AI-Optimized Cooling in Cloud Data Centers: A Feasibility Analysis

CS4991 Capstone Report, 2025

Gabe Silverstein
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
jgb9fz@virginia.edu

ABSTRACT

Data centers struggle with energy efficiency, particularly in their cooling systems, resulting in significant electricity consumption. To combat this energy challenge, researchers have proposed artificial intelligence (AI) and machine learning (ML) solutions to optimize data center cooling. The feasibility of this approach was examined by contrasting the potential increases in energy efficiency with the costs inherent to running the models. Specific methods utilized in the research include literature reviews, case studies, simulations, and statistical analyses. Through this meta-study, I determined that AI-optimized cooling is worth pursuing as the net energy savings produced generally outweigh the cost of training and running the models. Future research is required to determine which AI/ML solutions might produce the best energy efficiency, as well as investigate the broader socioeconomic and environmental impacts of these optimizations.

1. INTRODUCTION

The cloud is inherently energy-intensive. To meet consumer quality of service and availability demands, data centers are designed to be fault tolerant and rarely have outages. This emphasis on 24/7 availability introduces redundant infrastructure that can provide backups in case of system failures. Those extra computers are packed into already crowded server rooms, leading to considerable energy expenditure on cooling systems that prevent overheating. Coupled with other pronounced resource demands, data centers account for almost 2% of all electricity usage in the U.S., with 40% of that amount being used just for cooling (Wang et al., 2023). The scale and inefficiency of this energy usage pose a serious technical problem as the companies that provide power to these data centers are struggling to meet demand (Li & Zhang, 2024).

Given the current technology landscape, AI and ML solutions have been increasing in popularity when considering how to reduce the energy spent on cooling. Although there is extensive literature on this topic, researchers often fail to recognize that their proposed AI/ML models also require significant resources to train and run. This potential oversight may seem to produce a contradiction in which energy-intensive AI is used to remedy an energy efficiency problem. However, if an AI/ML solution can be found that results in a net reduction in overall energy usage, the costs of running the model would be worthwhile. To this end, I analyze the current literature and leverage statistical data to determine if that net reduction is feasible and whether further research into AI-optimized energy solutions is warranted.

2. RELATED WORKS

The main inspiration for my feasibility analysis comes from an article written by Geng et al. (2016) that describes the increasing

energy demands of data centers and how a particular ML approach is being used to improve cooling efficiency. The researchers were able to achieve a 10% energy savings for a real data center with their model but failed to recognize the costs of training and running the optimizer.

Tozzi (2024) provided the perspective that was missing in the above article, highlighting the costs associated with running AI in data centers. The discussed increases in energy consumption and heat production are referenced when evaluating the practicality of AI cooling solutions.

The aforementioned downsides of data center AI use are challenged by Lazic et al. (2018) who assert that recent developments in reinforcement learning (RL) have led to models that can improve cooling efficiency without requiring significant training data. Their paper focused on enhancements to a specific cooling setup, but the reported experiential data informed my analysis.

3. PROCESS DESIGN

To provide an unbiased assessment of the technical feasibility of AI-cooling optimizations, I developed an approach that considers multiple research methods and data from a variety of sources. For the purposes of my analysis, "technical feasibility" is defined as having a reasonable strategy to implement the technology such that the produced benefits outweigh the implementation costs.

3.1 Formulation

Before I even started collecting data for my analysis, I needed to construct a metric that could be used to evaluate the proposed AI/ML solutions. Although an array of factors play into what might be considered a practical technology, I opted to focus on net energy savings as that was the concept that inspired my research. With my framework determined, I set out to collect as much data as possible. I incorporated literature reviews, simulations, case studies, and statistical analyses to make the research process more thorough and address any problems with biased data.

3.2 Research

My meta-study started with a comprehensive literature review that analyzed existing research on AI-driven cooling optimizations in data centers. This process included examining prior studies on ML applications for energy efficiency, cooling system optimizations, and the energy costs associated with training and running AI models. Papers from peer-reviewed journals, conference proceedings, and industry reports were evaluated to gain insights into both the benefits and potential drawbacks of those cooling approaches.

The literature review was supported by simulation data that measured the potential energy savings that AI-optimized cooling systems claimed they could achieve. Researchers like Wang et al. (2023) simulated hypothetical data center configurations where AI models dynamically adjusted cooling parameters and compared those results against traditional, non-AI cooling strategies to measure efficiency gains.

To combat the above approaches' theoretical nature, I examined case studies of real-world applications of AI in data center cooling. While some studies noted increases in energy consumption, others showcased the direct opposite. One of the more surprising ones includes a deployment to a live Google data center where they achieved a 40% reduction in energy used for cooling (Evans and Gao, 2016).

Statistical analysis was used in addition to those case studies to provide further quantitative data for the feasibility analysis. Information from prior research and simulations was aggregated to calculate net energy savings while considering the energy

expenditure required to train and deploy various AI models. Significance tests were performed on these aggregations to determine whether the observed energy reductions were substantial enough to justify the adoption of an AI-driven cooling system.

3.3 Challenges

The main obstacle I encountered during my research was finding exact figures on energy expenditure when implementing AI/ML cooling technologies. Current literature often focuses on just the energy savings rather than the net change in energy consumption, so I had to perform extensive research and sometimes use nontraditional sources to find that data.

Another major challenge was the variability in AI model performance across different data center configurations. AI-driven cooling optimizations depend on numerous factors including data center size, type of cooling infrastructure, and efficiency of existing HVAC systems. As a result, a model that performs well in one environment may not necessarily perform as well in another. This variation made it difficult to generalize findings and required careful consideration when interpreting energy savings data.

Finally, I experienced challenges with broadly labeling certain ML approaches as feasible or infeasible. I decided to focus on net reductions in cooling energy expenditure, but AI can optimize other things in data centers like virtual machine (VM) allocation, water usage, and server load. Just because a certain model does not perform well at one type of task does not mean that AI-optimized energy usage as a whole is impractical. Furthermore, it is hard to know for certain if an approach is viable as my analysis focuses on a specific technical aspect of feasibility rather than the broader socio-economic and environmental factors that also impact energy usage in cloud data centers.

4. RESULTS

My analysis of AI-cooling optimizations revealed that, in most cases, the net energy savings outweigh the costs of running the models. Case studies, including Google's deployment of reinforcement learning for cooling (Evans and Gao, 2016), demonstrated reductions of up to 40% in cooling energy usage, while other research showed improvements in the 10–20% range (Wang et al., 2023). However, these gains varied depending on data center configurations, cooling infrastructure, and model efficiency. Statistical analyses confirmed that the observed energy reductions were significant enough to justify AI adoption, even when accounting for the energy required to train and operate the models (Tozzi, 2024). Advancements in reinforcement learning have lessened the effects of model training (Lazic et al., 2018), further confirming this result.

Despite my overall conclusion, this meta-study identified inconsistencies in reported energy savings due to differences in methodology and real-world implementation constraints. Some data centers saw diminishing returns when AI was applied to their cooling systems, while others experienced efficiency fluctuations based on workload variations. Additionally, the lack of standardized metrics for assessing net energy impact made direct comparisons between studies difficult. These findings suggest that while AI-enhanced cooling is generally beneficial, its success depends on careful implementation, continuous model refinement, and alignment with specific data center configurations to maximize energy efficiency.

5. CONCLUSION

The results of this meta-study contribute to the field of sustainable computing and provide insights into possible avenues to improve efficient resource management in cloud

infrastructure. My feasibility analysis highlights that the net energy savings from AI-optimized cooling generally justify its use. While the computational costs of the AI models are a valid concern, those cooling solutions have the potential to mitigate some of the harms of poor energy efficiency in data centers. However, there is no one-size-fits-all solution for each data center, and implementations need to be tailored to meet existing infrastructure and operational demands.

6. FUTURE WORK

To address the challenges that arose during my research, further studies are needed to explore metrics and AI models that can be generalized to multiple data center configurations. Additionally, more in-depth research is needed to evaluate the feasibility of AI not just in cooling, but in other areas of cloud resource management. Finally, while AI-optimized cooling may be technically feasible, future studies should investigate the broader socioeconomic and environmental impacts of its use as that may affect the technology's overall viability.

REFERENCES

Evans, R., & Gao, J. (2016, July 20). *DeepMind AI reduces google data centre cooling bill by 40%*. Google DeepMind. https://deepmind.google/discover/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-by-40/

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

Lazic, N., Boutilier, C., Lu, T., Wong, E., Roy, B., Ryu, M. K., & Imwalle, G. (2018). Data center cooling using model-predictive control. *Advances in Neural Information Processing Systems*, *31*. https://proceedings.neurips.cc/paper_files/paper/2018/file/059fdcd96baeb75112f09fa1dcc740cc-Paper.pdf

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

Tozzi, C. (2024, November 4). *Assessing AI's impact on data center heating and cooling needs*. Data Center Knowledge. https://www.datacenterknowledge.com/ai-data-centers/assessing-ai-s-impact-on-data-center-heating-and-cooling-needs

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