Biologically Inspired, Digitally Engineered

AKIDA™ Event-Domain Neural Processor

Ultra Low-Power Edge AI Solutions

Anil Mankar

Co-Founder and Chief Development Officer

BrainChip Inc

Agenda

- **Event-based Architecture advantage of AKIDA™ technology**
- **First Brainchip product based on AKIDA™ technology**
  - **AKD1000** Neural System On Chip (NSoC)
- **Example Neural Networks optimized to run on AKIDA™**
- **Discussion**

The Challenges of Edge Computing

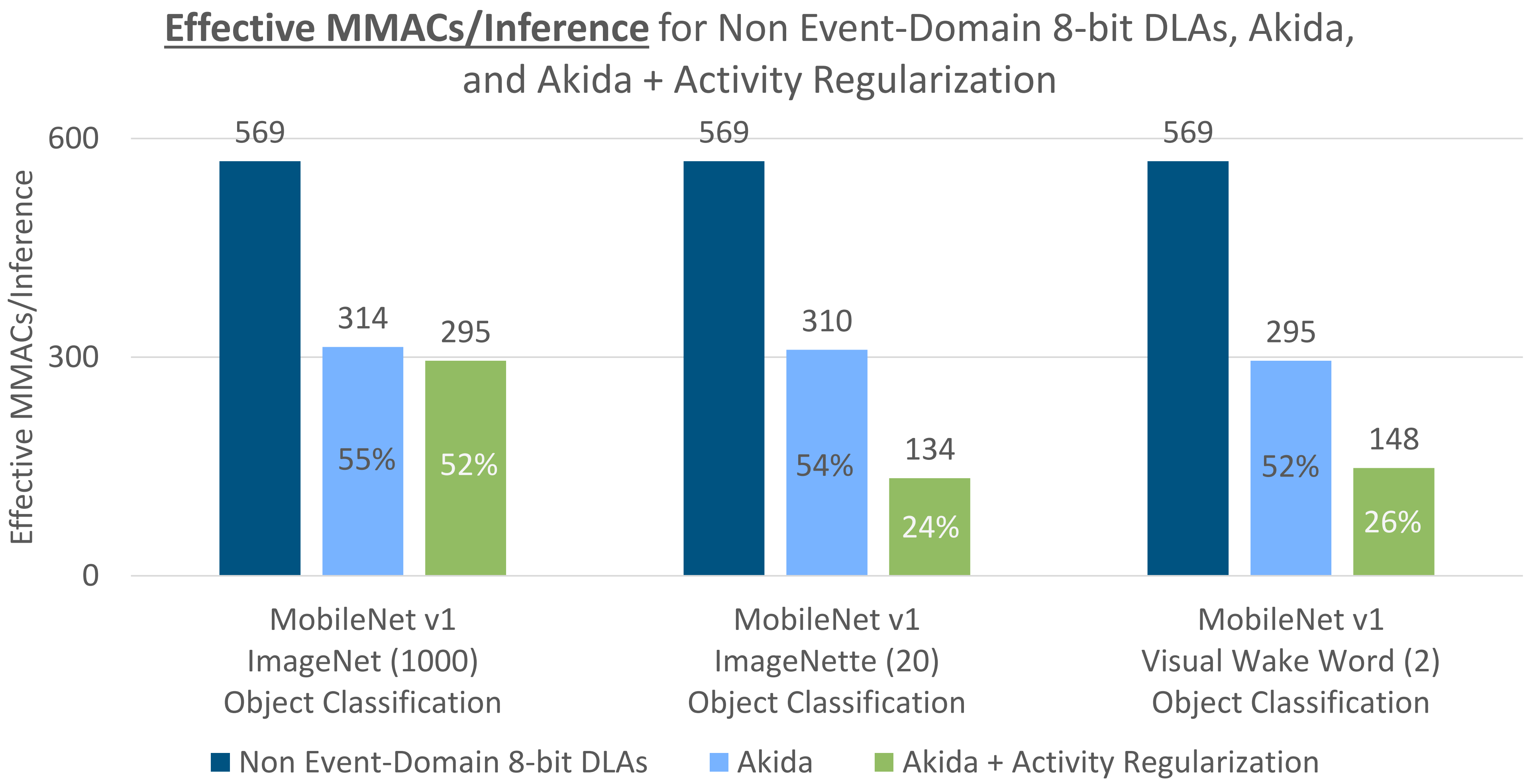
- AI at the Edge, computing requirements require a different solution:
  - Balance an extremely low power budget with real-time performance
  - Operate within severe constraints on memory capacity and bandwidth
  - Off-load tasks from the (limited) CPU
  - Real-time learning or rapid retraining at the edge

- **AKIDA™** overcomes these challenges by adopting a **Neuromorphic Architecture.**
  - **Neuromorphic processing:** event-based processing only consumes power when an event occurs
  - **Run the AI inference** by running it in **event domain**
  - **Reduced memory requirement, 1, 2 or** up to 4 bits for weights and activations
  - **On-chip learning in event domain,** using BrainChip's proprietary algorithms

BrainChip's AKIDA™ Neuromorphic Design Principles

- **Distributed Computation**
  - Computation spread across many cores (neural processing units – NPUs)
  - Each NPU has its own dedicated computational engine and memory, which reduces data movement
- **Event-Based Processing**
  - Non-zero activation map values are represented as multi-bit (1 to 4-bit) events
  - NPUs only perform computation on events, not activation maps
- **Event-Based Communication**
  - NPUs communicate by sending events to each other over a mesh network without host CPU intervention
  - Neural network connectivity is configurable in the field
- **Event-Based Learning**
  - AKIDA implements an on-chip, learning algorithm
  - No costly communication with cloud required

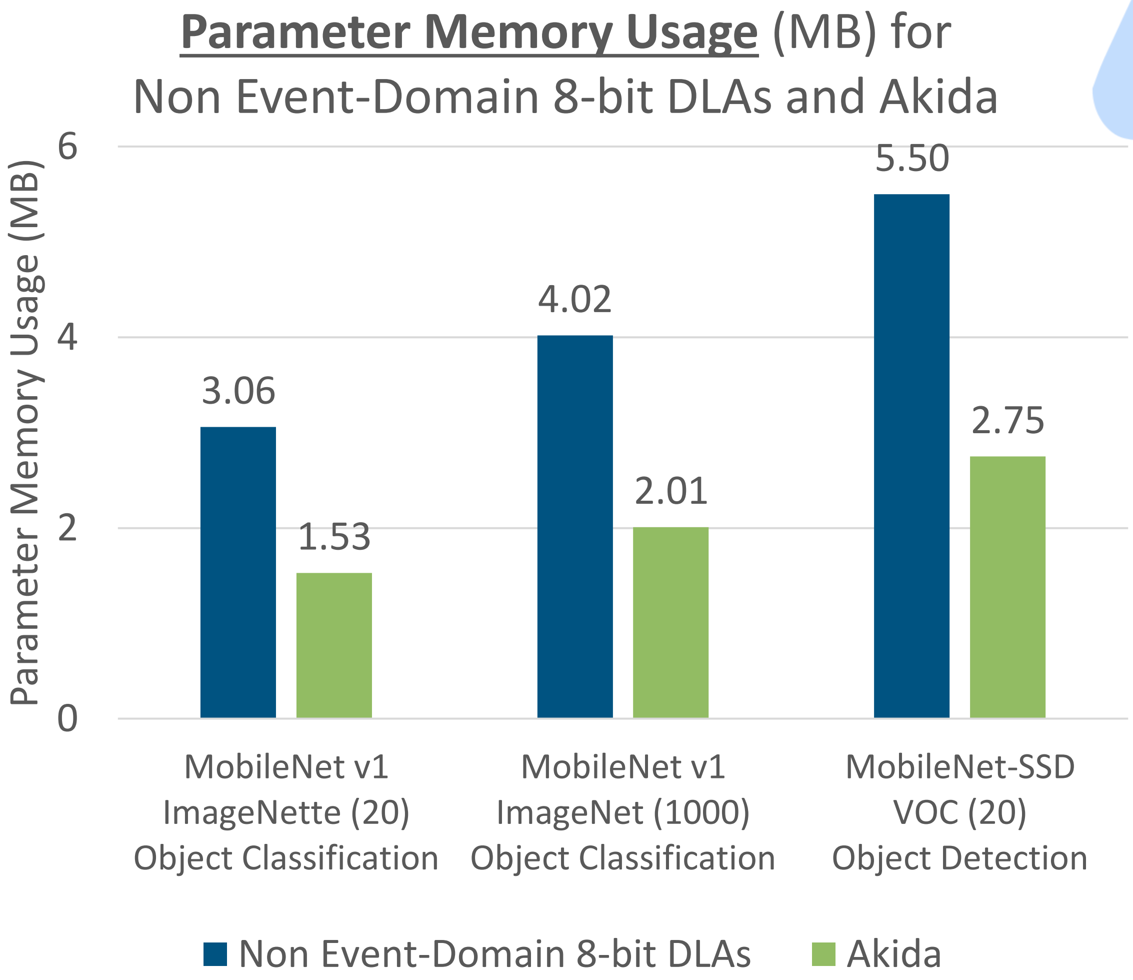
Operation Reduction Effect of Event-Based Processing



AKIDA™ Utilizes Low-Bit Precision to Reduce Memory/Bandwidth

- Akida uses 1-4 bits for activations and parameters
  - 50% (or greater) reduction in memory & bandwidth compared to 8-bit hardware
- We currently perform quantization-aware training to preserve accuracy
- Multiple research groups preserve accuracy with post-training 4-bit quantization*

*Banner, R., et al (2019) Advances in NIPS



https://www.technologyreview.com/2020/12/11/1014102/ai-trains-on-4-bit-computers/

Selected BrainChip Quantization Results

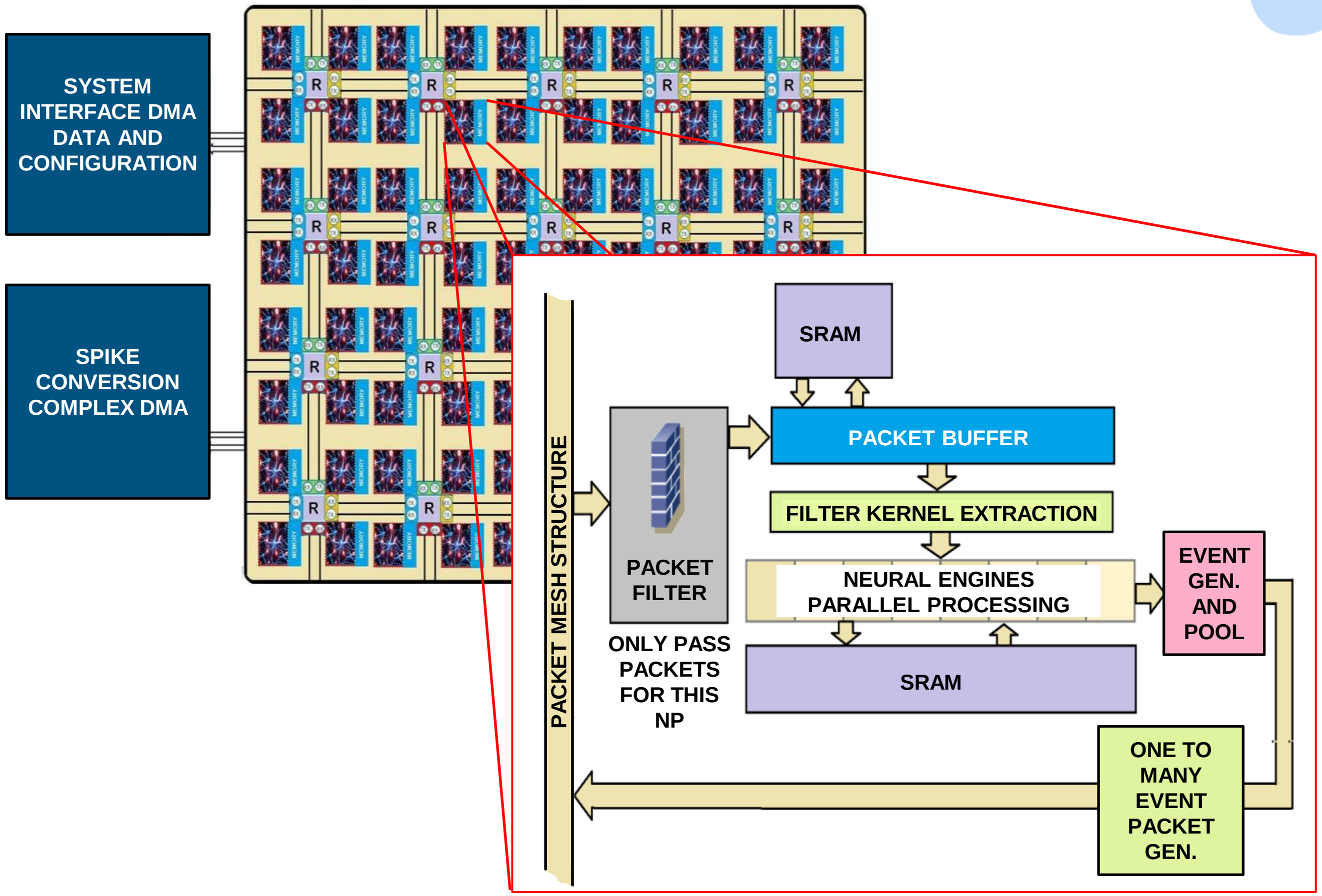
Model	Dataset	# Classes	Weight/Activation Quantization	Quantized Accuracy	32-Bit Float Accuracy
DS-CNN 47K parameters	Google Speech Commands	33	4/4	91.9%	93.4%
MobileNet 224 0.25 200K parameters	Visual Wake Word	2	4/4	89.7%	90.7%
MobileNet V1 2.7M parameters	CIFAR10	10	4/4	93.1%	93.5%
MobileNet V1 4.2M parameters	ImageNet 1000	1000	4/4	68.8%	71.4%
MobileNet SSD 300 5.8M parameters	VOC	20	4/4	65.4%	66.9%
VGG 14.0M parameters	CIFAR10	10	2/2	90.7%	93.2%

Performance for Small Models on Akida

Data Input Type	Model	Data Set	Num Class	Act. Spars. %	Number of Parameters	Required Parameter Memory	Clock (MHz)	FPS	Top-1 Acc. %
RGB Images	MobileNet v1 α=0.25 R224	Visual Wake Word	2	33	210.4 k	102.7 kB	127.84	10.0	89.7
Spatiotemporal 3D Point Cloud	BRN Hand Gesture CNN (4a/2w Mixed Prec.)	Custom DVS Hand Gesture	N/A	>90	1.7 M	418.3 kB	10.10	10.0	N/A
Spatial 3D Point Cloud	BRN MagikEye CNN	Custom MagikEye Hand Gesture	9	91	283.4 k	138.4 kB	7.49	10.0	~90
Audio MFCC	DS-CNNs	Google Speech Commands	33	61	47.2 k	26.6 kB	3.96	10.0	91.9
Resistance Time-Series	Fox 3000 Olfactory ANN	Fox 3000 Olfaction	20	0	138.2 M	16.4 MB**	0.68	10.1	98.2
Accelerometer Time-Series	BRN Custom Bearing Fault Detection CNN	CWRU Bearing Fault Detection	10	50*	139.8 k	68.3 KB	5.53	10	90.0

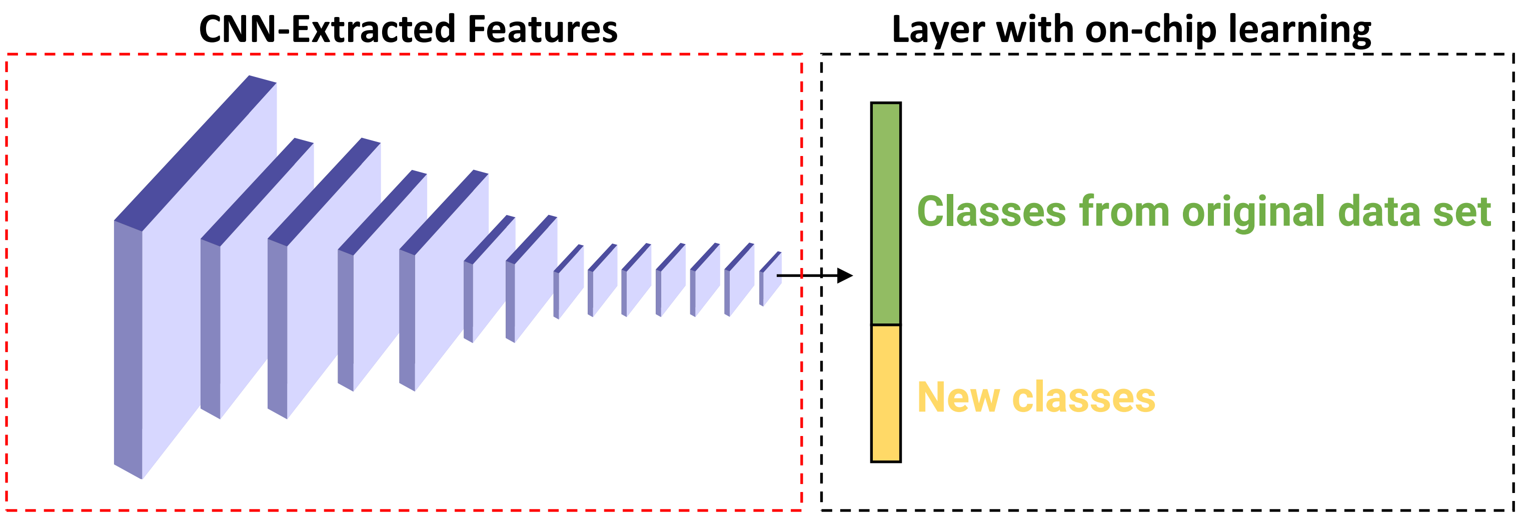
BrainChip's AKIDA™ NPU Architecture and IP solution

- NPUs communicate via a mesh network
- Layers distributed across multiple NPUs
- Each NPU has:
  - 100 KB of local SRAM for:
    - Parameters and activations
    - Internal event buffers
  - Eight compute engines running in parallel
  - Input event packet processing
  - Output event generation
  - Dedicated learning hardware
- Each NPU can be configured to process
  - 2D convolutional layers
  - Dense layers



Edge Learning with AKIDA™ On-Chip Learning

1. Train CNN feature extractor offline on original dataset
2. Replace last classifier layer with Akida layer capable of on-chip learning
3. Perform few-shot learning: learn from a few samples
    a) original classes (green)
    b) new classes (yellow) – should share similar features with original classes



- We have demonstrated edge learning for:
    - Object detection using MobileNet trained on the ImageNet dataset
    - Keyword spotting using DS-CNN trained on the Google Speech Commands dataset
    - Hand gesture classification using small CNN trained on a custom DVS events dataset

AKIDA™ Software Development Environment (ADE) and Training Workflow

Akida Software Development Stack

Akida™ Chip Simulator
pip install akida

Training tool (CNN2SNN)
pip install cnn2snn

Models
pip install akida-models

CNN2SNN Training Tool Workflow

Original NN Model

↓

Create Akida Compatible model with BrainChip TF/Keras API

↓

Perform Quantization-Aware Training (1 to 4-bit)

↓

Save trained network in Akida format

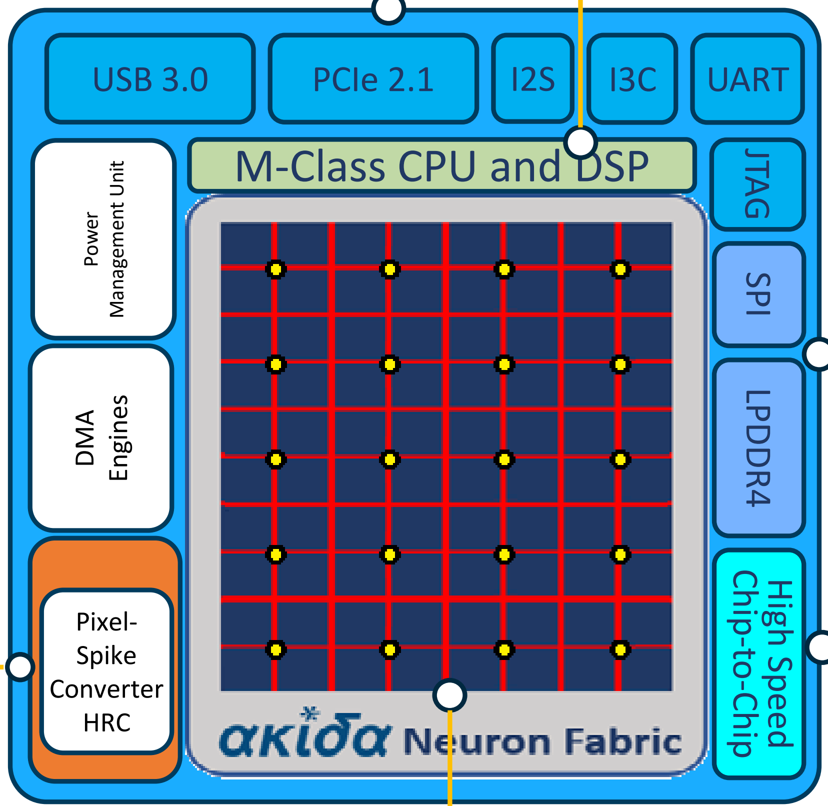
https://doc.brainchipinc.com/

AKIDA™ Based AKD1000 NSoC Chip Block diagram

Data Interfaces
- PCI-Express 2.1 x2Endpoint
- USB 3.0 Endpoint
- I3S, I2C, UART, JTAG

- Pixel-Spike Converter
- Generic Data-Spike Converter
  - Any multivariable digital data

Flexible Akida Neuron Fabric
- Event based mesh interconnect
- All-Digital logic with distributed SRAM, event-based Convolution
- User configurable, feed forward
- 20 nodes of 4 NPUs. 8MBytes
- 300 Mhz in TSMC28nm

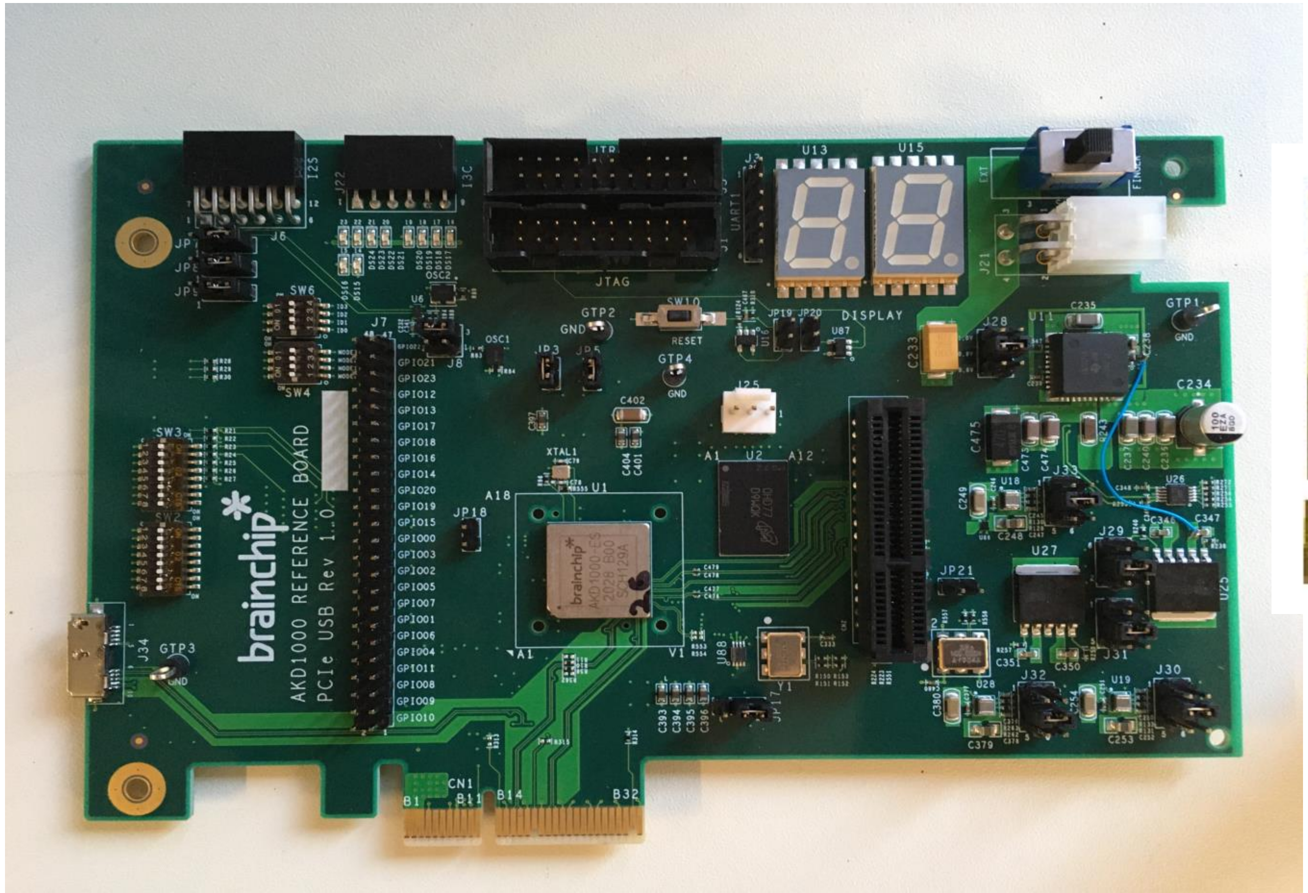


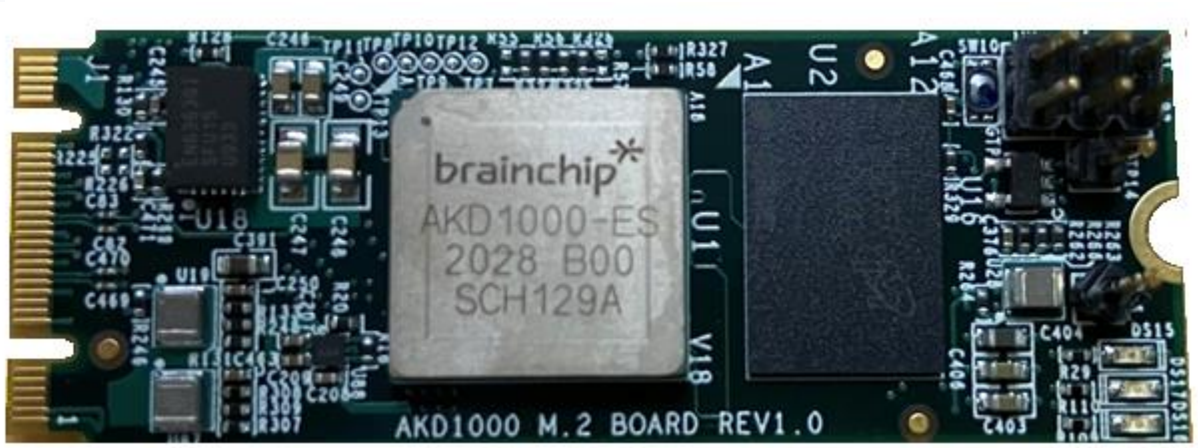
On-Chip Processor
- M-Class CPU with FPU & DSP
  - System management
  - RTOS, preprocessing

Memory Interfaces
- SPI FLASH for boot/storage
- LPDDR4 Additional Weights

Multi-Chip Expansion
- High-speed serial connection
- Connects up to 32 devices

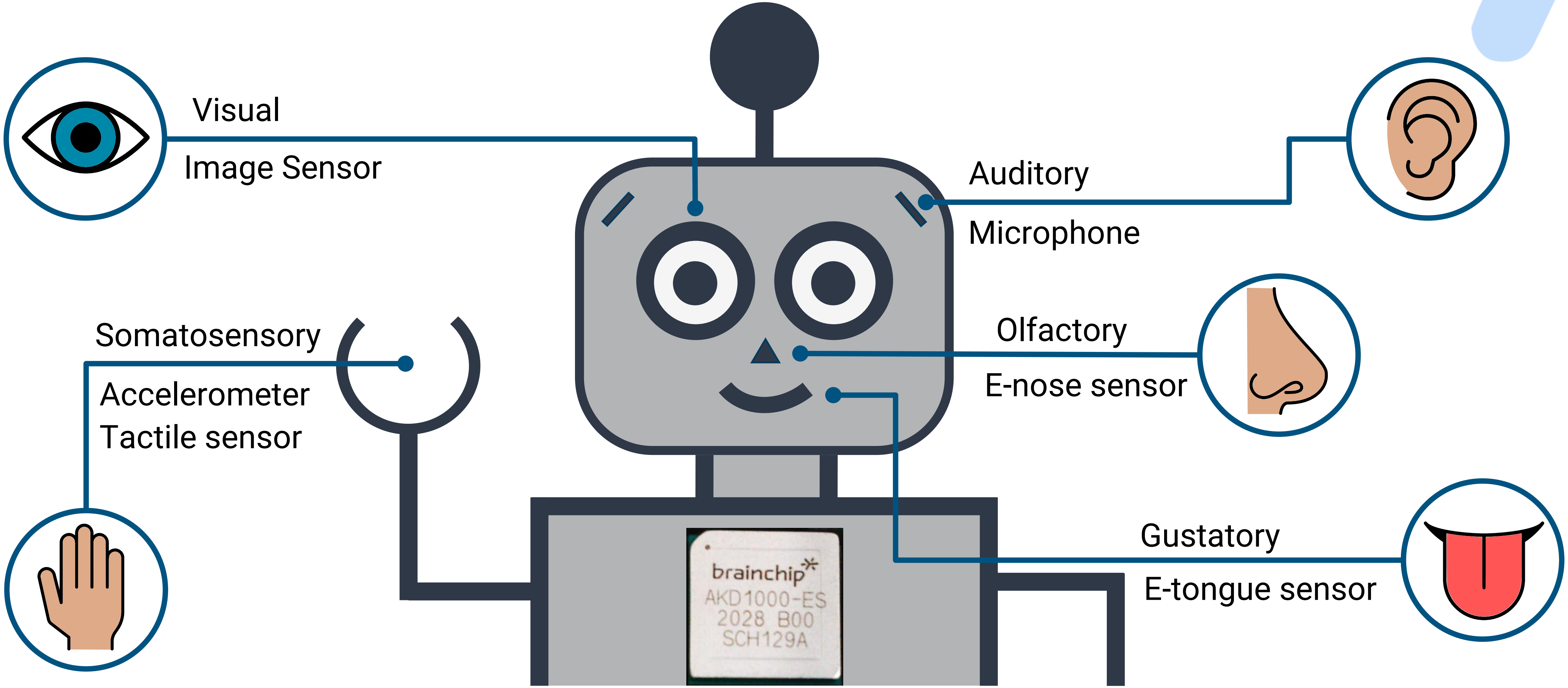
AKD1000 based PCIe Plug in CARD





Akida Applications

AKIDA™ Enables Efficient Processing of All Sensor Modalities



Keyword Spotting: Always on Listening to Microphone

- Google Speech Commands Data Set*
  - 65k 1-second long audio clips of 30 keywords
  - Each keyword has ~1,500 – 4,000 samples
  - Data set is split into training/validation/testing in an 80/10/10 ratio
- Class structure – 12 classes
  - 10 classes for 10/30 words
  - 1 class for silence
  - 1 class for 'unknown word' that represents the other 20/30 words

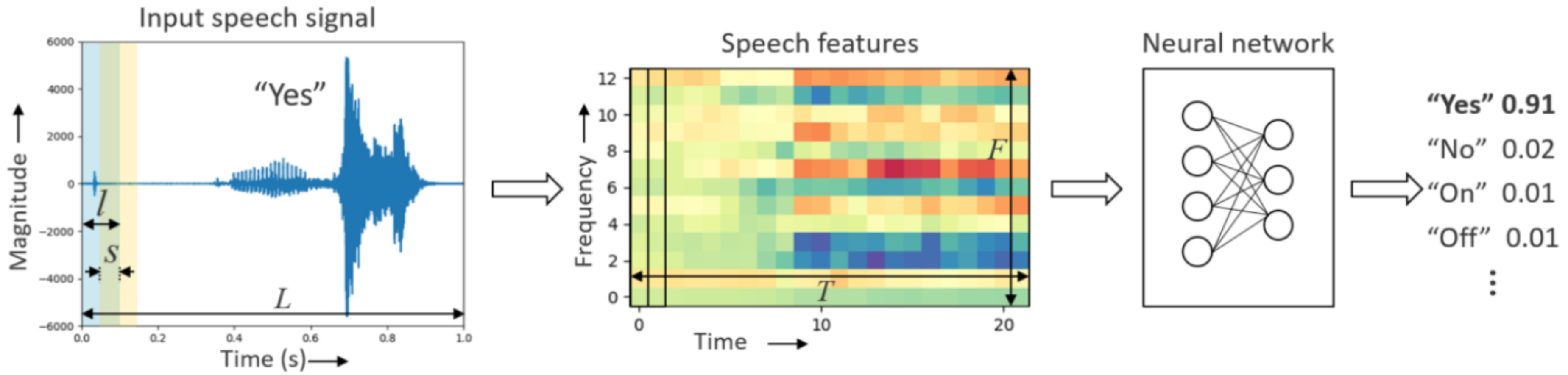


Figure 1: Keyword spotting pipeline.**

* Warden, Pete. 2018. *ArXiv:1804.03209 [Cs]*, http://arxiv.org/abs/1804.03209.
** Zhang, Yundong, Naveen Suda, Liangzhen Lai, and Vikas Chandra. 2018. *ArXiv:1711.07128 [Cs, Eess]*. http://arxiv.org/abs/1711.07128.

Keyword Spotting Network on AKIDA™ Chip: See Live Demo

- KWS Network
  - DS-CNN with 8 layers (4-bit act./wt.)
  - 47k parameters
  - 7 NPUs and ~55 kB of SRAM
- Results
  - Top-1 Accuracy
    - Floating point: 93.4%
    - 4-bit/4-bit weights/activations: 91.9%
    - Speed: 10 FPS and 100 ms latency
    - Activation Sparsity: 61%
    - Dynamic Power: 167 µW
    - Efficiency: 16.7 µJ/Inference

DS-CNN

Layer	Output Dim
Input	49x10x1
Conv MP 5x5	25x5x32
DWS Conv 3x3	25x5x64
DWS Conv 3x3	25x5x64
DWS Conv 3x3	25x5x64
DWS Conv 3x3	25x5x64
DWS Conv 3x3	25x5x64
GAP	1x1x64
DWS Conv 3x3	1x1x256
Dense	1x1x33

Total Params = 47,232

Person Detection: Always on Camera Input

- Visual Wake Word Data Set*
  - Person detection data set
  - Generated from COCO 2014 data set
  - 115k images for training and validation
- Class structure – 2 classes
  - Person class
  - Not-Person class

Some Example Images from the COCO training set*

Person

Not-Person

*Chowdhery, et al. 2019. "Visual Wake Words Dataset." *ArXiv:1906.05721 [Cs, Eess]*, June. http://arxiv.org/abs/1906.05721.

Visual Wake Word and MobileNet 0.25 on AKIDA™ Simulator

- Person Detection Model
  - MobileNet v1 0.25 at 96x96x3 (4-bit act./wt.)
  - 210k parameters
  - 14 NPUs and ~178 kB of SRAM
- Results
  - Top-1 Accuracy
    - Floating point: ~76%
    - 4-bit/4-bit weights/activations: 75.9%
  - Speed: 10 FPS and 100 ms latency
  - Activation Sparsity: 32%
  - Dynamic Power: 1.5 mW

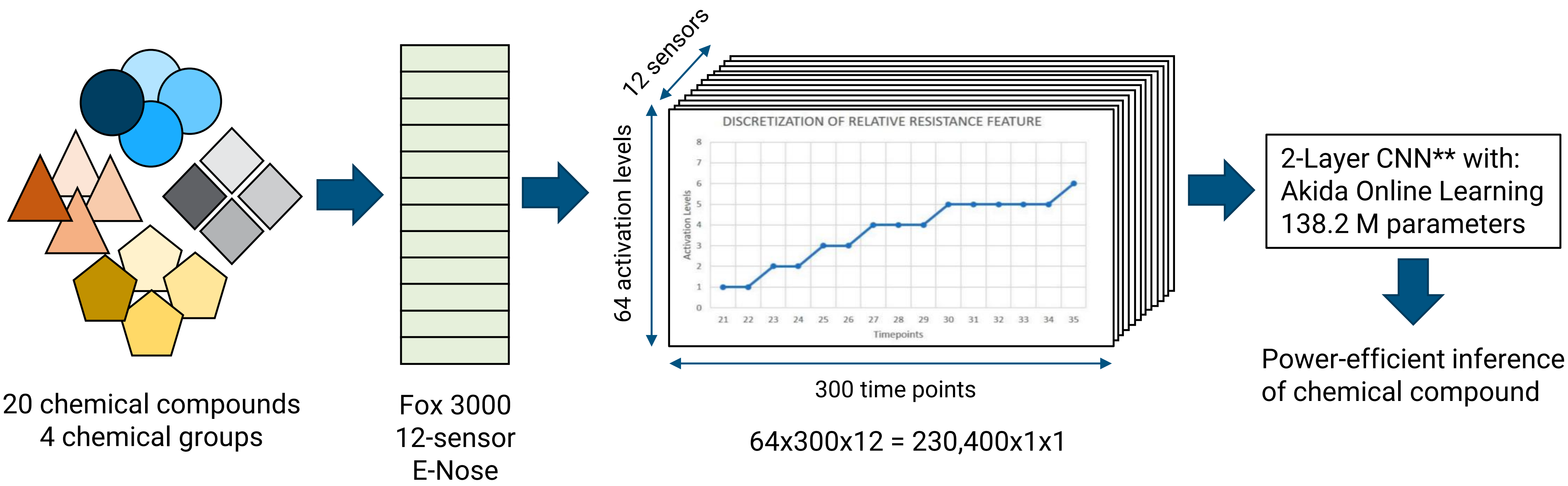
MobileNet 0.25 at 96x96x3

Layer	Output Dims	Filter Dims Stride	Number of Repeated Layers
Input	96x96x3	N/A	1
Conv 2D	48x48x8	3x3x3/2	1
DW Conv 2D	48x48x16	3x3x3/1	1
DW Conv 2D MP	24x24x32	3x3x3/2	1
DW Conv 2D	24x24x32	3x3x3/1	1
DW Conv 2D MP	12x12x64	3x3x3/2	1
DW Conv 2D	12x12x64	3x3x3/1	1
DW Conv 2D MP	6x6x128	3x3x3/2	1
DW Conv 2D	6x6x128	3x3x3/1	5
DW Conv 2D MP	3x3x256	3x3x3/2	1
DW Conv 2D	3x3x256	3x3x3/1	1
GAP	1x1x256	3x3x1/N/A	1
Fully	1x1x2	1x1x256/N/A	1

Total Params = 210,416

Fox 3000 E-Nose Olfactory Classification

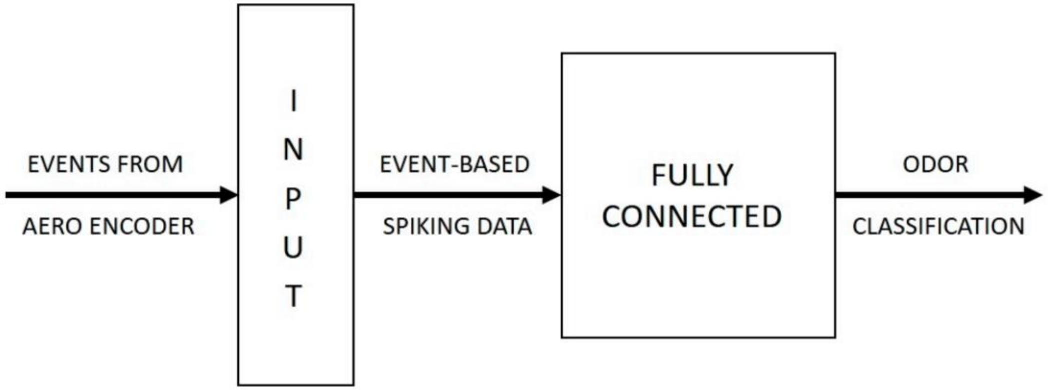
- Fox 3000 Olfactory Data Set*
  - 20 chemical compounds
  - 10 samples per compound
  - Each sample is 300 time points
  - 150 sec sampled at 2 Hz



*https://doi.org/10.4225/08/552C4424EE51E
**Vanarse et al. (2019) *Sensors*.

Fox 3000 E-Nose Olfactory Classification

- Training on Fox 3000 Olfactory data set*
  - 2-layer CNN with:
    - 1-bit weights
    - 1-bit activations
    - Single input layer connected to fully connected layer
  - Trained with Akida learning rule



20-Class 2-Layer CNN

Layer	Output Dim
Input	230,400x1x1
Dense	600x1x1

Total Params = 138.2 M

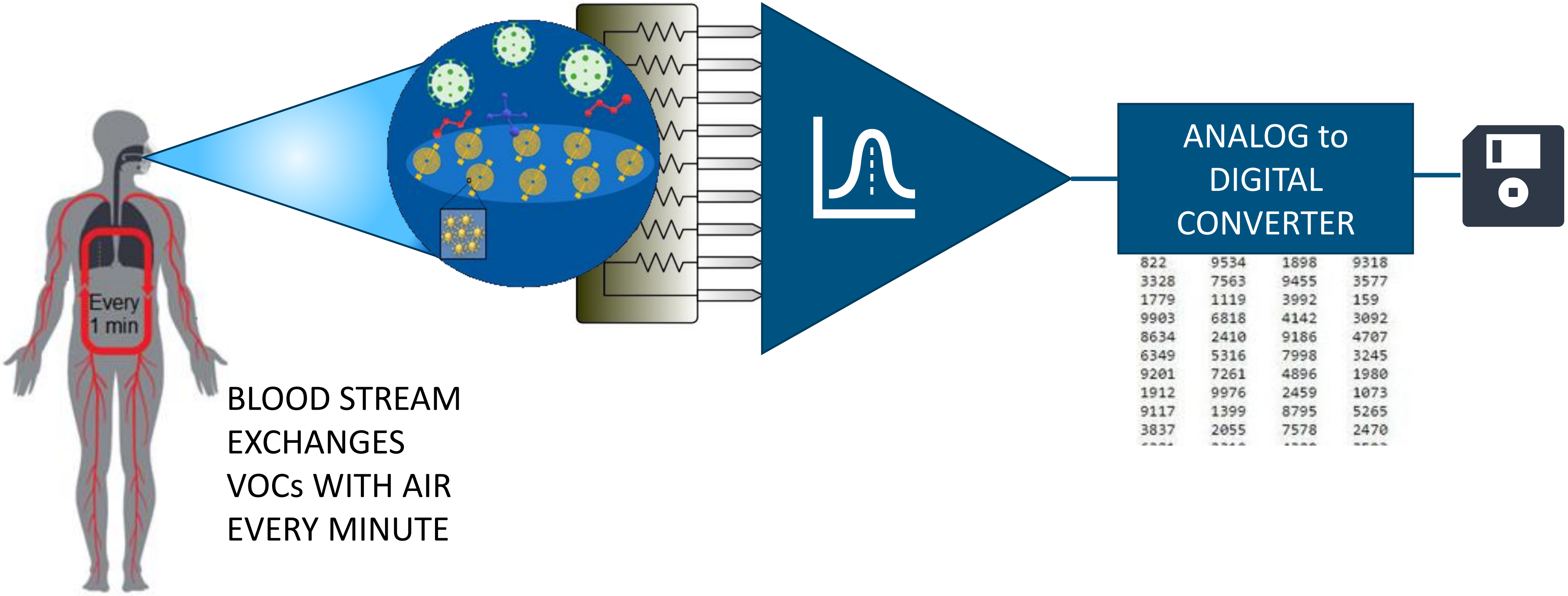
20-Class Results**

- 98.2% Top-1 accuracy
- 1 NPU/16.5 MB Ext. RAM
- 10 FPS/100 ms latency
- Dynamic power: 7.0 mW
- Clock Freq: 43.2 MHz
- Batch Size = 1

*https://doi.org/10.4225/08/552C4424EE51E
**Vanarse et al. (2019) *Sensors*.

COVID-19 and Akida, Collection Method

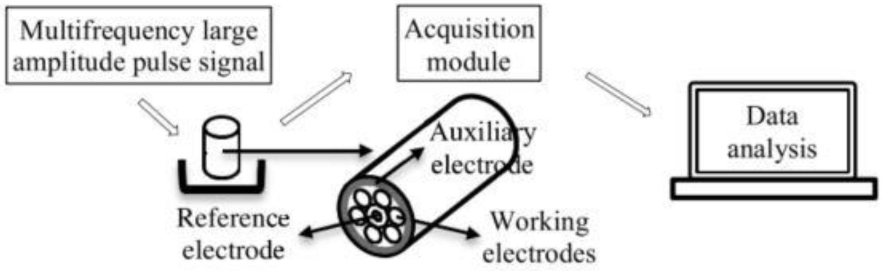
EXHALED BREATH -> MONOLAYER-CAPPED GNP SENSORS-> OPAMP -> ANALOG TO DIGITAL CONVERTER -> STORAGE



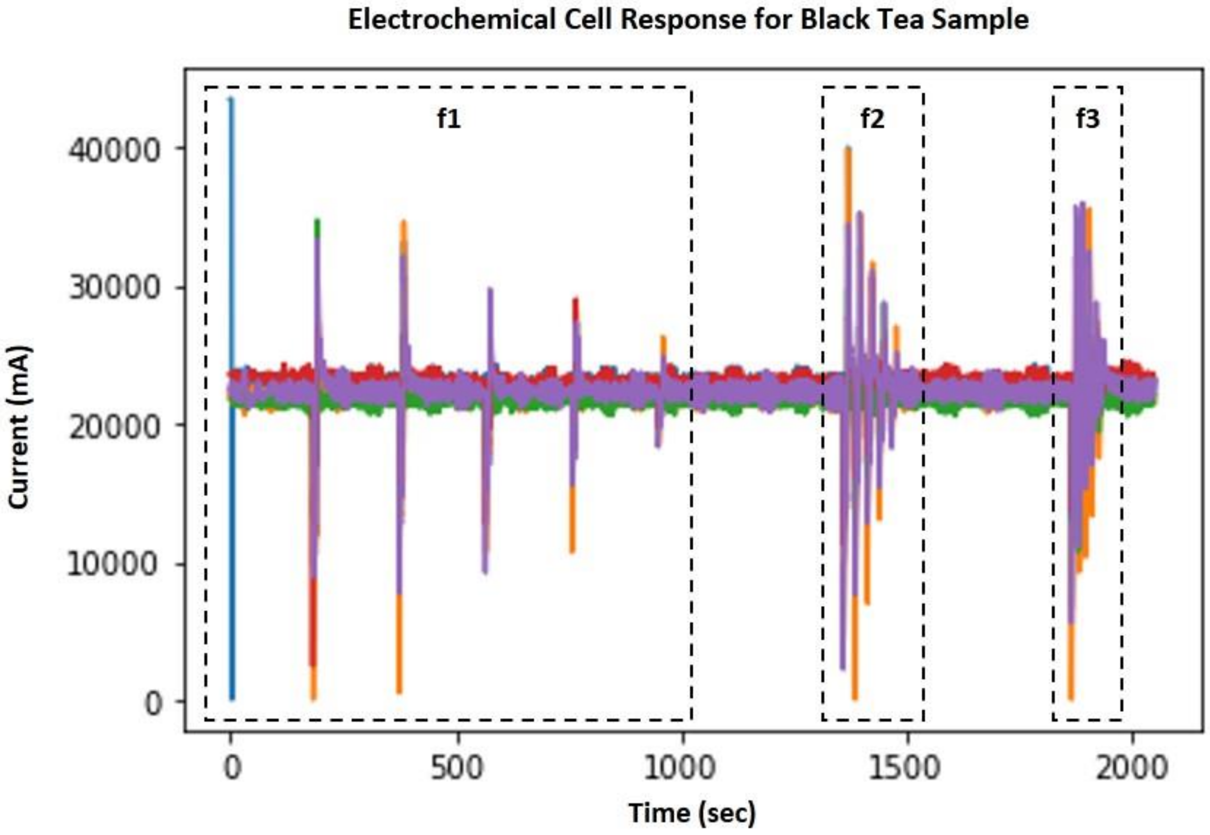
ACS Nano 2020, 14, 9, 12125-12132
Multiplexed Nanomaterial-Based Sensor Array for Detection of COVID-19 in Exhaled Breath

Taste Classification with E-Tongue Systems

- Voltammetry obtains information about a sample by measuring the current as the potential is varied
- Data set* composed of a series of pulse voltammetry waveforms
- Pulse signals comprised of:
  - Three frequencies
  - Five voltage amplitudes



Framework* for e-nose system

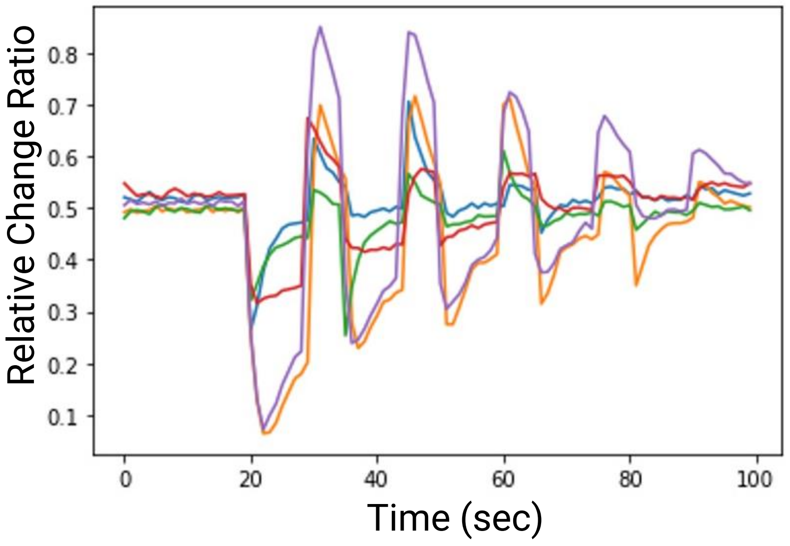


* Adapted from Zhang et al. IEEE transactions on cybernetics, vol. 49, no. 3, pp. 947-960, 2018.

Taste Classification with E-Tongue Systems

- E-Tongue Classification Data Set*
  - 114 measurements from five electrodes
  - 13 types of liquid samples
  - Each sample had 500 elements
    - Five sensors
    - 100 time-points per sensor
- Training
  - Trained with Akida learning rule
  - Train/test split: 70%/30% samples

E-Tongue Data Sample Example



BrainChip 2-Layer ANN

Layer	Output Dim
Input	100x16x5
Dense	1x1x260

Total Params = 2.1 M

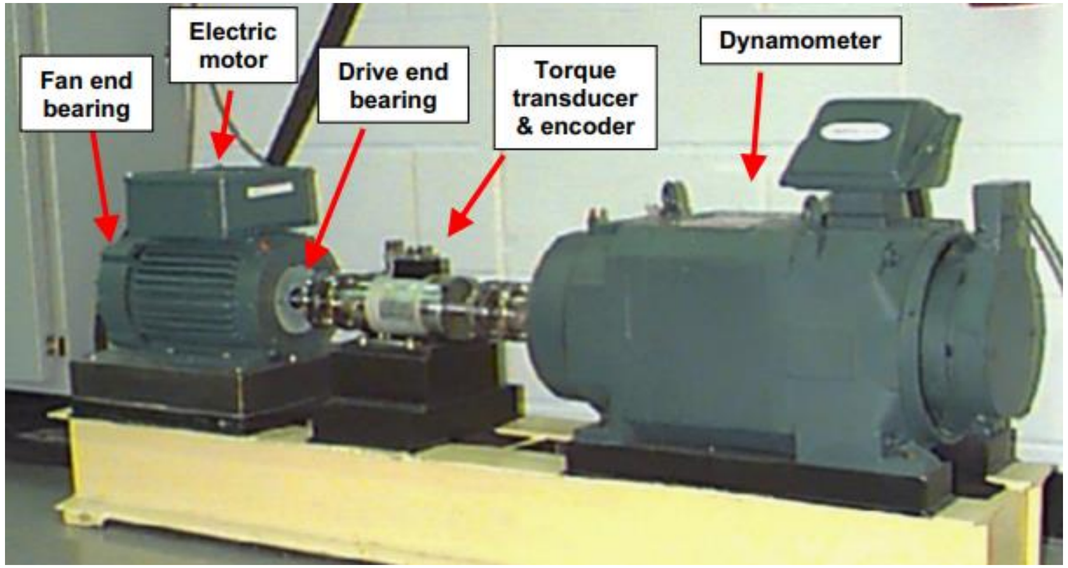
1-bit weights/activations

- Results
  - 95.8% Top-1 accuracy
  - 4 NPUs/254 KB Ext RAM
  - 10 FPS and 100 ms latency
  - Dynamic power = 1.1 mW
  - Batch Size= 1

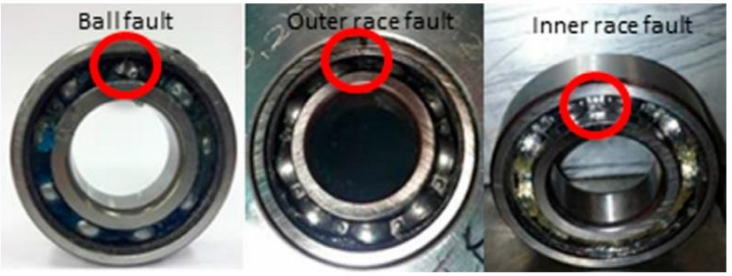
* Zhang et al. IEEE transactions on cybernetics, vol. 49, no. 3, pp. 947-960, 2018.

Electric Motor Ball Bearing Fault Diagnosis

- Condition-based maintenance (CBM):
  - Perform maintenance only when necessary
- This study focuses on diagnosis
- Motor bearings were seeded with faults
- Accelerometer data taken at locations near to and remote from the motor bearings



Taken from CWRU Bearing Data Center Website*



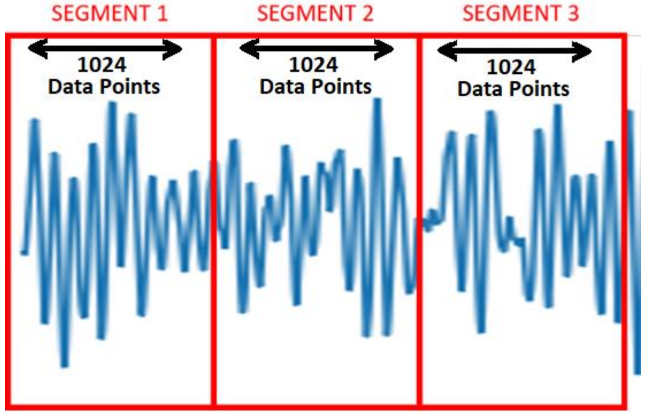
- There were three different fault types
  - Ball defect
  - Inner race fault
  - Outer race fault
- Each fault came in 3 sizes for a total of
  - 9 faulty classes
  - 1 normal class

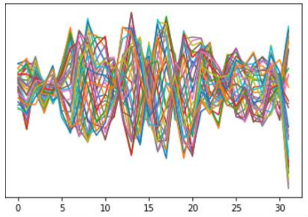
*CWRU Bearing Data Center Website: https://csegroups.case.edu/bearingdatacenter/pages/welcome-case-western-reserve-university-bearing-data-center-website

Ball Bearing Fault Detection

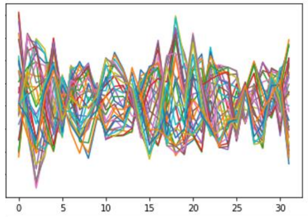
- Ball Bearing Data Set*
  - Accelerometer data collected at 48 kHz
  - 10 classes with 460 samples/class
  - Each sample had 1,024 elements
  - About 10 seconds of data per class
- Training
  - 6-layer CNN with 4-bit weights/ activations trained with back prop.
  - Train/test split: 3,600/1,000 samples

CWRU Bearing Data Example





Class 1



Class 2

BrainChip 6-Layer CNN

Layer	Output Dim
Input	32x32x1
Conv MP 3x3	32x32x32
Conv MP 3x3	16x16x32
Conv MP 3x3	8x8x64
Conv MP 3x3	4x4x64
Conv MP 3x3	4x4x128
GAP	1x1x128
Dense	1x1x10

Total Params = 140,448

- Results
  - 99% Top-1 accuracy
  - ~50% activation sparsity
  - 5 NPUs and 195 KB SRAM
  - 10 FPS and 100 ms latency
  - Dynamic power = 896 μW
  - Batch Size= 1

*CWRU Bearing Data Center Website: https://csegroups.case.edu/bearingdatacenter/pages/welcome-case-western-reserve-university-bearing-data-center-website

AKIDA™ Summary

- **Event-based computation benefits**
  - Run inference in ½ GOPS, ½ memory, and ½ memory bandwidth
  - 3 to 4 times lower power at same clock rate
- **Runtime software manages all configuration and network loading**
  - Application-level API similar to Tensor-flow/Keras
- **Incremental on-chip learning from few samples**
- **Available as a chip AKD1000 or Embedded IP in your SoC**

Questions

www.brainchip.com

@brainchip_inc

For additional information please contact:

Rob Telson

Worldwide Vice President of Sales and Marketing

rtelson@brainchip.com

Empowering Product Creators to Harness Edge AI and Vision

The Edge AI and Vision Alliance (www.edge-ai-vision.com) is a partnership of ~100 leading edge AI and vision technology and services suppliers, and solutions providers

Mission: To inspire and empower engineers to design products that perceive and understand.

The Alliance provides low-cost, high-quality technical educational resources for product developers

Register for updates at www.edge-ai-vision.com

The Alliance enables edge AI and vision technology providers to grow their businesses through leads, partnerships, and insights

For membership, email us: membership@edge-ai-vision.com

Join us at the Embedded Vision Summit
May 25-27, 2021—Online

The only industry event focused on practical techniques and technologies for system and application creators

- *“Awesome! I was very inspired!”*
- *“Fantastic. Learned a lot and met great people.”*
- *“Wonderful speakers and informative exhibits!”*

Embedded Vision Summit 2021 highlights:

- **Inspiring keynotes** by leading innovators
- High-quality, practical **technical, business and product talks**
- Exciting **demos**, **tutorials** and **expert bars** of the latest applications and technologies

Visit www.EmbeddedVisionSummit.com to sign up for updates

2021 embedded vision summit®
VIRTUAL | MAY 25-27