Neuromorphic Computing

SYNOPSIS

Submitted to:

Department of Electronics and Communication Engineering

Submitted by:

Name — Vedika Jain
Roll Number — 21BEC018

Under the Supervision of:
Prof. Sajad Ahmad Loan

Faculty of Engineering and Technology

Jamia Millia Islamia, New Delhi

List of Contents

Motivation

The trusty von Neumann architecture has given us systems which have allowed us to take giant leaps in technology. But since change is the only constant, it was about time that our present computational set-up started showing some cracks. The von Neumann bottleneck, caused due to shuttling continuously between processor and memory, stands as a major impediment to speed optimisation. It was believed that doubling the number of transistors every two years would help us keep up with our speed requirements.

But, it turns out that Moore's Law is nearing its end. In a bid to pack more transistors within the same area, chip architectures are becoming increasingly complex. This has led to a rise in foundry costs to such a point that it has now started influencing consumer pricing. Even this trend can't continue for long, as chip designers are very close to hitting the physical limits for transistor miniaturisation. To keep up with the law, chip size would have to go up, leading to bigger and bulkier devices. Power budgeting and cooling for these big chips would become difficult, leading to more power consuming devices.

Alternate technologies apart from our good old MOSFET are being looked at. Transistors like FinFET and CNTFET are gaining much fandom, but realisation of such intricate devices on a large scale would require economical synthesis techniques, which are still being looked for. Silicon is also being sought to be replaced by novel materials, like germanium alloys, InGaAs, and graphene, which often are already scarce, difficult to source/synthesise in operable quality, or need extensive study of properties to be considered viable for use.

GPUs are indispensable for computation heavy and high speed applications, like generative AI and gaming. But they can't be made the norm for daily computation needs. Their prohibitive pricing puts them out of reach for a majority of their customers. As the demand goes up and the supply fails to scale proportionately, the situation won't get any better. Moreover, the dead end encountered by general purpose processor manufacturers wouldn't evade the GPU architects for long. Accelerators await the same fate.

Meanwhile, our processing requirements continue to grow. As AI assimilates itself into our lives, thanks to cloud hosted applications, innovators wonder if it can penetrate deeper, to the lower strata of the society. As concerns over data security and privacy grow, and Internet blackouts are weaponised, a need is felt to disconnect from the cloud, and run these applications natively. With the gigantic power-hungry models we've built, this seems like a joke.

The inefficacy of these hacks suggests that AI is best run on hardware which is similar to the abstraction. The neurons in the brain do not discriminate between computation and memory. The connections between neurons keep evolving, lending us the ability to learn and unlearn. We don't think in binary, nature is analog. Only the active parts of the brain consume energy. The present models are very power intensive and inefficient. If we are to harness the power of AI, we need hardware which is high on speed, low on power consumption, and nimble with computations. The present von Neumann architecture has hit these roadblocks and can't support AI models for long.

Introduction

As the von Neumann system collapses, new ideas make their way into the processor discourse. One such idea is neuromorphic computing. Neuromorphic computing, a.k.a. brain-inspired computing, seeks to structure hardware after the most efficient processor known to us, the biological brain.

Its most utilised form, spiking neural network (SNN), is based on a unit inspired from the biological neuron. This “neuron” stores the data it needs to perform its computations. Thus, there is no need for separate memory and processing facilities, saving on time, power, and complexity. Two neurons are connected by a non-permanent “synapse”. A synapse is essentially an electrical signal. When the voltage level of the signal exceeds a threshold, it is treated as a synapse. The form and timing of this signal contains information, which is encoded and decoded by the neurons themselves, without using binary.

Many of these neurons are placed on a single chip, and only the communicating, or “spiking”, neurons consume power. Since dormant neurons become active only when sent a spike, there's no need for a clock to coordinate different instructions. Both of these features lead to enormous power saving. As the computations are performed, a neuron sends signals to the neurons in its vicinity. Stronger signals are sent to the neurons it needs to work with, thus creating synapses. As the signals sent out change, different neurons get connected and disconnected, leading to evolving synapses, lending the SNN enormous plasticity. Over time, connections which give the best results are adopted as defaults.

Since the inherent structure is so simple, neuromorphic systems are massively scalable. We can use as many neurons as we need, even connect two or more chips for higher processing capability. The neurons present on the same chip are independent. Their working does not affect others; they come together on a task only when a synapse is shared. This allows us to run multiple functions on the same chip at the same time, theoretically as many as there are neurons. This leaves no doubt about their parallel computing prowess. A neuron being a jack of all trades is capable of performing any instruction, and a piece of information is held by all the neurons that need it. Processing functions and memory locations are not under the confinement of specific transistors, which ensures that no loss of data or functionality is incurred in event of node failure. So, they’re highly fault tolerant.

Their power economy, absence of speed throttles, and adaptability allow us to deploy them for native AI applications which train on the job. Building these applications on top of supportive hardware would be more efficient. Robot learning can be improved, enabling us to design more sensitive and generalist robots. The computation capability of these systems makes them a good fit for use cases other than AI too, like stock markets, social media mapping, and epidemiologic studies. These are examples of Monte Carlo methods, which involve multiple random walks through a problem to find the optimum solution. This approach has a wide range of applications, and neuromorphic systems can help us solve these mammoth problems much more swiftly. They are also of utility in the discipline of neurology, by allowing us to simulate connectomes and understand how the brain works.

Present State

Neuromorphic computing is still a novel field, with constant research underway and innovations being broken every day. A lot of our research is focussed on understanding the biological brain, which influences our system design too. As our understanding of the brain widens, the architecture is also modified accordingly. Academia mainly looks for new opportunities in this domain, whereas industry giants are working on powerful neuromorphic chips. Two forerunners are IBM and Intel.

IBM, credited to many firsts of technology, built the TrueNorth chip in 2014. Its successor, NorthPole, was released last year. NorthPole is 4000 times faster than its predecessor, and can run tasks like ResNet-50 or Yolo-v4 image recognition about 22 times faster, with 25 times less energy and 5 times less space, when compared to GPUs (like Nvidia's V100) which use the same 12-nm node process that it was fabricated with. It even outperformed chips made with more advanced nodes. For example, compared with Nvidia's H100 GPU, implemented using a 4-nm node, NorthPole was five times more energy efficient. It includes 224 MB of RAM and 256 processor cores and can perform 2,048 operations per core per cycle at 8-bit precision, and 8,192 operations at 2-bit precision. It runs at between 25 and 425 MHz. It can be integrated into systems too as it appears as an active memory chip from outside the device.

Intel made the Loihi chip in 2017, followed by Loihi 2 in 2021. Advances in Loihi 2 provide up to 10 times faster processing, up to 15 times greater resource density with up to 1 million neurons per chip, and improved energy efficiency. Early evaluations suggest reductions of over 60 times fewer ops per inference compared to standard deep networks running on the original Loihi without loss in accuracy. Fully programmable neuron models and generalized spike messaging open the door to a wide range of new neural network models that can be trained in deep learning. It improves support for advanced learning methods, including variations of backpropagation. Greater programmability also allows a wider class of difficult optimization problems to be supported. This expands the scope of adaptation and data efficient learning algorithms that can be supported by low-power form factors operating in online settings. Loihi 2 addresses a practical limitation of its predecessor by incorporating faster, more flexible, and more standard input/output interfaces.

The Human Brain Project (HBP), funded by the European Union, boasts of two flagship supercomputers: SpiNNaker and BrainScaleS. SpiNNaker (Spiking Neural Network Architecture) is a massively parallel, many-core supercomputer architecture designed by the Advanced Processor Technologies Research Group at the Department of Computer Science, University of Manchester.

The Tianjic chip seeks to unify the divergent neuroscience and computational approaches towards the development of artificial general intelligence. It adopts a many-core architecture, reconfigurable building blocks and a streamlined dataflow with hybrid coding schemes, and can not only accommodate computer-science-based machine-learning algorithms, but also easily implement brain-inspired circuits and several coding schemes.

Opportunities

Some problems which warrant attention include:

- Absence of benchmarks to test performance
- No supporting software
- Lack of accuracy in computations
- Accessibility

This list is non-exhaustive. As further study leads to awareness regarding more intricacies and enlightenment about potential solutions, one of these concerns would be sought to be worked upon.

References

https://en.wikipedia.org/wiki/Von_Neumann_architecture

https://www.investopedia.com/terms/m/mooreslaw.asp

https://en.wikipedia.org/wiki/Moore%27s_law

https://www.allaboutcircuits.com/news/could-indium-gallium-arsenide-dethrone-silicon-race-smaller-transistors/

https://iopscience.iop.org/article/10.1088/1742-6596/423/1/012047

https://www.easytechjunkie.com/what-is-a-germanium-transistor.htm

https://en.wikipedia.org/wiki/Huang%27s_law

https://spectrum.ieee.org/the-accelerator-wall-a-new-problem-for-a-post-moores-law-world

https://www.techtarget.com/searchenterpriseai/definition/neuromorphic-computing

https://www.humanbrainproject.eu/en/science-development/focus-areas/neuromorphic-computing/

https://www.nature.com/articles/s43588-021-00184-y

https://en.wikipedia.org/wiki/Spiking_neural_network

https://spectrum.ieee.org/neuromorphic-computing-superconducting-synapse

https://spectrum.ieee.org/neuromorphic-computing-more-than-ai

https://spectrum.ieee.org/connectome-neuromorphic-chips

https://en.wikipedia.org/wiki/Cognitive_computer

https://spectrum.ieee.org/neuromorphic-computing-ibm-northpole

https://spectrum.ieee.org/neuromorphic-computing-with-lohi2

https://www.intel.com/content/www/us/en/newsroom/news/intel-unveils-neuromorphic-loihi-2-lava-software.html#gs.54ze8f

https://www.nature.com/articles/s41586-019-1424-8