Green Cloud Computing: A Technical and Policy Framework for the AI Era

Dorinn Bun[1*] and Movsun Kuy[2]

[1]Management of Information System, Paragon International University, Phnom Penh City, Cambodia

[2]Faculty of Computer Science, University of Namur, Namur, Belgium

*Corresponding author

doi: https://doi.org/10.21467/proceedings.8.1.3

Abstract

The global adoption of artificial intelligence is causing an unsustainable increase in the energy consumption of data centers, threatening to reverse a decade of efficiency gains. While legacy green computing strategies held power consumption steady through the 2010s, the unique, high-intensity nature of AI workloads renders these measures insufficient. Projections now indicate that data center electricity demand could triple by 2030, consuming a significant share of global electricity. This paper provides a framework for this challenge by reassessing legacy measures and examining four critical domains: advanced thermal management, security and data sovereignty, AI as a tool for energy optimization, and the economic and policy landscape. The analysis concludes that the exponential scaling of AI, which pushes server rack power densities from 15 kW to over 100 kW, requires a paradigm shift toward new infrastructure like liquid cooling and on-site clean power. Actionable recommendations are provided for industry and policymakers to ensure the AI revolution can proceed sustainably.

Keywords: Data center energy, green computing, AI workloads

Introduction

For over a decade, the data center industry successfully decoupled computational growth from energy consumption through innovations like server virtualization and improved cooling. Between 2014 and 2016, for instance, U.S. data center energy use remained stable despite a threefold increase in computational output. However, the widespread deployment of power-hungry hardware accelerators for artificial intelligence has reversed this trend. AI has become the primary driver of new electricity demand, with analysis showing U.S. data center electricity use more than doubled between 2017 and 2023 almost entirely due to AI.[7] Unlike the variable workloads of the past, AI training jobs pin specialized hardware at maximum power for extended periods, shifting the bottleneck from software efficiency to the physical limits of power and cooling. Projections are stark: data center electricity demand in the U.S. could grow from 4% of the national total in 2023 to over 9% by 2030[8], with some forecasts suggesting a tripling of consumption. Globally, the International Energy Agency (2024) warns that data centers, AI, and cryptocurrencies could consume up to 20% of the world's electricity by 2030.[1] This paper presents a framework for addressing this challenge, examining the necessary evolution in technology, security protocols, and policy required for the AI era.

Proceedings DOI: 10.21467/proceedings.8.1; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-989164-6-4

The Limits of Legacy Green IT

The foundational green IT strategies of the 2010s—virtualization, air-cooling optimization, and energy-aware scheduling—are necessary but no longer sufficient. Virtualization and workload consolidation delivered immense savings by increasing server utilization and reducing idle power; for every watt-hour saved at the server, an additional 1.9 watt-hours were saved in facility-level cooling and power distribution. Likewise, advanced air-cooling techniques, such as hot/cold aisle containment, were effective for racks consuming 5-15 kW.[10] However, they hit a physical ”thermal wall” when faced with modern AI racks that demand 60-120 kW. At such high heat densities, air is simply an inefficient medium for heat transfer. Finally, energy-aware scheduling techniques like Dynamic Voltage and Frequency Scaling (DVFS) are often incompatible with performance-sensitive AI training models that require hardware to operate at peak frequencies. The old toolkit, designed to optimize for idleness, is ill-equipped for the new reality of sustained, peak-level intensity.

Advanced Thermal Management: From Air to Liquid

The extreme power density of AI hardware necessitates a fundamental transition from air to liquid cooling. Air cooling is physically incapable of managing the thermal loads of high-density AI deployments, making liquid-based solutions the only viable path forward.

- **Traditional Air Cooling:** Relies on computer room air conditioners (CRACS) to circulate chilled air. It is cost-effective and simple but is limited to racks consuming below 20-30 kW.[10]
- **Direct-to-Chip (D2C) Liquid Cooling:** Circulates liquid through cold plates mounted directly on hot components like CPUs and GPUs. D2C is far more efficient, supporting rack densities up to 100 kW and improving hardware reliability by eliminating hot spots.[11]
- **Immersion Cooling:** Submerges entire servers in a non-conductive fluid, offering the most effective heat transfer possible. It can handle extreme densities of 100 kW and beyond, achieving near-perfect energy efficiency (PUE of ˜1.03), but at a high capital cost.[5]

Table 1: Comparative Analysis of Data Center Cooling Technologies

Cooling Technology	Typical PUE Range	Power Density Supported	Relative Capital Cost
Advanced Air Cooling	1.2-1.5	5-20 kW/rack	Low
Direct-to-Chip (D2C)	1.05-1.15	30-100 kW/rack	Medium
Full Immersion	1.03-1.10	100+ kW/rack	High

The Indispensable Layer: Security and Data Sovereignty

Optimizing for energy efficiency cannot be divorced from the complex legal and security landscape governing global data flows. The conflict between regulations like the European Union’s General Data Protection Regulation (GDPR) and the United States’ CLOUD Act creates significant challenges. GDPR strictly limits the transfer of EU citizens’ data outside the European Economic Area[12], while the U.S. CLOUD Act empowers U.S. law enforcement to compel access to data held by U.S. providers, regardless of its physical location.[13] This jurisdictional conflict necessitates the rise of "sovereign clouds"—architectures designed to keep data within a specific national or legal boundary.[14] This is achieved through technical controls like geo-fencing, region-aware scheduling, and robust encryption. Consequently, a modern workload scheduler must be a multi-objective engine that balances not only performance and energy but also data residency requirements and security policies.

Table 2: Risk Mitigation Strategies for Compliant Workload Scheduling

Scheduling Risk	Description	Mitigation Strategy
Cross-Border Data Transfer Violation	Moving regulated data to a non-compliant jurisdiction.	Region-aware scheduling, Geo-fencing, End-to-end encryption.
Unlawful Government Access	Foreign government compels data disclosure, violating local laws.	Sovereign cloud architecture, Customer-managed encryption
Co-location /Side-Channel Attack	Malicious tenant infers data from a co-resident VM.	Strict tenant isolation, Runtime monitoring, Dedicated hosts.

AI as a Solution and the Macro-Context

Paradoxically, AI itself is a key tool for mitigating its own energy impact. By analyzing vast streams of operational data, AI-driven control systems can achieve levels of efficiency unattainable by static rules. The benchmark case is Google's use of DeepMind AI to manage data center cooling, which yielded a 40% reduction in cooling energy and a 15% improvement in overall facility efficiency.[15] Beyond facility management, AI enables carbon-aware workload scheduling, shifting computation to times or locations with abundant renewable energy. This technological shift is occurring within a supportive economic and policy landscape. The data center market is experiencing unprecedented investment, with spending projected to grow significantly through 2025.[16] Simultaneously, governments are increasing regulatory oversight. The EU's updated Energy Efficiency Directive, for example, now mandates public reporting of data center energy performance.[17] In the U.S., the Inflation Reduction Act of 2022 provides powerful financial incentives, including a 30% investment tax credit for standalone energy storage systems, which encourages the development of on-site clean and resilient power.[18]

Conclusion

The AI era demands a paradigm shift from incremental efficiency gains to a fundamental re-architecting of data center infrastructure and policy. The strategies that successfully managed data center energy use in the past are no longer sufficient. The physical reality of AI workloads—their scale, density, and intensity—requires new infrastructure, new operational intelligence, and new policies. The path forward lies in an integrated approach that embraces advanced liquid cooling, embeds compliance into system architecture, leverages AI to optimize

Proceedings DOI: 10.21467/proceedings.8.1
ISBN: 978-81-989164-6-4

Series: AIJR Proceedings
ISSN: 2582-3922

operations, and is supported by a policy landscape that incentivizes investment in clean, resilient infrastructure. By treating sustainability as a first-class design criterion, the global community can ensure the digital future is as responsible as it is powerful.

Declarations

Competing Interests

The authors declare no conflict of interest regarding the publication of this paper.

Publisher's Note

The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. International Energy Agency, "Energy and AI," IEA, Paris, France, 2024. https://www.iea.org/reports/energy-and-ai
2. M. Kandemir, "Why AI uses so much energy, and what we can do about it," Penn State University, Institutes of Energy and the Environment, Apr. 2025.
3. McKinsey & Company, "How data centers and the energy sector can sate AI's hunger for power," 2024.
4. Dell'Oro Group, "Dell'Oro Group raises market forecast for data center liquid cooling and rack power distribution," Press Release, Feb. 5, 2025.
5. W. Wang, et al., "Immersion cooling technology development status of data center," Sci. Technol. Energy Transit., 2024.
6. A. Shehabi, S. Smith, D. Sartor, R. Brown, and M. Herrlin, "United States data center energy usage report," Lawrence Berkeley National Laboratory, Berkeley, CA, 2016.
7. A. Shehabi, et al., "2024 United States data center energy usage report," Lawrence Berkeley National Laboratory, Berkeley, CA, Rep. LBNL-2001637, Dec. 2024.
8. U.S. Department of Energy, Grid Deployment Office, "Clean energy resources to meet data center electricity demand."
9. U.S. Environmental Protection Agency, "Virtualize Servers," ENERGY STAR.
10. Uptime Institute, "Global data center survey 2024," 2024.
11. CoolIT Systems, "Energy efficiency in the data center with direct liquid cooling," White Paper, 2024.
12. J. D. Michels, C. Millard, I. Walden, and U. Wuermeling, "Cloud sovereignty and the GDPR, part one: US government access to European data," SSRN Electron. J., 2024.
13. G. La Scala, "United States and European Union clash over data protection," Fordham Int. Law J. Online, 2019.
14. A. A. Solanke, "Sovereign cloud implementation: Technical architectures for data residency and regulatory compliance," Int. J. Sci. Res. Arch., vol. 11, no. 2, pp. 2136-2147, 2024.
15. E. Gamble and J. Gao, "Safety-first AI for autonomous data centre cooling and industrial control," Google DeepMind, 2018.
16. Gartner, Inc., "GenAI boom: Datacenter spending forecast raised again," The Next Platform, Jan. 23, 2025.
17. European Union, "Directive (EU) 2023/1791 of the European Parliament and of the Council of 13 September 2023 on energy efficiency," Official Journal of the European Union, L 249, Oct. 2023.
18. U.S. Congress, Inflation Reduction Act of 2022, H.R.5376, 117th Congress, 2022.
19. Vertiv, "Quantifying data center PUE when introducing liquid cooling," White Paper.

Proceedings DOI: 10.21467/proceedings.8.1
ISBN: 978-81-989164-6-4

Series: AIJR Proceedings
ISSN: 2582-3922