brainchip™
Biologically Inspired, Digitally Engineered

Introducing Akida

NEURAL PROCESSING UNIT

Anil Mankar
Chief Development Officer & Co-Founder

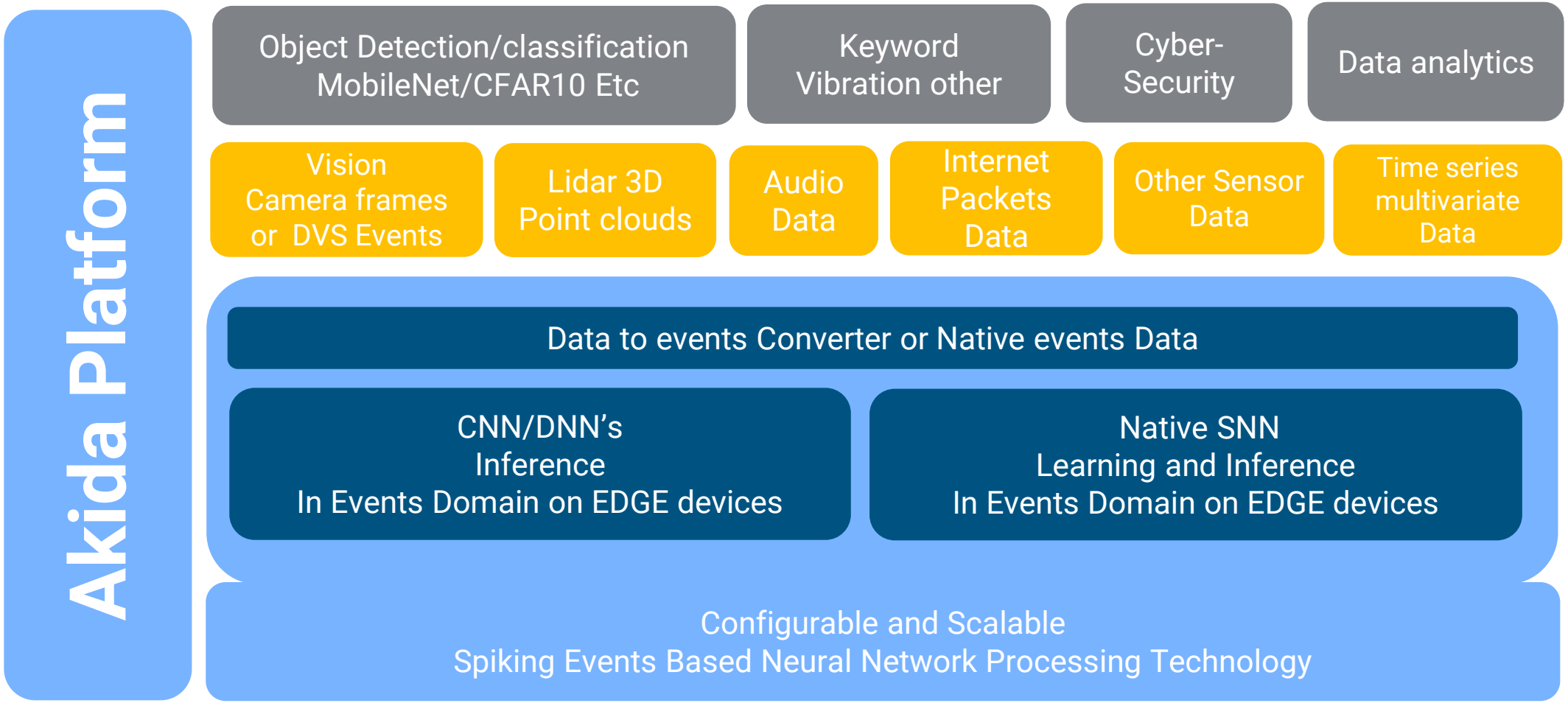
Linley Conference | October 24, 2019

About BrainChip

BrainChip is a global technology company that has developed Akida, a revolutionary advanced neural networking processor that brings artificial intelligence to the edge in a way that existing technologies are not capable. The solution is high-performance, small, ultra-low power and enables a wide array of edge capabilities that include local training, learning and inference.

Biologically inspired, digitally engineered

Akida NPU EDGE AI Technology

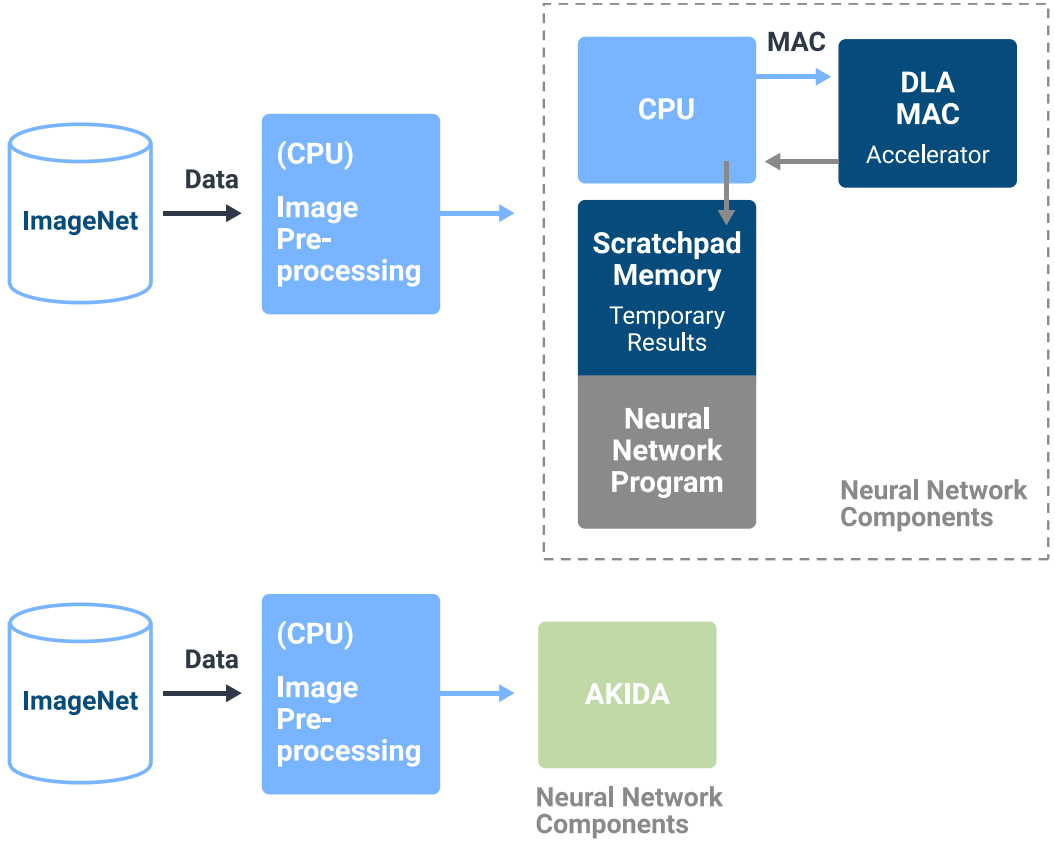


Why Use Spiking Event Based Processing

- Spiking Events are non-zero activations
  - We only process Events
- Events are inherently sparse leading to reduced number of operations
- Quantizing weights and activations to 1,2 or 4 bits reduces memory requirements
- Each Layer computations are done on allocated NPUs and all NPUs run in parallel
  - All intermediate results stored on chip memory : Eliminates off chip memory access
  - NPUs communicate over mesh network : No need for external host CPU
  - All layers of the Neural Network run in parallel : Akida Runs the entire network
- Akida allows inference and incremental learning on Edge Devices within a power, size and computation budget

Akida Architectural Differentiation

- Standard **D**eep **L**earning **A**ccelerators (DLA)
  - Perform a MAC function on arrays of numbers.
  - The neural network is executed on the host CPU.
  - Intermediate results are stored in host CPU memory.
- Akida Event Based Neural network Processing
  - Runs the entire network on NPU cores
- Only the preprocessing part is the same in both technologies



Akida NSoC: A Perfect solution for Edge AI Devices

- Single **low-power, hardware platform** to run:
  - **Conventional CNN/DNN** inference algorithms
    - **On-chip** incremental learning at EDGE devices
  - **Native SNNs** with event-based learning algorithm
  - **Entire Multi-layer Neural Network runs on AKIDA Fabric**
    - Preprocessed data in and Classified data result out in last layer
- Uses well-known Tensorflow Machine Learning ecosystem.
- Akida Technology offered as a NSoC Chip or Customized IP

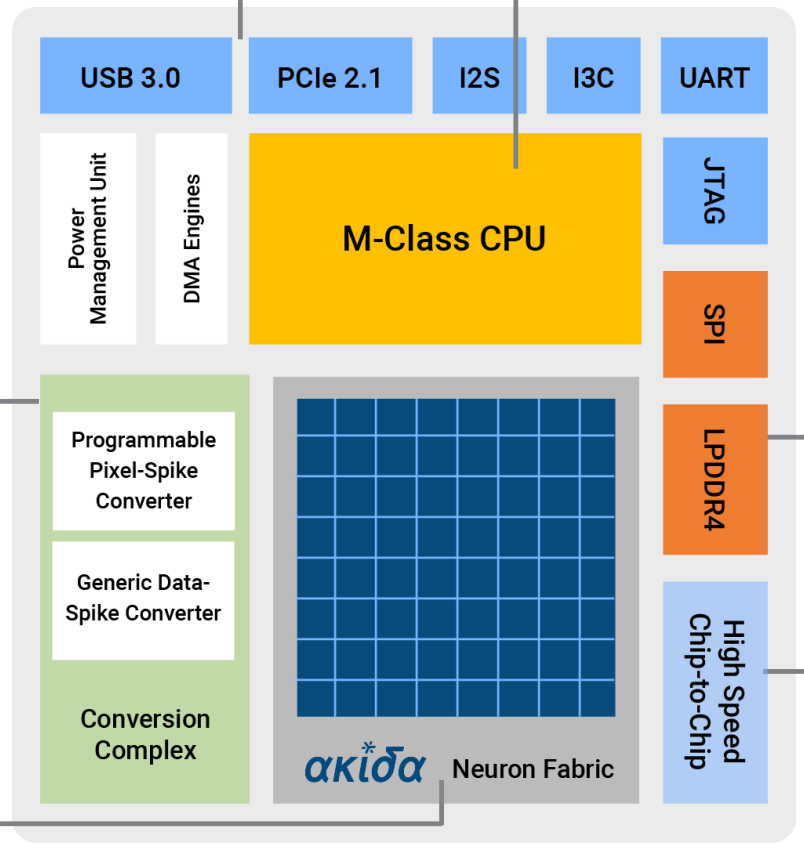
Akida NSoC Chip: Complete AI Edge Solution

Single Platform for CNN inference or Native SNN with on chip learning for EDGE Devices

Data input Interfaces
- PCI-Express 2.1 x2 Lane Endpoint
- USB 3.0 Endpoint
- I3C, I2S, UART, JTAG

Conversion Complex
- Pixel-Spike Converter
- SW Data-Spike Encoder
  - Any multivariable digital data
  - Sound, pressure, temp others

Flexible Akida Neuron Fabric
- Implements 80 NPUs
- All Digital logic with SRAM (8MB)
- Also Available as Licensed IP Core
- First Implementation : TSMC 28nm.



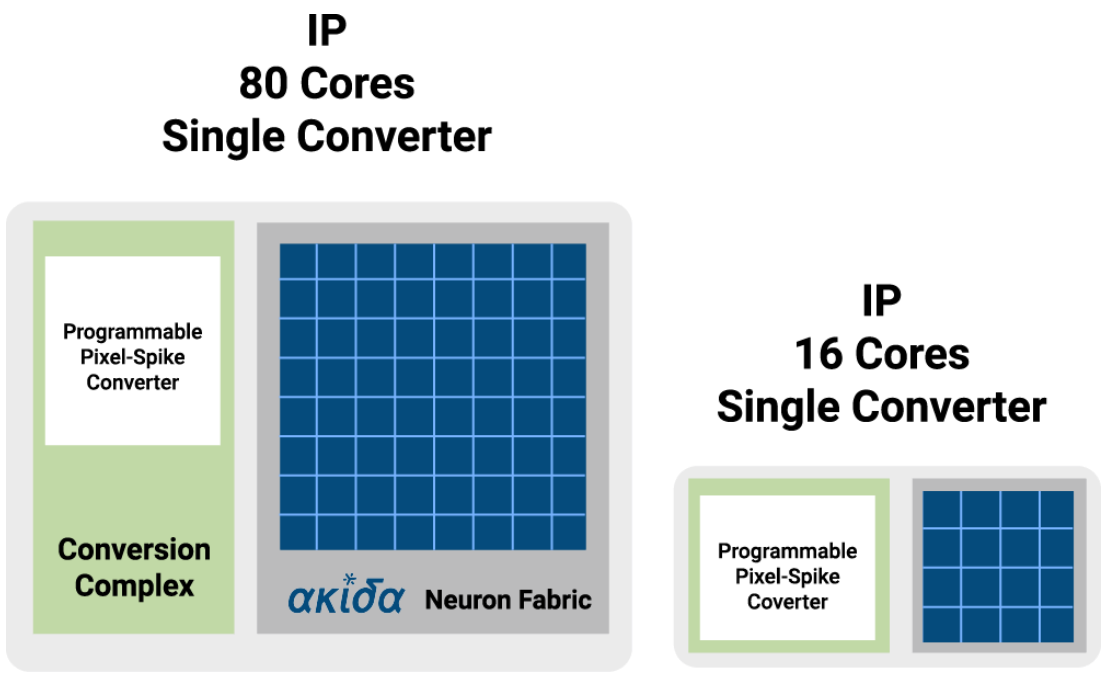
On-Chip Processor
- M-Class CPU with FPU & DSP
  - System management
  - Akida Configuration

External Memory Interfaces
- SPI FLASH for boot/storage
- LPDDR4 Program /Weights

Multi-Chip Expansion
- PCIe 2.1 2 lane root complex
- Connects up to 64 devices

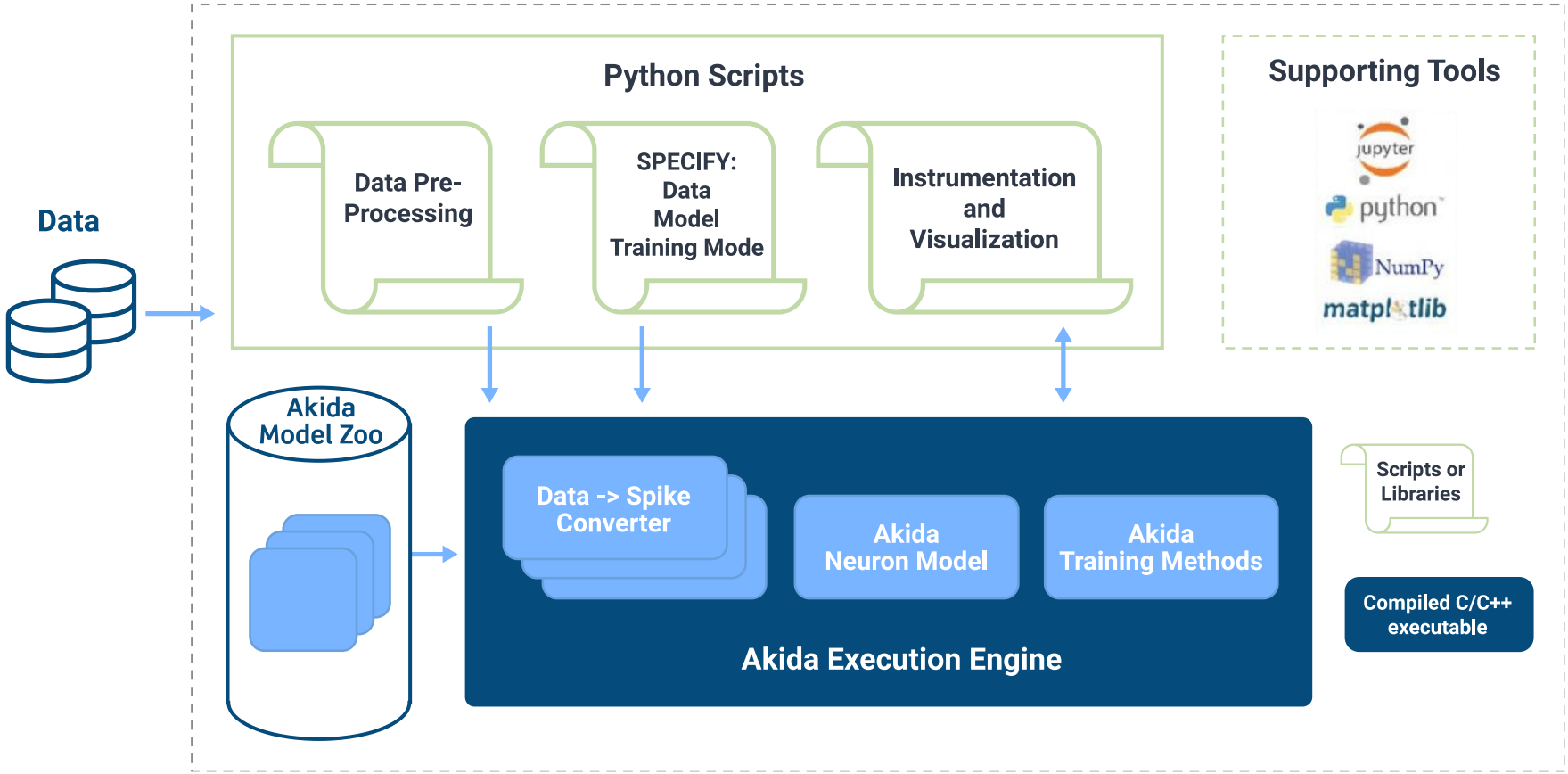
Akida : Configurable Intellectual Property Blocks :

- Configurable number of NPUs
- Programmable Data to Spike Converter
- Event Based Processing Engines running on a single clock
- Configurable on-chip SRAM memory
- Runs full Native SNN in Hardware
- Runs full converted CNN/DNNs in Hardware
- On chip communication Mesh network
- Process technology independent platform
- Network Size customized to application needs



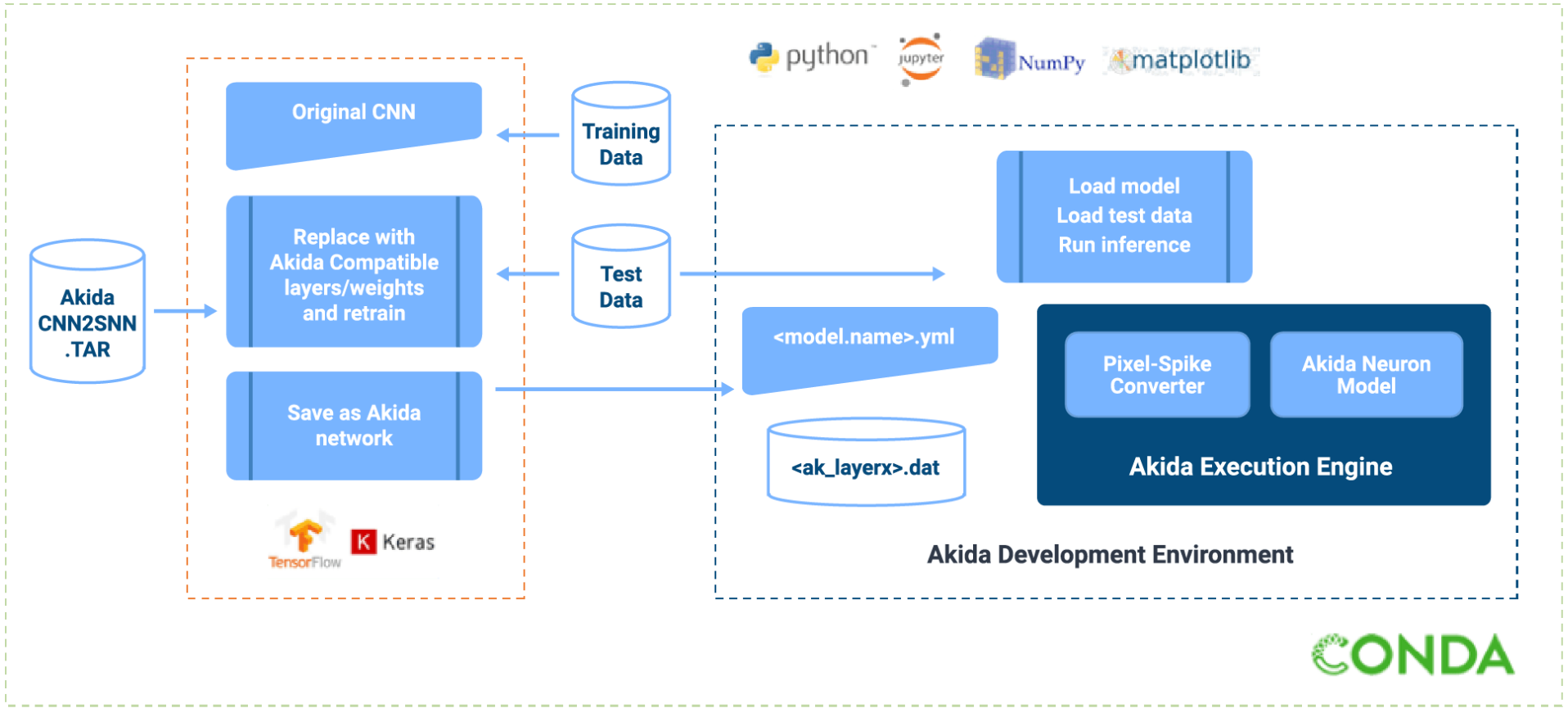
Akida

CHIP SIMULATOR



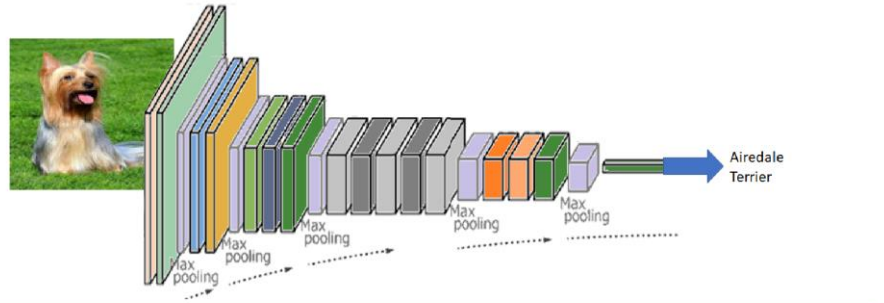
Akida

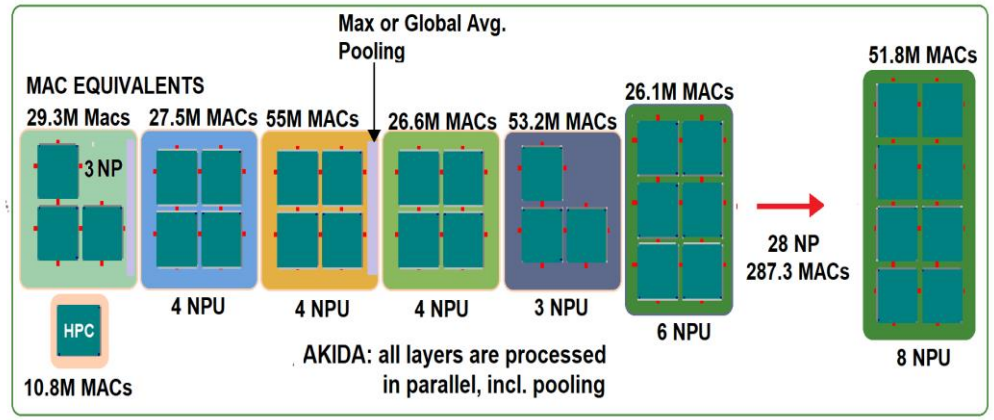
CNN2SNN FLOW



Akida : Mapping CNN to SNN for Inference

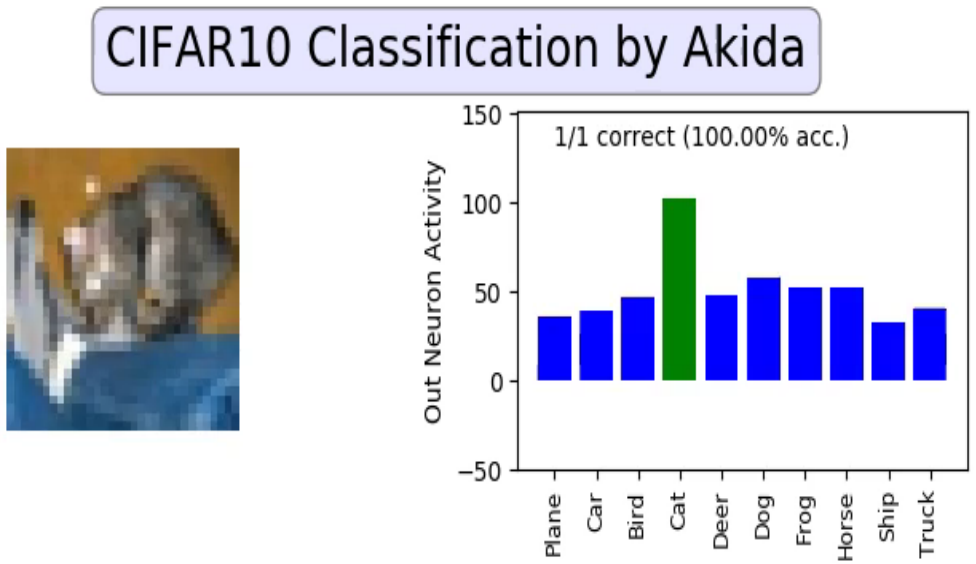
- MobileNet V1
  - 30 layers with many MAC operations
  - Entire network mapped onto separate Akida NPUs (4,208,224 parameters)
  - All layers run in parallel storing results within the distributed SRAM memory on chip
  - Performance from the chip simulator
    - Speed: 80 FPS
    - Power: 434 mw in 28nm
    - Performance: 184 FPS/W
    - **30 FPS @ 157 mw in 28 nm**
- **Power Performance Scales with Process node**
- All power estimates are for Akida NPUs only

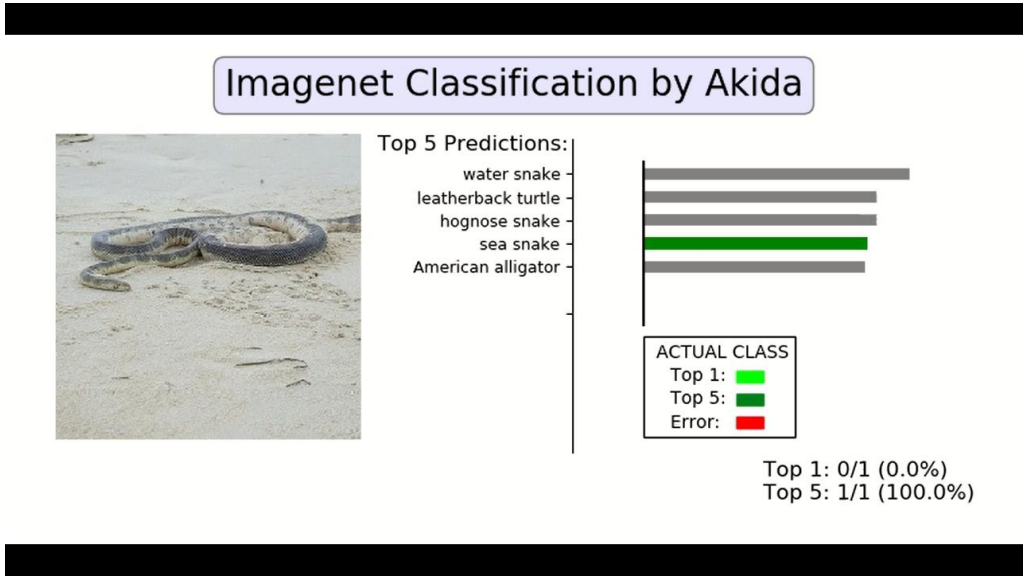




Akida : Edge Inference Simulated Demo

MobileNet and VGG-16 ported to AKIDA

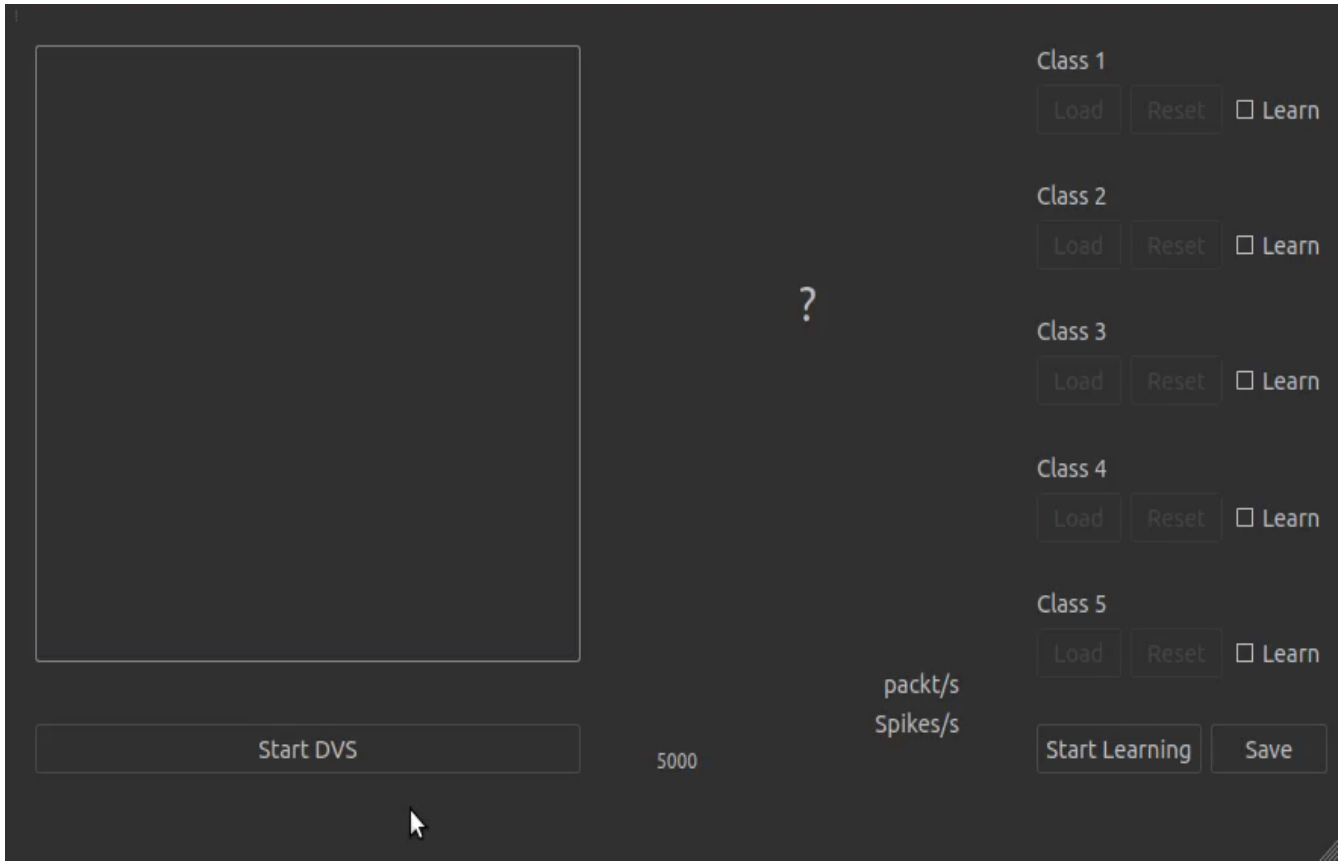




What does this mean? Any visual classification task that is done by a MobileNet or VGG DNN can be ported to AKIDA

Native learning on-chip

ON-CHIP LEARNING APPLICATION USING DVS CAMERA

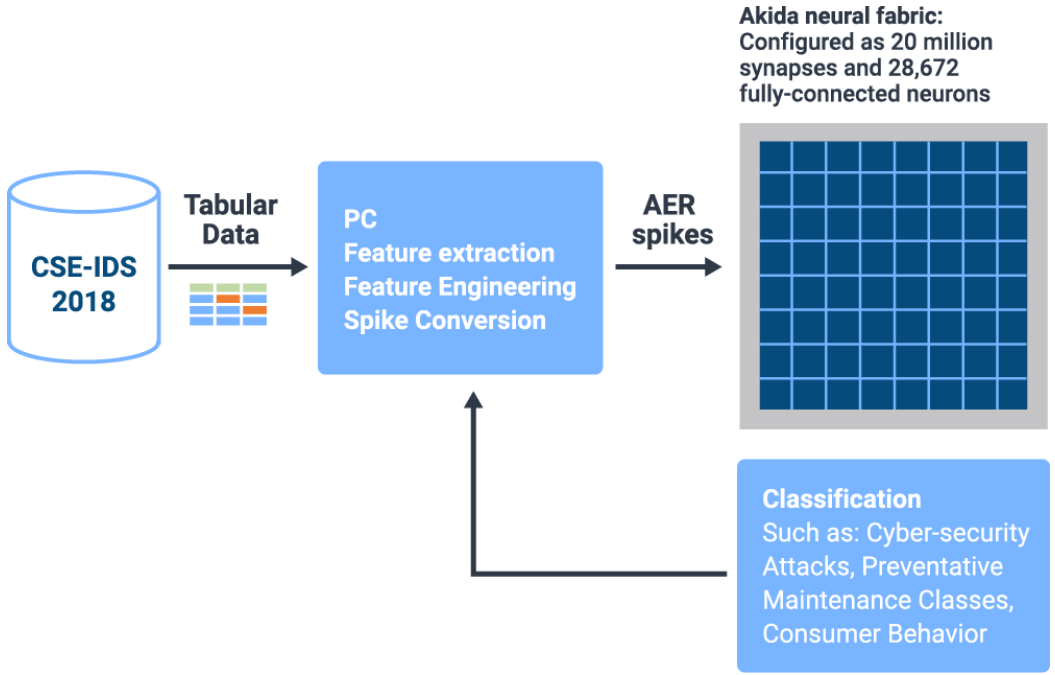


- Akida trained on hand-gestures using event-based camera (DVS)
- Real time learning and detection of patterns
- Uses our proprietary unsupervised learning rule

layer	filters	size	input	output
	(pre-processing)		128x128x1	64x64x1
conv	4	3x3 /1	64x64x1	64x64x4
max pool		3x3	64x64x4	22x22x4
conv	64 x 5 classes	7x7 /4	22x22x4	22x22x(64x5)
global max			22x22x(64x5)	1x1x(64x5)

Akida Native SNN Mode

- Example network is a cyber-security classification task using the CSE-IDS-2018 database (220.8 Gb)
- 15 classes (including ‘normal’)
- Training time is 2.2 hours, single epoch, including preprocessing on a simple i7 laptop with no GPU
- Automatic labeling of active neurons extracted from database
- Accuracy: 98% on test set
- Power: 20 mW at 30,000 inferences/s or 120 mW at 160,256 inferences/s
- 1 inference is one set of data points converted to spikes



Akida Summary - A Holistic Approach to Edge AI

- Changing the game for Edge AI
  - Run Full network on the HW efficiently
- Game changing FPS/W
  - Parallel execution , lower power and real time performance
- Lower minimum power
  - Avoid external memory accesses , reduce bit computation
- Fewer TOPS required for delivering the FPS
- Lower memory requirements, power and size
  - Quantization to 1,2 ,4-bit weights