies including the IEA, IPCC, Uptime Institute, and Deloitte. The descriptors used were: data center energy consumption, PUE efficiency, cooling techniques, renewable energy data centers, and hyperscale infrastructure, in both English and Portuguese. Priority was given to articles published between 2015 and 2025, without excluding seminal earlier works recognized in the field, such as Koomey (2008; 2011) and Smil (2017).

The analysis and synthesis of sources followed the approach of Lakatos and Marconi (2017), with results organized thematically into three categories: consumption landscape, energy efficiency techniques, and renewable energy integration.

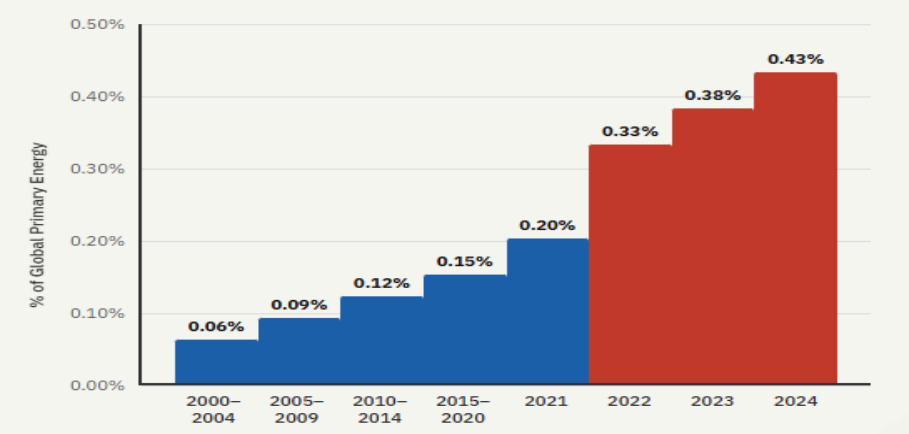
Part of the quantitative data mobilized in this review, particularly the primary energy series and geographic distribution figures, derives from Silva *et al*. (2025), a companion study by the same research group that established the methodological framework subsequently applied here.

4 RESULTS AND DISCUSSION

4.1 Global Energy Consumption Landscape

Data center energy consumption has grown steadily over the past two decades, with marked acceleration from 2021 onward. Between 2000 and 2004 the sector represented approximately 0.06% of global primary energy; that share rose to 0.09% between 2005 and 2009, and to 0.12% between 2010 and 2014, driven by the mainstreaming of cloud computing. From 2015 to 2020 growth stabilized at around 0.15%, reflecting a period of relative sectoral maturity. From 2021, the large-scale adoption of AI models pushed the share to 0.20%, reaching 0.33% in 2022, 0.38% in 2023, and 0.43% in 2024 as shown in Figure 1. (SILVA *et al*., 2025).

Figure 1 – Evolution of data center energy consumption as a share of global primary energy (2000–2024)



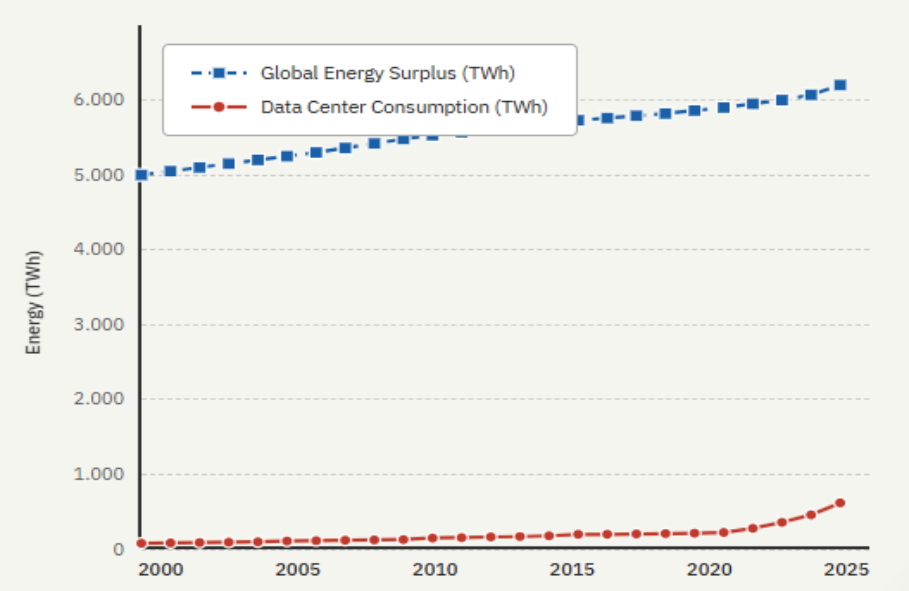
Source: Silva *et al*. (2025), adapted.

The geographic concentration of this consumption is striking: the United States and China together account for 58% of global data center energy use, at 31% and 27% respectively, followed by the United Kingdom (9%), France (8%), and Italy (6%) (Silva *et al*., 2025). This asymmetry carries significant geopolitical implications. Countries with limited generation capacity become vulnerable to the installation of large processing facilities without commensurate electrical infrastructure investment, while regions endowed with abundant clean energy attract capital flows that can entrench structural inequalities (Jones, 2018; SMIL, 2017).

In addition to the geographic concentration, a consistent upward trend in energy consumption associated with digital infrastructure can be observed. As shown in Figure 2, the energy consumption of data centers exhibited continuous growth between 2000 and 2024, accompanying the global expansion of cloud computing, artificial intelligence, and digital services. This increase occurs in parallel with the evolution of the global energy surplus, indicating growing pressure on power generation and distribution systems.

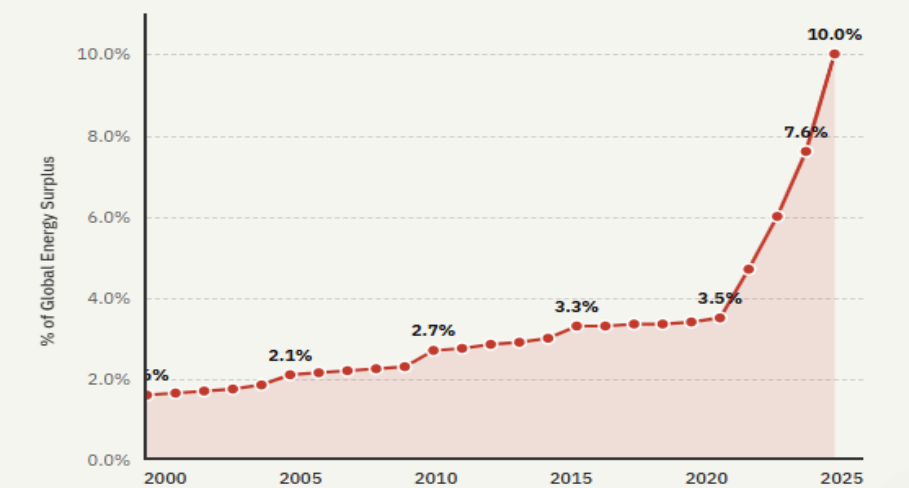
Even more significantly, Figure 3 shows that the relative share of data center energy consumption in the global energy surplus increased substantially over the analyzed period, rising from approximately 1.6% in 2000 to around 10% in 2024. This growth indicates that digital infrastructure has become one of the main drivers of energy demand in the context of the contemporary digital economy.

Figure 2 – Evolution of data center consumption and global energy surplus (2000–2024)



Source: Silva *et al*. (2025), adapted.

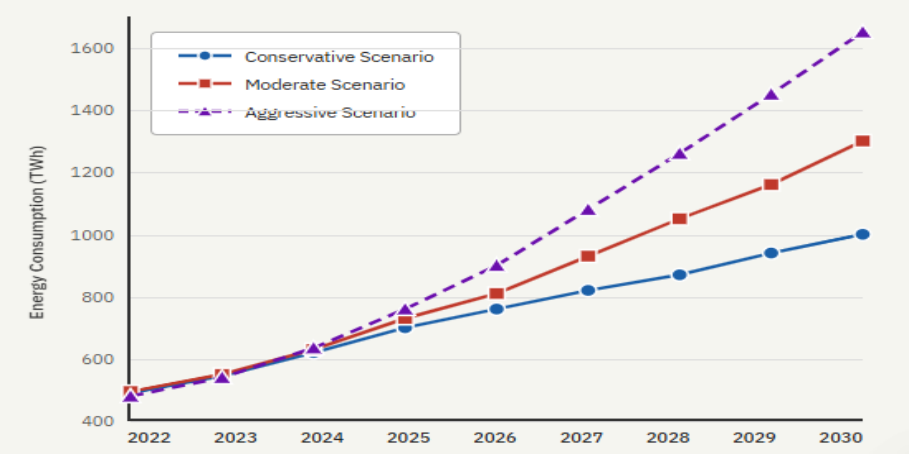
Figure 3 – Data center consumption as a percentage of global energy surplus (2000–2024)



Source: Silva *et al*. (2025), adapted.

Projections for 2030 indicate that consumption could range from 1,000 TWh (conservative scenario) to 1,600 TWh (aggressive scenario), depending on the pace of advanced AI adoption, the expansion rate of hyperscale infrastructure, and the effectiveness of energy efficiency policies (Silva *et al*., 2025; IEA, 2023).

Figure 4 – Data center energy consumption projections to 2030 (conservative, moderate, and aggressive scenarios)

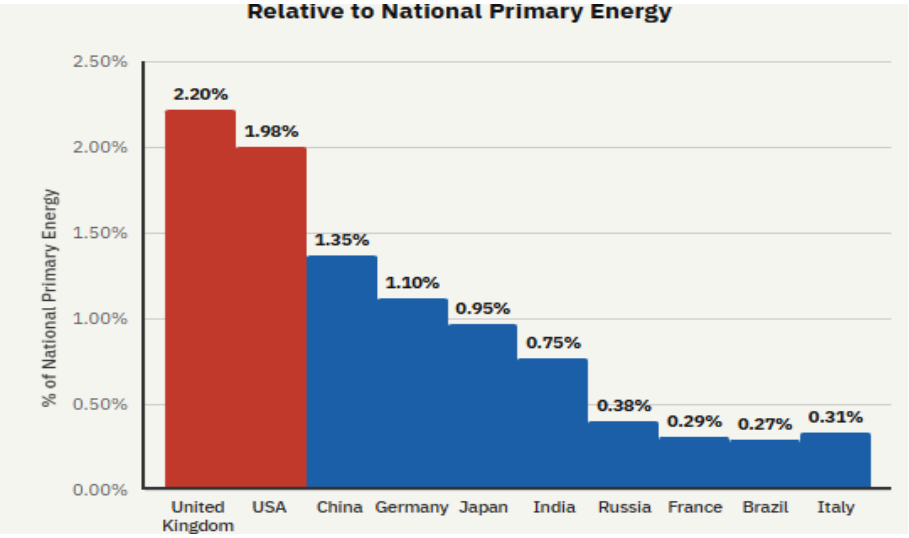


Source: Silva *et al*. (2025), adapted.

4.2 Regional Consumption Patterns

The regional analysis reveals significant disparities in both the pressure placed on national energy mixes and the share of global consumption. As illustrated in Figure 5, the United Kingdom (2.20%), the United States (1.98%), and China (1.35%) present the highest proportions of national primary energy consumption attributed to data centers. This pattern reflects the strong concentration of hyperscale digital infrastructure in these countries, which host a large share of global cloud computing services and generative artificial intelligence workloads. In contrast, Brazil represents a comparatively smaller share of national energy consumption related to data centers (0.27%), although its participation remains relevant within the global context.

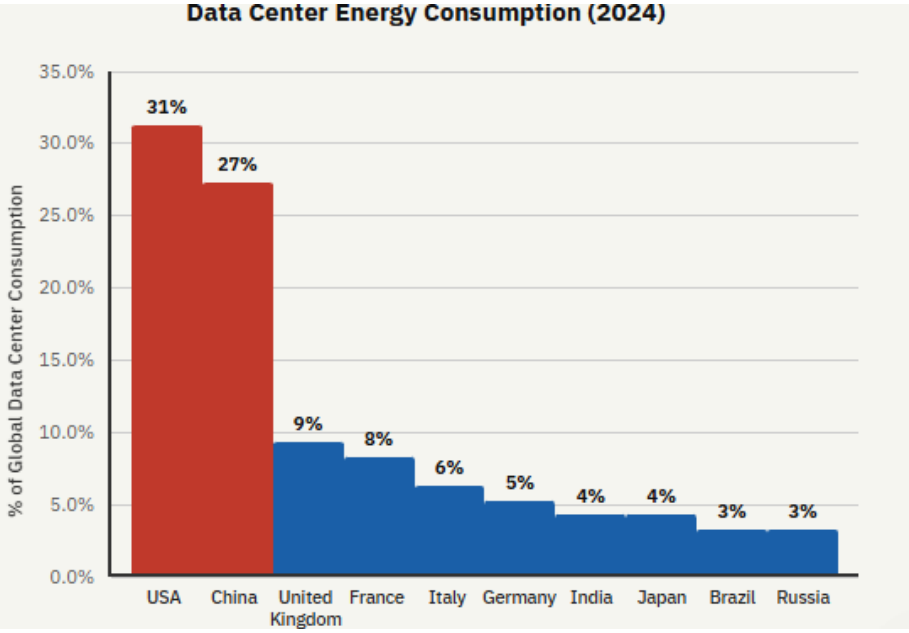
Figure 5 – Data center energy consumption in 2024 relative to national primary energy



Source: Silva *et al*. (2025), adapted.

From a global perspective, Figure 6 highlights the distribution of worldwide data center energy consumption, showing that the United States and China dominate the sector with shares of approximately 31% and 27%, respectively. These values reinforce the central role of these countries in the global digital infrastructure ecosystem. Meanwhile, countries such as the United Kingdom, France, and Italy contribute smaller shares but still represent important nodes in the international data center network. In this context, Brazil accounts for approximately 3% of global data center energy consumption (Silva *et al*., 2025).

Figure 6 – Each country's share of global data center energy consumption (2024)



Source: Silva *et al*. (2025), adapted.

4.3 Energy Efficiency Techniques

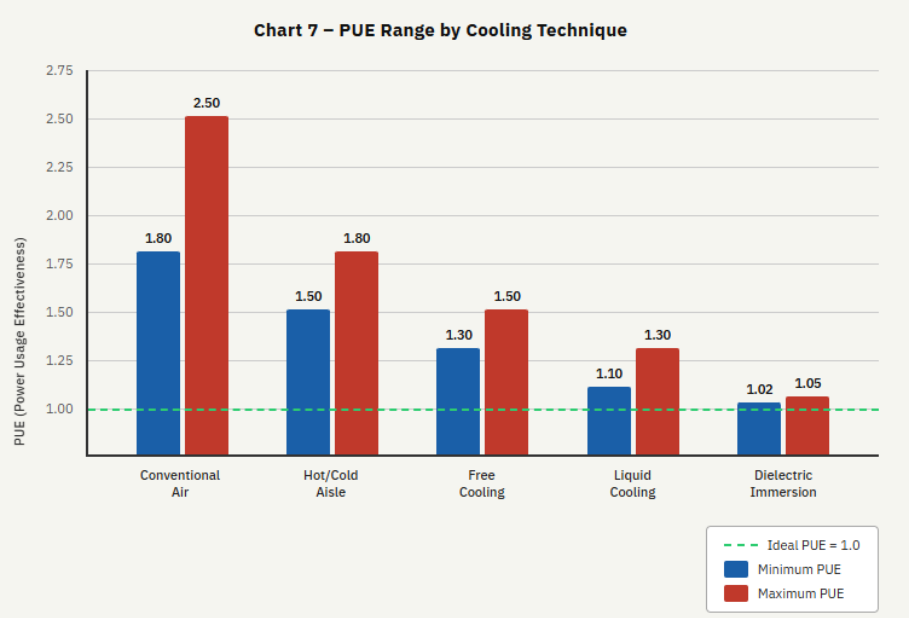
4.3.1 Cooling Systems

Cooling has historically been responsible for a substantial share of data center energy consumption. Studies indicate that cooling systems can account for 38% to 40% of a facility's total electricity use (Huang *et al*., 2024), making them the second-largest consumption component after IT equipment.

The literature distinguishes two broad design approaches: passive design (PD), which encompasses architectural strategies that reduce thermal load without additional energy expenditure, and active design (AD), which includes air-based, liquid-based, free cooling, and two-phase cooling systems. A comprehensive review of the literature from 2005 to 2024 concludes that passive design remains underexplored in the research community (Huang *et al*., 2024).

Hot aisle/cold aisle containment, which physically segregates hot and cold airflows between server racks, is one of the lowest-cost, highest-return interventions available, preventing airstream mixing and significantly improving cooling system performance (FEMP; NREL, 2024). As illustrated in Figure 7, different cooling techniques present distinct ranges of Power Usage Effectiveness (PUE), highlighting the efficiency gains achieved through advanced cooling technologies. Conventional air-based cooling systems typically operate at higher PUE levels, while techniques such as hot aisle/cold aisle containment and free cooling can significantly reduce overall energy consumption.

Figure 7 – PUE range by cooling technique



Source: compiled by the authors based on Koomey (2011), Huang *et al*. (2024), and FEMP/NREL (2024).

For high-density computing environments, liquid cooling and dielectric immersion cooling have emerged as leading solutions, achieving PUE values much closer to the theoretical ideal of 1.0. As shown in Figure 7, immersion cooling presents the lowest PUE range among the analyzed techniques, indicating superior energy efficiency. Consistent with these trends, transitioning from 100% air cooling to 75% liquid cooling can reduce a facility's total energy consumption by approximately 15.5% (Chen *et al*., 2024). The global direct liquid cooling (DLC) market is projected to grow from USD 1.85 billion in 2023 to USD 11.89 billion over the next decade, driven largely by the increasing computational demand of generative AI workloads.

4.3.2 Hardware and Infrastructure Efficiency

Data center energy efficiency increasingly hinges on hardware-level decisions, from processor architecture to server consolidation strategies and dynamic power management. This subsection presents the principal approaches identified in the literature, organized around three dimensions: processor architecture, server virtualization and consolidation, and dynamic power management techniques.

a) Processor Architecture: The Shift to ARM and Specialized Accelerators

Computing systems, including CPUs and GPUs, account for approximately 40% of a data center's electricity consumption, while networking and storage equipment contribute a further 10% (Huang *et al*., 2024). Historically dominated by the x86 architecture (Intel and AMD), the data center processor market is undergoing a significant transition toward ARM-based designs. The ARM RISC (Reduced Instruction Set Computing) model delivers superior performance per watt compared to the x86 Complex Instruction Set Computing (CISC) architecture (Anderson, 2025).

All three leading cloud providers have now commercially deployed their own custom ARM-based CPUs for data center use. Amazon Web Services (AWS) launched the fourth generation of its Graviton processor in 2023/2024: the Graviton4, featuring 96 Neoverse V2 cores, delivers up to 30% more compute performance, 50% more cores, and 75% greater memory bandwidth than the Graviton3, while consuming up to 60% less energy than comparable x86-based EC2 instances (AWS, 2024a; AWS, 2024b). In 2024, Amazon's internal adoption of Graviton resulted in an estimated reduction of 71,000 metric tonnes of CO_2 equivalent (Amazon Sustainability, 2024). Google Cloud launched the Axion processor in 2024, its first custom ARM-based CPU for data centers, built on ARM's Neoverse V2 architecture and reporting up to 60% better energy efficiency and up to 50% higher performance relative to comparable x86 instances (Google Cloud, 2024a; Google Cloud, 2024b). Microsoft made the Azure Cobalt 100 generally available in October 2024; based on the Neoverse N2 architecture, it is positioned as Microsoft's most energy-efficient compute offering, delivering up to 50% better price-performance than the previous generation of ARM-based VMs (Microsoft, 2024a; Microsoft, 2024b). In November 2025, Microsoft announced the Cobalt 200, with more than 50% performance gains over the Cobalt 100, with energy efficiency retained as a central design pillar (Microsoft, 2025).

Beyond general-purpose CPUs, the use of specialized accelerators for AI workloads is growing rapidly. AWS offers the Inferentia2 chip for AI model inference, delivering up to 50% better performance per watt than comparable EC2 instances, and the Trainium2 for AI model training, achieving up to 2× greater energy efficiency than the Trainium1 (Amazon Sustainability, 2024). Google launched the Ironwood TPU in 2025, with 10× greater peak performance than the TPU v5p and more than 4× more performance per chip than the TPU v6e, described as the most energy-efficient chip Google has ever built (Google

Cloud, 2025). According to the IEA (2023), combining ARM-based CPUs with purpose-built accelerators represents one of the most promising pathways to contain sector energy growth in a scenario of rapidly expanding AI adoption.

b) Server Virtualization and Consolidation

Server virtualization is one of the most established energy efficiency strategies in data centers, enabling multiple virtual machines (VMs) to share the resources of a single physical machine (PM). This approach, known as server consolidation, increases equipment utilization rates and reduces the total number of active machines, with a direct impact on energy consumption and cooling costs (IEEE Communications Society, 2025).

Average physical server utilization in data centers ranges between 10% and 50% of capacity, and most servers consume approximately 70% of their peak power draw even in idle states (Mostpolicyinitiative, 2025). Virtualization-based consolidation therefore not only eliminates redundant servers but also reduces energy waste from underutilized machines. Studies report energy savings of up to 50% through VM-based consolidation with optimized scheduling (IEEE Communications Society, 2025).

Live VM migration complements consolidation by enabling workload transfers between physical servers with minimal downtime, allowing idle servers to be powered down. A power-mode-aware consolidation algorithm proposed by Lin *et al*. (2024) demonstrated a 15.3% reduction in energy consumption compared to baseline algorithms, leveraging selective server shutdown and temperature-driven live migration. Deep Reinforcement Learning (DRL) algorithms have been applied to automate VM migration and placement decisions, simultaneously considering energy consumption, server temperature, and Service Level Agreements (SLAs). A systematic review published in 2025 identified 65 peer-reviewed articles applying RL/DRL algorithms to data center energy optimization, with more than 60% of the studies focused on IT systems (Kahil *et al*., 2025).

c) Dynamic Power Management: DVFS and Power Capping

Dynamic Voltage and Frequency Scaling (DVFS) is a technique that dynamically adjusts processor operating voltage and frequency in response to computational demand. Grounded in the physical relationship between dynamic power, voltage, and frequency, where dynamic power scales as the square of voltage multiplied by frequency, DVFS enables quadratic reductions in power draw relative to voltage reduction when workloads are light (Sciencedirect, 2024).

In data center deployments, DVFS yields modest savings of around 5% in underutilized servers, but can achieve peak consumption reductions of up to 20%, at the cost of a slight increase in response latency (Emergentmind, 2025). More sophisticated Service Level Objective (SLO)-driven frameworks that jointly tune CPU core frequency and memory controller frequency achieve 16% to 20% reductions in energy consumption while maintaining quality-of-service requirements. For GPU inference workloads, predictive DVFS models report 66% to 69% reductions in energy consumption relative to CPU-based reference implementations (Emergentmind, 2025).

Emerging approaches integrate Reinforcement Learning (RL) and meta-learning into DVFS, enabling continuous per-workload adaptation of voltage and frequency settings, outperforming traditional heuristic-based governors. Power capping is a complementary technique that enforces hard consumption ceilings on individual servers or server clusters, preventing demand spikes from overwhelming electrical infrastructure. When combined with VM consolidation and DVFS-aware scheduling, power capping reduces peak consumption and improves operational predictability, both critical enablers for integration with intermittent renewable energy sources (Kahil *et al*., 2025).

4.3.3 Workload Orchestration

Energy- and carbon-aware workload scheduling has become a strategically important research area. Thermal-aware scheduling approaches reduce heat generation and, consequently, cooling demand. Hybrid-aware workload scheduling integrates both thermal awareness and renewable energy availability, while strategies such as "follow the moon" migrate workloads to regions with higher renewable generation availability at any given time (Lin *et al*., 2023).

In cloud orchestration, carbon-aware scheduling on Kubernetes has emerged as a practical sustainability solution, enabling workload placement to be optimized according to the carbon intensity of power generation in each region. Researchers at Lehigh University developed a Mixed Integer Linear Programming (MILP) framework for workload scheduling optimization and energy management in AI data centers, achieving up to 20% in operational cost reductions through demand response strategies (Kishore *et al*., 2025).

4.4 Renewable Energy Integration

The decarbonization of data centers has advanced primarily through Power Purchase Agreements (PPAs), which are long-term contracts through which operators procure electricity directly from clean energy generators. In 2024, major internet companies accounted for 43% of all clean energy PPAs signed globally, establishing the technology sector as one of the primary drivers of renewable energy expansion (MCKINSEY, 2024).

Ambitions extend beyond annual carbon compensation. Companies such as Google and Microsoft are advancing toward a 24/7 clean energy model, which requires hourly matching of consumption with renewable generation. In early 2024, Google secured its largest-ever offshore PPA at 478 MW of capacity, achieving 90% hourly clean energy matching in its Netherlands operations. Microsoft has committed to becoming carbon-negative by 2030, with renewable energy powering all of its data centers (Mckinsey, 2024).

Progress in the European regulatory context has been particularly notable. In March 2024, the European Commission adopted a delegated act establishing a common framework for assessing the sustainability of data centers with a power

demand of 0.5 MW or more. From 2026, operators will be required to report sustainability indicators, including total electricity consumption and the share covered by renewable sources. Germany went further still: at the end of 2023 it enacted a new Energy Efficiency Act mandating that all operators source 50% of their consumption from renewables from January 2024, rising to 100% by 2027 (Aslan *et al*., 2025).

Despite these advances, full climate neutrality remains elusive. A 2025 analysis based on a German data center projects a 20% increase in energy consumption and a 13% increase in the total carbon footprint between 2020 and 2030, indicating that climate neutrality cannot be achieved through operational reduction measures alone. Scope 3 emissions arising from the hardware supply chain and facility construction are not fully eliminated even with 100% renewable electricity procured through PPAs (Aslan *et al*., 2025).

In the Brazilian context, the predominance of hydroelectric generation and the country's expanding wind and solar potential offer meaningful competitive advantages. Brazil's proportional data center consumption stands at 0.27% of national primary energy, with a 3% share of global sector consumption (Silva *et al*., 2025), indicating underutilized potential to attract investment in sustainable digital infrastructure. Public policies that reconcile foreign direct investment attraction with minimum energy efficiency and renewables requirements are essential for Brazil to capitalize on sectoral growth without compromising its energy security (Jones, 2018).

5 DISCUSSION

The literature analysis reveals that data center energy consumption is a multidimensional phenomenon encompassing technical, geopolitical, and regulatory dimensions. The efficiency techniques currently available, spanning cooling systems to workload management, already enable significant PUE reductions, yet their large-scale adoption continues to be hampered by implementation costs and the absence of mandatory regulations in many jurisdictions.

A critical issue identified in the literature is the divergence between the metrics in use. Analyses restricted to electricity consumption tend to overstate data centers' share of the global energy mix, generating both unwarranted alarmism and distortions in public policy design. The adoption of primary energy as the reference unit, as proposed by Silva *et al*. (2025) and endorsed by the IEA (2023) and the IPCC (2022), provides a more robust foundation for the formulation of sector-level strategies.

Emerging trends, such as the use of artificial intelligence for the operational energy management of data centers themselves, underwater facility concepts, and decarbonization via green hydrogen, signal that the sector is in accelerated transformation. Nevertheless, Smil (2017) and Jones (2018) caution that unchecked digitalization may generate a form of technological unsustainability disguised as innovation, rendering AI's energy demands a

challenge not only technical in nature, but political, environmental, and civilizational.

The geographic concentration of consumption, with the United States, China, and the United Kingdom accounting for approximately 80% of the global total (Silva *et al*., 2025), underscores the need for supranational regulatory mechanisms. Jones (2018) warns of the risk of so-called "computational colonialism," in which data processing is offshored to countries with cheap energy and weak regulation, deepening historical inequalities.

6 CONCLUSION

This literature review has demonstrated that data centers represent a growing, though still modest in absolute terms, share of global energy consumption. The sector has moved through distinct phases of expansion and stabilization, but the acceleration driven by generative AI from 2021 onward repositions the issue at the center of both energy and climate policy agendas.

The energy efficiency techniques currently available, particularly in cooling, hardware, and workload orchestration, already enable meaningful PUE reductions and operational cost savings, but their full adoption depends on public policy, regulatory incentives, and transparency from major technology corporations. Renewable energy integration is advancing, primarily via PPAs, but still faces challenges related to operational reliability and uneven geographic distribution.

Priorities for future research include the development of studies that incorporate carbon emission metrics across the full data center production chain, encompassing hardware manufacturing and facility construction, as well as comparative analyses of emerging regions such as Brazil within the context of the digital energy transition, both of which represent significant gaps in the current literature.

REFERENCES

AMAZON SUSTAINABILITY. **AWS Cloud Sustainability**. Amazon, 2024. Available at: https://sustainability.aboutamazon.com/products-services/aws-cloud. Accessed: jan. 10, 2026.

ANDERSON, M. The shift to ARM-based servers: introduction to a changing landscape in server architecture. **Medium**, Jun. 2025. Available at: https://medium.com/@mike.anderson007/the-shift-to-arm-based-servers-87aaa749ac19. Accessed: jan. 10, 2026.

ASLAN, T. *et al*. Toward climate neutral data centers: greenhouse gas inventory, scenarios, and strategies. **iScience**, v. 28, n. 1, p. 111637, Jan. 2025. DOI:

https://doi.org/10.1016/j.isci.2024.111637. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC11773490/. Accessed: jan. 10, 2026.

AWS – AMAZON WEB SERVICES. **ARM Processor** – AWS EC2 Graviton. 2024a. Available at: https://aws.amazon.com/ec2/graviton/. Accessed: jan. 10, 2026.

AWS – AMAZON WEB SERVICES. **Amazon EC2 R8g instances powered by AWS Graviton4 now generally available**. AWS What's New, Jul. 2024b. Available at: https://aws.amazon.com/about-aws/whats-new/2024/07/amazon-ec2-r8g-instances-aws-graviton4-generally-available/. Accessed: jan. 10, 2026.

CHEN, R. *et al*. Fiber membrane evaporative cooling for high-power electronics in data centers. **Joule**, 2024. DOI: https://doi.org/10.1016/j.joule.2024.02.020.

DELOITTE. As generative AI asks for more power, data centers seek more reliable, cleaner energy solutions. **Deloitte Insights**, 2024. Available at: https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2024/data-center-energy-consumption-demand-growth.html. Accessed: jan. 10, 2026.

EMERGENTMIND. **Dynamic voltage and frequency scaling (DVFS)**: overview and trends. 2025. Available at: https://www.emergentmind.com/topics/dynamic-voltage-and-frequency-scaling-dvfs. Accessed: jan. 10, 2026.

FEMP; NREL – FEDERAL ENERGY MANAGEMENT PROGRAM; NATIONAL RENEWABLE ENERGY LABORATORY. **Best practices guide for energy-efficient data center design**. Washington, D.C.: U.S. Department of Energy, Jul. 2024. Available at: https://www.energy.gov/sites/default/files/2024-07/best-practice-guide-data-center-design_0.pdf. Accessed: jan. 10, 2026.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 7th ed. São Paulo: Atlas, 2019.

GOOGLE CLOUD. **Introducing Google's new Arm-based CPU**. Google Cloud Blog, Apr. 2024a. Available at: https://cloud.google.com/blog/products/compute/introducing-googles-new-arm-based-cpu. Accessed: jan. 10, 2026.

GOOGLE CLOUD. **Try C4A, the first Google Axion Processor**. Google Cloud Blog, Oct. 2024b. Available at: https://cloud.google.com/blog/products/compute/try-c4a-the-first-google-axion-processor. Accessed: jan. 10, 2026.

GOOGLE CLOUD. **Ironwood TPUs and new Axion-based VMs for your AI workloads**. Google Cloud Blog, Nov. 2025. Available at: https://cloud.google.com/blog/products/compute/ironwood-tpus-and-new-axion-based-vms-for-your-ai-workloads. Accessed: jan. 10, 2026.

HUANG, J. *et al*. Data centers cooling: a critical review of techniques, challenges, and energy saving solutions. **Sustainable Computing: Informatics and Systems**, v. 42, p. 100989, 2024. DOI: https://doi.org/10.1016/j.suscom.2024.100989. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0140700724000458. Accessed: jan. 10, 2026.

IEA – INTERNATIONAL ENERGY AGENCY. **Data centres and data transmission networks**. Paris: IEA, 2023. Available at: https://www.iea.org/reports/data-centres-and-data-transmission-networks. Accessed: jan. 10, 2026.

IEA – INTERNATIONAL ENERGY AGENCY. **Electricity 2024**: analysis and forecast to 2026. Paris: IEA, 2024. Available at: https://www.iea.org/reports/electricity-2024. Accessed: jan. 10, 2026.

IEEE COMMUNICATIONS SOCIETY. **Energy efficiency in data centers**. IEEE Communications Society, 2025. Available at: https://www.comsoc.org/publications/tcn/2019-nov/energy-efficiency-data-centers. Accessed: jan. 10, 2026.

IPCC – INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE. AR6 **Synthesis Report**: Climate Change 2022. Geneva: IPCC, 2022. Available at: https://www.ipcc.ch/report/ar6/syr/. Accessed: jan. 10, 2026.

JONES, N. How to stop data centres from gobbling up the world's electricity. **Nature**, v. 561, n. 7722, p. 163-166, 2018. DOI: https://doi.org/10.1038/d41586-018-06610-y.

KAHIL, H. *et al*. Reinforcement learning for data center energy efficiency optimization: a systematic literature review and research roadmap. **Applied Energy**, v. 389, p. 125004, 2025. DOI: https://doi.org/10.1016/j.apenergy.2025.125004. Available at: https://www.sciencedirect.com/science/article/pii/S0306261925004647. Accessed: jan. 10, 2026.

KISHORE, P. *et al*. **MILP-based workload scheduling and energy management for AI data centers**. Energy and Buildings, 2025.

KOOMEY, J. G. Worldwide electricity used in data centers. **Environmental Research Letters**, v. 3, n. 3, 2008. DOI: https://doi.org/10.1088/1748-9326/3/3/034008. Available at: https://iopscience.iop.org/article/10.1088/1748-9326/3/3/034008. Accessed: jan. 10, 2026.

KOOMEY, J. G. **Growth in data center electricity use 2005 to 2010**. Oakland: Analytics Press, 2011.

LAKATOS, E. M.; MARCONI, M. A. **Fundamentos de metodologia científica**. 8th ed. São Paulo: Atlas, 2017.

LIN, H. *et al*. A novel virtual machine consolidation algorithm with server power mode management for energy-efficient cloud data centers. **Cluster Computing**, v. 27, n. 8, p. 11709–11725, 2024. DOI: https://doi.org/10.1007/s10586-024-04555-8. Available at: https://link.springer.com/article/10.1007/s10586-024-04555-8. Accessed: jan. 10, 2026.

LIN, Y. *et al*. Green-aware data centers: workload management, thermal management, and waste heat recovery. **Renewable and Sustainable Energy Reviews**, v. 187, p. 113761, 2023. DOI: https://doi.org/10.1016/j.rser.2023.113761.

McKINSEY & COMPANY. **How hyperscalers are fueling the race for 24/7 clean power**. McKinsey Insights, Dec. 2024. Available at:

https://www.mckinsey.com/capabilities/sustainability/our-insights/how-hyperscalers-are-fueling-the-race-for-247-clean-power. Accessed: jan. 10, 2026.

MICROSOFT. **Azure Cobalt processor-based virtual machines**. Microsoft Learn, 2024a. Available at: https://learn.microsoft.com/en-us/azure/virtual-machines/sizes/cobalt-overview. Accessed: jan. 10, 2026.

MICROSOFT. **How Azure Cobalt 100 VMs are powering real-world solutions**. Microsoft Azure Blog, Sep. 2024b. Available at: https://azure.microsoft.com/en-us/blog/how-azure-cobalt-100-vms-are-powering-real-world-solutions-delivering-performance-and-efficiency-results/. Accessed: jan. 10, 2026.

MICROSOFT. **Announcing Cobalt 200**: Azure's next cloud-native CPU. Microsoft Community Hub, Nov. 2025. Available at: https://techcommunity.microsoft.com/blog/AzureInfrastructureBlog/announcing-cobalt-200-azure%E2%80%99s-next-cloud-native-cpu/4469807. Accessed: Jan. 10, 2026.

MOSTPOLICYINITIATIVE. POWERING DATA CENTERS. MOST Policy Initiative, 2025. Available at: https://mostpolicyinitiative.org/science-note/powering-data-centers/. Accessed: jan. 10, 2026.

SCIENCEDIRECT. DYNAMIC VOLTAGE AND FREQUENCY SCALING. ScienceDirect Topics, 2024. Available at: https://www.sciencedirect.com/topics/computer-science/dynamic-voltage-and-frequency-scaling. Accessed: jan. 10, 2026.

SILVA, L. A. *et al*. Consumo energético dos data centers e a métrica da energia primária: análise global 2000–2024 e projeções até 2030. **Revista Observatório de la Economía Latinoamericana**, Curitiba, v. 23, n. 10, p. 01–28, 2025. DOI: https://doi.org/10.55905/oelv23n10-063.

SMIL, V. **Energy and civilization**: a history. Cambridge: MIT Press, 2017.

THE GREEN GRID. **Green grid metrics**: describing data center power efficiency. White Paper, 2008. Available at: https://www.thegreengrid.org. Accessed: Jan. 10, 2026.

UPTIME INSTITUTE. **Global data center survey results 2024**. New York: Uptime Institute, 2024. Available at: https://uptimeinstitute.com/2024-data-center-industry-survey-results. Accessed: jan. 10, 2026.