ActualTech Media

brainchip™
Train. Infer. Learn.

What Is the Akida Event Domain Neural Processor?

Brien M. Posey

CONTENTS

The Evolving Artificial Intelligence Model 2

What Is a Neural Network? 2

What Is the Akida Neural Processor? 3

How Does the Akida Event-Based Neural Processor Work at Ultra-Low Power? 4

How Does Akida Learn? 5

IN THIS PAPER

Previous generations of artificial intelligence and machine learning chips were useful, but their time is rapidly running out. The constraints on power and bandwidth imposed by edge devices mean that it's time for a new paradigm, a new design that fulfills the promise of AI and ML at the edge. It's time for the BrainChip Akida processor.

Although machine learning (ML) has existed for some time, the technology is still evolving. The BrainChip Akida processor overcomes many of the challenges that have long been associated with ML, particularly regarding deep learning neural networks.

The Evolving Artificial Intelligence Model

Before we jump into the guts of how BrainChip's Akida Neural Processor works, it's important to understand what it does and how it will transform artificial intelligence (AI).

BrainChip has focused the past 15 years on evolving the art of AI to overcome the shortcomings of today's deep learning technologies.

In utilizing AI, corporations are processing exabytes of data to extract information for a wide range of purposes, including surveillance and security, consumer behavior, advertising, language processing, video analysis, financial predictions, and many more.

These applications have spawned a monumental market for both software and hardware and have transformed nearly every industry. There can be no argument that the breakthroughs have been extraordinary, and the growth rate of applications has been explosive. Yet, this has represented, to date, only the tip of the iceberg for AI capabilities. With the expansion of the Internet of Things (IoT) comes a parallel expansion of AI into everyday appliances in the home, office, and industry.

Today's systems, although impressive, are merely first- and second-generation solutions relying on over-simplified and limited representations of how nature's intelligence—the brain—really functions. Today's systems have limited to no ability to learn without huge amounts of labeled data and many repetitions of deep learning and training cycles.

Deep learning systems recognize an object by statstically determining the number of features that match an image—features that were extracted from millions of images that it was previously trained on. These systems use several orders of magnitude more power than the brain.

Currently, ML systems rely on power-hungry CPUs and GPUs physically located in large data centers to ingest, process, and retrain data which is generated in a highly distributed fashion all over the globe. This drives an ever-growing and insatiable need for communication bandwidth to move the data to the data center.

> BrainChip has developed the Akida Neural Processor to solve the problems inherent in moving AI out of the data center and to the location where data is created: the edge.

BrainChip deems this model ripe for a revolution, and that AI needs to evolve to support intelligence at the location where the data is generated or sensed. It believes that the future of AI lies in the ability to achieve ultra-low power processing as data is being interpreted and transformed into information, and that continuous learning needs to be autonomous and continuous.

BrainChip has developed the Akida Neural Processor to solve the problems inherent in moving AI out of the data center and to the location where data is created: the edge, of which a large segment is often referred to as IoT.

This has several advantages. The most important one is privacy, and a sharp reduction of dependency on the internet. You would not want a device in your home that shoots images up to the internet, where they can be hacked and viewed by anyone—but a warning sent over the internet to your phone that an intruder or other unrecognized person enters your home would be an advantage.

What Is a Neural Network?

A neural network lies at the core of all AI. As its name implies, a neural network is modeled on the principles of neural processing—the cells that make up the brain network. However, today's technology (deep learning) is, at best, only loosely related to how the brain functions.

Neuromorphic computing is a field of computer science based on the study of the brain, and how the function of neural brain cells can be utilized in silicon to perform cognitive computing.

BrainChip has developed the Akida neural processor utilizing the fundamental concepts of Neuromorphic computing, in combination with the advances made in deep learning.

The Akida neural processor is a flexible, self-contained, event-based processor that can run today's most common neural networks, Convolutional Neural Networks in event-based hardware, as well as the next-generation Spiking Neural Networks.

The Akida neural processor is ultra-low power, requires only internal memory, and can perform inference and instantaneous learning within an AI solution. It represents the third generation of neural networking, and the next step in the evolution of AI.

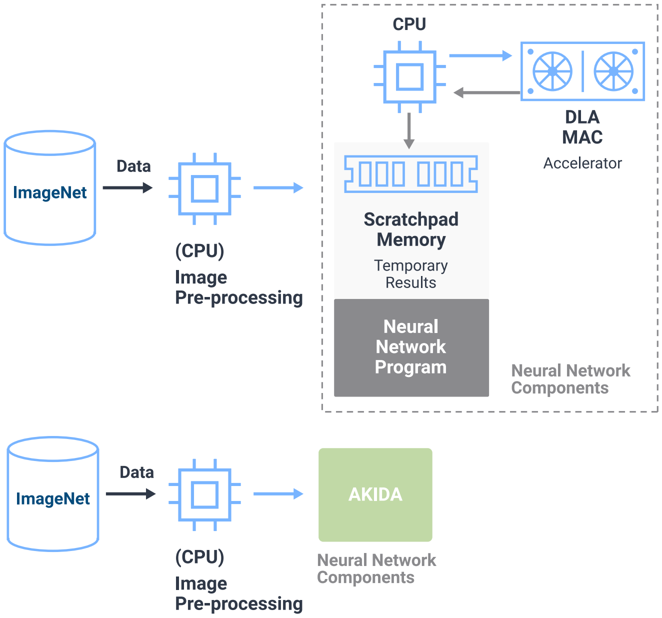


Figure 1: A traditional Neural Processing solution using a CPU, Deep Learning Accelerator, and external memory vs. the Akida solution as a fully integrated, purpose-built neural processor

What Is the Akida Neural Processor?

What makes Akida so different from first- and second-generation neural processors? Unlike those legacy processors, the Akida processor is event-based, which means it processes data in the form of events.

Events are the occurrences where things happen, such as a change of contrast in a picture, or a change of color. The human visual system encodes images in the same way.

An event is expressed as a short burst of energy. In Akida, the burst can have a value that indicates neural behavior. No events are generated where zero values occur in the network—for instance, where blank areas occur in a picture—making Akida's processing scheme intrinsically sparse. In other words, if no events exist or are generated, no processing needs to occur.

The Akida processor uses an encoding scheme called "rank coding," in which information is expressed as the time and place it occurs. Akida is not programmed in the traditional sense—it consists of physical neuron and synapse circuits configured for a specific task, defining the dimensions and types of network layers.

The entire network is mapped to physical neuron and synapse circuits on the chip. Synapses store weight values and are connected to neurons, which integrate the weight values when they're released by an incoming event. Each neuron can have thousands of synapses. Each reconfigurable core can contain the equivalent of tens of thousands of neurons.

Power is consumed only when inputs to a neuron exceed the predetermined threshold and generate an action potential to be processed by subsequent layers in the network.

No output event is generated when the sum of synaptic inputs is zero or negative, significantly reducing the processing requirements in all of the following layers. The neural and synapse functions in the Akida neural fabric are entirely implemented in digital hardware. Therefore, no computer code is running within any of the neural cores, resulting in a very low overall power consumption of approximately 3 pico-Joules per synaptic operation (in 28nm technology).

As stated previously, the Akida neural processor is a complete, self-contained, purpose-built neural processor. This is in stark contrast with traditional solutions, which utilize a CPU to run the neural network algorithm, a deep learning accelerator (such as a GPU) to perform, multiply, and accumulate mathematical operations (MACs), and memory to store network parameters (see **Figure 1**).

By integrating all the required elements into a consolidated, purpose-built neural processor, the Akida processor eliminates the excess power consumption associated with the interaction and communication between the three separate elements, as well as minimizing the physical footprint.

How Does the Akida Event-Based Neural Processor Work at Ultra-Low Power?

As described earlier, the Akida neural processor is differentiated from other solutions by two major factors:

1. It is a complete, fully integrated, purpose-built neural processor
2. It is an event-based processor

By fully integrating the neural network control, the parameter memory and the neuronal mathematics, the Akida neural processor eliminates significant compute and data I/O power overhead. This factor alone can save multiple watts of unnecessary power consumption.

The Akida event processor is constructed from event-based neurons, which work in a manner much more similar to the way the brain operates than the "perceptron" style neurons used in today's deep learned neural network hardware solutions.

> In the Akida event domain processor, "events" or "spikes" indicate the presence of information, eliminating wasted effort. This is a core principle.

All neural networks consist of some form of simulation or emulation of "neural cells" and the weighted connections between those cells. The connections between neural cells have memory, store a value, and are called "synapses" (see **Figure 2**).

In the end, only information is processed and consumes energy. In the Akida event domain processor, "events" or "spikes" indicate useful information, eliminating wasted effort. This is a core principle.

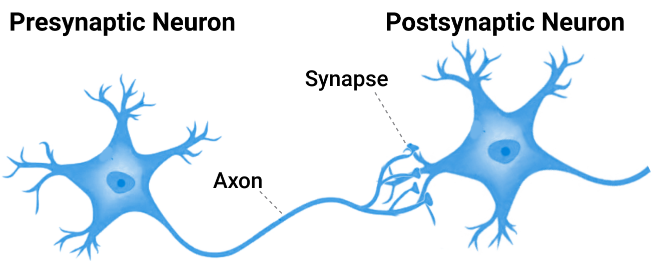


Figure 2: Biological neurons are cells that communicate with one another and store information in synapses. A neuron can have hundreds of thousands of synapses, the content of which is recalled by sensory input action potentials. The neuron integrates the values of active synapses and generates an action potential output when the integrated value reaches or exceeds a threshold value. Artificial Neural Networks model a similar behavior.

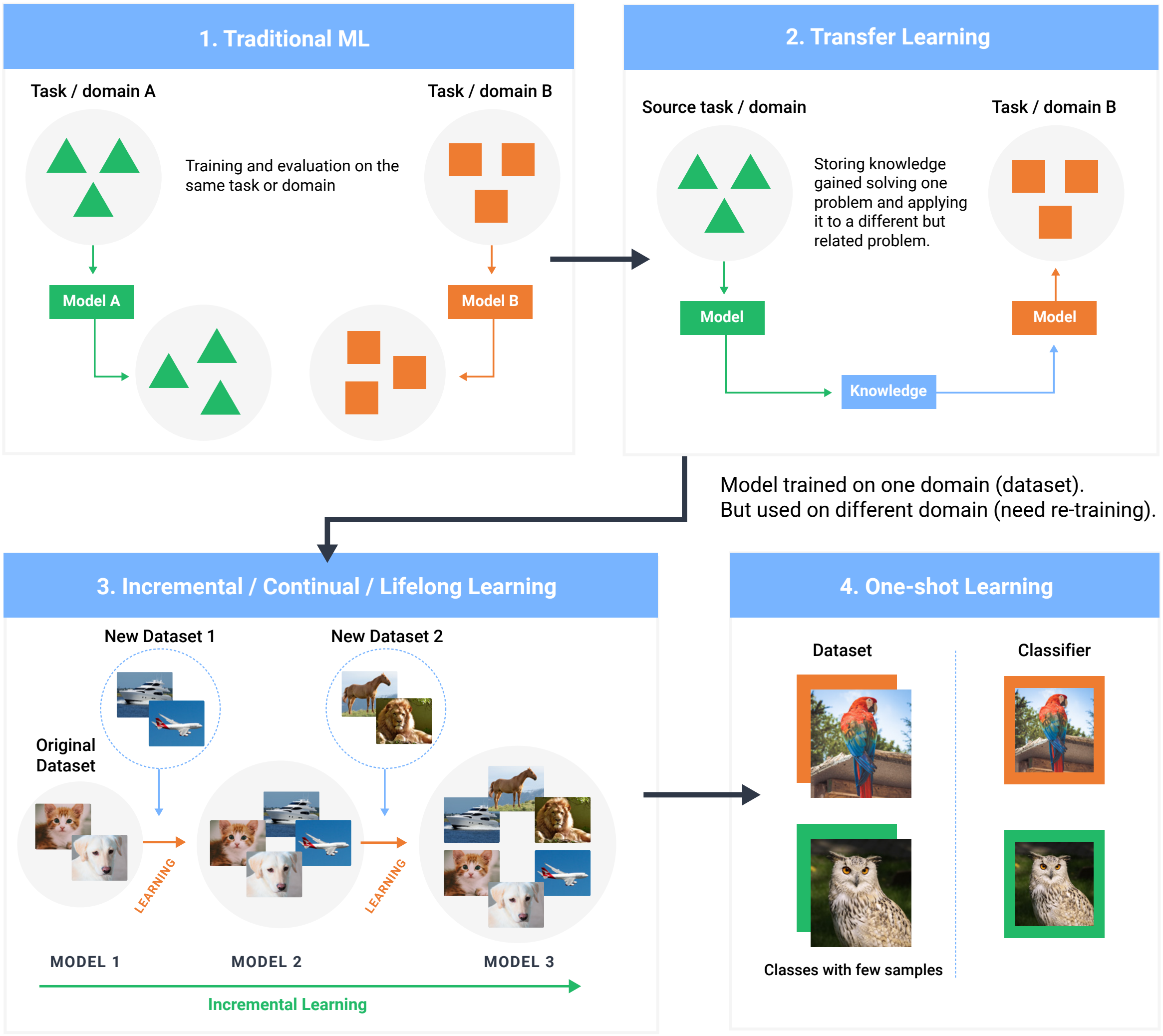
This is fundamentally different from the function of artificial neurons in Deep Learning Convolutional Neural Network hardware implementations, which process all information without discerning whether it contains valuable information or not.

Every pixel in an image is converted to data and processed, whether it contains any information or not.

To illustrate how this works, consider an extreme case. You could give a "standard" Convolutional Neural Network a blank page to process, and it will take every pixel and process it through millions of multiply-accumulate instructions to find out that the page is blank.

The Akida event-based processing method works similar to how the human brain would process a blank page: since there are no lines or colors on the page, it receives no events, so it does not need to process anything. It is this reduction of data that must be processed, known as "sparsity," that leads to significant power savings.

Combined with state-of-the-art circuit architecture and implementation, the Akida neural processor has demonstrated power reduction of up to 10x over the most power-efficient alternatives. In addition, power savings are up to 1,000x compared with standard data center architectures. For AI applications at the edge, where information is created, power budgets can be limited to micro-watts or milli-watts. The Akida platform, with its ultra-low power consumption, meets the power budget requirements for these applications.



Model trained on one domain (dataset).
But used on different domain (need re-training).

Learn new tasks/domains (datasets) without forgetting the old ones.

Learn with only a few samples.

Figure 3: The evolution of training and learning

How Does Akida Learn?

Training is an extremely time- and energy-consuming process in today's deep learning solutions, as it requires a tremendous amount of hand-labeled input data (datasets) and extremely powerful compute infrastructures to train a neural network.

All this has resulted in very useful and powerful solutions, but one with a significant drawback—once an AI solution is trained, it's not easy for the system to learn new things without going through the entire training process again, this time including the new information.

The Akida processor offers a solution that can take a deep learned neural network, run inference on that network, and then has the ability to learn things without going through a retraining. **Figure 3** shows the evolution of training and learning. Akida represents the third generation of AI, whereby instantaneous learning is enabled.

In native learning mode, event domain neurons learn quickly through a biological process known as Spike Time Dependent Plasticity (STDP), in which synapses that match an activation pattern are reinforced. BrainChip is utilizing a naturally homeostatic form of STDP learning in which neurons don't saturate or switch off completely.

STDP is possible because of the event-based processing method used by the Akida processor, and can be applied to incremental learning and one-shot or multi-shot learning.

The next generation of AI solutions will evolve by utilizing the concepts learned from studying the biological brain. The BrainChip Akida neural processor embodies this evolution by incorporating event domain neural processors in a practical and commercially viable way.

> Akida represents the third generation of AI, whereby instantaneous learning is enabled.

The ability to move AI to the edge depends upon a fundamental shift in how the core of AI solutions are built. The Akida neural processor provides the means. It's a self-contained, efficient neural processor that's event-based for maximum efficiency, ultra-low power consumption and real-time learning. Instantaneous learning reduces the need for retraining, and its processing capabilities eliminate the need for constant internet connectivity.

The BrainChip Akida neural processor is the next generation in AI that will enable the edge. It overcomes the limitations of legacy AI chips that require too much power and bandwidth to handle the needs of today's applications, and moves the technology forward in a significant leap, allowing AI to do more with less. Akida's time has come.