Hardware Solutions to Ensure Future Sustainable AI Developments

A White Paper Advocating the a Sustainable Development of AI for Researchers, Deeptech VCs, and Chip Manufacturers

Introduction: Artificial Intelligence and Climate Change

The world has transitioned into the era of artificial intelligence (AI) amid an explosive demand for better and larger models, driven largely by their numerous use cases in autonomous vehicles, automating tedious tasks, personal education, and more. There is a noticeable dramatic increase in the number of AI models trained from 2014 to 2024, as shown in Figure 2, contrasting with Figure 1, which illustrates the number of AI models trained from 1950 to 2014[1].

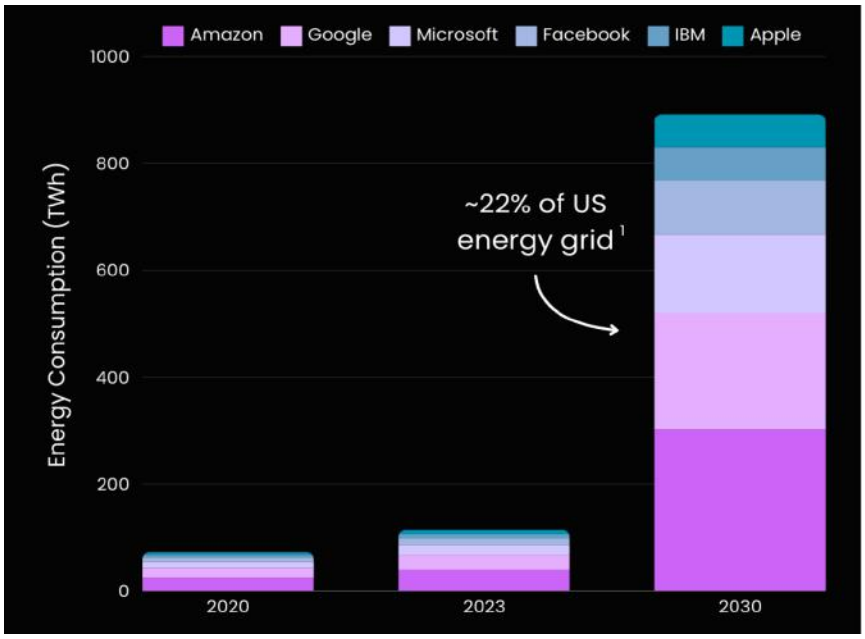

Figure 1. Number of AI Models Trained from 1950 to 2014.

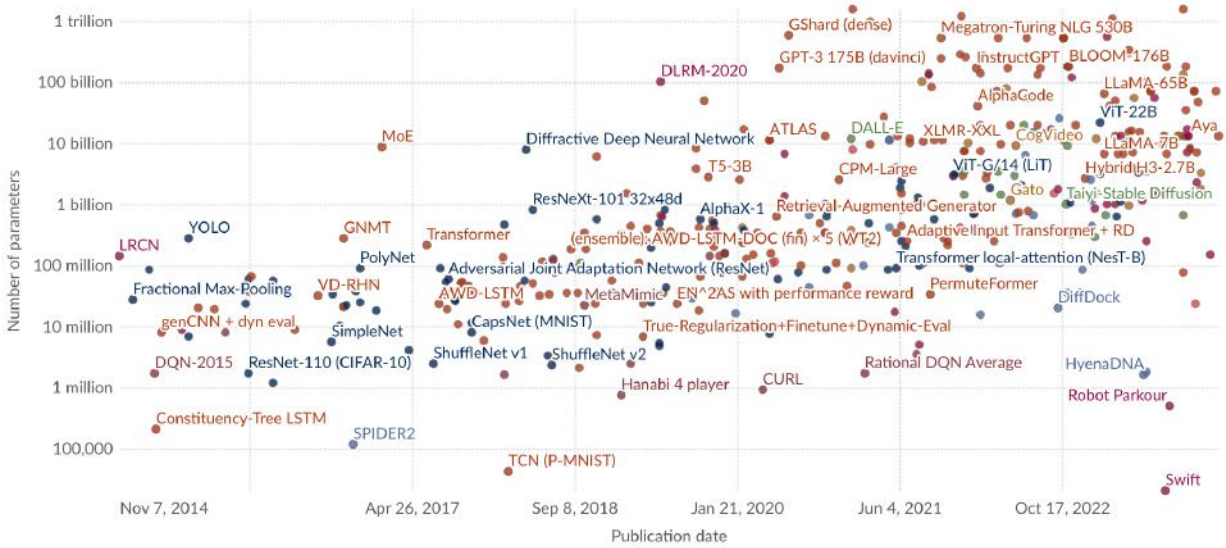

Figure 2. Number of AI Models Trained from 2014 to 2024.

However, this rapid advancement in AI technologies comes with a hidden consequence: an acceleration of global climate change.

—

[1] OurWorldInData, “Parameters in Notable Artificial Intelligence”, np.

Already, AI related usage has led to a significant growth in data center utilization – cloud computers that allow AI models to operate. Dr. Naveen Varma, a data center expert and Princeton professor, notes that data center usage has increased by approximately 1 million percent[2] and is projected to double every 6 months[3]. The global electricity consumption attributable to data centers has tripled from 2% to 6% between 2018 and 2020[4] and is forecasted to account for 8% and 21% of worldwide energy consumption by 2030[5].

OpenAI and the University of Massachusetts Amherst also highlight the resulting increase in CO2 emissions[6]. Training a moderately-sized language model—typically 40 billion parameters—results in approximately 300,000 kilograms of carbon dioxide emissions[7]. The GPT-4 and Google's Gemini Ultra models, which are 44 and 39 times larger, respectively, are driving further increases. Emerging data from Amazon already indicates a 15% increase in CO2 emissions due to AI[8].

AI's dependence on computational power means that a linear gain in performance requires an exponential increase in parameters[9] – the information that feeds and trains a specific model – and consequently, in carbon emissions and energy consumption[10]. Currently, the field of AI remains unsustainable. Researchers predict that if AI is operated as they are today, the resulting carbon emissions and energy use would lead to a greater than 2°C increase in global temperature – the limit set by the world during the Paris Agreement to prevent dooming the Earth to irreversible climate damage. **Thus, achieving AI growth that remains in line with the Paris Agreement requires it to be made more energy efficient.**

This whitepaper aims to address how AI will exponentially exacerbate the global climate crisis and highlight what current solutions chip manufacturers, researchers, and venture capitalists are working on or funding. Among them, this whitepaper will highlight biocomputing – the ability to perform AI tasks on biological cells – as an extremely promising solution. This paper emphasizes the emerging and exciting developments in the biocomputing field to academics that may be interested in pursuing or knowing about, the growing feasibility of designing biocomputing chips to chip manufacturing and design companies beginning to look for alternative climate-friendly solutions, and finally to venture capitalists who believe that the global tailwinds around AI and biocomputing would lead to the emergence of a $100 billion dollar valued startup.

—

[2] Princeton University, "New Chip Built for AI Workloads", np.
[3] AI Now Institute, "Computational Power and AI", np.
[4] IEA, "Electricity 2024 – Analysis", np.
[5] De Vries, "Growing Energy Footprint", 2191.
[6] OpenAI, "AI and Compute", np.
[7] Strubell, "Energy and Policy Considerations", 4.
[8] Dhar, "Carbon Impact of Artificial Intelligence", 425.
[9] Appenzeller, "High Cost of AI Compute", np.
[10] Dhar, "Carbon Impact of Artificial Intelligence", 425.

Background: The State of Silicon Hardware

AI model development across companies clearly indicates that these models will grow larger and will increasingly require more processing power and energy[11].

The prevailing strategy involves creating more energy-efficient silicon hardware chips. Historically driven by the trend that chips will progressively increase in memory capability (Moore's Law) and energy efficiency (Dennard's Law), these developments have produced silicon chips capable of efficiently training and operating AI technologies.

However, four primary reasons demonstrate why current silicon hardware cannot sustain energy-efficient AI demands:

1. Silicon Advancements Cannot Keep Up With AI Demands

The innovation in silicon chips as described by Moore's Law and Dennard's Law have decelerated[12-14]. Each individual transistor is so small – smaller than a virus[12] – that further miniaturization breaks the fundamental properties of physics[13]. Attempts for further miniaturization introduces novel quantum problems that only occur at the increasingly smaller scales. Consequently, progress in improving traditional silicon hardware to make better, denser, and more efficient chips has also dramatically slowed[14].

Taking into account that the number of transistors on a chip doubles every 3-4 years coupled with the fact that current AI models require the doubling of compute every 3-6 months, silicon computing can no longer keep up with the rapid requirements of AI.

2. Silicon Shortage Threatens Future Silicon Advancements

Another problem with silicon chips is the supply of silicon itself. According to estimates from the Semiconductor Research Corporation, the relentless pace of AI's data usage expansion could soon push the demand for silicon-based memory components beyond the annual global production capacity of the material[15]. This is because the process of mining and processing silicon is not keeping up with the demand for processed silicon. This scenario is projected to lead to significant shortages and create bottlenecks in the manufacturing of silicon chips, severely impacting the availability of essential components for a wide range of industries, from consumer electronics to critical infrastructure systems. Such constraints could not only inflate costs due to scarcity but also delay the production of new technology, stifling innovation and technological advancement in sectors dependent on these components[16].

—

[11] OurWorldInData, "Parameters in Notable Artificial Intelligence", np.
[12] Penn Today, "Hidden Costs of AI", np.
[13] Cai, "Brain Organoid Reservoir Computing", 1032.
[14] Mann, "Thermal Management", np.
[15] SRC, "Plan for Semiconductors", 5.
[16] Leffer, "Shocking Amount of Electricity", np.

3. Silicon’s Memory Bottleneck Limits Energy Efficiencies

Current silicon chips are also hampered by an inherent memory bottleneck[12-14], a limitation rooted in their core architecture, which dates back to the 1940s. This traditional design paradigm requires data to be stored in one location and computed in another, necessitating the constant shuttling of information between these two sites. As a result, data transfer between the storage and computing units within a single chip are slowing down the entire system. This leads to inefficiencies in both energy consumption and processing time. This separation exacerbates the energy inefficiency of silicon chips, leading to higher operational costs and increased carbon emissions, which contribute to the acceleration of global climate change. This architectural inefficiency not only affects the environmental footprint of these technologies but also limits their potential for scaling up to meet the growing computational demands of advanced AI systems.

4. Silicon’s Heat Output Demands Energy-Hungry Cooling

A final challenge with silicon hardware is the general release of heat as a byproduct of computing. Current Nvidia chips, for instance, generate kilowatts of heat as a byproduct. In a typical server rack, this results in a 21-24 kilowatts thermal load that must be dissipated in order to maintain optimal chip performance. Ultimately, this leads to massive energy costs just to cool the silicon hardware — typically 40-50% of a data center’s electricity consumption[15]. Moreover, the latest chips from AMD and Nvidia are seeing rapid increases in power consumption, which further exacerbates the cooling demand. Consequently, it is estimated that global data centers will experience a 50% increase in energy consumption solely for the purpose of cooling the silicon[16].

Silicon Is Not Capable for Achieving AI Sustainability

Given the mounting challenges faced by the current silicon-based computing paradigm, it is imperative for deeptech investors and chip manufacturing companies to recognize that continuing along this trajectory is unsustainable for future AI developments. The exponential growth in AI's complexity and resource requirements demands a radical shift in the global approach to computing. To sustain the rapid advancement of AI technologies and their applications, investors and companies must explore and invest in alternative computing architectures that promise greater efficiency, scalability, and environmental sustainability. A new paradigm of computing that can meet the ambitious demands of tomorrow's AI is required.

—

[12] Penn Today, "Hidden Costs of AI", np.
[13] Cai, "Brain Organoid Reservoir Computing", 1033.
[14] Mann, "Thermal Management", np.
[15] SRC, "Plan for Semiconductors", 5.
[16] Leffer, "Shocking Amount of Electricity", np.

Solutions: Silicon, Architecture, and Biocomputing

Silicon Advancements from Large Companies

Currently, major companies who design and manufacture chips, such as Intel, Nvidia, and AMD, continue to address the energy and carbon footprint problem of silicon chips with further advancements in silicon technology. For example, they claim that new materials combined with silicon, such as Germanium, would enhance the computing properties of using silicon. However, this approach stems from a short-sighted vision of sustainable AI developments and is fundamentally flawed for numerous reasons:

1. A silicon shortage will continue to exist as long as chip manufacturers rely on silicon as the base of their technology. In the near future, silicon supply will not keep up with the demand needed in AI development.
2. Although combining Germanium and other materials into silicon chips would theoretically improve the compute power of a chip, it does not address the memory bottleneck issue of traditional silicon chips. AI researchers project that the bottleneck in AI developments stems from this memory issue and not solely from generating necessary compute power.
3. The byproduct of heat from the silicon chips would remain in the kilowatt ranges, necessitating further energy consumption for cooling, and show no sign of slowing down.

Because these companies do not require venture capital investment, deeptech investors should not concern themselves with the emerging silicon advancements. However, researchers and chip manufacturers should greatly consider whether this path towards sustainability is a viable long-term solution, or is merely a band-aid over a growing problem.

Architecture Advancements from Startups

On the other hand, numerous startups involved in the chip sector have converged on a seemingly better solution: redesigning the architecture of chips that are used to train AI. For example, a new type of architecture called neuromorphic-based chip design is growing in popularity and being pursued by various startups, including RainAI and BrainChip. Other companies, such as Groq, are building “language processing units” that are specialized in AI conversing tasks. These approaches address the primary issue regarding sustainable AI development: the inefficiencies that stems from the memory bottleneck. These companies are redesigning how a chip is manufactured in order to bypass the memory bottleneck prevalent in existing chips, and thus creating more energy efficient chips.

Hardware architecture researchers and chip manufacturers should take a greater look at the advantages that are offered through architecture advancements over silicon advancements. Deeptech venture capitalists should also consider these “foreign” or alternative computing technologies within their portfolio, as the size and future growth of the AI market is poised for continued explosive growth, bringing need for these alternative technologies.

—

[17] Smirnova, “Reservoir Computing”, 934.

[18] Cai, “Brain Organoid Reservoir Computing”, 1033.

[19] Tang, “Bridging Biological and Artificial”, 1.

Biocomputing Paradigm from Researchers

With advances in bioengineering techniques[17], a new type of computing paradigm known as biocomputing is poised to emerge as an efficient and sustainable solution to the requirements of AI technologies[18]. The foundation of biocomputing involves utilizing cells within the human brain (neurons) for computation, rather than relying on traditional silicon hardware. This approach addresses each of the problems outlined with silicon:

1. Neurons are inherently energy-efficient due to its ability to adapt and learn[19] – a feature notably absent in traditional computing systems, which require pre-defined programming and fixed architectures, which is less energy efficient.
2. Neurons are also environmentally sustainable, as they can be infinitely sourced from biological models, such as self-replicating cells derived from rats[17-18].
3. Neurons bypass the limitations posed by the memory bottleneck prevalent in current silicon systems – important for the data-intensive tasks involved in AI training[19].
4. Neurons do not generate heat as a byproduct[17]. This characteristic eliminates the need for the kilowatts of energy consumed in cooling processes, dramatically reducing the overall energy and carbon footprint of computing operations[19].

The emerging biocomputing paradigm is an extremely promising solution to the current problem of AI sustainability. Already, current biocomputers have been used in rudimentary computer vision[17] and speech recognition tasks[18-19]. A future goal is to approach one-shot learning in neurons similar to human learning experiences. For example, traditional deep learning models are successful when there is a large, labeled, and unnoisy dataset. But humans are adept at learning from just one or few instances of data[19].

This emerging biologics-based approach offers extremely promising results that should attract the interests of researchers, venture capitalists, and chip manufacturers alike:

- AI and materials researchers should consider collaborating with primary neuroscience researchers (and vice versa) to explore this emerging field.
- Deeptech venture capitalists should keep note and analyze the emerging biocomputing industry as a long-term value proposition within their portfolios.
- Chip manufacturers should take the necessary steps to research this alternative computing technology, and if proven viable, be willing to rapidly adjust to this new paradigm through investments into wet-lab space and hiring neuroscientists.

—

[17] Smirnova, “Reservoir Computing”, 934.

[18] Cai, “Brain Organoid Reservoir Computing”, 1033.

[19] Tang, “Bridging Biological and Artificial”, 1.

Conclusion: The Current and Future State of AI

There is an urgent necessity for a paradigm shift in AI development to address the unsustainable energy consumption and carbon emissions linked to current silicon-based technologies. This whitepaper advocates strongly for the adoption of biocomputing, a promising and sustainable alternative form of computing. However, the transition to biocomputing also comes with challenges – including technology scaling, neuron maintenance, and supporting bio infrastructure. Addressing these challenges will be extremely challenging to solve, but if researchers, chip manufacturers, and deeptech investors are able to come together and fund research in this next wave of computing innovation, the AI industry can continue its rapid growth trajectory while adhering to environmental imperatives.

Bibliography

"AI and Compute." 2018. OpenAI. May 16, 2018. https://openai.com/research/ai-and-compute.

Appenzeller, Guido, Matt Bornstein, Martin Casado, Guido Appenzeller, Matt Bornstein, and Martin Casado. 2023. "Navigating the High Cost of AI Compute." Andreessen Horowitz. November 15, 2023. https://a16z.com/navigating-the-high-cost-of-ai-compute/.

Cai, Hongwei, Zheng Ao, Chunhui Tian, Zhan Hao Wu, Hongcheng Liu, Jason Tchieu, Mingxia Gu, Ken Mackie, and Feng Guo. 2023. "Brain Organoid Reservoir Computing for Artificial Intelligence." *Nature Electronics* 6 (12): 1032–39. https://doi.org/10.1038/s41928-023-01069-w.

"Computational Power and AI." 2023. AI Now Institute. October 11, 2023. https://ainowinstitute.org/publication/policy/compute-and-ai#:~:text=This%20trend%20has%20borne%20out,in%20only%205.7%20months%2012.

De Vries, Alex. 2023. "The Growing Energy Footprint of Artificial Intelligence." *Joule* 7 (10): 2191–94. https://doi.org/10.1016/j.joule.2023.09.004.

"Decadal Plan for Semiconductors - SRC." 2021. January 2021. https://www.src.org/about/decadal-plan/.

Dhar, Payal. 2020. "The Carbon Impact of Artificial Intelligence." *Nature Machine Intelligence* 2 (8): 423–25. https://doi.org/10.1038/s42256-020-0219-9.

"Electricity 2024 – Analysis - IEA." 2024. International Energy Agency. January 2024. https://www.iea.org/reports/electricity-2024.

Leffer, Lauren. 2024. "The AI Boom Could Use a Shocking Amount of Electricity." Scientific American. February 20, 2024. https://www.scientificamerican.com/article/the-ai-boom-could-use-a-shocking-amount-of-electricity/.

Li, Can, Miao Hu, Yuning Li, Hao Jiang, Ning Ge, Eric Montgomery, Jiaming Zhang, et al. 2017. "Analogue Signal and Image Processing With Large Memristor Crossbars." *Nature Electronics* 1 (1): 52–59. https://doi.org/10.1038/s41928-017-0002-z.

Mann, Tobias. 2023. "How thermal management is changing in the age of the kilowatt chip." *The Register*, December 18, 2023. https://www.theregister.com/2023/12/26/thermal_management_is_changing/.

Moon, J. W., Wen Ma, Jong Hoon Shin, Fuxi Cai, Chao Du, Seung Hwan Lee, and Wei Lü. 2019. “Temporal Data Classification and Forecasting Using a Memristor-based Reservoir Computing System.” *Nature Electronics* 2 (10): 480–87. https://doi.org/10.1038/s41928-019-0313-3.

“New Chip Built for AI Workloads Attracts $18M in Government Support.” 2024. Princeton University. March 6, 2024. https://www.princeton.edu/news/2024/03/06/new-chip-built-ai-workloads-attracts-18m-government-funding-revolutionary-tech.

“Parameters in Notable Artificial Intelligence Systems.” 2024. Our World in Data. March 4, 2024. https://ourworldindata.org/grapher/artificial-intelligence-parameter-count#sources-and-processing.

Smirnova, Lena, Brian Caffo, and Erik C. Johnson. 2023. “Reservoir Computing With Brain Organoids.” *Nature Electronics* 6 (12): 943–44. https://doi.org/10.1038/s41928-023-01096-7.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2019. “Energy and Policy Considerations for Deep Learning in NLP.” arXiv.Org. June 5, 2019. https://arxiv.org/abs/1906.02243v1.

Tang, Jianshi, Fang Yuan, Xinke Shen, Zhongrui Wang, Mingyi Rao, Yanlin He, Yuhao Sun, et al. 2019. “Bridging Biological and Artificial Neural Networks With Emerging Neuromorphic Devices: Fundamentals, Progress, and Challenges.” *Advanced Materials* 31 (49). https://doi.org/10.1002/adma.201902761.

“The Hidden Costs of AI: Impending Energy and Resource Strain | Penn Today.” 2023. Penn Today. March 8, 2023. https://penntoday.upenn.edu/news/hidden-costs-ai-impending-energy-and-resource-strain.

Yao, Peng, Huaqiang Wu, Bin Gao, Şükrü Burç Eryilmaz, Xueyao Huang, Wenqiang Zhang, Qingtian Zhang, et al. 2017. “Face Classification Using Electronic Synapses.” *Nature Communications* 8 (1). https://doi.org/10.1038/ncomms15199.

Logical Outline

Premise

- (Given) that AI has undergone recent explosive growth in the past 5 years
- (Given) that AI development shows no sign of slowing down and is actually accelerating
- (Given) that AI research requires computers (GPUs) to train, which requires energy
- (Given) that climate change is a major problem which is mainly caused by carbon emissions from human activities

Proposition

- (Thus) Achieving AI growth that remains in line with the Paris Agreement requires it to be made more energy efficient.

Reasons & Evidence

1. (Problem, Why) Current silicon chip design should be abandoned in favor of newer technologies
    a. (Because) Silicon chips are not inherently energy efficient due to the memory gap problem
    b. (Because) Silicon is unsustainable as the world enters into a silicon shortage
    c. (Because) Silicon chips requires massively amounts of energy in order to cool
    d. (Because) Silicon technology is reaching the limits of physics and can no longer be advanced at a rate suggested by Moore's Law
2. (Solution, Why) Silicon, Chip Architecture, and Biocomputing are current attempts to solve the problem
    a. (Because) Silicon advancements might offer greater speed, but are hard to attain and are likely unfeasible in the future.
    b. (Because) Chip architecture fixes the main issues with AI training, but still runs into some of the fundamental problems associated with silicon.
    c. (Because) Biologics-based chips do not generate heat and thus use less energy and is a very promising solution for AI training and deployment in the future.

Rhetorical Outline

- Proposition: Thus, achieving AI growth that remains in line with the Paris Agreement requires it to be made more energy efficient.
- Audience: Chip manufacturing companies, deeptech venture capitalists
- Genre: White paper
- Motive of the Author: To promote biologics-based computing for investment and R&D
- Motive of the Reader: To recognize the climate problems caused by AI & to fund biologics-based computing and support climate tech solutions in the long-term
- Plan: Publish as a company white paper research (similar to Bitcoin's whitepaper: https://bitcoin.org/bitcoin.pdf), publish in TechCrunch or related site
- Rhetorical Strategies: Reference expert sources and research, use charts and data to support claims logically, identifying a problem and providing solutions, and then highlighting a single and most promising solution. A call to action that appeals to companies and VCs to make a profit, while also creating a sense of urgency in solving this problem.
- Keywords: AI, climate, energy, biologics-based computing, biocomputing, silicon.