Faster Than the Speed of Thought

by
Andi Qu

Envisioning the Future of Computing Prize
Social and Ethical Responsibilities of Computing
Massachusetts Institute of Technology

Summary

Neural networks have the remarkable ability to learn from vast amounts of data. This ability has enabled countless technological innovations in recent years. However, this great power also comes at a great cost to our fragile climate. Data centers housing the computers for training neural networks account for 2% of the United States' electricity usage and greenhouse gas emissions, and this figure will only increase as neural networks grow larger and more complex.

A significant contributor to neural networks' energy consumption is the CMOS hardware on which they operate. Although CMOS is well-suited for consumer applications like personal computers, it has failed to keep up with the computing demands of the AI industry.

This essay argues that instead of relying exclusively on CMOS, we should turn our attention toward superconducting computing. In particular, the essay explores two leading approaches to classical superconducting computing – adiabatic quantum flux parametron (AQFP) and rapid single flux quantum (RSFQ) logic. AQFP and RSFQ circuits have demonstrated an order of magnitude improvement over CMOS circuits in both speed and efficiency – a feat unmatched by any other hardware platform.

They are also uniquely suited for neural networks because they use pulses of electricity to transmit signals, just like biological neurons. On top of all that, the technology for fabricating and operating these circuits is mature enough for large-scale use. We can use this technology to create specialized circuits to accelerate neural network computations and drive down the AI industry's carbon emissions.

Of course, superconducting computing is not without drawbacks. For example, creating the circuits requires niobium, and operating them usually requires liquid helium for cryogenic cooling. However, niobium is extremely rare and mined in the Brazilian Amazon, while helium is completely non-renewable and constantly running out. The technology may also entrench inequality and empower objectionable industries like Bitcoin mining. But ultimately, superconducting computing's benefits (both environmental and social) far outweigh the harms.

The infrastructure is maturing, the advantages over CMOS are clear, and the need for faster and more efficient hardware is more salient than ever. Now is the best time to further develop this marvelous technology.

Faster Than the Speed of Thought

Superconductivity's Vital Role in the Computing Landscape of Tomorrow

Introduction	**2**
The Problems With CMOS	**3**
Superconductivity to the Rescue	**4**
Integration into the Computing Landscape	**6**
Environmental Effects	**6**
Social Effects	**7**
A Catalyst for Change	**8**
Conclusion	**9**
Acknowledgments	**9**
References	**10**

Introduction

Neural networks are computing systems designed to mimic the human brain, with the remarkable ability to learn from vast amounts of data. This ability to learn has enabled recent breakthroughs across almost every field of science and industry, from medical diagnosis to self-driving cars. In exchange for this ability, however, they require enormous amounts of computation to train and run.

For example, training the first iteration of ChatGPT (OpenAI's flagship AI model) took 34 days of non-stop computation, even with the power of 1024 computers. Run on silicon-based hardware, this training alone consumed 1.287 gigawatt hours of energy – enough to power 120 households in the United States for a year [1]. All this before even accounting for the competing (and similarly energy-intensive) AI models or the millions of daily queries made to ChatGPT.

The AI industry's growing demand for computation poses an imminent threat to our already fragile climate. Data centers housing the thousands of computers used to train and run neural networks account for about 2% of the United States' electricity usage and greenhouse gas emissions [2], and this number will only rise as neural networks grow exponentially larger to solve increasingly complex problems.

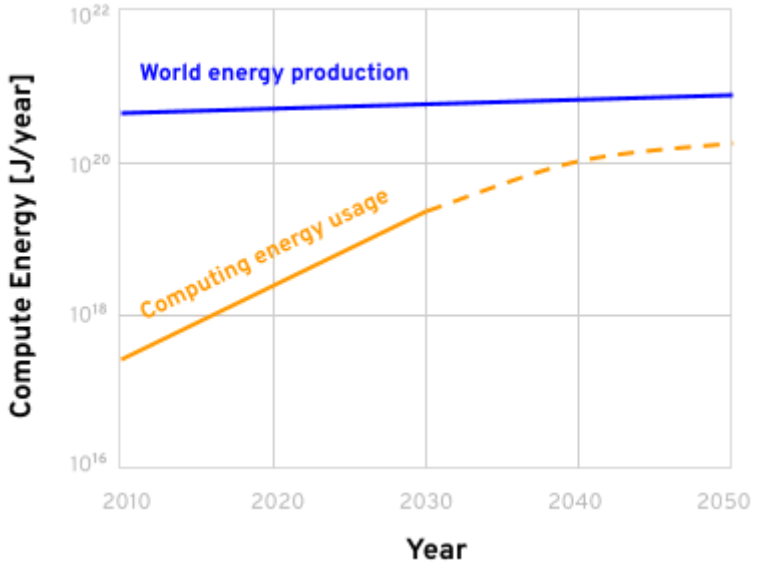

Fig. 1: Projected annual energy usage from computing. Adapted from [24].

How did such a marvelous technology, often touted as a solution to the world's most important problems, become such a force of environmental destruction? The problem lies in the type of hardware we use. Almost all of today's digital infrastructure uses complementary metal-oxide semiconductor (CMOS) electronics, from the smallest

microcontrollers to the biggest supercomputers. CMOS is performant enough for consumer applications, but it is too slow and inefficient to keep up with the computational demands of modern neural networks.

This essay argues that the solution lies in superconducting computing – a hardware platform that is an order of magnitude faster and more efficient than CMOS. By using superconducting circuits to accelerate neural network computations, we can dramatically decrease the amount of time and energy required by the AI industry.

The Problems With CMOS

CMOS uses transistors (electrically controlled switches) to control voltage levels, which then encode information as ones (high voltage) and zeroes (low voltage). Transistors dissipate energy as heat when switching between voltage levels, and this "switching energy" generally dominates CMOS's total energy consumption. A transistor's switching energy is relatively low (on the order of 10^{-16} joules per switch) but still about a hundred thousand times greater than the theoretical minimum predicted by information theory [3]. This difference quickly adds up when considering the billions of transistors on each microprocessor chip.

Furthermore, high switching energy limits the maximum operating speed of microprocessors. The heat generated from switching prevents transistors that switch too fast from being used in densely populated chips, resulting in microprocessor clock speeds plateauing at around five gigahertz since 2004 [3, 4]. All that heat also needs to be removed, so data centers consume millions of gallons of water daily for cooling [5]. This water consumption often threatens to drain local water supplies and intensify droughts.

Another problem facing CMOS's application to neural networks is that it does not mimic the human brain very well. Instead of using binary high/low voltage levels, biological neurons use short voltage pulses to transmit signals. This different signaling method is believed to be one reason why human brains are so much more efficient than computers [6], and initial studies of circuits using pulse-based signaling support this hypothesis [7]. Although it is possible to create CMOS circuits that use pulses, this modification requires additional circuitry that wastes energy. As such, the performance gain is not significant enough to justify this modification.

Until recently, engineers have countered these problems by simply shrinking the sizes of transistors, which naturally improves their performance. However, as transistors reach their physical size limits, it is becoming clear that we must develop a completely different hardware platform to overcome these problems.

Superconductivity to the Rescue

Many beyond-CMOS platforms have been proposed over the past few decades, but few are as promising as superconducting computing. Its speed and efficiency are unrivaled by any other existing digital technology, and its fabrication processes are mature enough to produce entire microprocessors. (In contrast, most other beyond-CMOS platforms are still limited to simple circuits containing just a few devices.)

Superconductors are materials that exhibit zero electrical resistance under cryogenic temperatures, which enables ultra-fast signal transmission (near light speed in many cases) and ultra-low energy dissipation [8]. While CMOS uses transistors, superconducting computing uses Josephson junctions – superconducting devices that switch between superconductive and resistive states depending on the applied electrical current. When a Josephson junction switches, it emits a short voltage pulse that can then encode information; for example, as a one if a pulse is present and zero otherwise. These pulses are typically about one picosecond short (thanks to an effect known as "flux quantization"), so Josephson junctions can switch hundreds of billions of times per second – much faster than practical CMOS transistors [8].

The two leading approaches to superconducting computing are adiabatic quantum flux parametron (AQFP) and rapid single-flux quantum (RSFQ) logic [8–10]. These two approaches are classical computing technologies (not quantum computing, despite their names), and both are much more performant than CMOS.

AQFP's most significant advantage is its extraordinarily low switching energy, typically tens of thousands of times lower than CMOS's switching energy [10]. This switching energy is so low that AQFP is at least an order of magnitude more efficient than CMOS, even accounting for the energy overhead from cryogenic cooling. Indeed, an AQFP microprocessor developed in 2021 demonstrated 80 times less power consumption than a comparable CMOS microprocessor [11]. In addition to this efficiency, AQFP is faster than CMOS, capable of operating at tens of gigahertz [10].

RSFQ loses some of AQFP's efficiency but gains a massive speed boost. Its switching energy is only about 50 times lower than CMOS's, but it can operate at hundreds of gigahertz – up to 770 gigahertz in one instance [8, 12]. Because RSFQ operates much faster than AQFP, less time and energy are needed to maintain cryogenic temperatures to complete a computation, often making RSFQ more efficient than CMOS too.

AQFP and RSFQ naturally mimic the human brain better than CMOS because they use voltage pulses to transmit information, no additional circuitry required. And just like in the

human brain, these voltage pulses allow RSFQ to perform fast and efficient analog calculations like multiplication in an otherwise digital setting [13]. Although this style of superconducting mixed-signal computing (aptly named bioSFQ) is still in its infancy, initial experiments show promising results in accelerating neural networks.

A less overt advantage of superconducting computing is its relatively mature infrastructure. From fabrication processes to circuit architecture, much of the supporting infrastructure needed to make the technology viable is much more advanced than other competing technologies. This infrastructure arose from superconductors' widespread use outside of computing, from medical imaging to high-precision sensors. Because of these commercial applications, the supporting infrastructure has consistently improved over decades of steady industrial research.

Cryogenic cooling, historically the most significant barrier to commercial use for any superconducting technology, has dramatically improved in cost and efficiency and is no longer a major concern in modern superconducting systems [8]. Fabrication technology, another barrier for most emerging hardware platforms, is also not a major concern because superconducting circuits are simpler than CMOS circuits, require similar fabrication techniques, and are resistant to variations in the manufacturing process. Some foundries, such as the MIT Lincoln Laboratory, even have dedicated processes for superconducting circuits, capable of producing chips containing almost a million Josephson junctions [14].

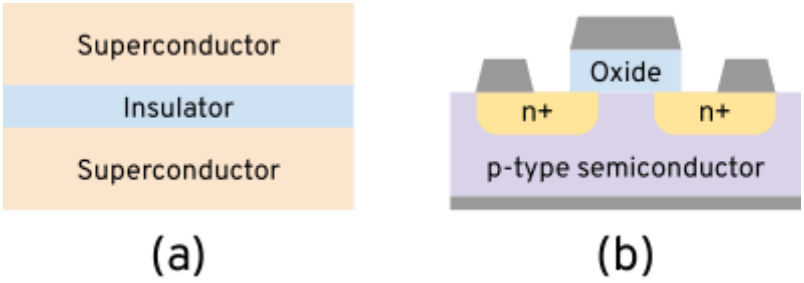

Fig. 2: The typical device structures of (a) a Josephson junction and (b) a CMOS transistor. Layer thicknesses are not drawn to scale.

Because AQFP and RSFQ are classical computing technologies, they can also leverage existing classical architectures and algorithms, such as matrix multiplication and the backpropagation algorithm for training neural networks. And because they, like CMOS, use voltage to encode signals, it is possible to integrate all three technologies into a single system [15]. Such a system would have the best of all worlds – CMOS's versatility, AQFP's efficiency, and RSFQ's speed and analog capabilities.

Integration into the Computing Landscape

Given these advantages of superconducting computing – its speed, efficiency, neuromorphic nature, and commercial feasibility – the technology is ideally suited as specialized circuits for accelerating neural network computations. Such circuits are already an established part of the AI industry (usually as graphics processing units) but currently only use CMOS. By using superconducting computing instead, we can dramatically increase the hardware's speed while also lowering the industry's massive carbon emissions.

Likewise, any high-performance computing system would benefit from superconducting computing, particularly those that repeatedly perform the same types of computation. Such systems have a broad range of applications, from cryptography to scientific research.

However, the technology would probably not replace CMOS in consumer electronics like personal computers because CMOS is already plenty good enough for those applications. And importantly, CMOS is much cheaper and more portable. Thus, instead of fully replacing it, superconducting computing should complement CMOS in the computing landscape, with each technology playing to its strengths.

Environmental Effects

On paper, superconducting computing's environmental benefits may seem rather straightforward. Being able to spend orders of magnitude less time and resources to solve the world's most critical problems – what is not to love about it? However, energy usage alone does not fully quantify a technology's environmental impact. Instead, we must also account for the materials used to develop and operate the technology. In superconducting computing, the most consequential materials come from cryogenic cooling and circuit fabrication.

Cryogenic cooling of superconductors usually involves cooling helium to its liquid state [16]. Despite advances in cooling technology, using helium is problematic because it is completely non-renewable, and the earth is constantly running out of it [17]. Setting aside enough helium to cool massive data centers would leave much less available for more critical applications like healthcare and increase our risk of depleting the resource.

Luckily, liquid helium is no longer the only option in cryogenic cooling. In many cases, liquid nitrogen reaches temperatures cold enough for superconductivity and is much more readily available. Even more promising are pulse-tube refrigerators – cryocoolers that do not require cryogenic liquids at all but can still reach liquid-helium temperatures [18]. Although pulse-tube refrigerators are currently not as efficient as conventional liquid-based

cryocoolers, there is much ongoing research to improve the technology and lead the way to a helium-free future.

Increased demand for superconducting materials may also become problematic because niobium is the most common superconductor used to make superconducting circuits [8]. Niobium is rare, making up only 20 parts per million of the earth's crust; in contrast, silicon makes up more than a quarter. In addition to its rarity, it is mined primarily in Brazil, where niobium mining operations have destroyed large swathes of the Amazon rainforest. This destruction threatens to displace thousands of indigenous Amazonians and irreversibly damage the climate. Furthermore, niobium mining often involves forced child labor because of the widespread poverty in Brazil [19]. A significant increase in demand for niobium would undoubtedly exacerbate these two problems, ultimately doing more harm than good.

However, superconducting computing is unlikely to strain niobium reserves or mining operations to this extent. Superconducting circuits actually contain relatively little niobium because the metal only appears in them as ultra-thin layers. Each layer is generally a few hundred nanometers thick [14]; in contrast, silicon wafers used as base layers in these circuits are about 2500 times thicker. And because Josephson junctions have such simple structures, fabrication processes only deposit around ten such layers for a single chip [14]. Altogether, niobium makes up a tiny fraction of the materials used to make a superconducting circuit, and the little niobium used can also be recycled because it is not doped with foreign atoms.

Less than 1% of the niobium mined today appears in electronics, including the large electromagnets used in industrial machines. The rest (over 70 thousand tons) is used to manufacture steel and other structural materials [20]. It is thus safe to assume that superconducting computing will not threaten the world's niobium supplies.

Social Effects

In addition to environmental effects, innovations in computing also often entrench inequality when groups of people lack the means to access the newest technology. This inequality would thus be especially severe for superconducting computing because it offers such an extreme performance boost but costs so much to build and run. As a result, poorer nations would risk falling even farther behind their wealthier counterparts when the technology matures.

Those with access to the technology may also use it at the expense of the rest of society. For example, high-frequency trading (HFT) and Bitcoin mining are two industries that profit

from trading stocks and processing transactions as quickly as possible; even a slight speed advantage makes a massive difference. So armed with millions of dollars and locked in an arms race for speed, firms in those industries would almost certainly want to (and have the means to) exploit superconducting computing to boost their profits.

Whether HFT and Bitcoin benefit society is still debatable; after all, HFT provides liquidity in the financial markets, while Bitcoin enables easier online transactions. However, some evidence suggests that the arms race for speed ultimately wastes resources and harms the consumers that the industries claim to benefit [21]. And unfortunately, superconducting computing would likely further drive this arms race.

Several other objectionable applications would also benefit from superconducting computing: hackers may use it to crack passwords and steal personal information, hostile militaries may use it to wage deadlier wars, and oppressive governments may use it to establish or expand mass surveillance operations.

Yet despite all these malicious uses of superconducting computing, the benefits – to the climate, economic growth, and scientific progress – far outweigh the harms. Poorer nations are usually the least insulated from climate-related catastrophes, so the reduced carbon emissions due to the technology would protect them the most. Cloud computing would also enable some degree of global access to the technology until advances in manufacturing eventually allow those nations to build their own superconducting infrastructure. Even pushing HFT and Bitcoin mining toward more efficient hardware would be a net benefit, as they currently consume enough energy to power entire nations [22].

A Catalyst for Change

A more positive social effect of superconducting computing would be the renewed interest it would bring to the broader field of beyond-CMOS electronics. The technology entering the mainstream would be more than just a technological breakthrough; it would catalyze changes in how we think about and design computing systems.

Because Josephson junctions and transistors operate on fundamentally different physics, demonstrating that superconducting computing is not only commercially viable but far more performant than CMOS could mark the biggest shift in computing since the invention of the microprocessor in 1971. Such a shift would likely create public excitement about the technology and solidify the need for investment in future research.

This excitement and investment would benefit the field of beyond-CMOS electronics in two key ways. Firstly, it would attract new talent by helping to create and highlight rewarding

career paths in electronics research. (It also helps that superconductors are exotic-sounding materials explored extensively in science fiction.) Secondly, it would diversify the computing landscape and allow new paradigms of computing to emerge – paradigms that could similarly challenge our assumptions about the limits of computing.

These benefits could also spill over into adjacent fields like physics and materials science. Superconducting computing may very possibly spur the invention of practical quantum computing or even a room-temperature superconductor. In short, it would breathe new life into some of our most fundamental scientific fields.

Conclusion

In many ways, the story of neural networks mirrors that of superconducting computing. When neural networks were invented in the 1940s, they were also overlooked because the supporting infrastructure (powerful computers and vast amounts of training data) still needed to be created. So, for decades, the world relied on a convenient yet flawed computing model, the von Neumann architecture, for all its computing needs [23]. Only in the late 2000s did the supporting infrastructure finally catch up, allowing neural networks to enter the mainstream and transform the world into what we know today.

History repeats itself, and I believe that superconducting computing holds similar significance in the future of computing. Now is the best time to further develop this technology – the infrastructure is maturing, the advantages over CMOS are clear, and the need for faster and more efficient hardware is more salient than ever.

Of course, superconducting computing is not perfect and faces several important environmental and social challenges. But the technology does not exist in a vacuum, and its success would spur the next generation of computing paradigms. Superconducting computing may not be an ideal solution to all our computing problems, but it could certainly lead us down the right path toward one.

Acknowledgments

I thank [redacted] and [redacted] for introducing me to the fascinating world of superconducting computing, as well as [redacted] for stylistic suggestions during editing. I am also grateful to the Schwarzman College of Computing and everyone involved with the Future of Computing Prize for making the competition possible.

(Names redacted for anonymity.)

References

[1] Patterson, David, et al. “Carbon emissions and large neural network training.” *arXiv preprint arXiv:2104.10350* (2021).

[2] Office of Energy Efficiency & Renewable Energy. “Data Centers and Servers.” *United States Department of Energy*, www.energy.gov/eere/buildings/data-centers-and-servers. Accessed 23 Jan. 2024.

[3] Mukhanov, Oleg A. “Energy-efficient single flux quantum technology.” *IEEE Transactions on Applied Superconductivity* 21.3 (2011): 760-769.

[4] “CPU DB - Looking At 40 Years of Processor Improvements.” Stanford University VLSI Research Group, cpudb.stanford.edu/. Accessed 3 Feb. 2024.

[5] Osaka, Shannon. “A New Front in the Water Wars.” *The Washington Post*, 25 Apr. 2023, washingtonpost.com/climate-environment/2023/04/25/data-centers-drought-water-use/. Accessed 3 Feb. 2024.

[6] Luo, Liqun. “Why Is the Human Brain so Efficient?” *Nautilus*, 3 Apr. 2018, nautil.us/why-is-the-human-brain-so-efficient-237042/. Accessed 15 Jan. 2024.

[7] Feldmann, Johannes, et al. “All-optical spiking neurosynaptic networks with self-learning capabilities.” *Nature* 569.7755 (2019): 208-214.

[8] Likharev, Konstantin K., and Vasilii K. Semenov. “RSFQ logic/memory family: A new Josephson-junction technology for sub-terahertz-clock-frequency digital systems.” *IEEE Transactions on Applied Superconductivity* 1.1 (1991): 3-28.

[9] Fourie, Coenrad J., et al. “Results from the coldflux superconductor integrated circuit design tool project.” *IEEE Transactions on Applied Superconductivity* (2023).

[10] Chen, Olivia, et al. “Adiabatic quantum-flux-parametron: Towards building extremely energy-efficient circuits and systems.” *Scientific Reports* 9.1 (2019): 10514.

[11] Ayala, Christopher L., et al. “MANA: A monolithic adiabatic integration architecture microprocessor using 1.4-zj/op unshunted superconductor Josephson junction devices.” *IEEE Journal of Solid-State Circuits* 56.4 (2020): 1152-1165.

[12] Chen, Wei, et al. “Rapid single flux quantum T-flip flop operating up to 770 GHz.” *IEEE Transactions on Applied Superconductivity* 9.2 (1999): 3212-3215.

[13] Semenov, Vasili K., Evan B. Golden, and Sergey K. Tolpygo. “BioSFQ circuit family for neuromorphic computing: Bridging digital and analog domains of superconductor technologies.” *IEEE Transactions on Applied Superconductivity* (2023).

[14] “Superconducting Integrated Circuits.” MIT Lincoln Laboratory Microelectronics Lab, https://www.ll.mit.edu/sites/default/files/facility/doc/2018-09/GOMAC_spotlight_supconducting_ic_2018.pdf. Accessed 3 Feb. 2024.

[15] China, F., et al. “Study of Signal Interface between Single Flux Quantum Circuit and Adiabatic Quantum Flux Parametron.” *2015 15th International Superconductive Electronics Conference (ISEC)*. IEEE, 2015.

[16] Ganni, Venkatarao, and James Fesmire. “Cryogenics for superconductors: Refrigeration, delivery, and preservation of the cold.” *AIP Conference Proceedings*. Vol. 1434. No. 1. American Institute of Physics, 2012.

[17] Brumfiel, Geoff. “The World Is Constantly Running Out of Helium. Here’s Why It Matters.” NPR, NPR, 8 Nov. 2019, www.npr.org/2019/11/01/775554343/the-world-is-constantly-running-out-of-helium-heres-why-it-matters. Accessed 29 Jan. 2024.

[18] Green, M. A. “The cost of coolers for cooling superconducting devices at temperatures at 4.2 K, 20 K, 40 K and 77 K.” *IOP Conference Series: Materials Science and Engineering*. Vol. 101. No. 1. IOP Publishing, 2015.

[19] Williams, Lee. “Why Do Children Work in Mining.” *Minespider*, 4 Apr. 2019, www.minespider.com/blog/why-do-children-work-in-mining. Accessed 2 Feb. 2024.

[20] “Niobium, 2023.” United States Geological Survey. pubs.usgs.gov/periodicals/mcs2023/mcs2023-niobium.pdf. Accessed 3 Feb. 2024.

[21] Budish, Eric, Peter Cramton, and John Shim. “The high-frequency trading arms race: Frequent batch auctions as a market design response.” *The Quarterly Journal of Economics* 130.4 (2015): 1547-1621.

[22] “Cambridge Bitcoin Electricity Consumption Index.” Cambridge Centre for Alternative Finance. ccaf.io/cbnsi/cbeci. Accessed 3 Feb. 2024.

[23] Roberts, Eric. “Neural Networks History: The 1940’s to the 1970’s.” cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/History/history1.html. Accessed 3 Feb. 2024.

[24] “The Decadal Plan for Semiconductors.” Semiconductor Technology Leadership Initiative. www.src.org/about/decadal-plan/. Accessed 4 Feb. 2024.