The Answer to the End of Moore's Law:

Neuromorphic Architecture

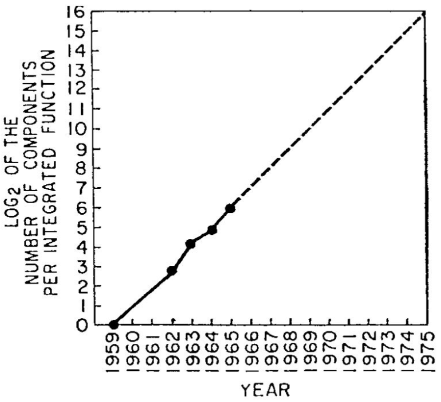
Ryan Darras

University of Colorado at Colorado Springs

Introduction

Co-founder of Intel, Gordon Moore made a prediction in 1965 that integrated circuits, or chips were the path to cheaper electronics. Moore's law states that the number of transistors, the tiny switches that control the flow of an electrical current that can fit in an integrated circuit, will double every two years, while the cost will halve (Moore, 1965). This exponential growth has brought massive advances in computational power, hence why the tiny computers in our pockets that are as powerful as the computers on the *Eagle*, the lunar module which landed on the moon in 1969. Chips today can contain billions of transistors that are about 14 nm wide. Experts are claiming that the trend is slowing down and Intel recently disclosed that it's becoming more difficult to roll out smaller transistors in a two year timeframe while also being affordable (ExtremeTech.com, 2017). To power the next wave of electronics, there are a few promising options in the works; one of which is quantum computing, another is neuromorphic computing. Computer chips that are built using neuromorphic architecture are modeled after our own brains. They're capable of learning and remembering at the same time at an incredibly fast rate. The human brain has billions of neurons, each of which forms synapses which are connections to other neurons. Synaptic activity relies on ion channels, which control the flow of charged atoms like sodium and calcium that make your brain function and process properly. A neuromorphic chip copies that model by relying on a densely connected web of transistors that mimic the activity of ion channels. Each chip has a network of cores, with inputs and outputs that are wired to additional cores, which all operate in conjunction with each other. This connectivity allows neuromorphic chips to be able to integrate memory, computation, and communication all together. These chips are an entirely new computational design.

What is Moore's Law and where is it going?



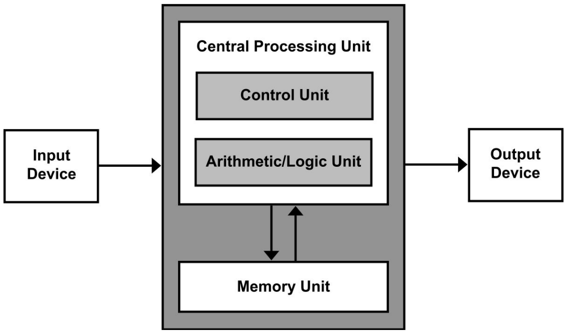
In 1965, co-founder of Fairchild Semiconductor and Intel Gordon E. Moore authored a paper describing how the number of components per integrated circuit would double every year (Moore, 1965). As a result, chip performance would increase due to the increased amount of transistors and the transistors being faster. However, Moore knew that this rate wouldn't last forever, and he revised his forecast in 1975 to state that the number of components per integrated circuit would double every two years. In 2015, Moore stated, "I see Moore's law dying here in the next decade or so" (IEEE Interview with Gordon Moore, 2015). In 2017, Intel released information on the company's next generation processor to be unveiled in 2018 and they stated that they expect to reach the 10 nm node which would result in a three-year cadence, showing that the rate of progress is approaching saturation (ExtremeTech.com, 2017). Due to fundamental laws of physics and thermodynamics most semiconductor industry forecasters, including Gordon Moore, expect Moore's law will end by around 2025 (Kumar, 2015). Moore's law was coined by Caltech professor Carver Mead in 1975.

What is Neuromorphic Architecture?

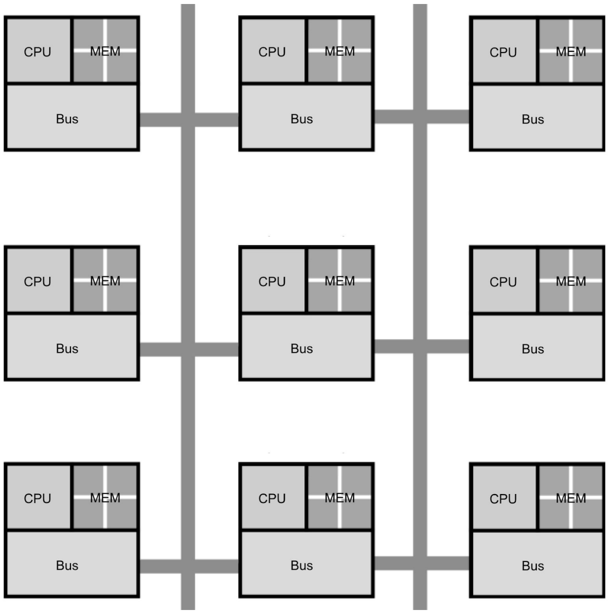
From the same man that coined Moore's Law comes the concept of neuromorphic architecture. In the late 1980's, Carver Mead described the use of very large scale integration (VLSI) systems containing electronic analog circuits to mimic neuro-biological architectures present in the nervous system (Mead, 1990). However, recently the term has been used to describe analog, digital, mixed-mode analog/digital VLSI, and software systems that implement models of neural systems such as perception, motor control, and multisensory integration. At the hardware level a chip designed with neuromorphic architecture is comprised of the same components as a chip designed with von Neumann architecture, but it also includes memristors. A memristors electrical resistance is not constant; they keep a history of the current that had previously flowed through them. More specifically, a memristors current resistance depends on how much current has flowed in which direction through it in the past (Chua, 2010). Memristors are non-volatile, which means they can retain information without power. Imagine a pipe that carries water. When the water is flowing in one direction, the pipe’s diameter expands and allows the water to flow quickly, but when the water starts flowing in the opposite direction, the pipe’s diameter contracts and slows the water down. In the case that the source controlling the flow of water is shut down, the pipe retains its diameter until the water is turned back on. Because of this, memristors also have the feature that if a computers plug was pulled, the computer would remember everything that was in memory at the time the CPU stopped receiving power. By using memristors, chips designed with neuromorphic architecture can be developed in a way that simulates the neurons, axons, synapses, and dendrites which are the components that our brains are comprised of, hence allowing the chip to remember behaviours and patterns that modern day chips simply cannot do efficiently.

How does neuromorphic architecture solve the computing problem?

The architecture behind modern computers was proposed by John von Neumann. In the von Neumann architecture, programs and data are held in memory while the processor takes that information to calculate results. Latency is unavoidable in this architecture because the processor and memory are separate and data moves between the two. The image below is a depiction of a chip designed using the von Neumann architecture.



The memory unit is connected directly to the CPU by bus. In computer architecture, a bus is a communication system that transfers data between components inside a computer, or between computers. Although buses are fairly good at what they do, they come with an inherent latency whenever you want to send or receive data through it which is a limitation on throughput. Neuromorphic architecture solves this issue by integrating a processor and memory on a single core. The inputs and outputs of these cores are connected to other cores which allows them to operate in conjunction with one another. This connectivity allows neuromorphic chips to be able to integrate memory, computation, and communication all together which bypasses many of the bottlenecks inherited from a von Neumann system. Another benefit of using neuromorphic architecture over von Neumann



architecture is that by placing a processor and memory on a single core, you reduce the amount of heat dissipating from the chip by an enormous amount. As a CPU runs, it allows electrical current to pass through which causes it to heat up. IBM’s TrueNorth neuromorphic chip consumes less than 100 milliwatts while simulating complex recurrent neural networks and is capable of 46 billion synaptic operations per second, per watt. Compared to the Intel Core i7-7700k which consumes more than 100 watts on full load (Dharmendra, 2014).

Drawbacks

Our brains are fantastic organic entities capable of doing amazing things in an instant, so of course you would think that modeling a computer chip based on the human brain would lead to a chip with no limits. Unfortunately, researchers have discovered that while neuromorphic architecture will be a huge advancement in technology, it comes with a huge fault that is present in our own brains. Have you ever had a moment where you remember something, but then second guess yourself? Or perhaps you claim something that you have seen as fact, only to find out later that you were incorrect? On modern computers, any turing recognizable problem can be solved, but this is because the computer will algorithmically brute force every answer if it needs to. Contrast that vs a neuromorphic chip that uses its memory to solve problems it remembers, similar to our brain; it has the possibility of being incorrect. An extreme example of this would be that while you are playing a video game, you start rendering textures on objects that aren’t correct, but the neuromorphic architecture remembers seeing that specific texture on that specific object. Due to this nature, it is very unlikely that neuromorphic chips will simply replace von Neumann chips. We are more likely to see CPU’s shipped with a neuromorphic unit on the side, or neuromorphic chips as a 3rd party “farm” for the CPU to use similar to how modern CPU’s interact with GPU’s. Regardless to this drawback, putting the power of a neuromorphic chip into the hands of the consumers will greatly increase computational efficiency of home machines.

Alternatives

It is 2018, and almost every person has heard of quantum computing and the stories about how it will end the world as we know it. This is, however, not the case. Quantum computing is an up-and-coming technology that will be magnitudes faster than anything currently available today for a set of specific tasks. Unfortunately, quantum computers don’t come without their faults. Quantum computers will likely never replace your home PC or your work laptop because it doesn’t do so well at some of the more primitive tasks that current architecture does incredibly well. We are more likely to see quantum data farms in which you can buy computation power from for incredibly difficult problems than a quantum unit built nearby and working in tangent with a modern household CPU. Quantum computers also rely on a cooling system that would get your chip to as close to absolute zero as possible because any extra energy in the system can produce extraneous results, which isn’t exactly feasible in an average household.

Related Work

Neuromorphic architectures are still fairly new, yet we are starting to see companies take a shot at developing neuromorphic chips. Companies like IBM, Qualcomm, Hewlett Packard, International Business Machine Corporation, Samsung Electronics, and Brain Corporation are taking a stab at creating a neuromorphic architecture. IBM’s TrueNorth is a neuromorphic chip with 4096 cores, with each core simulating 256 programmable silicon neurons, totalling at just over one million neurons. Following, each neuron has 256 programmable synapses that convey signals between the neurons. Hence, the total number of programmable synapses is just over 268 million (2^{28}). This very efficient chip bypasses the bottlenecks of the von Neumann architecture and has a power density that is 1/10,000th of conventional microprocessors (Hsu, 2014). Every single core in TrueNorth includes the entire computing package: memory, computation, and communication, which helps bypass a bottleneck in traditional von Neumann architecture where program instructions and operation data cannot pass through the same route simultaneously. Dharmendra Modha, IBM's chief scientist for brain-inspired computing at IBM in San Jose, California claims, "This is literally a supercomputer the size of a postage stamp, light like a feather, and lower power like a hearing aid" (Hsu, 2014).

Conclusion

Standard chips to day are built based on von Neumann architecture where the processor and memory are separated and the data moves between them through bus. A CPU runs commands that are fetched from memory to execute tasks. This is what makes computers really good at computing but not as efficiently as they could be. Neuromorphic chips completely change that model by having both storage and processing connected within neurons that are all communicating and learning together. The hope is that these neuromorphic chips could transform computers from general purpose calculators into machines that can learn from experience and make decisions. We'd leap into a future where computers wouldn't just be able to crunch data at high speeds but could process sensory data in real time.

References

Moore, G. (1965) Retrieved from
http://www.monolithic3d.com/uploads/6/0/5/5/6055488/gordon_moore_1965_article.pdf

IEEE Interview with Gordon Moore (2015) Retrieved from
https://spectrum.ieee.org/computing/hardware/gordon-moore-the-man-whose-name-means-progress

ExtremeTech.com (2017) Retrieved from
https://www.extremetech.com/computing/254209-details-leak-intels-upcoming-ice-lake-cpu-10nm-schedule

Kumar, S (2015) Retrieved from
https://arxiv.org/ftp/arxiv/papers/1511/1511.05956.pdf

Mead, C (1990) Retrieved from
https://ieeexplore.ieee.org/abstract/document/58356/

Chua, L (2010) Retrieved from
https://link.springer.com/content/pdf/10.1007%2Fs00339-011-6264-9.pdf

Dharmendra, M (2014) Retrieved from
http://www.research.ibm.com/articles/brain-chip.shtml

Hsu, J (2014) Retrieved from
https://spectrum.ieee.org/computing/hardware/how-ibm-got-brainlike-efficiency-from-the-truenorth-chip