**FEASIBILITY OF AI-OPTIMIZED ENERGY USAGE IN CLOUD DATA CENTERS**

**ENVIRONMENTAL STRATEGIES OF THE CLOUD COMPUTING INDUSTRY**

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Gabe Silverstein
November 8, 2024

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Caitlin D. Wylie, Department of Engineering and Society

Rosanne Vrugtman, Department of Computer Science

## Introduction

The cloud computing industry is facing a pivotal moment in its history as the rate of technological innovation draws attention to critical issues surrounding sustainable resource management. One significant problem is that the "Cloud now has a greater carbon footprint than the airline industry." Its energy usage is also so large that "data centers collectively devour more energy than some nation-states," with a single data center using the "equivalent electricity of [up to] fifty thousand homes" (Monserrate, 2022, Cloud the Carbonivore section). These concerning statistics, among others, have led to resource conservation and effectiveness being the main research priority in the field (Panwar et al., 2022). Furthermore, if action isn't taken soon, the effects of poor resource management will worsen as emerging technologies like artificial intelligence (AI) strain cloud infrastructure even further (Yan et al., 2024).

To investigate the pressing issue of sustainable cloud resource management, this paper proposes that research be conducted in two parts. The technical portion of the research would focus on the significant energy demands of cloud infrastructure, particularly in data centers, and highlight the need for scalable, energy-efficient cloud solutions. As using AI to optimize energy usage is a popular research topic in this field, the technical paper would further assess the practicality of that approach. In comparison, the STS research would explore corporate social responsibility and consumer expectations by analyzing the sustainability decisions of major cloud providers. It would also address commonalities or gaps in policy and outline the sociotechnical factors that drove the cloud industry to reach its current state. By distilling the overarching sociotechnical problem into two prominent

areas of interest, the proposed research papers hope to provide a holistic perspective on the challenges of sustainable resource management.

**Technical Topic**

Data centers account for almost 2% of all electricity usage in the U.S., with 40% of that amount being used just on the cooling of cloud computing systems (Wang et al., 2023). The scale and inefficiency of this energy usage is a serious technical problem as the companies that provide power to cloud data centers are struggling to meet demand (Li & Zhang, 2024). To combat this energy challenge, engineers within the field have primarily focused on improvements to cooling efficiency. In particular, there is extensive literature on using AI and machine learning (ML) to optimize data center cooling. Researchers like Geng et al. have used empirical evidence to demonstrate that AI/ML algorithms can have marked impacts on data center efficiency, even achieving 10% cooling energy savings when compared to standard systems (2016). While I agree with Geng et al. that research into AI-cooling solutions is warranted due to its potential to increase energy efficiency, current literature on this topic often fails to recognize that the AI models themselves require energy to run. This potential oversight may seem to produce a contradiction in which energy-intensive AI is used to reduce an energy efficiency problem. However, if an AI model can be found that results in a net reduction in overall energy usage, the costs of running the AI would be worthwhile. Therefore, it is important to research the feasibility of AI-optimized energy usage in cloud data centers.

In order to analyze whether AI models can act as a viable energy-saving measure in data centers, the technical paper will incorporate the results of evidence obtained through multiple research methods. Those methods include document analysis, data mining, model evaluation, and literature review. Document analysis and data mining will be tailored towards reports with energy-usage data from companies and researchers experimenting with AI-based cooling. Comparing electricity usage before and after implementing an AI-cooling model will allow for patterns within the data to emerge that might point toward the feasibility of incorporating AI into data center systems. Literature review and model evaluation will be used on academic articles authored by AI and cloud experts. Current literature on AI-enabled cooling often evidences case studies, simulation data, mathematical proofs, and statistics that outline reasons why an AI-optimized cooling system could help tackle the energy demands of data centers. Analyzing the tradeoffs of various optimization algorithms through those sources will provide further context on why and how researchers continue to investigate AI solutions despite their inherent energy usage. For instance, Wang et al. informed me of the different applications for model-free and model-based algorithms, which will help me in my future research to determine if some AI approaches are more feasible than others in achieving a reduction in energy utilization (2023). The technical research proposed above seeks to use the discussion on AI-optimized energy usage to underscore the importance of energy-efficient practices in cloud resource management. The paper will also aim to further conversation on the role of AI in technical systems.

**STS Topic**

It is widely known that cloud computing results in the consumption of vast amounts of electricity and the emission of large quantities of greenhouse gases. Despite these environmental impacts, the demand for cloud resources continues to grow as emerging technologies like Big Data and artificial intelligence necessitate an increase in data centers (Yan et al., 2024). This balancing act between demand for compute resources and its environmental ramifications poses a clear sociotechnical issue that needs to be addressed. Yan et al. highlight the urgency of efficient resource management in this setting, and I agree that environmental sustainability will be in jeopardy if current trends continue (2024). A peer-reviewed MIT article by Monserrate further underscores the critical nature of this problem as the cloud industry is mostly self-regulated (2022). I learned that there is no overarching federal agency to impose data center guidelines, which means that the proliferation of data centers could go unchecked and result in significant damage to the environment. To investigate this pressing issue, I propose research regarding how and why major cloud providers adopt their environmental strategies. By understanding the factors that play into the sustainability policies of cloud providers, it becomes easier to pinpoint where changes need to be made in order to strike a balance between technological innovation and preserving the environment.

In order to compare the key considerations that go into developing cloud sustainability strategies, the STS paper will utilize evidence obtained through a variety of research methods. The main research methods that will be employed in the STS paper are document, policy, and ethical analysis. Document analysis will be used on the

environmental reports of leading cloud providers to compare emissions, water stewardship, and electricity usage statistics. These primary source documents can also be used to find commonalities or gaps across the different policies. For instance, it would be interesting to explore the sociotechnical factors behind why Amazon Web Services (AWS) has the least ambitious carbon emissions policy of the top 3 cloud providers and how that affects its marketing. Specifically, AWS hopes to achieve net-zero carbon emissions by 2040 (Venkatesan & Karibandi, 2024) compared to Google Cloud Platform's goal to reach net-zero carbon emissions by 2030 (Hölzle, 2022). Microsoft Azure's policy goes a step further and outlines how its strategy is the most aggressive, trying to be carbon-negative by 2030 (Nakagawa & Smith, 2023). Cloud scholars and climate activists know that operational costs and public image are often some of the driving sociotechnical factors in these policy decisions (Shan et al., 2024), but their research can be expanded upon using the Social Construction of Technology (SCOT) theory (Bijker et al., 2012). Investigating corporate social responsibility and consumer expectations in the cloud computing industry under this framework will highlight how cloud computing reached its current state. In addition to document analysis, policy analysis can be used on state/federal codes and the few regulations that exist for data centers. Marwah et al. taught me that it is very difficult for cloud service providers to balance environmental sustainability with consumer demand for quality of service and availability (2010), so it is essential to explore how the lack of federal regulation and any new proposed bills affect the strategies that cloud industry leaders adopt. Finally, ethical analysis can be used on statements and reports from cloud industry executives to determine how much of a role ethics plays in these strategies. I learned from Lucivero that the policies of cloud providers tend to focus

on physical sustainability, rather than the socioeconomic environment that the data centers affect (2019), so it would be interesting to explore deontological or utilitarian viewpoints (Johnson, 2020) to analyze what cloud providers should cover in their environmental strategies.

**Conclusion**

Sustainable resource management is an essential topic to explore in today's digital age. Researching this sociotechnical issue in the cloud computing industry can drive further discussion and highlight potential innovative solutions for an industry that is struggling to keep up with the rate of technological innovation. In particular, the technical and STS research posed in this paper aims to grow awareness of the challenges of cloud resource management and propose new paths forward that can help reduce operational and environmental costs. The anticipated deliverable from the technical research is an artifact outlining the feasibility of incorporating AI algorithms into cloud computing systems to conserve energy. Potential technical findings from this research include a ranking of AI resource optimization algorithms (if any are found to be practical) as well as the identification of the largest culprits of inefficient energy usage in cloud data centers. These technical findings can help cloud providers make informed decisions on how they should handle resource management. The expected result of the STS research is a detailed analysis of the environmental strategies of major cloud service providers, which will improve readers' understanding of the ethical and policy decisions surrounding cloud infrastructure. Possible conclusions from this research are distinguishing which cloud providers have the most ambitious environmental goals and pinpointing gaps in certain

policies. By combining the outputs of the technical and STS research, readers will have the knowledge and tools to start new dialogues and research to improve sustainable cloud resource management.

## References


Bijker, W. E., Hughes, T. P., & Pinch, T. J. (2012). *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. The MIT Press.

Cao, Z., Zhou, X., Wu, X., Zhu, Z., Liu, T., Neng, J., & Wen, Y. (2023). Data Center Sustainability: Revisits and Outlooks. *IEEE Transactions on Sustainable Computing*, *9*(3), 236–248. https://doi.org/10.1109/tsusc.2023.3281583

Chen, J., Du, T., & Xiao, G. (2021). A multi-objective optimization for resource allocation of emergent demands in cloud computing. *Journal of Cloud Computing*, *10*(1). https://doi.org/10.1186/s13677-021-00237-7

Dhar, P. (2020). The Carbon Impact of Artificial Intelligence. *Nature Machine Intelligence*, *2*(8), 423–425. https://doi.org/10.1038/s42256-020-0219-9

Geng, H., Sun, Y., Li, Y., Leng, J., Zhu, X., Zhan, X., Li, Y., Zhao, F., & Liu, Y. (2024). TESLA: Thermally Safe, Load-Aware, and Energy-Efficient Cooling Control System for Data Centers. *In Proceedings of the 53rd International Conference on Parallel Processing*, 939–949. https://doi.org/10.1145/3673038.3673144

Johnson, D. G. (2020). *Engineering Ethics: Contemporary and Enduring Debates*. Yale University Press.

Hölzle, U. (2022, November 21). Our commitment to climate-conscious data center cooling. *Google.* https://blog.google/outreach-initiatives/sustainability/our-commitment-to-climate-conscious-data-center-cooling/

Li, X., & Zhang, S. (2024). Management Mode and Path of Digital Transformation of Power Grid Enterprises Based on Artificial Intelligence Algorithm. *International Journal of Thermofluids*, *21*. https://doi.org/10.1016/j.ijft.2023.100552

Lucivero, F. (2019). Big Data, Big Waste? A Reflection on the Environmental Sustainability of Big Data Initiatives. *Science and Engineering Ethics*, *26*. https://doi.org/10.1007/s11948-019-00171-7

Marwah, M., Maciel, P., Shah, A., Sharma, R., Christian, T., Almeida, V., Araújo, C., Souza, E., Callou, G., Silva, B., Galdino, S., & Pires, J. (2010). Quantifying the sustainability impact of data center availability. *ACM SIGMETRICS Performance Evaluation Review*, *37*(4), 64–68. https://doi.org/10.1145/1773394.1773405

Monserrate, S. G. (2022). The Cloud Is Material: On the Environmental Impacts of Computation and Data Storage. *MIT Case Studies in Social and Ethical Responsibilities of Computing, Winter 2022*. https://doi.org/10.21428/2c646de5.031d4553

Nakagawa, M., & Smith, B. (2023, May 10). On the road to 2030: Our 2022 Environmental Sustainability Report. *Microsoft*.

https://blogs.microsoft.com/on-the-issues/2023/05/10/2022-environmental-sustainability-report/

Panwar, S. S., Rauthan, M. M. S., & Barthwal, V. (2022). A systematic review on effective energy utilization management strategies in cloud data centers. *Journal of Cloud Computing*, *11*(1). https://doi.org/10.1186/s13677-022-00368-5

Shan, L., Sun, L., & Amin Rezaeipanah. (2024). Towards a novel service broker policy for choosing the appropriate data center in cloud environments. *Computer Communications*, *228*. https://doi.org/10.1016/j.comcom.2024.107939

Venkatesan, V., & Karibandi, A. (2024, September 16). Embracing Modernization with a Sustainability Focus. *Amazon Web Services*. https://aws.amazon.com/blogs/migration-and-modernization/embracing-modernization-with-a-sustainability-focus/

Wang, R., Cao, Z., Zhou, X., Wen, Y., & Tan, R. (2023). Green Data Center Cooling Control via Physics-Guided Safe Reinforcement Learning. *ACM Transactions on Cyber-Physical Systems*, *8*(2). https://doi.org/10.1145/3582577

Yan, D., Chow, M.-Y., & Chen, Y. (2024). Low-Carbon Operation of Data Centers With Joint Workload Sharing and Carbon Allowance Trading. *IEEE Transactions on Cloud Computing*, *12*(2), 750–761. https://doi.org/10.1109/tcc.2024.3396476