Learning How to Learn:

Neuromorphic AI Inference at the Edge

Q&A with Peter van der Made, BrainChip Founder and Chief Technology Officer

brainchip
Essential AI

Contents

Introduction

In 1990, Carver Mead observed that neuromorphic compute architectures are often many orders of magnitude more effective than conventional systems. Indeed, the semiconductor industry has long struggled to bypass Von Neumann bottlenecks, recalibrate Moore's Law, and overcome the breakdown of Dennard Scaling.

The impact and acceleration of these trends recently prompted Gartner analysts to warn traditional computing technologies will hit a "digital wall" in 2025 and force a shift to new paradigms such as neuromorphic computing. To be sure, advanced edge AI applications are fast approaching the limits of conventional silicon and cloud-centric learning models. With enormous amounts of targeted compute power available in cloud data centers, AI training and inference models leveraging GPU and TPU hardware accelerators continue to increase in both size and sophistication.

This reflects a larger industry trend which has seen compute power increase over the past decade as networks grow larger and more complex. In parallel, cloud-based streaming video AI solutions are demanding ever-more internet bandwidth. Clearly, these trends cannot continue without severe consequences including unmanageable latency, rapidly expanding carbon footprints, and security exploits that could potentially intercept and target raw data sent to cloud data centers.

In this Q&A with BrainChip Co-Founder and Chief Technology Officer Peter van der Made, we discuss the evolution of neuromorphic computing, talk about the limitations of current compute models for edge AI, and explore how neuromorphic silicon is driving a more intelligent and sustainable future.

Chapter 1:
Overcoming analog limitations with digital engineering

Q: Peter, as a silicon pioneer, you've accomplished an incredible amount over the years, such as inventing a computer immune system, developing advanced graphics technology, and designing neuromorphic processors. You've also written a book titled "Higher Intelligence: How to Create a Functional Artificial Brain," which examines how the human brain functions – not as a control center – but as a learning, interacting machine.

With that in mind, can you tell us what inspired you to pick up where Carver Mead left off, move away from a Von Neumann compute model, and build an exciting neuromorphic future?

Peter: I was troubled by the huge amount of circuitry in CPUs and GPUs that is required for AI processing. All these power-hungry processors are doing very little compared to what the human brain accomplishes with 86 billion neural cells and over 100 trillion synaptic connections. The brain is faster than a giant supercomputer and runs on the energy equivalent of just 20 watts.

However, before designing a viable neuromorphic processor, I had to overcome some basic limitations. This is because Carver Mead's initial research was based on analog neurons to mimic human neurons. Like their biological counterparts, analog neurons only function properly within a very limited and stable temperature range. As well, data stored on analog neurons cannot be easily copied or used by other chips.

That's why I started by building a very comprehensible model of the behavior of a brain cell using digital circuits. I patented that design in 2008. Why use software on a computer to simulate a processor, a neuron cell, which is totally different from a computer? It makes more sense to build a computer that works like the processors in the brain. Hence the neuromorphic – which means "like the brain" – design. A digital model is far more stable than an analog circuit. We've seen the evolution from analog to digital in other fields, like mobile phones.

Although a giant leap forward, I also needed to efficiently scale neuron count to develop a commercially viable solution. Early FPGA-based prototypes first contained seven digital neurons, then 64, and the next iteration increased this number to 256. To get over a million neurons on a single chip, I had to simplify the neuron model without compromising its computational power. So, I eliminated extraneous biological-inspired elements such as neurotransmitters and exponential decays while preserving the neuromorphic functions that are essential in its computational function.

Chapter 2:
Designing the world's first commercial neuromorphic processor

Q: Adopting a biologically inspired, digitally engineered neuromorphic model was visionary, because it ultimately enabled you and the BrainChip team to effectively design, prototype, and launch AKIDA. Since you started your experimental circuits and design well before others in the semiconductor industry, can you highlight some of the major differences between BrainChip's AKIDA, IBM's TrueNorth, and Intel's Loihi?

Peter: We designed a learning function into our chip right from the start because learning is an essential function in AI. We also implemented on-chip convolution, a method of using the same neuron multiple times in different locations. We designed the AKIDA chip to be easy to configure so that engineers, not neuroscientists, can deploy it. The on-chip learning function and on-chip event-based convolution make AKIDA a very compact, low-power solution for edge AI use cases.

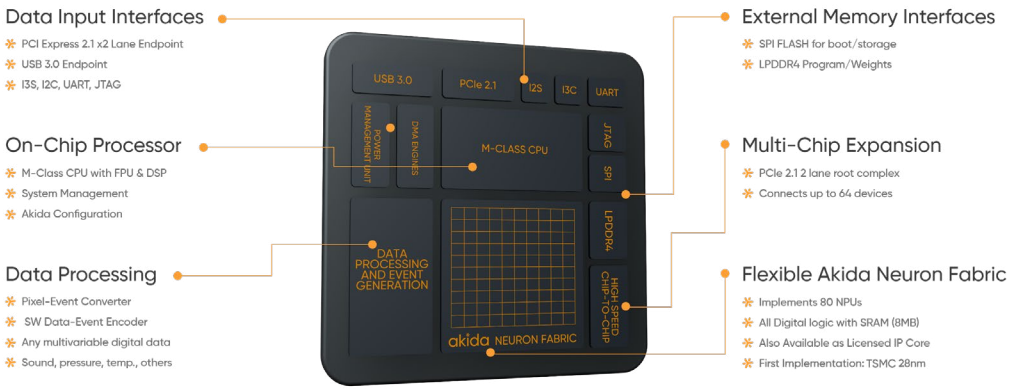
IBM's TrueNorth wasn't designed to support on-chip convolution, or on-chip learning. It is a static design consisting of 'corelets' that are programmed using a Fortran-like language to connect corelets together to create functions. In contrast, AKIDA neurons are arranged in layers and configured with MetaTF which supports Python and TensorFlow.

Although system designers can program their own learning methods, Intel's Loihi processor doesn't have the same on-chip learning and convolutional framework as AKIDA. In addition, Loihi's architecture seems to be targeted at neuroscientists rather than engineers, as the former are generally more familiar with the complex process of connecting individual neurons together (rather than layers).

I'd like to briefly elaborate on how BrainChip's MetaTF is making it easy for engineers to create, train, and test neural networks on AKIDA. Firstly, we designed MetaTF to automatically convert TensorFlow models and leverage Python, along with associated tools and libraries including Jupyter notebooks and NumPy. Secondly, with MetaTF, the data exchanged between layers is not the usual dense multidimensional arrays, but rather sets of spatially organized events modeled as sparse multidimensional arrays.

Q: The benefits of on-chip learning and on-chip convolution were recently highlighted by two olfactory (odor classification) applications. One was developed for AKIDA, and the other for Loihi. The application developed on Loihi required 80 chips, while BrainChip's olfactory application used a fraction of one chip, identified odors with high accuracy (97%), and consumed significantly less power. Although both AKIDA and Loihi are neuromorphic processors, there are major architectural differences between them. Can you highlight these differences?

Peter: By only processing meaningful data that represents relevant data, AKIDA efficiently performs up to trillions of operations per second within a minimal power envelope.



Take, for example, a security camera pointed at a warehouse scene. Most of the time nothing is happening, the scene is static. Conventional AI computes every pixel all the time, just to see that nothing has changed. Event-based neuromorphic processing is far more efficient. The scene is static and does not generate any events. Events are generated once someone walks into the scene. At that point, AKIDA processes only the relevant events to detect whether it is a person or an insect crawling across the lens.

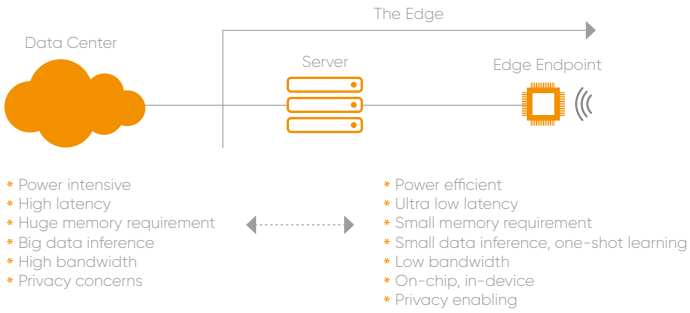
Another major architectural difference between AKIDA and other neuromorphic processors is scalability and configurability. We work closely with customers to achieve the most cost-effective solutions by optimizing node configuration to balance performance and efficiency. That's why AKIDA scales down to two nodes for ultra-low power applications - and scales up to 256 nodes for complex use cases. Every node consists of four neural processing Units (NPUs), each with scalable and configurable SRAM. Within each node, the NPUs can be configured as either convolutional or fully connected.

BrainChip's IP fabric can also be placed either in a parallelized manner that would be ideal for ultimate performance, or space-optimized to reduce silicon utilization and further reduce power consumption. This means entire neural networks can be placed into the fabric, removing the need to swap weights in and out of DRAM. This model significantly reduces power consumption while increasing throughput. Additionally, users can modify clock frequency to further optimize performance and power consumption.

Chapter 3:
Untethering edge AI from cloud data centers

Q: With AKIDA's neuromorphic architecture, BrainChip is enabling the semiconductor industry to untether edge AI from cloud data centers. This is quite timely, because conventional AI silicon and cloud-centric inference models aren't performing efficiently at the edge, even as the number of edge-enabled IoT devices are expected to hit 7.8 billion by 2030. Can you elaborate on the notion of untethering?

Peter: Increasing internet congestion is increasing latency as more edge devices upload their data. The power consumption and heat production of massive parallel Von Neumann type processors is also increasing linearly with the computing power required by AI applications. That's why untethering edge AI from the cloud with AKIDA is a critical step to designing faster and more environmentally sustainable endpoints.



Differentiating intelligent endpoint requirements

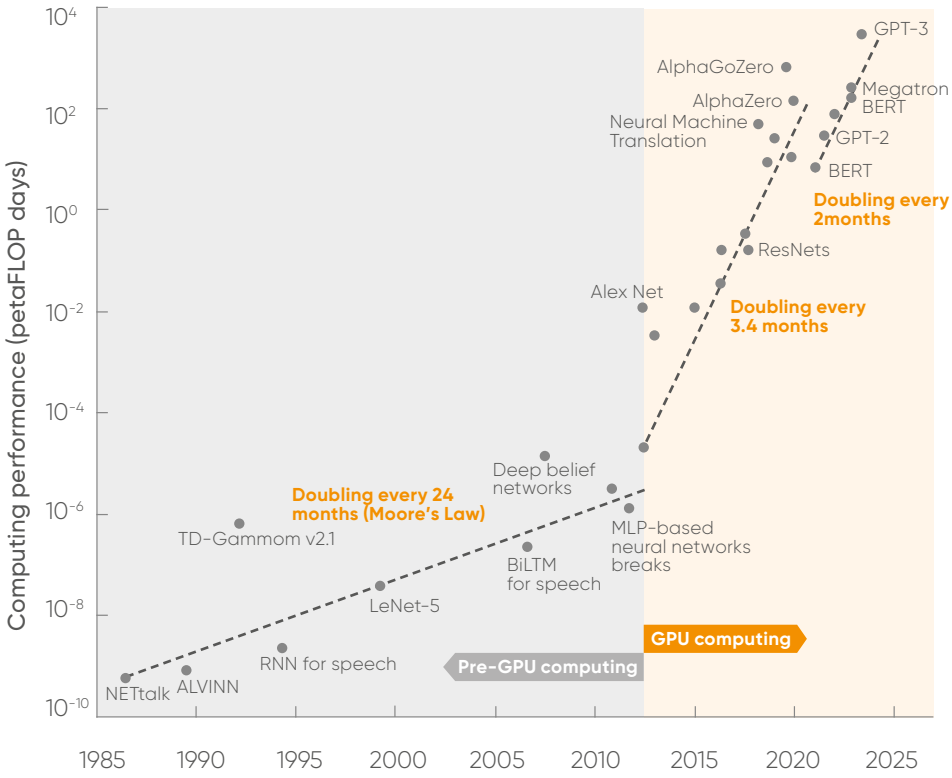
Data centers hosting cloud-based workloads emitted an estimated 600 megatons of greenhouse gases in 2020 alone, more than the consumption of the entire United Kingdom (GB). Unless something radically changes, data centers will consume over 20% of the world's energy by 2050! With its on-chip learning and low power, high throughput inference capabilities, we believe AKIDA can help reduce data center carbon emissions by 98% by decentralizing AI processing. Intelligently analyzing data on-chip will help put an end to the yottabytes of raw, unprocessed, and mostly irrelevant data sent to cloud data centers by millions of endpoints, solving the impeding internet congestion problem.

Using image recognition as an example, we can quantify the power savings enabled by AKIDA's on-chip capabilities compared to a GPU in today's data center. Specifically, AKIDA can efficiently analyze and categorize the 1.2 million images of the ImageNet dataset with a minimal power budget of 300 milliwatts. A GPU performing this task consumes up to 300 watts! This huge difference illustrates why simply scaling down conventional AI hardware to meet the unique requirements of edge endpoints is insufficient.

Other “low power” edge AI solutions such as DSP-based chips cannot match the performance of the AKIDA chip - and the scaled-down versions of GPU technology suffer from significantly higher power consumption. The AKIDA design is very flexible because the neural fabric can be reconfigured within milliseconds for many different network architectures.

Q: Gartner’s Bob Gill recently described a heterogenous approach to the edge. Dubbed “edge-in,” this model would potentially see companies build edge applications optimized for low-latency and low-bandwidth autonomous connections - and tap the cloud for other tasks. Do you agree with this assessment?

The exponential increase in computing power demands



Computational demands are increasing rapidly (Nature Portfolio, Brain-inspired computing needs a master plan)

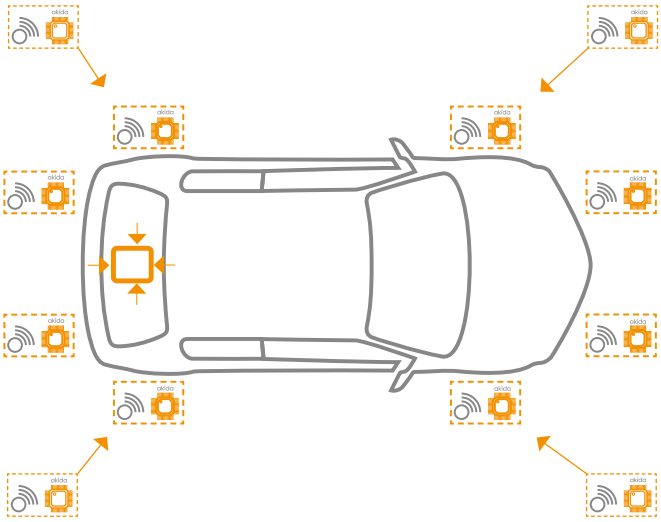
Peter: I believe data centers will take on more of a cloud-based warehousing role, storing massive amounts of information that can be leveraged over time for different purposes. Distributed processing in edge AI devices using AKIDA will only need to upload meta-data, not complete video streams. Data can be analyzed on edge devices for trends before the results are shared or uploaded. This will reduce the pressure to build ever larger, more power-hungry data centers in the future.

What we are seeing is a repeat of what happened in the 1980s when IBM launched its first PC. People bought computers that processed data locally and not on a remote mainframe computer. Distributed edge AI - not data centers - will drive a new, efficient, and faster model of distributed computing optimized to meet the requirements of intelligent endpoints.

Chapter 4:
Revolutionizing AI inference at the edge

Q: BrainChip customers are already deploying smarter endpoints with independent learning and inference capabilities, faster response times, and a lower power budget. Can you give us some real-world examples of how AKIDA is revolutionizing AI inference at the edge?

Peter: Automotive is a good place to start. Using AKIDA-powered sensors and accelerators, automotive companies can now design lighter, faster, and more energy efficient in-cabin systems that enable advanced driver verification and customization, sophisticated voice control technology, and next-level gaze estimation and emotion classification capabilities.



Enabling advanced LiDAR with AKIDA

In addition to redefining the automotive in-cabin experience, AKIDA is helping enable new computer vision and LiDAR systems to detect vehicles, pedestrians, bicyclists, street signs, and objects with incredibly high levels of precision. We're looking forward to seeing how these fast and energy efficient ADAS systems help automotive companies accelerate the rollout of increasingly advanced assisted driving capabilities.

In the future, we'd also like to power self-driving cars and trucks. But we don't want to program these vehicles. To achieve true autonomy, cars and trucks must independently learn how to drive in different environmental and topographical conditions such as icy mountain roads with limited visibility, crowded streets, and fast-moving highways.

With AKIDA, these driving skills can be easily copied and adaptable by millions of self-driving vehicles. This is a particularly important point. AKIDA driving updates will be based on real-world knowledge and skills other cars have learned on the road. We want to avoid boilerplate firmware updates pushed out by engineering teams sitting in cubes. A programmed car lacking advanced learning and inferential capabilities can't anticipate, understand, or react to new driving scenarios. That's why BrainChip's AKIDA focuses on efficiently learning, inferring, and adapting new skills.

Q: Aside from automotive, what are some other multimodal edge use cases AKIDA enables?

Peter: Smart homes, automated factories and warehouses, vibration monitoring and analysis of industrial equipment, as well as advanced speech and facial recognition applications. AKIDA is also accelerating the design of robots using sophisticated sensors to see, hear, smell, touch, and even taste.

Chapter 5:
Inferring the future of neuromorphic computing

Q: In addition to identifying, extracting, analyzing, and inferring only the most meaningful data, AKIDA supports on-chip incremental and one-shot learning. How will these learning models help shape the future of neuromorphic computing and enable the design of more intelligent machines at the edge? As well, how does on-chip incremental and one-shot learning differ from deep learning convolutional neural networks (CNNs)?

Peter: On-chip incremental and one-shot learning models excel at the edge, allowing cognitive systems to be configured in the field, within seconds, to recognize the owner's face or voice. Deep learning CNNs are mostly static, that is, they are trained once at the lab and cannot learn anything new thereafter. Put simply, conventional AI silicon and cloud-based learning models have not yet learned how to learn. They are trained, using a cumbersome successive approximation method.

CNNs running on GPUs can be thought of as a 1950s model paired with a 2011 learning method. These models need to be exposed to millions of images to recognize a cat, a person, or any other object in the real world. The human brain, and AKIDA, learn by different principles. Incremental learning works by exposing AKIDA to a single image, which is then recognized anywhere. We have highlighted this capability in one of our demos where we show the AKIDA chip can accurately identify and classify a toy elephant and give it the label "elephant." AKIDA then recognizes elephants in photographs or videos in the wild.

With regards to the future of neuromorphic computing, I'd like to emphasize that while Hollywood has made a lot of people scared of AI, the human brain has a total of 86 billion neurons, with 69 billion neurons located in the cerebellum alone. Current AI hardware and learning models are nowhere near achieving human levels of intelligence. In fact, standard GPU cause-and-effect systems only have the approximate intelligence of a frog - as they are limited to repeating the same action for every stimulus. That is not intelligence.

It is important to note that Von Neumann-based systems were already beating humans at chess back in the 1960s. Recently, a trained Google AI called AlphaGo beat human champions in a game of Go. But that is all that those AI systems can do. They cannot follow instructions to complete any other task, even a simple task, as a human child can.

Even the most advanced silicon-based neuromorphic systems lack adaptability, intelligence, and creativity, a combination that makes people unique. Humans can adapt to living in cities, jungles, frozen tundra, and even deserts. We also create new synapses - the storage areas of the brain - as we assimilate new information. That's why I strongly disagree with people who say "Oh, yes. We can build systems that are smarter than humans." No, we can't. Not with our emotional and creative intelligence. And certainly not by 2040.

To increase the capabilities of AI, we must break away from the traditional layered architecture of neural networks. The brain is a complex system consisting of many modules, each with unique architecture and function. At the BrainChip Research Institute we are studying this complex architecture of the brain and building models to construct future AKIDA systems that we expect will mimic at least some aspects of human intelligence.

The simple layered structure of today's AI networks is limiting the ability of AI to progress beyond basic classification. BrainChip's advanced research is pushing the boundaries of artificial intelligence by researching neuromorphic models of the human cortex, the seat of intelligence. The frontal lobes of the cortex are significantly associated with intelligence. The aim of our research is to construct in hardware a temporal neuromorphic cortical network that is as modular and flexible as the human brain.

Advanced artificial intelligence systems that result from this research must be capable of recognizing partially obscured objects, anticipating expected outcomes, and recognizing behavior. This is the next generation of neuromorphic computing, and I hope to continue contributing to the research and science that will make this possible.

Peter van der Made

Founder and Chief Technology Officer at BrainChip

Always at the forefront of innovation, Peter invented a computer immune system and founded vCIS Technology where he served as CTO and Chief Scientist when it was acquired by Internet Security Systems and subsequently IBM. He also founded PolyGraphic Systems and designed a high resolution, high-speed color Graphics Accelerator board and chip for IBM PC graphics.

At BrainChip, Peter designed the first generations of digital neuromorphic devices on which the AKIDA chip is based and published a book, Higher Intelligence, which describes the architecture of the brain from a computer science perspective. Peter is actively involved in the development of new AKIDA IP and continues his research in advanced neuromorphic architectures based on the human neocortex, the cerebellum, and its interactions with the hippocampus.