The Linley Group

MICROPROCESSOR report

Insightful Analysis of Processor Technology

BrainChip Akida Is a Fast Learner

Spiking-Neural-Network Processor Identifies Patterns in Unlabeled Data

By Mike Demler (October 28, 2019)

BrainChip will offer its spiking-neural-network (SNN) technology in a new processor called Akida, which it optimized for low-power edge AI. *Akida* is the Greek word for *spike,* so it's an appropriate description of the temporally sparse activations the processor uses to enable event-driven computation. The company previously offered its SNN technology as an FPGA accelerator, but at the recent Linley Fall Processor Conference, it announced plans to by the end of the year tape out the chip for TSMC's 28nm technology and to start sampling in early 2020. It also sells Akida's core architecture as configurable intellectual property (IP).

The new chip is built around a mesh-connected array of 80 neural processor units (NPUs), as Figure 1 shows. The interconnect is a proprietary design that supports core-to-core packet transfers as well as intrachip communications among the various function blocks. The cores employ digital logic that mimics the spiking behavior of biological neurons, but the chip also integrates spike converters, which enable it to run popular convolutional neural networks (CNNs) such as MobileNet. The converters can generate spikes from audio, image, lidar, pressure, temperature, and other sensor data, as well as from Internet packets and multivariate time-series data. Although Akida is primarily an inference engine, its native SNN mode allows it to learn new spiking patterns, too.

Aggressive Quantization Saves Memory

Unlike its FPGA-based predecessor (see *MPR 5/21/18,* "BrainChip Aims to Spike Neural Nets"), Akida is a standalone processor. It integrates a Cortex-M4 CPU that includes DSP and FPU functions. The M4 boots from SPI-connected flash memory, loads the neural-network computational graph, and manages I/O functions. The 32-bit LPDDR4 PHY allows the chip to access up to 4GB, but most of the intended applications keep the network weights and intermediate data in the 8MB SRAM, distributed in 100KB portions to the 80 cores.

Akida's DMA controller connects to the mesh network, enabling the NPUs to directly fetch image frames and other sensor data from DRAM. Data transfers from the PCIe 2.1 and USB3.0 interfaces also go directly to DRAM. The two-lane PCIe interface delivers 1.0GB/s, which is more than sufficient for the low-power applications that Akida targets. A separate two-lane PCIe 2.1 root complex allows connection of up to 64 devices to support larger neural networks. The chip can stream sensor data directly from the I^2S, I^3C, and UART ports as well. The I^3C interface is a new MIPI standard that combines I^2C and SPI.

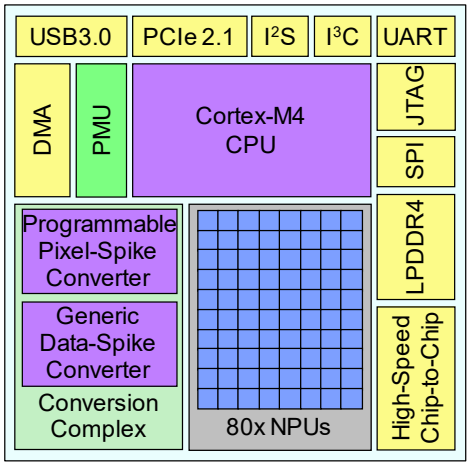

Figure 1. BrainChip Akida processor. The 28nm chip's AXI-based mesh network connects 80 event-based neural processor units (NPUs). The NPUs implement a spiking neural network, but Akida also integrates spike converters that allow it to run convolutional neural networks.

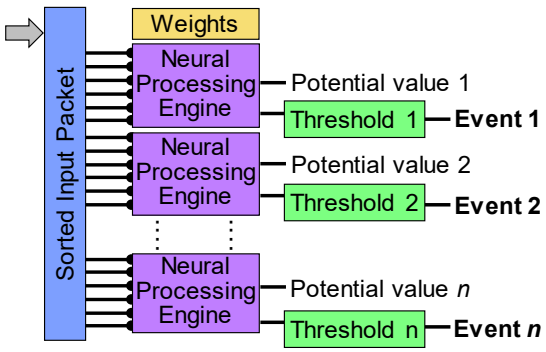

Figure 2. Akida neural processor unit. Each NPU includes eight neural processing engines (NPEs) that run concurrently. They handle convolutions and pooling-layer operations, along with ReLU activations.

As Figure 2 shows, Akida's NPUs include eight neural processing engines (NPEs), which run simultaneously. The NPEs perform event-based convolutions, handling 1x1, 3x3, 5x5, and 7x7 kernels. Each can perform the equivalent of four simultaneous multiply-accumulate (MAC) operations, but since they run asynchronously on the basis of sparse spiking events, a comparison with conventional clocked MAC units is inappropriate. Nevertheless, adding Akida's 640 PEs at the maximum 300MHz clock frequency totals 1.5 trillion operations per second (TOPS), although it'll use much less in real-world applications.

The PEs also support 2x2 and 3x3 max-pooling operations, along with ReLU activations. Conventional CNNs normally employ INT8 or higher-precision MAC units, but Akida trades accuracy for lesser storage requirements by aggressively quantizing input activations and weights using 1-, 2-, or 4-bit precision for each layer. Depending on the network, preprocessing, and training, the loss in moving from INT8 to INT4 CNNs is typically 1–3%, according to BrainChip's measurements. Single-bit precision reduces the multiplications to simple additions. Activations are unsigned, but weights include a sign bit.

The engines synchronize their outputs to a system clock, but they're idle until the sparse spiking events (i.e., activations) arrive at their inputs, allowing the chip to minimize active power. The NPEs output a result to the mesh network only when the calculations exceed the pretrained spiking threshold. The 28nm design can clock at up to 300MHz, but keyword spotting and other simple tasks can save power by running much slower. The NPUs employ about 40KB of SRAM for the incoming-event buffer and a further 40KB to store weights. The remainder provides storage for intermediate results. The event storage is a ping-pong buffer, allowing the NPUs to receive input packets from other NPUs or the spike converter while the NPEs operate on data from the previous event window.

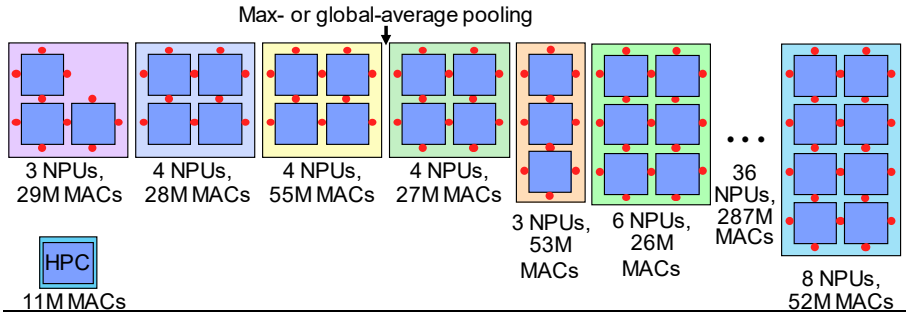

Figure 3. MobileNet running on Akida. M=millions. To handle an object-recognition network such as MobileNet, the chip first converts pixels to spikes using high-precision convolution (HPC). It can run all 30 layers on just 68 of the 80 NPUs, storing intermediate results in on-chip SRAM.

Spiking the Net

Designers can use Akida as a native SNN processor, or they can use the BrainChip CNN2SNN software to retrain their CNNs, reducing power by changing convolutions to event-based computations. Customers can directly retrain networks built with TensorFlow and Keras. To run object-classification networks such as MobileNet, Akida's pixel-to-spike converter implements a proprietary algorithm that the company calls high-precision convolution (HPC). This algorithm's INT8 convolutions generate events that serve as the inputs to the NPUs. Although the company withheld details, the technique involves detecting intensity variations in small regions of the image, converting the result to a spike pattern.

For example, to process video, the converter first generates spiking events by applying convolution filters over an entire frame. If the filtered pixel values exceed a predefined threshold, the converter outputs a 1- to 4-bit spike. The chip can either store the events in external DRAM first or write the data directly to the event buffers in the NPUs that constitute the next neural-network layer. For the latter method, the converter packetizes the events, transmitting them to the mesh in individual event windows for each NPU. When it finishes processing, it issues an end-of-frame command; from that point, all NPUs in each layer execute in parallel. As images feed forward through the network pipeline, all the layers run simultaneously.

Google designed MobileNet to enable efficient computer-vision applications in mobile devices. The network achieves its efficiency by employing 1x1 point-wise- and depth-wise-separable convolutions to reduce the total number of operations (see *MPR 9/17/18,* "DeePhi Accelerates Xilinx AI Strategy"), a technique Akida supports. In its maximum-resolution configuration, MobileNet v1 comprises 28 layers with about 4.2 million parameters, equivalent to 2.1MB at the chip's maximum 4-bit resolution.

As Figure 3 shows, Akida can load that entire network, running all

Price and Availability

BrainChip withheld pricing. Akida IP is available for licensing now, and the company expects to begin sampling the processor in early 2020. For more details, access *www.brainchipinc.com/products/akida-neuromorphic-system-on-chip*.

the layers in parallel and storing intermediate results in SRAM. It performs classification at up to 80fps using MobileNet's 224x224 images while the cores consume just 434mW, but at 30fps, they consume just 157mW, according to BrainChip's estimates. The company also simulated MobileNet on a 16nm Akida, reducing power to 79mW at 30fps.

Works Well With Little Supervision

In addition to its low power consumption, Akida's spiking neural network offers the ability to perform incremental learning. BrainChip tested this capability by first training MobileNet to recognize 64 classes in 84x84 images from the mini-ImageNet database. That process runs offline, using 32-bit floating-point operations to achieve 96% accuracy.

The next step is to replace the last fully connected layer with learning PEs that represent 100 neurons, then train five new classes on a few samples for each class. At just one sample per class, Akida created a new class that delivers an average of 64% accuracy on 35 test samples per class. Increasing to 20 samples per class improved average classification accuracy to 82%.

Akida beat that incremental-training result on a test of keyword spotting. In the first-pass training, it learned to recognize 10 words and reject 15 unknown words. Converting the last layer to self-learning mode delivers 89% accuracy. Performing on-device training, Akida learned three new words, achieving 94% accuracy. The chip can classify seven words per second when running at 500kHz, with the cores consuming just 200 microwatts.

In its native spiking mode, Akida can also perform unsupervised learning, identifying patterns in unlabeled data sets. In one example application, it recognized 128x128 hand-gesture images produced by an event-based camera employing a dynamic vision sensor (DVS). Such devices reduce bandwidth and processing requirements by only outputting pixels that sense a change in their photon levels. The sparse pixel output constitutes an event-based spike pattern. Akida can autonomously recognize different patterns, such as the number of fingers in each gesture, enabling the customer to assign each pattern to a unique class.

Because Akida's SNN can recognize patterns in any time series, it's useful for data analytics and security. When training on the Communications Security Establishment and Canadian Institute for Cybersecurity intrusion-detection 2018 (CSE-CIC-IDS2018) data set, the company's software automatically labeled the active neurons by extracting features from the data, converting the patterns to spikes. Using the spikes as input, Akida achieved 98% accuracy on 30,000 inferences per second while consuming only 20mW.

An Efficient Hybrid Engine

Akida is a low-power hybrid inference engine that natively runs SNNs in addition to supporting low-precision (INT4) CNNs. BrainChip's proprietary data-to-spike and pixel-to-spike converters are a much more efficient solution than Tsinghua's experimental Tianjic hybrid AI processor, which implements two different compute methods to handle CNNs and neuromorphic functions in each compute core (see *MPR 9/23/19,* "Tsinghua Pedals Hybrid AI Processor").

For keyword spotting, Akida competes with Syntiant's NDP10x speech-recognition processor (see *MPR 3/18/19,* "Syntiant Knows All the Best Words"). The NDP10x uses older 40nm ULP technology, but it can classify up to 64 words while consuming 200 microwatts, matching Akida's low-power performance. Syntiant also cuts storage requirements by using 4-bit weights, but it offers higher precision by using 8-bit activations. Akida requires off-chip calculation of Mel-frequency components, whereas the NDP10x implements that function on chip. Because the NDP10x handles only audio data and lacks incremental-learning capability, however, it can't do on-device training.

Akida will also compete with Eta Compute's ECM3531 (Tensai), but that chip is more a general-purpose low-power MCU with minimal machine-learning capabilities than an edge-AI processor (see *MPR 10/29/18,* "Eta Compute MCU Puts AI in IoT"). Tensai is manufactured in 55nm embedded-flash technology, combining asynchronous logic and subthreshold analog-design techniques to reduce power consumption to less than 1mW at 48MHz. It supports neural-network operations by complementing the Cortex-M3 CPU with NXP's CoolFlux DSP. Tensai can store weights in its 512KB flash memory, but it only integrates a 128KB SRAM that the CPU and DSP must share. Nevertheless, it can handle basic AI tasks, such as 32x32-pixel CIFAR-10 text recognition. It uses the DSP to run SNNs for object and wake-word detection, but it lacks the horsepower to perform classification.

BrainChip has developed a unique product that's more versatile than its low-power competitors. OEMs can use Akida in smart cameras or to power gesture-based and voice UIs in edge devices. Its audio-pattern-detection capabilities are well suited to industrial tasks such as vibration monitoring, and its ability to recognize data patterns associated with common Internet exploits is ideal for protecting IoT devices. Even as researchers continue to investigate SNN applications, Akida offers a solution that addresses several types of problems. ♦