TENN/IP and Akida Overview

At the heart of BrainChip's offering is the TENN IP, which encodes sensory data into sparse, time-stamped "events" rather than dense, clock-driven activation vectors. Each neuron in a TENN processes inputs only when an event (e.g., a pixel change or audio spike) arrives, preserving temporal context through adjustable synaptic time constants. Because events occur only when something meaningful happens, the network remains largely quiescent during idle periods, drastically reducing dynamic power. The licensed TENN blocks can be stitched together on a mesh of Neural Processing Units (NPUs), each housing up to tens of thousands of digital "spiking" neurons and their synaptic weights. When instantiated in silicon, this becomes the Akida neuromorphic processor. Akida integrates up to 1.2 million spiking neurons and 10 billion synapses across multiple NPUs, all communicating via an on-chip event-routing fabric. Crucially, Akida supports on-chip incremental learning, so edge devices can continuously adapt models (for example, adding new voice commands) without offloading data for retraining.

Comparison to GPUs

GPUs excel at dense linear algebra, it is common to batch hundreds of image frames or audio windows through large convolutional kernels. However, this model requires fetching and multiplying entire weight matrices at every inference step, even if most pixels or audio samples carry no new information. As a result, an edge GPU (e.g., NVIDIA Jetson) may draw several watts (even while mostly idle) because its pipelines and DRAM remain active. In contrast, BrainChip's event-driven architecture only "wakes up" neurons when real events arrive. Inference power consumption can drop into the tens of milliwatts, making Akida far more suitable for always-on tasks like wake-word detection or anomaly monitoring. Latency is also deterministic: instead of waiting for a batch to fill and a GPU kernel to launch, Akida processes spikes in hardware as soon as they occur, often within sub-millisecond timescales.

Comparison to Other Neuromorphic Chips

Several neuromorphic platforms exist (Intel's Loihi, IBM's TrueNorth, and research projects like SpiNNaker) but BrainChip distinguishes itself in three main ways:

1. Digital, CMOS-Compatible IP
    - *Intel Loihi* also uses digital spiking neurons and on-chip learning, but it remains largely a research platform with limited commercial integration.
    - *IBM TrueNorth* pioneered large-scale spiking inference but lacks on-chip incremental learning: once a TrueNorth network is trained offline, its weights are fixed in silicon.

   - BrainChip's TENN/IP is manufacturable in standard foundries, and its license model lets semiconductor partners embed Akida cores in custom SoCs (e.g., in security cameras or industrial sensors).

2. On-Chip Incremental Learning

   - While Loihi supports learning rules like STDP, throughput is limited by how many neurons and synapses can update in real time; moreover, scaling Loihi for commercial volumes remains uncertain.

   - TrueNorth has no field-programmable learning: updates require re-fabricating the chip.

   - Akida can add new patterns or adjust weights on the fly at the edge (e.g., recognizing a new face in a camera), then periodically back up weight snapshots to the cloud. Enabling a true "train-as-you-go" workflow.

3. Edge-Focused Power Profile

   - *SpiNNaker* uses clusters of low-power ARM cores to simulate millions of spiking neurons, but it consumes hundreds of watts for large-scale experiments. Unsuitable for battery-powered devices.

   - Analog neuromorphic approaches (e.g., SynSense) can drive power further down but often require specialized "neuromorphic" fabrication processes and face yield challenges.

   - Akida's fully digital, CMOS-based NPUs strike a balance: the chip runs on sub-100 mW for small TENN models, while still using familiar digital design flows.

Benefits of TENN for Edge AI

- Sparsity and Temporal Encoding
  Because events carry both "where" (which neuron fired) and "when" (timestamp) information, TENNs can represent continuous streams (like audio or video) without heavy buffering. This contrasts with CNN-on-GPU pipelines that must accumulate multiple frames or audio windows before processing, incurring memory and latency penalties.

- Scalable Mesh of NPUs
  Akida's mesh architecture lets designers choose network size based on application: from a few thousand neurons for simple audio tasks to over a million for richer

vision models. Each NPU holds local synaptic weights, reducing the need for off-chip DRAM.

- Incremental Learning
  Edge devices in dynamic environments (e.g., changing lighting conditions, new speakers in a room) benefit from continual model refinement. Akida’s on-chip training engine uses local spike histories and backprop-in-time approximations to update weights without interrupting inference. This capability is rare among neuromorphic platforms and absent in GPU-based edge deployments, where retraining typically requires sending data to a server or data center.