Enabling Ultra-Low Power Edge Inference and On-Device Learning with Akida

Nandan Nayampally

CMO

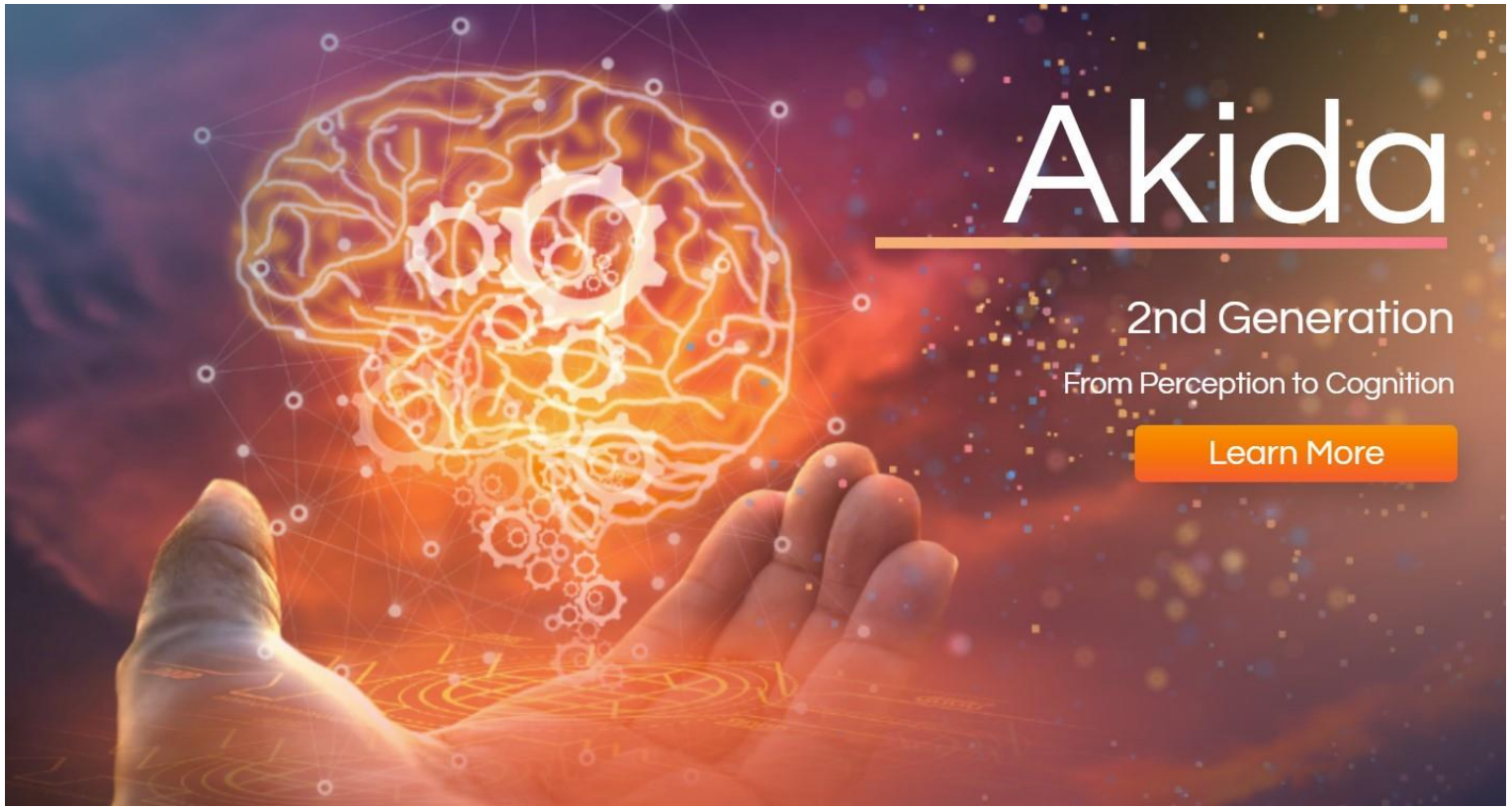
Brainchip Inc.

brainchip

Essential AI

Agenda

- The Challenge
- The Approach
- The Delivery
- The Results
- Akida in action



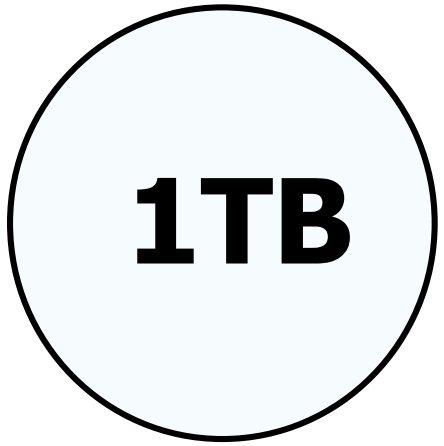
The Challenge

brainchip
Essential AI

The Problem

Costs of training a single High-end model [1]

Annual losses in Manufacturing due to unplanned downtime[2]



Data generated by 1 Car per day[3]

Healthcare cost and lost productivity due to preventable chronic disease[4]

[1] Courtesy: Spectrum.Ieee.org. "The cost of training, made retraining the model infeasible"

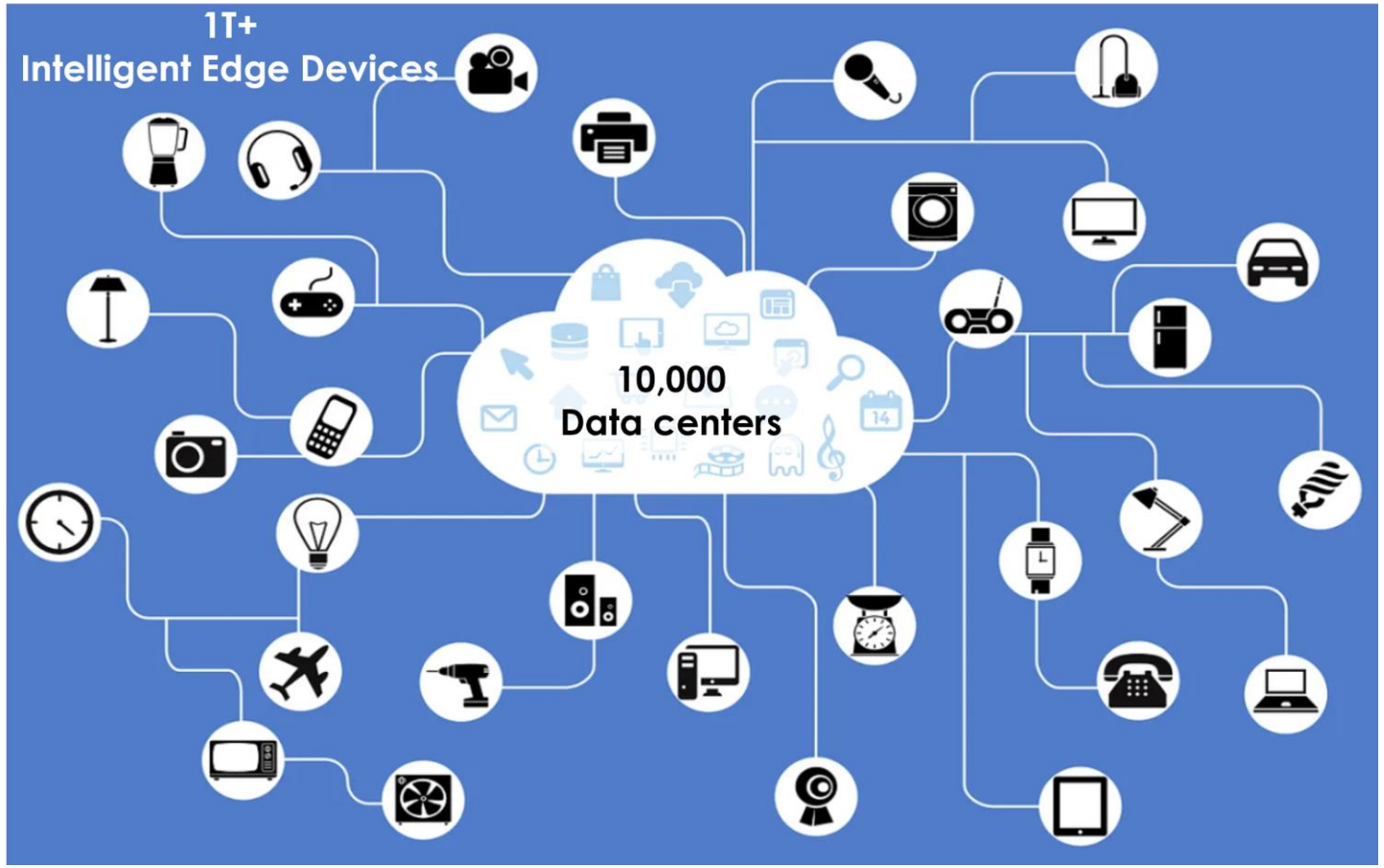
[2] Courtesy: Deloitte.com

[3] Courtesy: Forbes.com

[4] Courtesy: fightchronicdisease.org

brainchip
Essential AI

The Opportunity



$15.7T
Global GDP Benefit from AI in 2030*

1T+
Edge devices in 2030**

$1.2T
AIoT revenue in 2030**

Image: Courtesy Pixabay (From Mckinsey AIoT 2030 forecast)

* PWC analysis report

** Forbes Business Insights

The Challenge

- Cost of cloud services
- Responsiveness & reduced latency
- Scalability & efficiency
- Privacy protection & security

The Challenge

- Cost of cloud services
- Responsiveness & reduced latency
- Scalability & efficiency
- Privacy protection & security

The Solution

- Reduce cloud inference cost
- Minimize cloud retraining
- Rapid computation at edge
- Real-time compute for critical tasks
- Efficiency within thermal and power budgets
- Reduced memory and system cost
- Prevent exposure of sensitive data
- Minimize raw data being sent to cloud

The Approach

brainchip
Essential AI

The Neuromorphic Advantage

- Compelling high-performance
- Extreme efficiency
- Continuous learning
- Secure communication

Event-based processing

Advanced spatial-temporal capability

Event-based communication

At-memory computation

Event-based learning

brainchip
Essential AI

akida

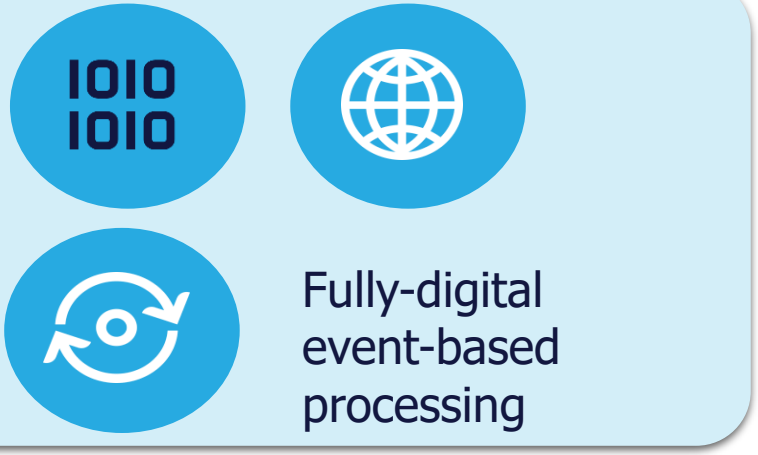
2nd Generation

- Fully-digital, neuromorphic, event-based AI
- Unique ability to learn on device without cloud dependency

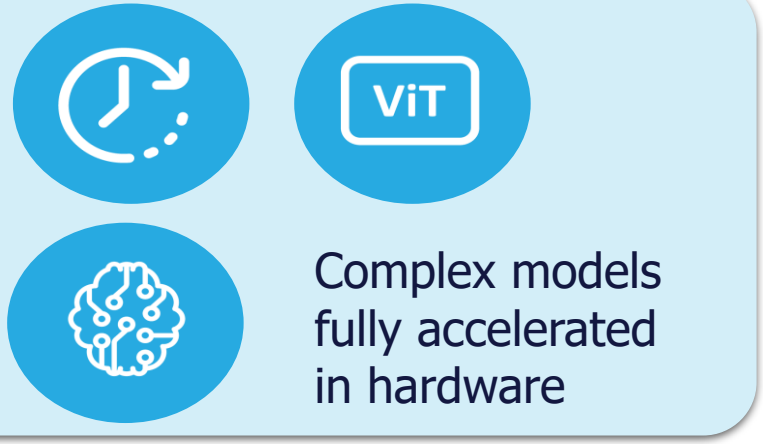
What's New:

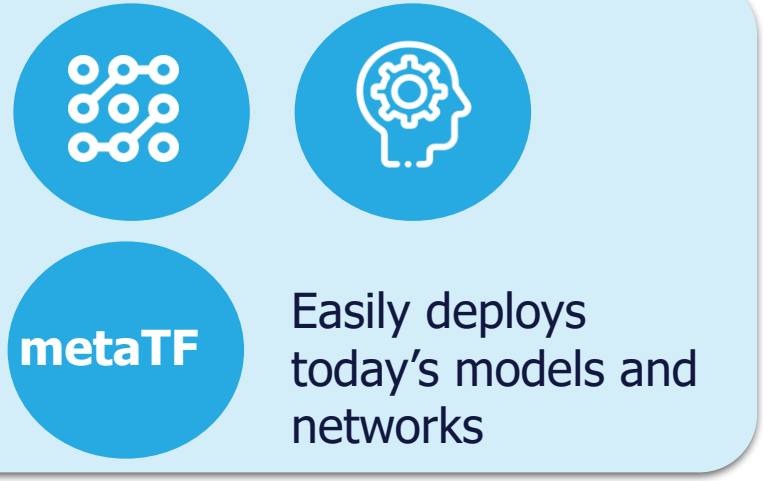
- High performance compute with extreme energy-efficiency on complex models
- Spatial-temporal convolutions that enable innovative handling of 3D and 1D data with Temporal Event-based Neural Nets (TENN)
- Low-power support for vision transformers in edge AIoT
- Improved accuracy with efficiency for production devices

Neuromorphic Principles in Real Solutions



akida

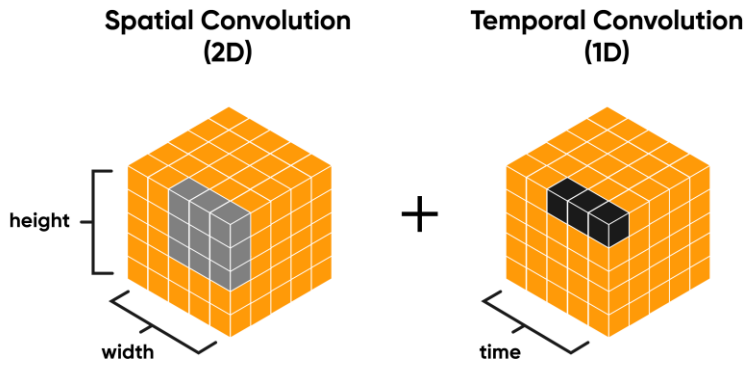




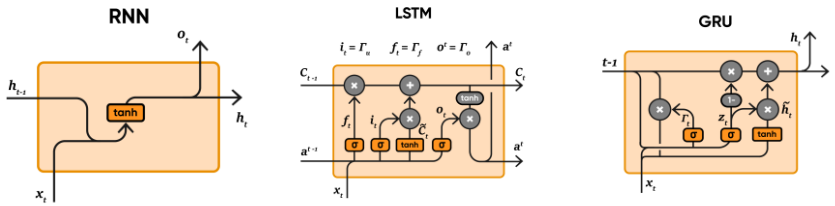
brainchip
Essential AI

Temporal Event Based Neural Nets

Extremely efficient 3D convolutions



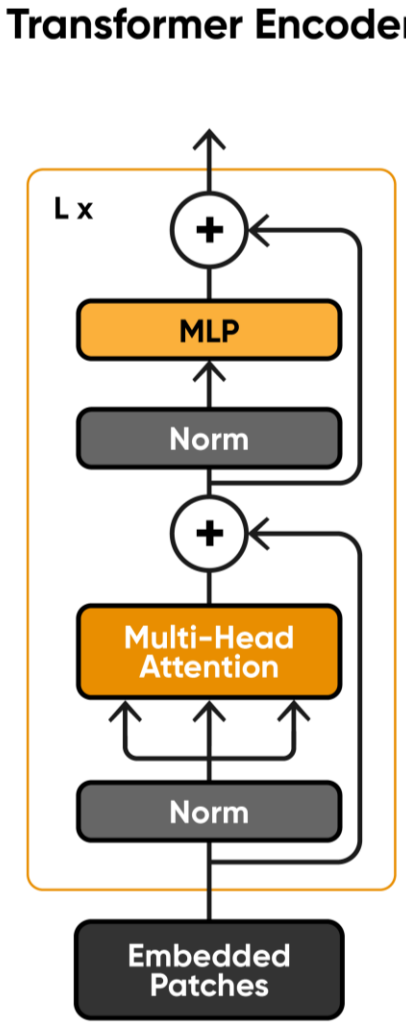
TENNs deliver the benefits of and are much more efficient to train than RNNs



- Easy to train and extremely data-efficient
- 3D data has spatial (frames) and temporal (time) components
  - TENN trains with backpropagation like a CNN
  - Extracts spatial (2D) + temporal (1D) kernel
  - Inference in recurrent mode
- 1D time series data focused on temporal components
  - Training 1D data with backpropagation
  - Extracts temporal kernels
  - Inference in recurrent mode

brainchip
Essential AI

Vision Transformer (ViT) for Efficient Performance Boost



- Vision transformer encoder block functionality fully supported in hardware
  - Optional configurations from 2 Node to 12 Nodes
  - Builds on at-memory compute benefits
  - Encoder block is fully self contained
  - Managed through DMA and runtime

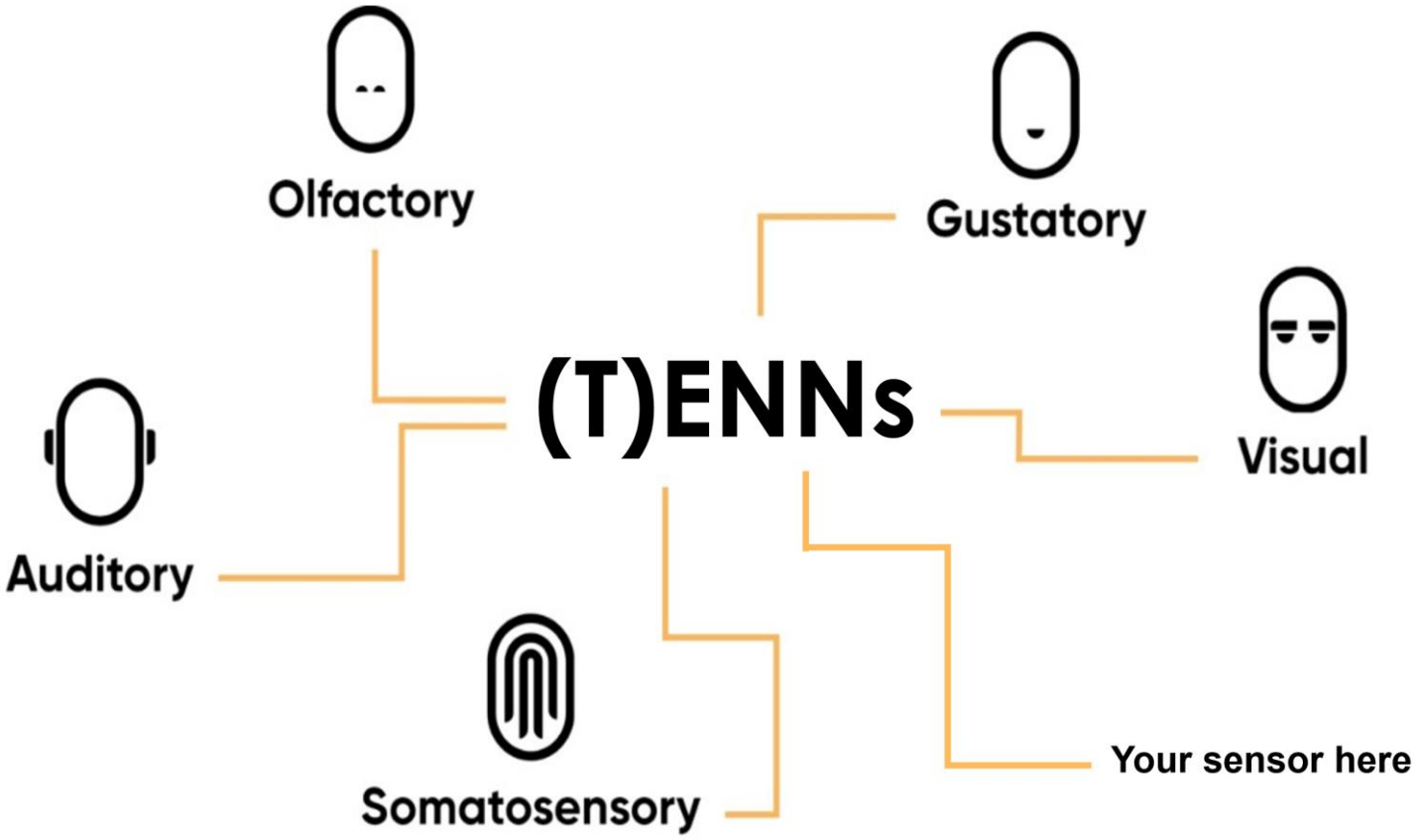
- Delivers power and system efficient performance
  - 2 Nodes running at 800 MHz give 30 FPS performance (224x224x3)
  - No external memory bandwidth needed once layers are loaded
  - Size incremental to standard event-based node.

brainchip
Essential AI

The Delivery

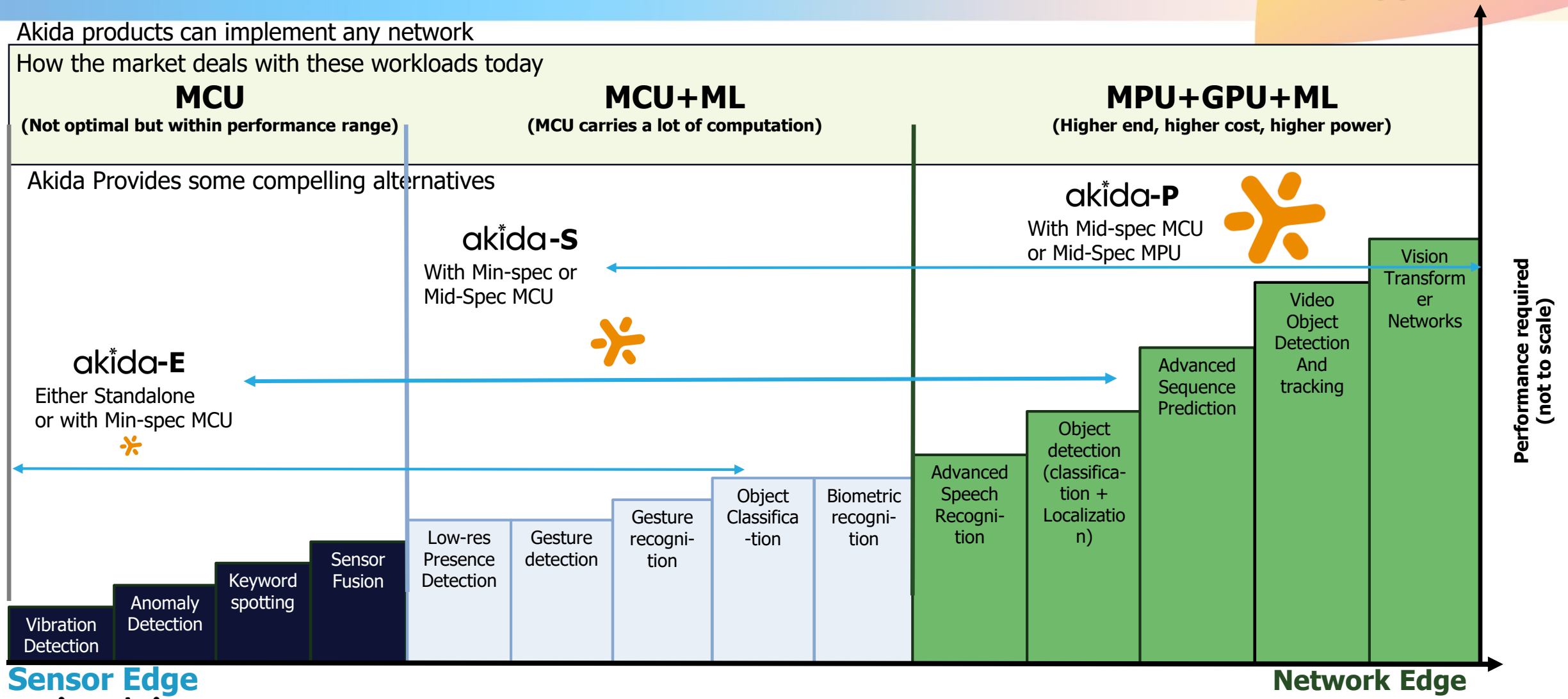
brainchip
Essential AI

Sensor Agnostic AI Efficiency



- Optimal for any network model
- Self-managed operation
- Reduced system Load
- Extreme efficiency
- Simplified development process

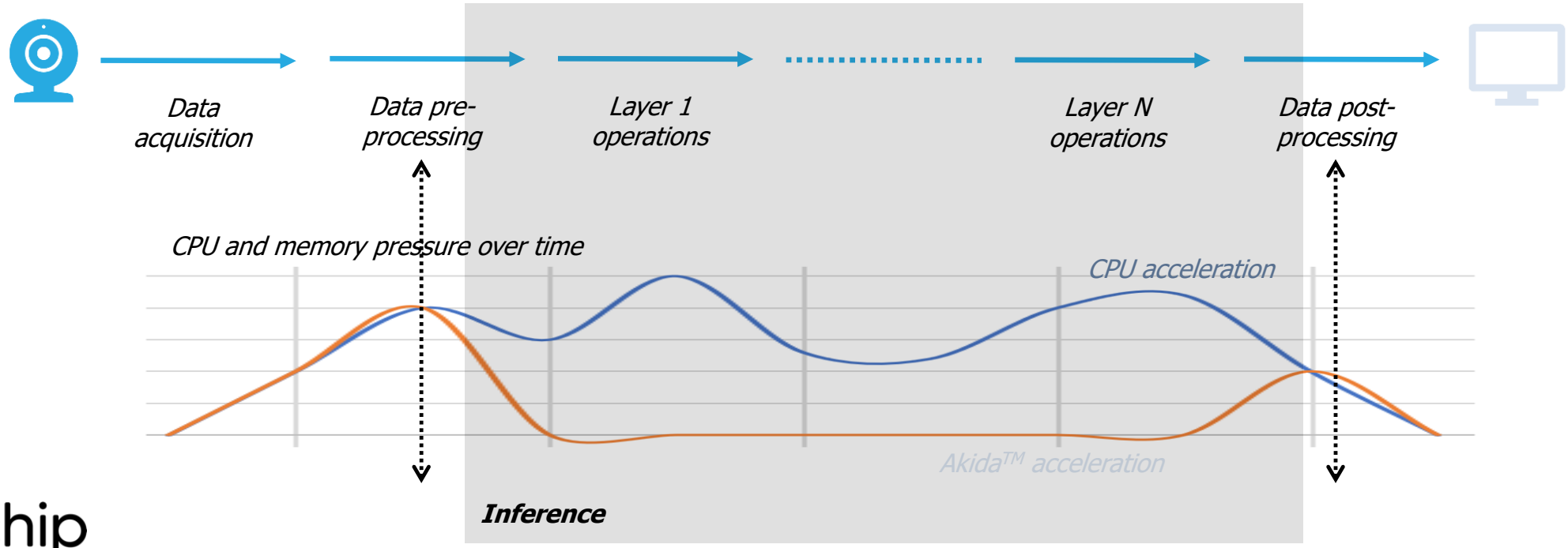
Enabling Compelling Edge AI (example)



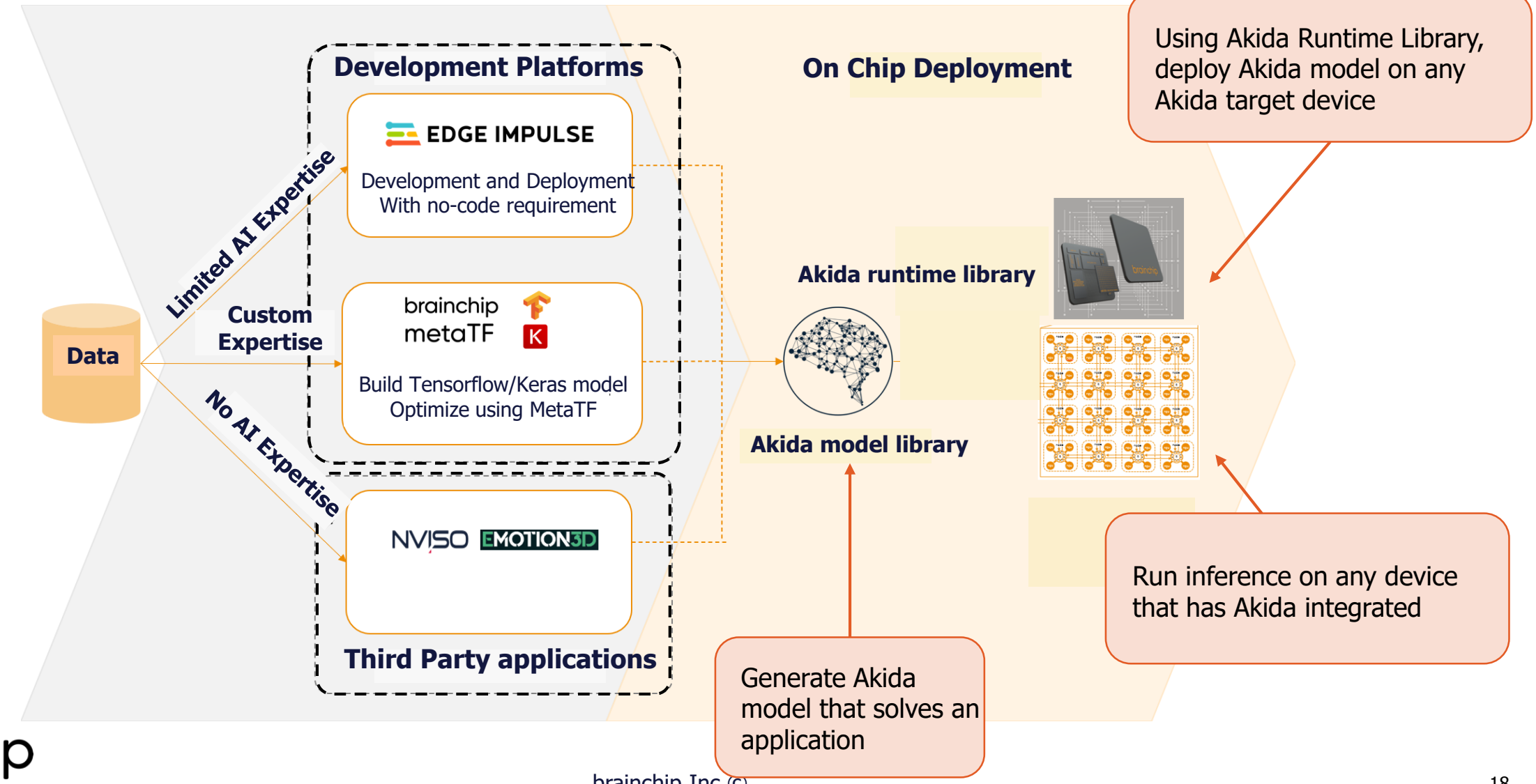
Simplifying Deployment with Akida

Akida IP benefits

- Underlying AI operations offloaded to Akida IP
- CPU and Akida IP running in parallel
- Model's parameters sitting in Akida IP local memory
- No memory congestion during inference



BrainChip Akida Model Development



The Results

brainchip
Essential AI

KITTI 2D dataset

Network	mAP	Parameters (millions)	MACs / sec (Billions)
Sim CLR (Resnet50)	0.57	26	82
Akida TENN* + CenterNet	**0.576**	**0.57**	**18**

Benchmark

Equivalent Precision | **50x fewer Parameters** | **5x fewer Operations**

- Akida TENN matches the benchmark precision*
- Improved performance
- Substantially smaller model – less memory and system load
- Much greater efficiency

< 75mW
For 30FPS in 16nm**

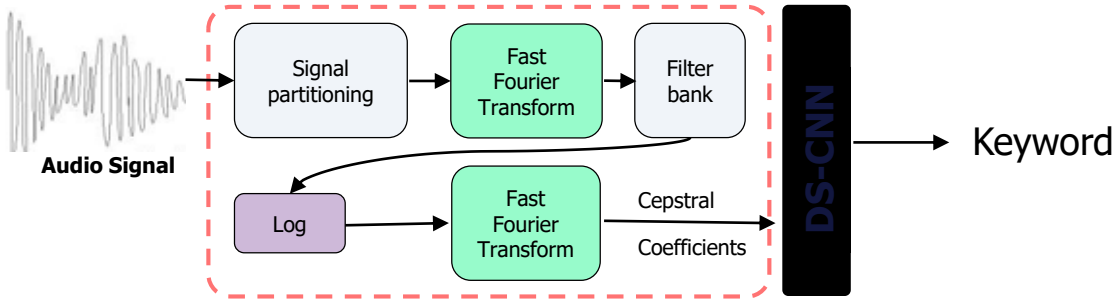
Resolution:1352x512

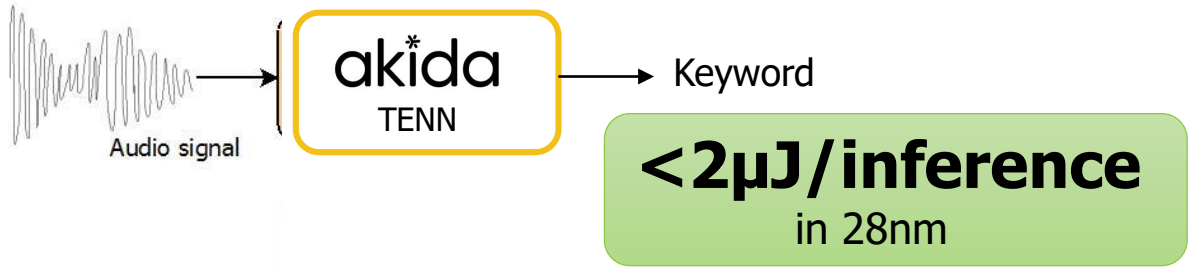
SimCLR with a RESNET50 backbone is the benchmark in object detection
Source: SiMCLR Review
Reference: SimCLR overview
* Can be further tweaked to improve mAP
** Estimates for Akida neural processing scaled from 28nm

brainchip
Essential AI

Simplifying Raw Audio

Disruptive solutions for consumer audio, hearing aids and more





- **Today's generic solution: MFCC + DSCNN**
  - Hardware filtering, transforms and encoding.
  - Memory intensive
  - Heavier software load

Model	Accuracy	Parameters	Total Memory (KB)	MACs (M/sec)
MFCC+DSCNN	92.43%	21 k	93.61	320

- No additional filtering or DSP hardware
- Memory efficient, smaller models, fewer ops
- Much faster and more power efficient

Model	Accuracy	Parameters	Total Memory (KB)	MACs (M/sec)
Akida TENN	97.12%	52 k	26	19

Better accuracy

Lower memory, BOM cost

16x fewer Ops

brainchip
Essential AI

Vital Signs Prediction: Heart Rate

Beth Israel Deaconess Medical Center Dataset

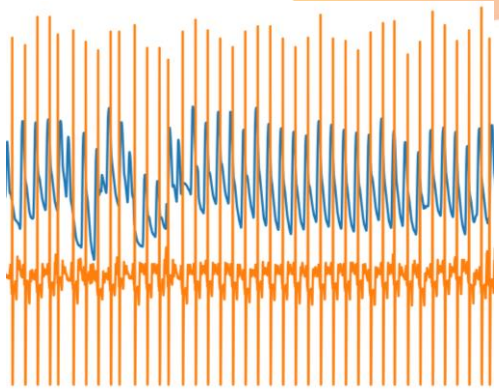
Model	Heart Rate Error (RMSE*)	Number of Parameters (million)	Billion MACs / sequence
S4 (SOTA)	0.332	0.3	11.2
Akida TENN**	**0.4721**	**0.063**	**0.021**
ExpRNN	1.87	0.127	~0.51

~SoTA Accuracy | **5x fewer Parameters** | **500x fewer Operations**

acceptable

RMSE=1.0

unacceptable



- Akida TENN substantially more efficient than current State of The Art** and very close in accuracy.
- Works on raw data. No pre-processing required
- Better than any current industry standard models
- Enables compelling mobile/portable edge devices to be create

S4 is a state of art algorithm that hasn't yet made it to production.
LSTM based models have been used, but not accurate enough for
* Root Mean Square Error (lower is better)
** Accuracy can be further improved
Source: 2206.11893.pdf (arxiv.org)

brainchip
Essential AI

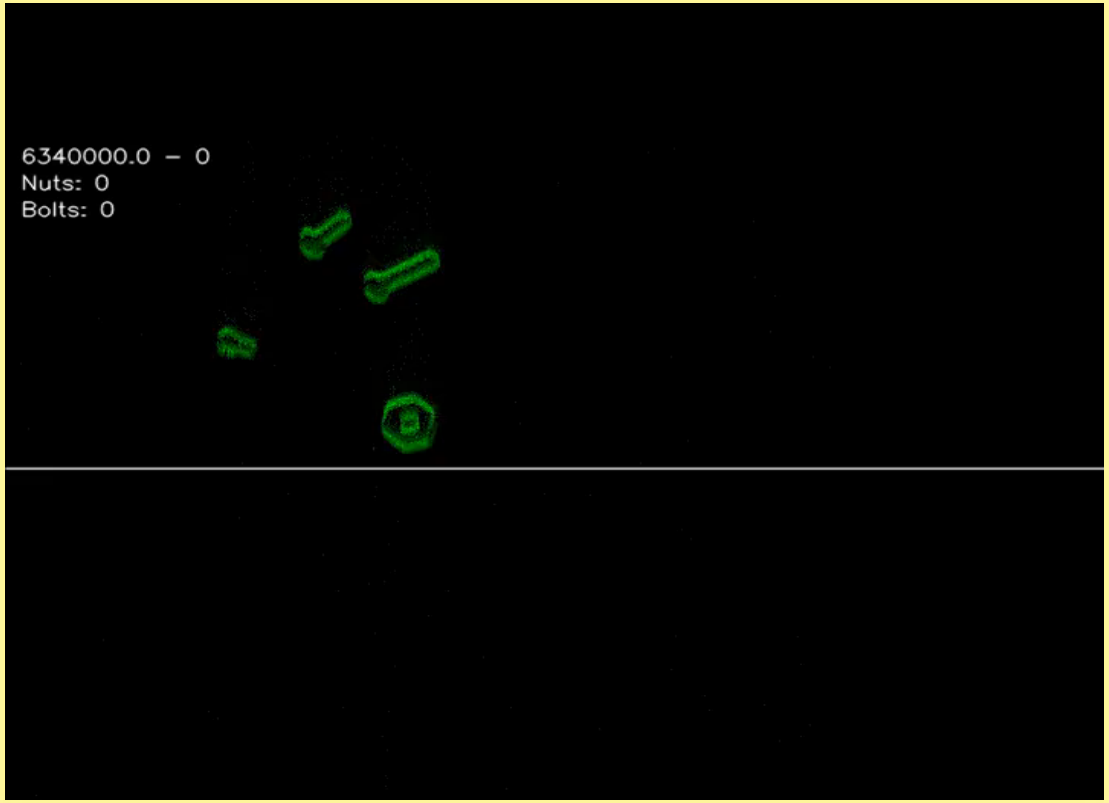
Akida in Action

brainchip
Essential AI

Akida in Action

FOMO – Faster Objects More Objects

Nuts and Bolt Classification
(with DVS Camera)



From Concept to Delivery

Ready to take your ideas to fruition

Evaluate

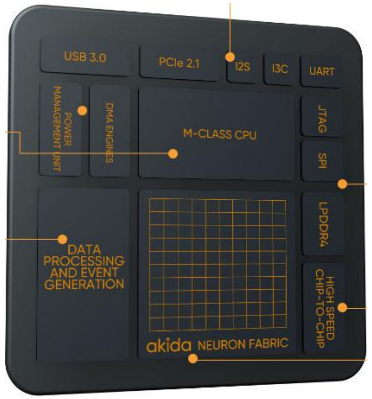
metaTF

Models Zoo

EDGE IMPULSE

Design

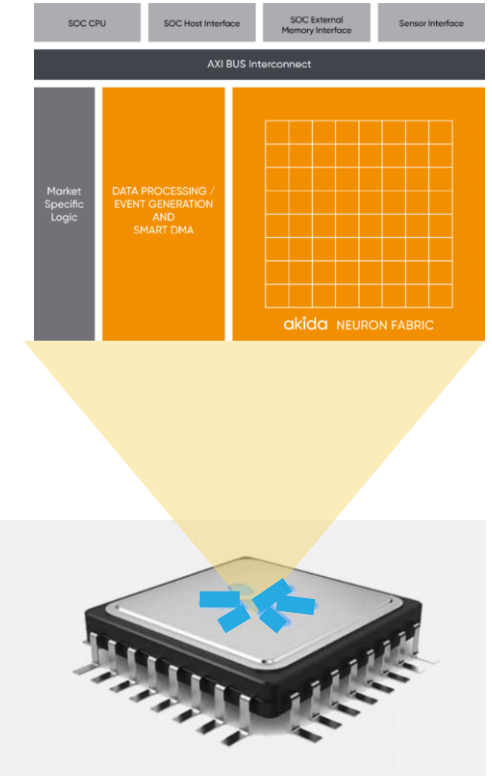
Develop



metaTF

EDGE IMPULSE

Scale



brainchip
Essential AI

Ecosystem & Partnerships

Early Adopters

Valeo

NASA

Mercedes-Benz

VORAGO TECHNOLOGIES
Opening up new possibilities

Licensees

MegaChips

RENESAS
BIG IDEAS FOR EVERY SPACE

Partnerships

Technology

arm

SiFive

intel foundry services

PROPHESEE

Enablement

EDGE IMPULSE

NVISO

EMOTION3D

Ai Labs

Integration

TEKSUN
CULTIVATING TECHNOLOGY

MegaChips

Ready When You Are!

- Akida platform boosts performance and efficiency in disruptive edge AIoT solutions
- Ready for today's complex models and future-proofed for the new ones
- Secure learning and intelligent customization at the edge without need for cloud retraining of model
- Easy to deploy with a fast-growing ecosystem
- We're ready! Are you?


brainchip
Essential AI

Resources

- BrainChip main site: https://www.brainchip.com
- BrainChip white papers: https://www.brainchip.com/white-papers-case-studies
- BrainChip MetaTF: https://docs.brainchipinc.com
- Akida models zoo: https://doc.brainchipinc.com/user_guide/akida_models.html
- Edge Impulse Support: https://docs.edgeimpulse.com/docs/development-platforms/officially-supported-ai-accelerators/akd1000