Relative Performance and Efficiency of Apple Silicon for Training Deep Neural Networks

The Rise of Generative Artificial Intelligence as A Technological System

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
Tao Groves

November 8, 2024

Technical Team Members: Tao Groves

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

Rider Foley, Department of Engineering and Society
Felix Lin, Department of Computer Science

Introduction

In the last few years, generative artificial intelligence (GenAI) has metamorphosed from a technology that professionals rely on to optimize and support infrastructure, products, and academic research (Jordan, 2015) to being synonymous with chat bots and homework help. This transformation entered the public eye in 2022 when OpenAI, the current AI technology leader, released ChatGPT, the first free and public-facing Large Language Model (LLM). However, this transformative event was predated by years of technological and infrastructural development behind the scenes which laid the groundwork for the capabilities seen in today's GenAI applications. Since the release of ChatGPT, LLMs have been applied across a broad range of applications, which in turn has driven unprecedented interest and investment in AI technology. With companies competing to push the limits of AI's capabilities, the focus has predominantly been on scaling model sizes and enhancing performance, often with limited attention to the environmental costs. This rapid pace has exacerbated infrastructure strain, as AI workloads now account for a significant portion of global data center energy use, contributing to broader concerns around sustainability (Patterson et al., 2022; Wu et al., 2022).

Fortunately, there exist a multitude of potential avenues to improving AI sustainability, the most promising of which involve designing novel computing hardware with efficiency as its primary focus. This research project combines original research on the challenges faced by current AI computing hardware, along with a sociotechnical exploration of the exigence and trajectory of GenAI, in an effort to develop practical recommendations for engineers to develop and implement these systems sustainably.

Relative Performance and Efficiency of Apple Silicon for Training Deep Neural Networks

Most industry professionals believe there are multiple avenues through which GenAI could become more sustainable (Patterson et al., 2022; Wright et al., 2023; Wu et al., 2022). These include selecting efficient model architectures like sparse models to decrease computation time, leveraging cloud computing for better energy management, sourcing energy more locally and sustainably, and developing specialized processors which can compute the calculations needed for large models while using less energy. There is evidence to suggest that a combination of these strategies can be applied to great effect even with current technology; Google began applying all of them in its data centers starting in 2019 and by 2021 was able to reduce its carbon footprint per training iteration per server by a factor of 747 (Patterson et al., 2022). However, this holistic strategy is only achievable when designing a self-contained system. Google builds its own data centers to train its own models on its own data and has control over every stage of the pipeline. Most organizations wishing to research or implement GenAI technology must rely on many outside factors. They do not have the resources to design a custom sparse model and must rely on a large general-purpose baseline, cannot move their datacenters to a location with more sustainable energy sources, and must purchase computing hardware from a third party. With such constraints, attempting to balance economic, compute, carbon, and energy efficiencies becomes extremely difficult even with the optimistic assumption that every organization has sustainable intentions (Wright et al., 2023).

I believe the most pragmatic path to widespread progress is improving the efficiency of computing hardware by designing it from the ground up with AI in mind. A chip which can perform the same computation while using less energy is inherently more compelling to those

who pay for that energy. It can be implemented at scale without requiring social or political pressure.

While AI hardware has evolved extremely rapidly in recent years, the driving force behind progress has been increasing speed, not reducing energy cost (Wright et al., 2023). A notable exception is Apple Silicon, a proprietary architecture developed by Apple which is primarily designed and marketed around efficiency (see Figure 1) (Vena, 2022).

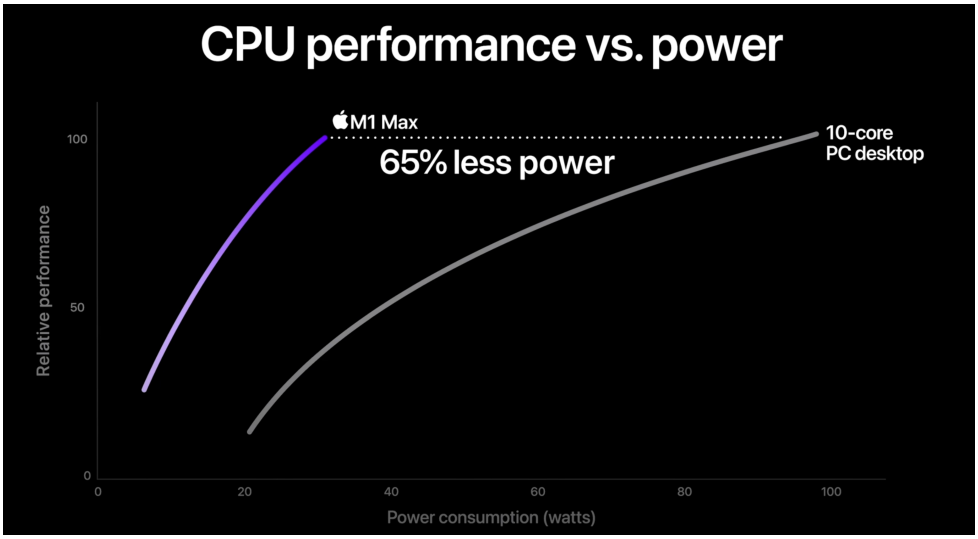

Figure 1: Apple marketing graph highlighting the efficiency of Apple Silicon vs. its competitors (Apple, 2022)

The design details of Apple Silicon are not public, and it is not clear what makes it so efficient nor how it compares to other prevailing AI accelerators across the wide variety of tasks required for AI training and inference. Answering this question could give engineers valuable insight when designing their own efficiency-focused chips.

The goal of this technical project is to explore the potential and limitations of current processor architectures for generative AI training, to determine the limiting factor of the training pipeline, and to suggest specific hardware design principles for maximum efficiency. The project will begin with review of current leading hardware (Muralidhar et al., 2022; Wang et al., 2020)

as well as existing model design principles (Patterson et al., 2021; Wu et al., 2022). It will then proceed to original research measuring the relative performance and efficiency of Apple Silicon against its competitors across a variety of tasks to determine its specific strengths and weaknesses at a hardware level. These findings, along with a thorough analysis of the reverse salients preventing existing designs from achieving similar efficiency, will be compiled in a peer-reviewed academic research paper and published in spring 2025.

The Rise of Generative Artificial Intelligence as A Technological System

In 2017, Google released a landmark paper detailing a novel machine learning model architecture, the "transformer," which was the first sequential model capable of effectively translating text between languages without requiring an impractical amount of computing power (Vaswani et al., 2023). Researchers quickly began to discover that this efficiency could be leveraged to build much larger models capable of a broader range of language tasks. In 2018, OpenAI released GPT-1, the first LLM, which could be trained to a high level of language proficiency with very little data and then finetuned to a specific task (Radford et al., 2018). Development on GPT models continued at a rapid rate, and they began to be applied throughout the industry – for example, Google Translate switched to a transformer architecture in 2020 (Caswell & Liang, 2020). However, LLMs did not enter the public eye until the release of ChatGPT, a free online tool that allowed anyone to interface with a powerful LLM in a familiar "chat" interface. ChatGPT became an overnight sensation, gaining over 100 million users in two months and becoming the fastest-growing web application in history (Milmo, 2023). Tech corporations were quick to jump on this trend, and LLMs are now integrated into everything from Google search to the iPhone.

According to the theory of technological momentum as defined by Hughes (1987), a newly born technology is plastic, able to be shaped and controlled by society to suit our needs. Gradually, as the technology grows and becomes entrenched in a wider system, it begins to harden, and becomes more and more difficult to steer. Eventually it gains such momentum that no one can shift it from the path it now lies on, and it begins to influence society in turn, often in unforeseen ways. I believe the rise of GenAI reflects this shift. Before the release of ChatGPT, when LLMs were open source, abstracted behind the tools they powered, and trained for specific tasks, their applications could be much more varied. Since the models themselves were smaller and free to use they could be tweaked and tuned for any purpose. The model behind ChatGPT in contrast, the largest of its kind and first with its design kept secret, was trained specifically to become a chatbot, with a rigid "prompt to answer" pipeline (OpenAI, 2022). Whoever used it must abide by the same interface, making it more difficult to design novel applications. However, this chat-based design was far easier to use and more capable than anything before and was the first to truly break into public culture. It quickly spread across the tech world and remains the only lens through which most people understand LLMs.

Just as technological momentum dictates, GenAI has developed more deterministic effects on society as its trajectory has become more rigid. Some believe there is a growing disconnect between the capabilities being developed and the practical, ethical, and social needs of the broader population (Weidinger et al., 2021) and that a constant thirst for progress has encouraged the industry to prioritize rapid development over risk management and environmental sustainability (Wu et al., 2022). A 2024 survey of the UK public showed a generally optimistic view of the technology's potential, but many participants reported feeling uncertain or uninformed about the risks associated with AI (Bright et al., 2024). This uncertainty

is just as prevalent among AI developers. In June 2024, current and former OpenAI employees published an open letter condemning the company's reckless disregard of ethical and safety concerns (Field, 2024). Additionally, the AI industry has begun to attract the attention of national governments, which are incentivized to bend the rules and suppress discourse even further in pursuit of becoming the leader in the space. A large-scale analysis of China's public sphere found that AI-related discourse was being manipulated by its government to create a more positive image of the technology and shut down criticism (Jing Zeng & Schäfer, 2022), and experts in and out of government have argued that the US should leverage its AI dominance to gain a foreign-policy advantage (Frank, 2024).

Even without the ethical risks involved, there exists a gargantuan environmental cost of building and operating so much AI infrastructure. As an example, Microsoft reported a 2.5-fold increase in its overall energy consumption between 2018 and 2023, an increase it attributes mostly to GenAI (see Figure 2).

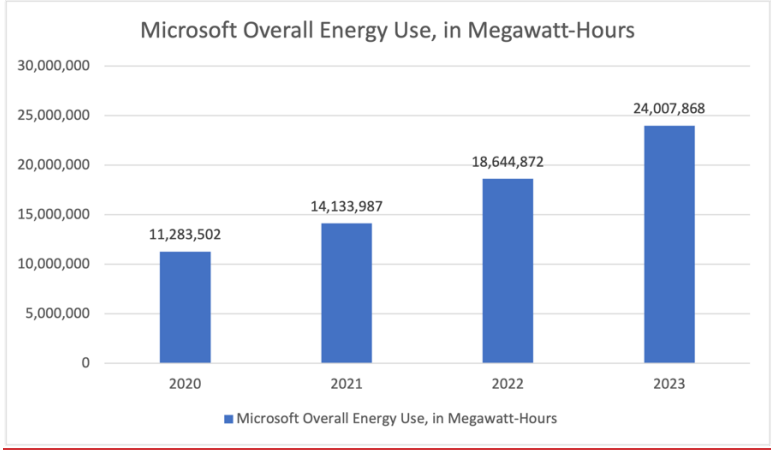

Figure 2: Microsoft Energy Use Over Time (Microsoft, 2024)

This rate of increase in power consumption, on an international level, is unprecedented, and its effects are already being felt. Grids are being strained, resources are being redirected, and retirement of fossil fuel power plants is being delayed. (Halper, 2024). Global data center water

consumption is likely to reach more than half that of the entire United Kingdom by 2027 (Li, 2023). It is clear that the prioritization of speed and expansion over environmental and long-term economic stability has put the GenAI industry on a destructive and unsustainable path.

Research Question and Methods

GenAI has transcended its original bounds to become a political weapon, the driving factor behind growth of the AI industry as a whole, and a major influence on society. This begs the question: Why did GenAI take on this greater purpose? What makes it unique from other technological trends, and how will it shape the future of computing? Most importantly, how can politicians and engineers work together to minimize its social and environmental harms and work toward a sustainable future of AI for everyone?

To investigate this problem, I will conduct a meta-review on the rise of GenAI in the United States, looking at public sphere heuristics such as Google search trends and social media engagement alongside industry messaging (AI product releases, venture capital investment) and infrastructure expansion (financial reports, energy consumption data). By considering these statistics through the lens of technological momentum, how they shaped the trajectory of GenAI at its inception and how it now influences them in turn, I hope to gain valuable insight into the past, present, and future of this technology. To determine whether social and political factors might become increasingly more influent, I will compare public and private perception and adoption between the United States and China, evaluating the potential for GenAI to become a techno-political artifact as defined by Langdon Winner (Dyson et al., 2021).

Conclusion

Generative Artificial Intelligence has expanded rapidly in both technical capabilities and public impact, requiring massive infrastructure and resource inputs for training and deployment. This growth, if unchecked, risks imposing significant environmental costs, exacerbating social divides, and potentially overlooking broader ethical and public concerns. The objective of my research is to investigate how generative AI can be developed sustainably by exploring both the technical challenges of AI efficiency and the social imperatives that must be addressed to guide responsible AI adoption. By evaluating the problem from both sides, I hope to develop a grounded and nuanced perspective from which to draw potential solutions. These may take the form of social change, ways we can change our perception or application of this technology to reduce its potential harm, or technological change, avenues to more robust and efficient AI infrastructure. Through this research, I aim to make a meaningful contribution to a sustainable future of AI for everyone.

REFERENCES

2024 Environmental Sustainability Report. (2024). Microsoft.

AI Index Report 2024 – Artificial Intelligence Index. (2024). Stanford AI Index. https://aiindex.stanford.edu/report/

Apple. (2022, March 8). Apple Event – March 8. [Video recording].

Bright, J., Enock, F. E., Esnaashari, S., Francis, J., Hashem, Y., & Morgan, D. (2024). Generative AI is already widespread in the public sector (arXiv:2401.01291). arXiv. https://doi.org/10.48550/arXiv.2401.01291

Caswell, I., & Liang, B. (2020, June 8). Recent Advances in Google Translate. Google Research Blog. http://research.google/blog/recent-advances-in-google-translate/

Hughes, T. P. (1987). The Evolution of Large Technological Systems. In W. E. Bijker, T. P. Hughes, & T. Pinch (Eds.), *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology* (pp. 51–82). MIT Press.

Field, H. (2024, June 4). Current and former OpenAI employees warn of AI's "serious risks" and lack of oversight. CNBC. https://www.cnbc.com/2024/06/04/openai-open-ai-risks-lack-of-oversight.html

Frank, M. (2024, September 22). US Leadership in Artificial Intelligence Can Shape the 21st Century Global Order. The Diplomat. https://thediplomat.com/2023/09/us-leadership-in-artificial-intelligence-can-shape-the-21st-century-global-order/

Hughes, T. P. (1987). The Evolution of Large Technological Systems. In W. E. Bijker, T. P. Hughes, & T. Pinch (Eds.), The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology (pp. 51–82). MIT Press.

Jing Zeng, C. C., & Schäfer, M. S. (2022). Contested Chinese Dreams of AI? Public discourse about Artificial intelligence on WeChat and People's Daily Online. Information, Communication & Society, 25(3), 319–340. https://doi.org/10.1080/1369118X.2020.1776372

Milmo, D. (2023, February 2). ChatGPT reaches 100 million users two months after launch. The Guardian. https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app

Muralidhar, R., Borovica-Gajic, R., & Buyya, R. (2022). Energy Efficient Computing Systems: Architectures, Abstractions and Modeling to Techniques and Standards. ACM Comput. Surv., 54(11s), 236:1-236:37. https://doi.org/10.1145/3511094

OpenAI. (2022, November 30). Introducing ChatGPT | OpenAI. https://openai.com/index/chatgpt

Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2022). The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink (arXiv:2204.05149). arXiv. https://doi.org/10.48550/arXiv.2204.05149

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon Emissions and Large Neural Network Training (arXiv:2104.10350). arXiv. https://doi.org/10.48550/arXiv.2104.10350

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (n.d.). Improving Language Understanding by Generative Pre-Training.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention Is All You Need (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Vena, M. (2022, March 15). The 3 Most Important Marketing Messages About Apple Silicon. LinkedIn. https://www.linkedin.com/pulse/3-most-important-marketing-messages-apple-silicon-mark-vena/

Wang, Y., Wang, Q., Shi, S., He, X., Tang, Z., Zhao, K., & Chu, X. (2020). Benchmarking the Performance and Energy Efficiency of AI Accelerators for AI Training. 2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID), 744–751. https://doi.org/10.1109/CCGrid49817.2020.00-15

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Gabriel, I. (2021). Ethical and social risks of harm from Language Models (arXiv:2112.04359). arXiv. https://doi.org/10.48550/arXiv.2112.04359

Wright, D., Igel, C., Samuel, G., & Selvan, R. (2023). Efficiency is Not Enough: A Critical Perspective of Environmentally Sustainable AI (arXiv:2309.02065). arXiv. https://doi.org/10.48550/arXiv.2309.02065

Wu, C.-J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Behram, F. A., Huang, J., Bai, C., Gschwind, M., Gupta, A., Ott, M., Melnikov, A., Candido, S., Brooks, D., Chauhan, G., Lee, B., Lee, H.-H. S., … Hazelwood, K. (2022). Sustainable AI: Environmental Implications, Challenges and Opportunities (arXiv:2111.00364). arXiv. https://doi.org/10.48550/arXiv.2111.00363