akida

Product Brief

BrainChip's first-to-market neuromorphic processor IP, Akida™, mimics the human brain to analyze only essential sensor inputs at the point of acquisition, processing data with unparalleled efficiency, precision, and economy of energy. Keeping AI/ML local to the chip and independent of the cloud dramatically reduces latency while improving privacy and data security.

Infer and Learn at the Edge

Akida is a fully customizable event-based AI neural processor. Akida's scalable architecture and small footprint boosts efficiency by orders of magnitude, supporting up to 256 nodes that connect over a mesh network.

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. The Akida neural processor leverages data sparsity, activations, and weights to reduce the number of operations by at least 2X.

Benefits

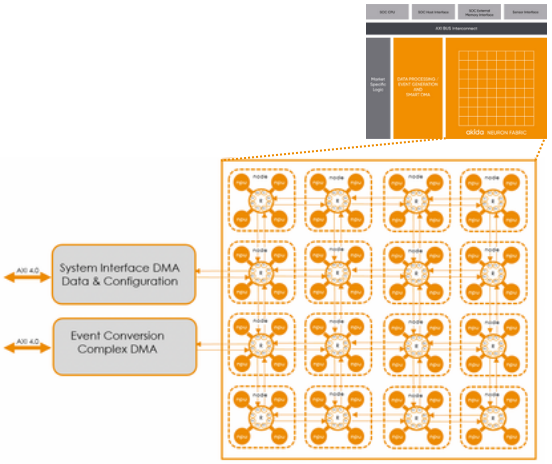
Distributed computation and event-based action delivers unparalleled performance and efficiency.

- Ultra-low latency
- Runs multiple networks in real-time
- Performs one-shot learning
- Remarkably power efficient
- Cloud independent
- Flexible and quick to deploy
- Privacy and security protected

Akida is Uniquely Essential

BrainChip's IP fabric can be placed either in a parallelized manner that would be ideal for ultimate performance, or space-optimized in order to reduce silicon utilization and further reduce power consumption.

Entire neural networks can be placed into the fabric, removing the need to swap weights in and out of DRAM resulting in reduced power consumption while increasing throughput. Additionally, users can modify clock frequency to optimize performance and power consumption further.



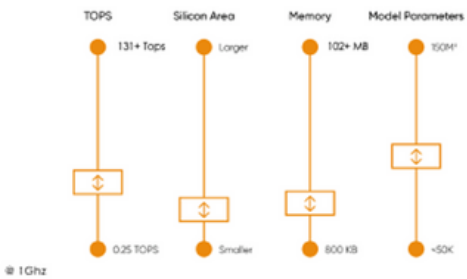
Highly Configurable IP Platform

Akida is flexible and scalable for multiple edge AI use cases. Achieve the most cost-effective solution by optimizing the node configuration to the desired level of performance and efficiency.

Scale down to 2 nodes (@ 1Ghz = 1 TOPS) for ultra low power or scale up to 256 nodes (@ 1Ghz = 131 TOPS) for complex use cases.

Multi-pass processing provides flexibility to process complex use cases with fewer nodes increasing power efficiency.

Quantization in MetaTF converts model weights and activations to lower bit format reducing memory requirement.

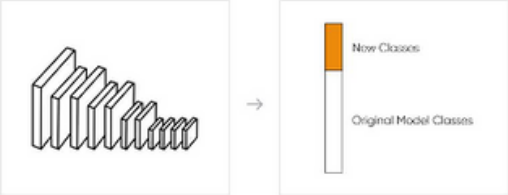


brainchip
Essential AI

akida

Multi-Pass Processing Delivers Scalability

One-Shot, On-Chip Learning



Tech Foundations

Distributed Computation

Each NPU has dedicated compute and memory, reducing data movement

Event-Based Processing

NPUs perform computationally on events (non-zero values)

Event-Based Communications

Send events over mesh network without host CPU intermediation

Event-Based Learning

On-Chip learning algorithm

Key Features

- Brain inspired Neuromorphic Hardware Architecture.
  - Neural Processing Unit (NPU) at memory compute architecture implementing Integrate and fire neuron.
  - Emulate multiple neurons with configurable Synapses.
  - Compute only when events occur.
  - Up-to 4 bits for weights and activation.
  - 4 NPU interconnect to make up 1 node.
  - Multiple Node interconnect via packet switched mesh network.
- Same hardware supports standard CNN after conversion to SNN, and native SNN.
- MetaTF software framework to convert CNN to SNN.
- Configurable architecture for size, power, speed:
  - Number of nodes (2-256).
  - Internal memory per NPU (50 to 100 KB per NPU).
- Integrated DMA and image-to-event converter.
- Standard AXI 4.0 interface for off-chip communication.
- Single clock design speed depending upon technology choice: 1 Ghz in 14nm. Higher in 7nm.

IP Delivery

- Fully synthesizable RTL.
- IP deliverables package with standard EDA tools.
- Complete test bench with simulation results.
- RTL synthesis scripts and timing constraints.
- Customized IP packaged targeted for your application.
- Run time software C++ library.
- Processor and OS agnostic.
- Runs networks in two possible modes.
- Power efficient mode:
  - Optimize power with limited NPUs for sufficient inference performance.
- High performance mode:
  - Use maximum number of NPUs per layer to get highest performance and efficiency.

brainchip
Essential AI