UNIVERSITY OF CALIFORNIA

SANTA CRUZ

COMPARING THE INTRINSIC HARDWARE EFFICIENCY OF THE HUMAN BRAIN WITH SILICON BASED COMPUTERS

A thesis submitted in partial satisfaction
of the requirements for the degree of

MASTER OF SCIENCE

in

PHYSICS

by

Jet Widjaja

June 2023

The Thesis of Jet Widjaja
is approved:

Professor Joshua Deutsch, Chair

Professor Anthony N. Aguirre

Professor Alexander Sher

Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

Abstract

Comparing the Intrinsic Hardware Efficiency of the Human Brain with Silicon Based Computers

Jet Widjaja

The human brain and modern computers are both extremely efficient computational machines, executing large amounts of operations at incredibly fast speeds. Such performance costs energy, with the theoretical minimum being Landauer's limit. We attempt to investigate and compare the efficiency of computation between our brain and the modern computer, i.e. cost of energy per bit of information. Every machine, including our brain, specialize in certain types of computation. We are more adept at survival tasks such as spotting colors versus mathematical operators, which the modern computer is so efficient at. Therefore, we will avoid the algorithm and architecture aspect of these computational machines. We instead quantify what a bit of information is and estimate the energy cost in both machines. We find that modern computers are extremely efficient, requiring approximately 10^{-19} J/bit which is only 10^{3} away from the Landaeur's limit while our brain requires 10^{-18} J/bit. Despite the lack of precision, the disparity is great enough to establish that modern computers are more efficient at executing a bit of computation.

Introduction

The relentless advancements in technology have transformed the landscape of computation, leading to an ever-increasing demand for energy-efficient computing systems. From microprocessors that power our personal computers to the intricate neural networks of the human brain, all forms of computation require energy. Beyond this similarity however, the human brain and modern computers are individually unique, each with its own unique architecture, algorithm and hardware. As we continue to explore the potential of artificial intelligence and machine learning, understanding the fundamental limitations of energy consumption during computation becomes paramount. The Landauer Theory states that the theoretical physical limit for computation for a single bit operation is $k_BT \ln2$ which approximates to 10^{-21} J. [1] Unfortunately, as we will discuss, our brain and computers are both orders of magnitude away from this limit.The hope is that by better understanding our brain's computation, we will be able to incorporate the lacking aspects of our technology, thereby improving our computers which is plateauing. This thesis aims to investigate and compare the efficiency of energy consumption in the human brain and modern computers.

Brain

The human brain, a marvel of biological complexity, serves as an efficient and intricate computing system capable of performing a myriad of tasks. Through millions of years of evolution, our brain, similar to the relatively compact history of computers, has also grown more efficient and powerful. With approximately 88 billion neurons and a power consumption of approximately 20 W, the human brain demonstrates remarkable energy efficiency. At the core of this functionality lies a vast network of interconnected neurons that communicate through electrical signals known as action potentials. Understanding the energy dynamics of action potentials is crucial for uncovering the mechanisms that enable the brain's remarkable energy efficiency and offers valuable insights into the development of efficient computing technologies.

Action potentials, or nerve impulses, are the fundamental means of communication between neurons in the brain. These electric signals are generated through the orchestrated interplay of ion channels and pumps that maintain and manipulate the transmembrane potential, enabling the propagation of information along axons. While the generation and propagation of action potentials are essential for the brain's functionality, they also entail significant energy costs. A detailed understanding of the energy dynamics of action potentials is not only critical for grasping for the brain's energy efficiency but also may have broader implications for the development of energy-efficient computing systems. We will investigate the energy consumption of action potentials in human brain axons through two different methods:

1) Through the induced voltages and current from an action potential
2) By calculating the capacitance across the axon membrane

By thoroughly investigating these aspects, we aim to develop a comprehensive understanding of the energy dynamics of action potentials in human brain axons.

Modern Computers

Modern Central Processing Units (CPUs), the primary components responsible for executing instructions in computers, consume a significant portion of a system's energy. Modern CPUs, characterized by their advanced architectures and billions of transistors, are responsible for performing a wide range of tasks, from executing complex calculations to managing system resources. The energy consumption of CPUs depends on various factors, including clock speed, transistor count, manufacturing process, and power management techniques. In recent years, there has been a continuous push for improved performance, often accompanied by an increase in energy usage.

Despite the increase in total energy consumption, the average cost of computation, defined in terms of floating point operations (FLOPS), has decreased significantly over the years, by as much as 10^{12} since its inception in the 1940s.

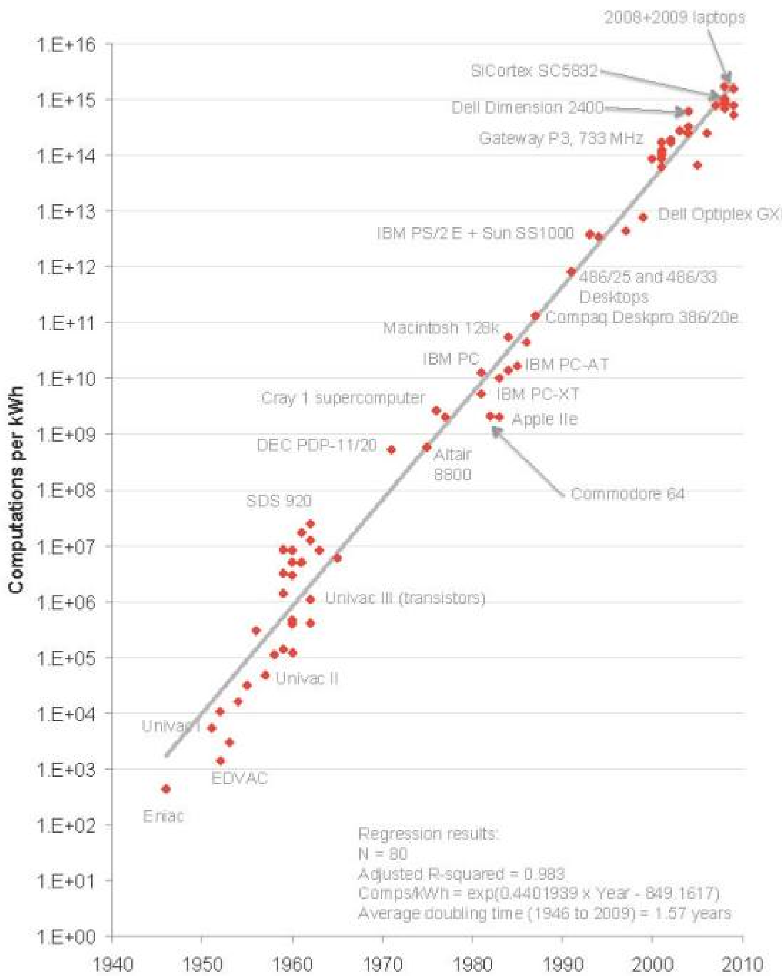

Figure 1: Computations per kWh over recent history. Taken from Dr Jon Koomey, [http://www.koomey.com/post/14466436072, CC BY-SA 3.0](http://www.koomey.com/post/14466436072)

The Electronic Numerical Integrator and Computer (ENIAC) was the first programmable general-purpose electronic digital computer built during World War I, utilizing vacuum tubes. It was first proposed by physicist John Mauchly in 1942, which was then used by the US Army to calculate complex wartime ballistics tables.

This 1800 ft^2 computer was capable of computing 5000 additions per second with a clock speed of 100 kHz. However, this technology was extremely energy inefficient, requiring 160 kW, most of which was lost as heat.

The first transistorized computer, Transistor Digital Computer (TRADIC), was built in 1954 by Jean Howard Felker of Bell Labs. An improvement from the ENIAC, TRADIC had a clock speed of 1 MHz, which unfortunately was still slower than the vacuum tube computers of the day. However, shifting from vacuum tubes to transistors significantly reduced the power consumption down to less than 100 W and shrank it to a mere 3 ft^3. Since then, as Moore's law predicted, the rate of computing power and efficiency has since exponentially increased, with a mere 800 transistors in TRADIC to billions today. Combined with an improvement in architecture, the average computer today has a clock speed of approximately 5 GHz and a processing speed of 1 teraflop with just approximately 100 W, while the world's fastest supercomputer performs at 124.5 petaflops. Our computing power and efficiency has since come a long way which begs the question if computers or the human brain, through millions of years of evolution, is more efficient?

Comparison

Many argue that our brain's computing efficiency is superior to our most advanced modern computers. For instance, OpenAI's ChatGPT, an artificial intelligence that serves to answer any of queries, thus mimicking an extremely smart human being, took megawatts and several months to train. Another similar Large Language Model (LLM), the BigScience Large Open-science Open-access Multilingual Language Model (BLOOM) with 176 billion parameters, took 118 days to train, using a staggering 433196 kWh or 1.56 TJ of energy, several orders of magnitude larger than any human would consume in a lifetime. [2] Yet, in many ways, humans are superior to these LLMs, which often still make many minor mistakes that the average human does not. However, this is most likely due to the underlying algorithm. Afterall, the human brain has had millions of years to evolve and adapt to be more efficient at such tasks. It is therefore important to distinguish the difference between the "software" and "architecture" of these computers which includes algorithms and "hardware".

Our focus here will be on comparing the efficiency of the "hardware", or the material through which signal is being transmitted in both the brain and computers instead of the algorithm. We will be comparing the energy required for the fundamental processing of 1 bit of information, which we consider to be an action potential for the brain and the flipping of a transistor switch in a modern computing processor.

How our brain processes these bits of information depends on its algorithm and thus software and architecture which we will not be investigating in this thesis.

We are assuming a fairly standard model of neuronal processing which is quite similar to the way artificial neurons are modeled. Namely that the output of axons are generated by first doing a weighted sum over time averaged inputs, and then put through a nonlinear activation function. This involves order N computations, so with N inputs there is of order one computation per input. It is also possible that the processing is more complex, resulting in an underestimation of the efficiency of computation in the brain. However, we have yet to find any experimental evidence to support this.

The overall computing process in a brain is far more complex than mere action potentials and transistor bit flips. Neurons in the brain communicate with many others, up to thousands through synaptic potentials. Similarly, multiple transistors work in conjunction with one another during computational processes. It is therefore extremely difficult, if not impossible to quantify the amount of information being processed, i.e. what 1 bit of information is in our brain and CPUs. We believe that our assumption is reasonable for our comparison in energy efficiency between the brain and modern computers.

Calculating energy consumption

Brain

The brain is an extremely complex system we are unable to fully comprehend until today. Therefore, we will be taking two different approaches towards calculating the computational energy of our brain to verify our estimation. We will specifically be estimating the energy consumption per action potential along the length of our brain neurons.

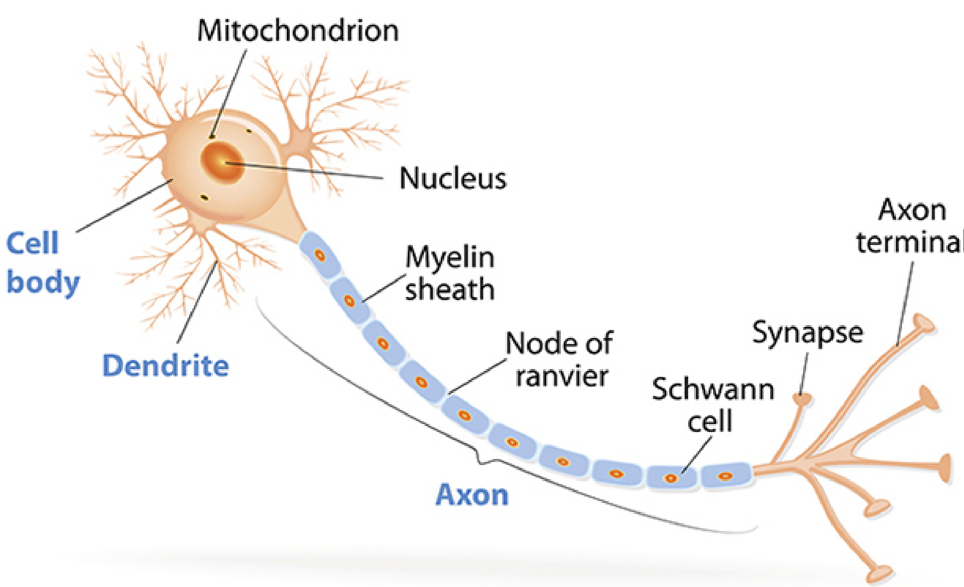

Figure 2: Anatomy of a neuron in the human brain

https://qbi.uq.edu.au/files/24323/Axon-neuron-brain-QBI.jpg

Action potentials are essentially electrical signals that facilitate communication between neurons,the primary cells of the nervous system. They are fundamental for most bodily processes like cognition, sensation, and motor control.

The process starts with a stimulus that causes the neuron's membrane to depolarize in some region, meaning it becomes less negative compared to the resting potential. If this depolarization reaches a certain threshold, an action potential is triggered. The voltage-gated sodium channels open, allowing sodium ions to rush into the neuron, causing further depolarization.

This rapid influx of sodium ions causes the inside of the neuron to become positively charged relative to the outside, which is the "upstroke" of the action potential. This process is referred to as depolarization.

Once the local membrane potential reaches around +30 mV, the sodium channels close and the voltage-gated potassium channels open, allowing potassium ions to flow out of the neurons. This returns the neuron to a more negative internal charge, a process known as repolarization.

Sometimes, the efflux of potassium ions can cause the neurons to become more negative than the resting potential, a state known as hyperpolarization. Eventually, the membrane potential in a local region returns to the resting state and an area close to it becomes depolarized. It is however worth noting that there is a wave propagating down the axon. The whole axon does not simultaneously change its state.

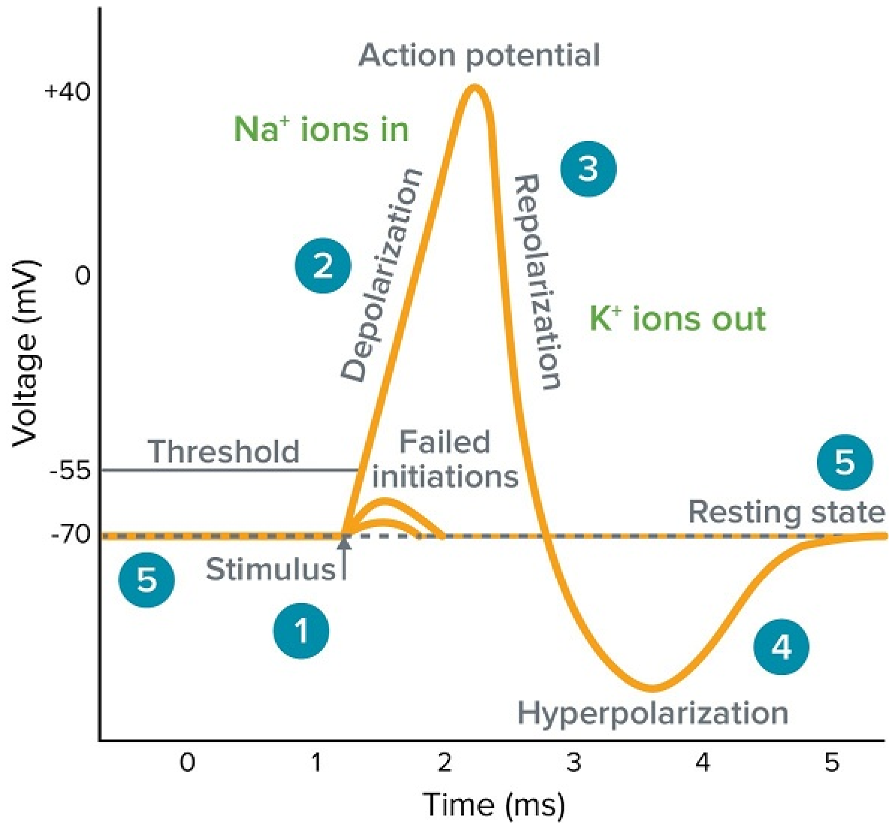

Figure 3: Voltage vs Time plot of an action potential in a neuron

https://www.moleculardevices.com/sites/default/files/images/page/what-is-action-potential.jpg

The entire process propagates along the length of the axon in a wave-like fashion, serving as a mechanism for transmitting information from the neuron's cell body to the synapse, where it can influence other neurons via neurotransmitter release. The “all-or-nothing” nature of the action potential ensures that the signal can travel long distances without losing strength.

The brain can be divided into white brain matter (myelinated axons) and gray brain matter (unmyelinated axons). Myelin, a type of fatty tissue, wraps around the axons of many neurons in a series of layers. This myelination serves a couple of important functions, including speeding up signal transmission and insulating the axon. Myelinated axons transmit action potentials much faster and more efficiently than unmyelinated axons due to a process called saltatory conduction, where the electrical impulse jumps from one node of Ranvier (exposed gaps between the myelin sheath), to the next. In contrast, the entire length of an unmyelinated axon needs to be depolarized during an action potential, thus requiring significantly more energy. The only upside to unmyelinated axons is that they take up less space which can be crucial in areas of the brain where space is extremely limited.

However, the function of white brain matter, as currently understood, serves to speed up transmission of signals between gray brain matter, where the bulk of computation and processing occurs. Since we are comparing the computational energy between the human brain and modern computers, we will be focusing on the gray matter. Estimation of the brain’s energy consumption will be done so through two methods, the peak voltage/current of an action potential, and by obtaining the capacitance of an unmyelinated axon, coupling it with the peak voltage of an action potential. A third method was attempted using ATP molecules. We attempted to search for data on the number of ATP molecules consumed per action potential by an unmyelinated axon. However, the current literature on this is extremely limited, with the availability of some data only on myelinated axons.

Peak Voltage/Current

The first method we will be using is by obtaining the peak change in voltage and current during an action potential. Due to the absence of real human data, we will be reviewing data from other mammalian species due to our similar genetic composition, in particular, cats and mice.

An average of 0.62nA of peak current was delivered to the soma (cell body) of a neuron during the visual stimulation of a cat's visual cortex. [3] Pairing that together with the average change in potential during an action potential of approximately 70mV.

$$P = IV = 0.62nA * 70mV = 43.4pW$$

This voltage/current spike lasts for approximately 1ms which gives us an approximate average energy consumption of

$$E = Pt = 43.4pW * 1ms = 43.4fJ$$

A patch clamp experiment on the axon initial segment (AIS) of the cortical layer 5 pyramidal neurons in Wistar rats (2-4 weeks old) showed a similar reading of 500 pA or 0.5nA. [4] The induced change in voltage here is about 60 mV within 1 ms. This gives an estimated energy consumption of

$$E = IVt = 0.5nA * 60mV * 1ms = 30fJ$$

Both experiments gave similar results for the estimated energy consumption for action potential in mammalian brain neurons, approximately within the 10^{-14} J range.

These results are admittedly somewhat inconclusive and debatable, due to the extremely complex nature of the brain and use of data collected from other mammalian species.

Capacitance

To determine the surface area of the axon that should be used in this calculation, let us first consider the length along the axon that the potential is elevated, during an action potential moving down the axon. The average velocity of an action potential in an unmyelinated axon is about 0.5 m/s to 10 m/s. [5]. Taking a conservative estimate of 1 m/s, coupled with the spike time of 1 ms, this gives us an approximate of about 1 mm for the axon's length. Another approach would be to look at the density of cortical neurons which is approximately $n = 4 * 10^4$ neurons/mm^3. [6]. The average distance between neurons is $1/n^{1/3}$ which is approximately 0.03 mm. On the other hand, there are of order 10^3 connections (many of which are on a single axon however). This might then give us an average distance 10 times the distance between neurons, which is 0.3 mm. In this case, we will be using 1 mm, which we consider to be the upper bound for the length of an axon, in our subsequent estimations.

The average capacitance per unit area of a cortical pyramidal neuron is $0.92 \pm 0.05\ \mu\mathrm{F/cm^2}$. [7] With an approximate diameter and length of 1 micron and 1 mm respectively, we can estimate the average capacitance of axons in a neuron to be 28.9 pF. The average diameter of the soma is about 21 μm. [8] This gives us a radius of 10.5 μm and an average somatic capacitance of 12.7 pF. The total average estimated capacitance of a neuron is therefore 41.6 pF. An experiment which analyzed the somatic capacitance of varying sizes of Wistar rats shows that this estimation is reasonable. [9] Coupled with the average peak action potential of approximately 70 mV, we can estimate the total energy of this action potential to be

$$E = \frac{1}{2}CV^2 \simeq 100fJ$$

assuming a length of 1 mm, and about 52.5 fJ, taking the length to be more typical of two connected neurons in the gray matter. The white matter is a completely different issue. Axons in that region are often many centimeters long. However fMRI studies show that most of the oxygen uptake, and therefore energy consumption is in the gray matter. Altogether, the capacitance calculation we described is surprisingly close to the voltage/current calculation described above, both being at the 10^{-13} to 10^{-14} J range, depending on the length and size of the axon and soma respectively.

Discussion

We have established that an action potential requires approximately 10^{-14} to 10^{-13} J of energy, Another approach to estimating this would be to take the total number of 86 billion [10] neurons in the brain and multiply that by the average 1000 synapses it has. Even though synapse potentials are smaller than action potentials, they are still within the same order of magnitude, 10^{1} mV and are therefore relevant for this estimation. From Figure 3, we can see that an action potential, coupled with its refractory period, takes approximately up to 5 ms. We can estimate the number of action potentials per second by finding the inverse of that duration. Running at a frequency of a few hundred hertz, and assuming that we only use about 10% of our brain, we arrive at approximately 10^{-14} J. This is within the bounds of reasonable and realistic parameters. However, since synapses are not actually responsible for processing bits of information, but rather for neurons to communicate with each other, we will not investigate this method further. Rather, this should support that our 10^{-14} J approximation is highly plausible.

Modern Computer

There are several measurements of computational speed in a modern central processor unit (CPU), mainly floating point operation per second (FLOPs) and clock speed. FLOPs refer to the speed at which the CPU can perform numerical calculation or floating point operations per second. It is most often used in scientific computing. Due to the nature of its specific mathematical application, we will avoid using this measure of speed. We will instead focus on clock speed.

Clock speed refers to the number of clock cycles a CPU undergoes in a second. To simplify it, a clock cycle in a CPU is one complete round of sequence of events in which a single instruction is processed by the CPU. During each clock cycle, the CPU can perform, or partially perform a certain number of operations, such as reading data from memory, performing a mathematical operation, or writing data back to memory which is in contrast to FLOPs. All of these operations are determined by bit flips in the billions of transistors encased within an average 16 mm x 16 mm modern CPU chip. A bit flip in a transistor corresponds to a change in its state, from 0 to 1 or vice versa, otherwise known as an operation on a bit. This change in state is fundamental to the operation of digital circuits, including the CPU in a computer.

The number of operations a CPU can perform per clock cycle depends on its architecture and design. Thus, while clock speed can be used as a gauge of computing performance, there are other factors such as architecture and number of cores to consider. While we can include the number of cores by simply multiplying it with the clock speed, as discussed above, we will not attempt to consider any measure that has to do with the CPU's architecture in our comparison, as we believe that this will give us a more direct comparison of the efficiency of the underlying hardware.

We will be using the following equation:

$$Energy/bit = \frac{P_{max}}{\#cycle/s \times \#transistors}$$

where P_{max} refers to the maximum wattage rating of the CPU, # cycles/s refers to the number of clock cycles the CPU undergoes every second and # transistors refers to the total number of transistors in a chosen CPU.

Advanced Micro Devices' (AMD), one of the world's leading CPU manufacturers, commercial grade chip AMD Ryzen 7 7800X3D released in 2023, has a maximum power rating of approximately 120 W with 6.57 billion transistors.

Physical	
Socket:	AMD Socket AM5
Foundry:	TSMC
Process Size:	5 nm
Transistors:	6,570 million
Die Size:	71 mm²
I/O Process Size:	6 nm
I/O Die Size:	122 mm²
Package:	FC-LGA1718
tCaseMax:	61°C
tJMax:	89°C

Processor	
Market:	Desktop
Production Status:	Active
Release Date:	Jan 4th, 2023
Retail Availability:	Apr 6th, 2023
Launch Price:	$449
Part#:	100-000000910
Bundled Cooler:	None

Performance	
Frequency:	4.2 GHz
Turbo Clock:	up to 5 GHz
Base Clock:	100 MHz
Multiplier:	42.0x
PBO Curve:	up to 50.5x
Multiplier Unlocked:	No
TDP:	120 W
PPT:	162 W

Figure 4: Specifications of the AMD Ryzen 7800X3D CPU

https://www.techpowerup.com/cpu-specs/ryzen-7-7800x3d.c3022

Assuming only 10% of the transistors changing their state under “maximum load”, an extremely conservative estimate, we will get an average output of $4.22*10^{-18}$ J or 4.22 aJ/bit, approximately 3 orders of magnitude away from the Landauer’s limit, a theoretical threshold for the minimum amount of energy required per bit of information. In essence, our current computing hardware is extremely advanced, nearing its theoretical limit. Any further improvements have diminishing returns and new types of computer, such as quantum computers are required for us to make any giant leaps.

Conclusion

As our estimations have shown, modern computers' hardware are several orders of magnitude more efficient than our human brain at computation. More specifically, computers use approximately 10^{-18} J per bit of computation versus our brain which requires about 10^{-14} J. Our brain is a technological marvel we cannot truly comprehend despite our other scientific accomplishments, including putting men on the moon. Several sources argue that the brain, being a mere 20 W "machine", computes tremendous amounts of information at once, including processing all 5 of our sensory organs. However, the efficiency is more likely due to our brain's architecture, programming and algorithm, which we have lumped together to call "software", to distinguish it from its underlying hardware, namely, neurons communicating via action and synaptic potentials.

The "hardware" of computers, including the copper wires and transistors are extremely efficient at transmitting information, beating out our axons and dendrites by several orders of magnitude. Seeing that computers are nearing its theoretical limit, unfortunately, any further engineering improvements bring minimal gains. As can be seen from major technological companies such as Google and Microsoft, efforts are already being focused on researching quantum computers, although the technology is still at its infancy with several limitations such as its physical size, a problem our modern computers faced in the early days too.

Seeing how efficient our brain is with overall computation leads us to believe that other factors, mainly the “software” of our brain, are extremely well adapted to minimal energy consumption and rightfully so through millions of years of evolution. However, even then, it is difficult to pinpoint which algorithm is superior since everything that computes, ranging from mobile devices, to humans and even insects are specialized. For instance, our modern computers are extremely fast at doing numerical operations, particularly floating point operations and the fastest human is orders of magnitude slower than a computer despite on average only consuming 10% the energy of a modern computer. In contrast, we have fantastic motor skills and the most advanced human-like robots are far from mirroring our physical movements, and as of 2023, we still have intellectual capabilities that the most advanced AI algorithms do not have, such as being able to come up with relativity and quantum mechanics or prove Fermat’s last theorem.

Nevertheless, having a deeper understanding of our brain mapping can aid us in developing more efficient computing architectures and programming for various specialized applications. I reviewed approximately 10 papers that I could find about energy cost of action potential. All of them had flaws their made their conclusions unreliable.

An example would be a paper we studied regarding the amount of ATP molecules converted per action potential. [11] This highly cited paper calculated the energy cost from ATP based on what appears to be a flawed analysis. In particular, below their equation 8, they state, without explanation, a novel variant of the Nernst equation, which instead of comparing potential differences between different points in space, involves differences in time. This is clearly incorrect. If one were to slowly increase the potential of the solvent holding the neurons, by an arbitrary amount, this would not affect charges or any of the physics on the neuron themselves. This tells us that the energy analysis of action potentials require more work and increases the significance of my work here.

While we are extremely far away from being able to mimic the brain for all potential computing technology, we should constantly strive to continue studying and learning more about the brain, while contributing genuine and well written scientific papers to the community so that we can make better and quicker progress in the field of neuroscience and possibly computing.

Bibliography

[1] R. Landauer, Irreversibility and Heat Generation in the Computing Process, in IBM Journal of Research and Development, vol. 5, no. 3, pp. 183-191, July 1961, doi: 10.1147/rd.53.0183.

[2] Luccioni, A.S., Viguier, S. and Ligozat, A.-L. (2022) Estimating the carbon footprint of bloom, a 176B parameter language model, arXiv.org. Available at: https://arxiv.org/abs/2211.02001 (Accessed: 26 May 2023).

[3] Ahmed, B. *et al.* (1998a) 'Estimates of the net excitatory currents evoked by visual stimulation of identified neurons in cat visual cortex', *Cerebral Cortex*, 8(5), pp. 462–476. doi:10.1093/cercor/8.5.462.

[4] Kole, M.H. *et al.* (2008) 'Action potential generation requires a high sodium channel density in the axon initial segment', *Nature Neuroscience*, 11(2), pp. 178–186. doi:10.1038/nn2040.

[5] Purves D, Augustine GJ, Fitzpatrick D, et al., editors. Neuroscience. 2nd edition. Sunderland (MA): Sinauer Associates; 2001. Increased Conduction Velocity as a Result of Myelination. Available from: https://www.ncbi.nlm.nih.gov/books/NBK10921/

[6] Leuba, G. and Garey, L.J. (1989) 'Comparison of neuronal and glial numerical density in primary and secondary visual cortex of man', *Experimental Brain Research*, 77(1). doi:10.1007/bf00250564.

[7] Gentet, L.J., Stuart, G.J. and Clements, J.D. (2000) 'Direct measurement of specific membrane capacitance in neurons', *Biophysical Journal*, 79(1), pp. 314–320. doi:10.1016/s0006-3495(00)76293-x.

[8] Yan, C. *et al.* (2013) 'Automated and accurate detection of Soma location and surface morphology in large-scale 3D neuron images', *PLoS ONE*, 8(4). doi:10.1371/journal.pone.0062579.

[9] Limón, A. *et al.* (2005) 'Ca2+-activated K+-current density is correlated with soma size in rat vestibular-afferent neurons in culture', *Journal of Neurophysiology*, 94(6), pp. 3751–3761. doi:10.1152/jn.00177.2005.

[10] Azevedo, F.A.C. *et al.* (2009) 'Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain', *The Journal of Comparative Neurology*, 513(5), pp. 532–541. doi:10.1002/cne.21974.

[11] Attwell D, Laughlin SB. An energy budget for signaling in the grey matter of the brain. J Cereb Blood Flow Metab. 2001 Oct;21(10):1133-45. doi: 10.1097/00004647-200110000-00001. PMID: 11598490.