IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Enhancing Energy Efficiency in AI-Powered Data Centers: Challenges and Solutions

Sahana Shankar

Vista Del Lago High School, 1970 Broadstone Pkwy., Folsom, CA 95630, USA; sahe.shankar@gmail.com

ABSTRACT: The rapid proliferation of AI-powered data centers has led to significant increases in electricity consumption, with projections suggesting that these facilities could account for up to 9% of the United States' total electricity generation by 2030. This paper explores the factors contributing to the high energy demands of AI data centers, such as model training and online/offline inference. Estimates for the power and energy consumption for AI infrastructure are presented here, along with the analysis of existing energy efficiency techniques deployed in the industry. Additionally, the paper discusses the role of regulatory measures, corporate governance, and public awareness in supporting energy-efficient practices. The paper also highlights the potential of revolutionary technologies like quantum computing. By addressing these challenges, the technology industry can mitigate the environmental and economic impacts of growing energy demands, ensuring sustainable growth and development in the AI domain.

KEYWORDS: Robotics and Intelligent Machines; Machine Learning; AI Infrastructure; Image Classification; Edge Computing.

Introduction

AI-powered data centers are critical infrastructures that leverage artificial intelligence to optimize data processing and storage. Typically, they are utilized in applications such as cloud computing, IoT, big data analytics, and AI/Machine Learning. As the technology industry rapidly expands, demand for these applications has surged, resulting in immediate growth and sophistication in AI data centers. However, this expansion came with a substantial rise in energy consumption. Considering the sudden boom with AI-powered search engines, many of these applications were estimated to require ten times the electricity of a standard Google query: "0.3 watt-hours were necessary for a traditional Google search to 2.9 watt-hours for a ChatGPT query"[1] This sizable growth in energy demand poses serious issues along with economic and environmental challenges.

This research addresses the question: What can be done to improve the energy efficiency of AI-powered data centers? While several studies have explored various aspects of data center efficiency, few have explicitly focused on the unique challenges AI workloads pose. This review aims to fill that gap by examining the specific factors contributing to the high energy demands of AI data centers and evaluating the effectiveness of existing and emerging energy efficiency techniques. This review addresses a critical gap in the literature by providing a comprehensive analysis of these strategies. It offers actionable insights for researchers, policymakers, and industry stakeholders seeking to reduce the environmental impact of AI technologies.

Examination of factors contributing to high energy in AI data centers:

AI-powered data centers consume significant amounts of energy due to several key factors. Understanding these factors is crucial for developing strategies to enhance energy efficiency. There are two main phases or stages in the AI data pipeline that are energy intensive - 1. Model development or AI training, where we set up and teach the model how to behave, and 2. AI Inference, where the trained model is deployed in production and can take prompts or inputs to produce original responses. The primary needs of AI are vastly different from existing data centers due to their unique workload, which requires more complex hardware.

AI Training:

AI training is done on cured or prepared data. There is a pre-training step for processing raw data, cleaning it, and transforming it into input for model training.[2] As AI models get complex, the pre-training step will require sophisticated algorithms and lead to more time.[3] This would add to higher energy consumption. With the prepared data, the AI models are typically trained for weeks or months, depending on the amount of data or complexity of the model. In the recently published Llama[3] research paper by Meta, it is estimated that "pre-training took nearly 57 days using 16K H100 NVIDIA GPUs and training for around 80+ days."[4] The training process is not latency sensitive, so AI training clusters need not be near any populous city; they can be anywhere in the world. However, AI training workloads are extremely power-hungry, and AI hardware operating at power levels is always at its maximum thermal design power (TDP). Using a simple calculator, we could estimate the power and energy consumption to train complex AI Large Language Models (LLMs) such as Meta's Llama and Open AI's Chat GPT to be around ~24 MW and ~560 MWh.

Meta shares the detailed hardware infrastructure for AI training.[5,6] The AI infrastructure includes combining multiple

DOI: 10.36838/v6i11.15

GPUs (8 or 16) into Compute Nodes; multiple nodes are then connected by network fabric with access to the Storage array. NVIDIA compute nodes and network fabrics are widely adopted, and their Thermal Design Power (TDP), their max-rated power, can be accessed via a datasheet.[7-9] The network and storage fabric power per server node is assumed to be ~10% of the total power.[10] Using the above information, the total power and energy per hour for developing AI models can be estimated, as shown in Table 1. As noted, the power and energy for Llama 3 (405B parameter) is higher than Llama 2 (70B parameter).

Table 1: Power and Energy estimate for AI training models, demonstrating that larger models require significantly more energy to train due to their increased complexity.

AI Training model	Compute Node	TDP per Compute node (KW)	Network fabric Power per node (KW)	Storage power per node (KW)	# nodes	Total Power (MW)	Energy per day (MWh)
Llama2 (70B)	NVIDIA DGX A100	6.5	0.73	0.18	2000	14.82	355.7
Llama3 (405B)	Equivalent of DGX H200	10.2	0.73	0.18	2000	22.22	533.3
chat GPT 4	NVIDIA DGX A100	6.5	0.73	0.18	3125	23.16	555.8

Total power = TDP of Compute Server nodes + Network servers + Storage server nodes.
Energy = Total Power x 24 hr.

The hardware infrastructure capability has been expanding for the last few years, propelled by massive investments by big tech companies. This trend is expected to continue, supported by the recent announcements of Meta4 and X. Table 2 shows the AI infrastructure expansion shared by Meta and X.

Table 2: Power and energy trends for AI infrastructure expansion at big tech firms highlight the increase in power and energy demand as these companies build large AI clusters.

AI infrastructure	Compute Node	# nodes	Total Power (MW)	Energy per day (MWh)
Meta -AI RSC phase 1 (2020)	NVIDIA DGX A100	760	5.63	135.2
Meta - AI RSC phase 2 (2022)	NVIDIA DGX A100	2000	14.82	355.7
Meta - AI RSC beyond phase 2 (2024)	Equivalent of DGX H200	2000	22.22	533.3
Meta infrastructure build-out plan for future	Equivalent of DGX H200	43750	486.06	11665.5
X (2024)	NVIDIA DGX H200	12500	138.88	3333.0

AI Inference:

Inference is the application of the trained model to new data. This process is usually less resource-intensive than training but is performed much more frequently, especially in real-time applications. This process, however, is very latency-sensitive. For this reason, the AI inference clusters are distributed (not centrally located) and must be capable of handling high volumes of prompts/requests. Typically, CSPs (Cloud Service Providers) such as AWS, Google Cloud, and Microsoft Azure host the Inference model and make them available via APIs for enterprise and commercial usage. Some enterprises may also decide to host the inference on their dedicated data centers instead of the cloud for faster latency. The hardware components required for AI inference are mostly the same as those required for AI training (Compute node, Network, Storage). Still, the scale and availability (in multiple locations) could drastically differ.

Though AI training has been the main driver for AI servers so far and is expected to continue growing, as noted in Table 2, the growth of AI inference servers is now poised to take off as AI models move into production. Some predict that inference servers will soon outpace training servers.[11] The daily energy consumption for AI training and inference in 2023, with 1.5 million AI servers sold, is estimated at 0.45 TWh, nearly equivalent to Denmark's daily energy consumption of 0.478 TWh.[12] However, compared to US energy consumption in 2023, AI servers account for less than 1%. Based on the estimated demand growth of AI servers, which is expected to double in volume every four years, with each operating at an average of ~10 to 12 KW TDP, the total power consumption by the 4 million AI servers (AI training + AI inference) alone in 2030 could range in 45GW and exceed 1 TWh of energy consumption daily. This will still account for ~<2% of US energy consumption. The AI servers accounting for an estimated 9% of US energy consumption in 2030[13] means not just the volume of AI servers but, more importantly, the computing power or energy is expected to increase by ~4 to 5x by the end of this decade. This presents a significant opportunity for research and development to optimize AI server energy efficiency.

Given the explosion of AI models and the industry's thirst for adopting Generative AI use cases, AI training and Inference processes are expected to grow exponentially in the future. This includes but is not limited to ultra-large-size models trained across modalities and agentic capabilities. Given the above, optimizing power will continue to be a key area of research in AI training.

Existing energy efficiency measures - benefits and gaps:

Table 3 comprehensively analyzes current techniques and methods to improve energy efficiency across AI-powered data centers. The sources used to compile these energy efficiency methods include, but are not limited to, white papers, technical articles, real-world case studies, and performance data from industry and academic sources. These approaches were selected based on their proven real-world application and measurable impact on power consumption. By combining these diverse and reliable sources, Table 3 reflects the state-of-the-art methods used in both industry and academic research to enhance the energy efficiency of AI data centers.

Table 3: Analysis of current methods and techniques to improve energy efficiency across hardware, software, and data center infrastructure levels, revealing the benefits and gaps associated with each approach.

Category	Techniques	Benefits	Gaps or challenges	Real-world example
Data center facility	Liquid cooling - Direct to-chip Immersion [14]	Reduces energy requirements, increases cooling efficiency, supports higher-density servers	High initial setup cost, potential for hardware compatibility issues, complexity in maintenance	Google's Power Usage Effectiveness (PUE) reduction- a Google-owned and -operated data center is 1.8 times as energy efficient as a typical enterprise data center. This is achieved by keeping the temperature to 80°F, using outside air for cooling, and building custom servers. Google also has shared detailed performance data to help move the entire industry forward. [15]
	Scheduling AI training in colder climates/downtime	Utilizes lower ambient temperatures to reduce cooling costs and increase efficiency during off-peak hours.	Limited flexibility in scheduling, possible delays in processing, and dependency on specific weather patterns	
	Geographical location, green energy generation	Reduces carbon footprint, utilizes renewable energy sources (e.g., solar, wind)	Availability and consistency of renewable energy sources, higher setup costs, and potential geographic limitations	

Hardware	Power capping devices	Limits power consumption to prevent overloading, optimizes power usage and manages peak loads[16]	It may impact performance during peak loads, complexity in implementation, and potential compatibility issues with existing infrastructure	IBM Power systems energy scale, HPE Power capping, Dell Edge Power Capping.[17]
	Using application-specific processors for AI tasks instead of General-purpose processor	Increases computational efficiency, reduces energy consumption, and optimizes performance for specific AI workloads	Higher initial cost, limited flexibility for general computing tasks, and potential underutilization of specialized hardware	The Groq Language Processing Unit, the LPU, is a new type of hardware that delivers speed and energy efficiency on a scale. Fundamentally different from the GPU – originally designed for graphics processing – the LPU was designed for AI inference and language.[18]
	Storage: Using AFA Storage systems made entirely of SSDs, eSSDs with low-power states	Improved performance/Watt, less heat than HDDs	higher cost per GB	Dell PowerStorage, NetApp AFF, Pure Storage FlashArray/C, HPE Nimble Storage All Flash Arrays.
	Network: Using switches, routers, and NIC with EEE[19]	Reduces power consumption	High initial setup cost, compatibility issues with existing configuration	Cisco Catalyst 9300 Series Switches, D-Link DSR-250 routers
	using the lowest process node	Reduces power consumption, increases performance	Higher manufacturing costs, increased risk of component failure, and limited	The **Apple A17 Pro** chip, used in the iPhone 15 Pro models, is one of the first consumer products built on TSMC's advanced **3nm** process node, providing improved performance and energy efficiency
		efficiency, and lowers thermal output	availability of cutting-edge process technologies	
AI Training & Inference	Batch processing	Reduces energy consumption per task, increases throughput, and optimizes resource utilization	Potential increase in latency, complexity in data handling, and risk of data congestion or bottlenecks	Solutions such as AWS batch, Google Kubernetes Engine (GKE)
	Edge computing[20]	Reduces latency, saves bandwidth, and decreases energy consumption by processing data closer to the source	Limited computational power at the edge, security and privacy concerns, and complexity in data synchronization	NVIDIA Jetson Xavier NX is a compact, energy-efficient AI-edge computing platform. Lenovo ThinkEdge SE450 is an edge server optimized for AI and IoT workloads
	Effective Prompting	Reduces computation by guiding AI models towards efficient processing paths and improves response time.	It requires sophisticated AI model training, has the potential for reduced accuracy if not properly implemented and depends on specific AI models.	Few shot prompting, Chain-of-thought (CoT) prompting, Frameworks such as LangChain
AI Algorithm efficiency	Model compression - Pruning, Quantization, Distillation	Reduces model size and complexity, decreases energy consumption, and maintains or improves processing speed	Possible loss of model accuracy, complexity in implementing compression techniques, and potential incompatibility with existing infrastructure	Snorkel AI/Google's step-by-step distillation,[21] PyTorch quantization libraries
	Fine-tuning	It improves model accuracy with minimal data and reduces energy consumption by focusing on specific tasks	It may require significant computational resources, potential overfitting, and challenges in managing diverse data sources	State-of-the-art fine-tuning techniques such as Supervised fine Tuning, PETS (Parameter Efficient Tuning), and LoRA (Low-rank adaptation) using libraries such as Hugging face, Lamini AI
	Efficient algorithms - Lightweight neural networks, Depth separable convolutions[22,23]	Reduces computational complexity, lowers energy consumption, and maintains performance for specific tasks	The trade-off between model simplicity and accuracy, limited applicability for complex tasks, and challenges in scalability	On-device ML models that are smaller in size such as Google (Gemini Nano, Gemma 2 models), Meta (Llama3), Microsoft (Phi-3), etc.
Software	Load balancing and dynamic scaling - distributing workloads evenly across available resources[24,25]	Optimizes resource utilization, reduces energy waste, and improves system reliability	Complexity in implementation, potential performance overhead, and challenges in real-time monitoring and adjustment	Solutions such as AWS ELB (Elastic load balancing) and Google Cloud load balancing.

Methods

Out of the above existing solutions, a research experiment was conducted on power efficiency using smaller models - "ODML" (on-device machine learning). The objective of the experiments is to measure energy efficiency by adopting simplified AI algorithms and techniques, such as batch processing for image classification inference tasks.

This study used two pre-trained AI models - MobilenetV3 and ResNet 50. Experiments were conducted on an Orange Pi 5 Pro device to measure the Inference time (latency) and power consumption; Energy is calculated based on the time and power. Fifty-nine personal images were used for data validation, and a YOJACK USB power meter was used to measure power, as shown in Figure 1. Though the number of images used as a dataset could be small, the latency per image is not expected to vary significantly. Hence, the experiment results are valid in conclusion. The experiment setup for the two models is shown in Table 4.

Table 4: Experiment methods for AI inference evaluation.

Model	MobilenetV3	ResNet50
Framework	Tensor flow lite	Pytorch
Model size	20 MiB	103 MiB
Operating system	Debian Linux	Debian Linux
Hardware	Orange Pi – 8-core ARM CPUs	Orange Pi – 8-core ARM CPUs
Application	Computer Vision (Image classification)	Computer Vision (Image classification)

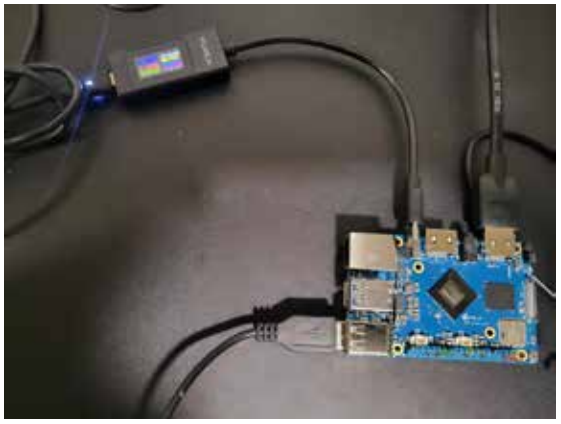

Figure 1: Experiment setup - Orange Pi 5 Pro with Yojack USB power meter.

Table 5: Sample data set and image classifications – comparison of output for MobileNet and ResNet showing slight variations in classification accuracy.

Image	MobileNet detected image	ResNet detected image
	Bell cote	Tile roof
	Vase	Vase
	Ram	Tup
	Volcano	Alp
	Promontory	Seashore, Coast

DOI: 10.36838/v6i11.15

Results and Discussion

Figure 2 below illustrates the inference time, peak power, and energy consumption for two AI models, MobileNet and ResNet, with varying batch sizes running on an Orange Pi 5 Pro device. The x-axis lists the AI models and ResNet batch sizes, while the y-axis on the left represents inference time (in seconds) and peak power (in watts), and the y-axis on the right has energy consumption (in watt-seconds, Wsec).

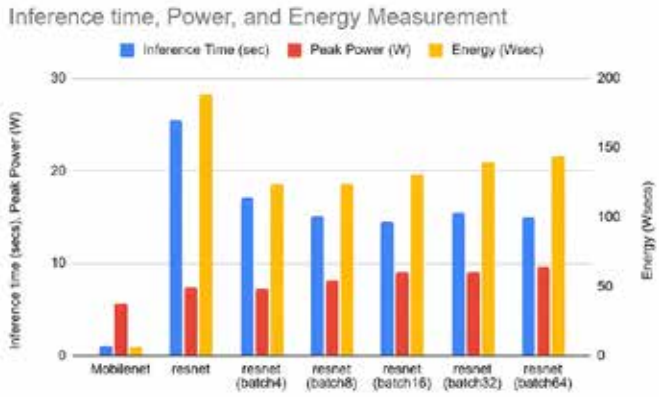

Figure 2: Inference time, power, and energy measurement illustrate the efficiency of MobileNet compared to ResNet and the benefit of batch processing.

The analysis highlights that MobileNet is significantly faster than ResNet—completing tasks 23 times quicker—and far more energy-efficient, using 30 times less energy overall. This efficiency is evident in the correlation between power and time, as energy consumption (which equals power multiplied by time) is minimized by MobileNet's lower power usage and shorter inference times. For ResNet, while batching improves inference speed by about 1.7x, particularly with batch 16, this comes with a corresponding increase in power consumption. This trade-off means that even though tasks are completed faster, higher power usage can offset the time gains, leading to higher overall energy consumption. However, batch processing with eight images strikes a more favorable balance, effectively managing energy consumption by optimizing time and power usage through parallel processing on multi-core devices like Orange Pi.

Interestingly, batch processing does not significantly enhance MobileNet's energy efficiency, so they are not represented in Figure 2. This is likely due to its already optimized algorithm. Still, ResNet leads to substantial energy savings by reducing the total energy consumed through a balanced reduction in time and power.

The experiments in this research were conducted on smaller models, such as MobileNetV3 and ResNet50, which are less complex than the large-scale models in Table 1 (e.g., Llama2, Llama3, and ChatGPT4). However, the core principles of energy consumption apply similarly across models of varying sizes. For instance, in Table 4 and Figure 2, ResNet50 consumed significantly more power. It took longer to process compared to MobileNetV3, reflecting how larger models like Llama3 and ChatGPT4 require more energy than Llama2 due to their increased complexity. Although my experiments involved smaller models, they effectively demonstrate the scalable relationship between model size, complexity, and energy consumption, as seen in the larger AI models in Table 1.

Recommendations

AI data centers worldwide must pursue energy efficiency techniques across all categories listed in Table 3 to make substantial gains. Though hurdles are associated with initial investment cost and compatibility with existing infrastructure, that should not be a deterrent. Given that AI workloads already require greenfield or new hardware computing infrastructure like GPUs and TPUs, the energy-efficient measures on the rest of the hardware infrastructure (Storage, Networking) and Data center facility (cooling techniques, geography, etc.) can be absorbed as part of the initial capital investment. Despite all available techniques deployed, they are primarily evolutionary, and they may not be enough over a long period due to ever-increasing AI computational demand.[26] A revolutionary technology solution would be required for sustainability. Quantum computing, based on the principles of quantum mechanics, uses quantum bits or qubits to store and process information. Unlike classical computing, which uses bits in either the '0' or '1' state, qubits can exist in multiple states simultaneously, allowing for parallel and faster computations. Quantum AI,[27] which refers to the application of quantum computing to the field of AI, is a compelling future state with the possibility to significantly enhance energy efficiency in AI applications, particularly in data centers and other energy-intensive environments.

We can do more in the regulatory (Government - Climate control), Media (News coverage), and corporate governance (Good global corporate citizen). Governments can implement energy tax credits for data centers with energy efficiency techniques. Governments can also cap the maximum energy supply from nonrenewable energy sources.[28] The environmental impact of data centers is not discussed or covered in the media.[29] It would be worthwhile to have a metric like GDP that tracks the energy efficiency index of data centers and tracks them publicly to bring awareness and to the extent it can impact the corporation's stock price. Lastly, corporations are responsible for reporting energy consumption statistics; PURE Storage is an example of a corporation that explicitly calls the energy statistics of their products.[30] In the recently disclosed 92-page Llama3 white paper by Meta,[5] there is no mention of KWh statistics for pre-training and training stages. In the age of open-source information that fosters collaboration growth, sharing the energy cost/statistics will be considered a good role model/good corporate citizen of the world.

Conclusion

While AI development primarily focuses on advancing hardware, software, and algorithms, energy consumption is increasingly gaining attention due to its growing environmental and economic impacts. Companies like Google, Meta, and NVIDIA have begun prioritizing energy-efficient solutions, reflecting a shift in how AI's power demands are being approached. Though AI's growth is still early, optimizing energy efficiency will be essential as models continue to scale. The growth of AI models, such as Llama and ChatGPT, which de

mand enormous amounts of power, highlights the urgency of addressing energy consumption in tandem with computational efficiency. This study has shown that while current strategies such as optimizing AI algorithms, batch processing, and advanced cooling technologies offer measurable improvements in energy efficiency, they may not be sufficient to keep pace with AI's rapidly increasing computational demands.

To truly achieve sustainable growth, the industry must not only implement current efforts to improve existing infrastructure but also invest in revolutionary technologies such as quantum computing, which holds the potential to dramatically reduce energy consumption in AI applications. Additionally, a coordinated effort involving regulatory bodies, corporate governance, and public awareness is crucial. Governments should incentivize energy-efficient practices through tax credits and stricter regulations, while companies must take responsibility for transparent reporting of their energy consumption and actively pursue greener alternatives.

The future of AI and data centers hinges on our ability to innovate beyond the constraints of current technologies. The following steps in research should focus on exploring and developing these groundbreaking solutions to ensure that the advancement of AI does not come at an unsustainable environmental and economic cost. The time for action is now—both industry leaders and policymakers must work together to secure a future where AI can thrive in harmony with our planet's resources.

Acknowledgments

I would like to thank my family for their support and encouragement in exploring my passion for technology and the environment. I also would like to thank Mr. Neeraj Sharma for his guidance and feedback.

References

1. https://www.wpr.org/wp-content/uploads/2024/06/3002028905_Powering-Intelligence_-Analyzing-Artificial-Intelligence-and-Data-Center-Energy-Consumption.pdf (accessed August 7, 2024).
2. Western Digital. Infographic: The AI Data Cycle. https://documents.westerndigital.com/content/dam/doc-library/en_us/assets/public/western-digital/collateral/white-paper/white-paper-how-5G-AI-and-IoT-are-shaping-future-of-storage.pdf (accessed August 5, 2024).
3. Forbes. Why Does AI Consume So Much Energy? https://www.forbes.com (accessed August 5, 2024).
4. Meta. The Llama 3 Herd of Models. AI at Meta. https://www.ai.meta.com (accessed August 5, 2024).
5. Meta. Building Meta's GenAI Infrastructure. Engineering at Meta. https://www.fb.com (accessed August 5, 2024).
6. Meta. Introducing the AI Research Supercluster. AI Research. https://www.ai.meta.com (accessed August 5, 2024).
7. NVIDIA. NVIDIA DGX A100 Datasheet. https://www.nvidia.com (accessed August 5, 2024).
8. NVIDIA. NVIDIA DGX GH200 Datasheet. https://www.nvidia.com (accessed August 5, 2024).
9. NVIDIA. NVIDIA DGX H200 Datasheet. https://www.nvidia.com (accessed August 5, 2024).
10. Semi Analysis. AI Datacenter Energy Dilemma - Race for AI Datacenter Space. https://www.semianalysis.com (accessed August 5, 2024).
11. Tom's Hardware. AI Becomes Leading Server Workload as Shipments of Other Servers Drop. https://www.tomshardware.com (accessed August 5, 2024).
12. Enerdata. World Energy Consumption Statistics. Enerdata Yearbook, 2023. https://yearbook.enerdata.net/total-energy/world-consumption-statistics.html (accessed August 11, 2024).
13. Electric Power Research Institute. Data Centers and Artificial Intelligence Could Significantly Increase Electricity Demand by 2030, According to a New EPRI Study. Electric Power Research Institute, June 14, 2023. https://www.epri.com/about/media-resources/press-release/q5vU86fr8TKxATfX8IHf1U48Vw4r1DZF#:~:text=According%20to%20a%20new%20study,supply%20challenges%2C%20among%20other%20issues (accessed August 11, 2024).
14. Supermicro. Liquid Cooling for Supermicro Servers. https://www.supermicro.com (accessed August 5, 2024).
15. Google. Data Centers: Efficiency. https://www.google.com/about/datacenters/efficiency/ (accessed August 30, 2024).
16. Dell. Viewing and Configuring Power Cap Policy. https://www.dell.com/support/manuals/en-us/idrac9-lifecycle-controller-v3.0-series/idrac_3.00.00.00_ug/viewing-and-configuring-power-cap-policy?guid=guid-8b420aae-ecdb-4d50-9cc9-13ac6669afd5&lang=en-us (accessed August 30, 2024).
17. Groq. About Us. https://groq.com/about-us/ (accessed August 10, 2024).
18. Hsieh, C.-Y.; Liu, J.; Yang, Z.; Chen, W.; Tang, C.; Tan, C.; Lin, J.; Shen, M.; So, D.; Yuan, X.; Li, C.-L.; Yeh, C.-K.; Nakhost, H.; Fujii, Y. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. arXiv, 2023. https://arxiv.org/abs/2305.02301 (accessed August 30, 2024).
19. McDonald, J.; Bayram, I.; Bove, M.; Deng, A.; Gao, X.; Hashimoto, T.; Levenberg, J.; Liu, R.; Min, J.; Shakeri, S.; Siskind, J.; Li, B.; Frey, N.; Tiwari, D.; Gadepally, V. Great Power, Great Responsibility: Recommendations for Reducing Energy for Training Language Models. Findings of the Association for Computational Linguistics: NAACL 2022, 2022. https://doi.org/10.18653/v1/2022.findings-naacl.151.
20. IEEE 802. An Overview of Energy-Efficient Ethernet. https://www.ieee802.org (accessed August 5, 2024).
21. The Next Platform. Running AI and Edge Workloads on CPUs. June 11, 2024. https://www.nextplatform.com/2024/06/11/running-ai-and-edge-workloads-on-cpus/ (accessed August 7, 2024).
22. Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv, 2017. https://arxiv.org/pdf/1704.04861v1 (accessed August 11, 2024).
23. Rybczak, M.; Kozakiewicz, K. Deep Machine Learning of MobileNet, EfficientNet, and Inception Models. Algorithms 2024, 17 (3), 96. https://doi.org/10.3390/a17030096.
24. Sharma, H. HPC-Enhanced Training of Large AI Models in the Cloud. ResearchGate, 2024. https://www.researchgate.net/profile/Himanshu-Sharma-197/publication/383229361_HPC-Enhanced_Training_of_Large_AI_Models_in_the_Cloud/links/66c377f2d1d8790f9e61ba6f/HPC-Enhanced-Training-of-Large-AI-Models-in-the-Cloud.pdf (accessed August 11, 2024).
25. Patil, K.; Desai, B. Intelligent Network Optimization in Cloud Environments with Generative AI and LLMs. Preprints.org, 2024. https://www.preprints.org/manuscript/202406.0578 (accessed August 11, 2024).
26. World Economic Forum. How to Manage AI's Energy Demand — Today and in the Future. https://www.weforum.org (accessed

DOI: 10.36838/v6i11.15

August 5, 2024).
27. Baratz, A. It's Time to Plan for an AI and Quantum Future. Forbes Technology Council. Forbes, May 16, 2024. https://www.forbes.com/councils/forbestechcouncil/2024/05/16/its-time-to-plan-for-an-ai-and-quantum-future/ (accessed August 5, 2024).
28. Scientific American. The AI Boom Could Use a Shocking Amount of Electricity. https://www.scientificamerican.com (accessed August 5, 2024).
29. Guzman, T. AI's Insatiable Appetite for Energy Is Driving Environmental Destruction. Jacobin, June 12, 2024. https://jacobin.com/2024/06/ai-data-center-energy-usage-environment/ (accessed August 5, 2024).
30. Pure Storage. Product Carbon Footprint FY'24 | FlashArrayFamily. https://www.purestorage.com/docs.html?item=/type/pdf/subtype/doc/path/content/dam/pdf/en/white-papers/wp-greening-the-data-center.pdf (accessed August 5, 2024).

Author

Sahana Shankar is a senior at the Vista Del Lago High school in California. She has a passion for engineering, technology, and environmental science. She is interested in pursuing an engineering degree and wants to apply her knowledge to create sustainable methods of technology use.