The Akida event-based neural processor is a high-performance, low-power SoC designed for edge applications, distinguishing itself from traditional deep learning accelerators (DLAs) through 2 key features:

**Feature 1: Low-power CNN Inference Using Event-Based Processing**
- Akida runs CNNs/DNNs in the event domain, which enables:
  - A reduction in computation (40-60%) when compared to non-event-based designs.
  - Novel speed/power/accuracy trade-offs
- Akida distributes network computation across ~60 small neural processing units (NPUs), each with its own collection and processing memory, resulting in:
  - A more granular distribution of computation to layers that need it most.
  - A reduction in data movement and SRAM reads.
  - Eliminating the need for off-chip memory access or external host CPU (in many cases).
- Akida utilizes primarily 1-bt weight and activity quantization and depthwise separable convolution, which:
  - Reduces the total required memory.
  - Reduces the computational cost of each event processed.

**Feature 2: On-Chip Learning**
- Akida incorporates a bio-inspired learning algorithm, adapted from spike timing-dependent plasticity (STDP).
- Combined with pre-trained feature extractor networks, this allows us to perform learning directly on the chip.

### Akida SoC

**Figure 1** below shows the Akida SoC, which includes 80 NPUs connected via a mesh network, a pixel-converter, on-chip M-Class CPU, and data input/output memory interfaces.

### The Akida Event-Based Neural Processor

The Akida SoC efficiently processes only events, which are non-zero activation outputs from DLAs utilizing a single, large matrix multiplication engine (e.g., systolic array) to perform dense computational operations very efficiently. The Akida architecture distributes computation across ~60 NPUs. Each NPU sends only events to other NPUs for processing via the mesh network. Below we show some trade-offs between the Akida architecture and a traditional DLA architecture.

**Akida Architecture**
- Becomes more efficient as activation sparsity increases.
- Computation can be more evenly distributed through short latency.
- Each layer can be tuned for a mix of weight and activation bit-widths (1-4).
- Clock speed runs low enough to allow low-leakage memory.
- Most suitable for small to moderately large CNNs that fit completely on-chip.

**Traditional DLA Architecture**
- Cannot take advantage of activation sparsity.
- Latency scales linearly with side length of systolic array.
- An NPU has additional memory overhead (2xCMN).
- Most suitable for moderate to very large CNNs/DNNs.

The event-based computations (e.g., convolution) Akida performs are algorithmically identical but become increasingly efficient as the activation sparsity increases. Figure 2 shows a comparison between frame-based and event-based convolution.

**Figure 2.** Comparison of frame-based convolution and event-based convolution.

Many CNNs utilize Rectified Linear Units (ReLU) as the network activation function shown as the blue line in Figure 3 below. Because ReLU produces zero output for input values less than zero, many popular CNN models have a mean activity of sparsity of 40-50% (Albericio et al., 2016).

Akida takes advantage of this activity sparsity by processing and sending only non-zero activations (events). Instead of processing/branching 32-bit floating point activations, we use 1, 2, and 4-bit quantized ReLUs (QReLU).

We have observed activation sparsity ranging between ~40% and ~60% in our CNNs and have successfully integrated activation sparsity into our training loss functions in TensorFlow Keras.

### Low-Power Pretrained Networks for Inference

Our ‘cnn2snn’ toolkit enables easy preparation of trained networks for low-power inference tasks. The conversion flow is shown in Figure 4.

**Figure 3.** ReLU and QReLU Activations.

### Learning on the Edge

- **Base CNN Feature Extractor (pre-trained)**
- **CNN top layer replaced with Akida Native Learning Layer**
- Enables:
  - Few-Shot Learning
  - Continuous Learning
- Easily incorporate many state-of-the-art innovations in CNN training, e.g. dense classification, LEVitas (et al., 2019).

### Mini-ImageNet on the Edge

**Model Preparation**
- Quantized MobileNet V1
- Trained on the miniImageNet training set (84 classes, see Vinyals et al., 2016)
- Top layer replaced with Akida Native Learning Layer

**Edge Learning**
- 5-way (520 classes, cross-validated x 100)
- 1-to-20-shot learning tested
- Simple data augmentation only
- Conclusions:
  - Competitive Performance
  - Small network
  - Fully Edge-compatible learning

### Audio Keywords on the Edge

**Model Preparation**
- 6-layer CNN (see (et al., Test Case 2))
- Pre-training on baseline 10 keywords (see Figure below)
- Top layer replaced with Akida Native Learning Layer

**Edge Learning**
- Novel classes learned and tested per individual
- 15 novel classes selected based on sufficient repeats per subject
- 4-shot, 3-way learning

**Conclusions**:
- Quantized edge-compatible Akida layer can be as good as pre-trained CNN
- Edge-trained subject-specific novel classes classified with similar accuracy
- Novel classes do not disrupt base class accuracy

### References