AI HARDWARE & EDGE AI SUMMIT

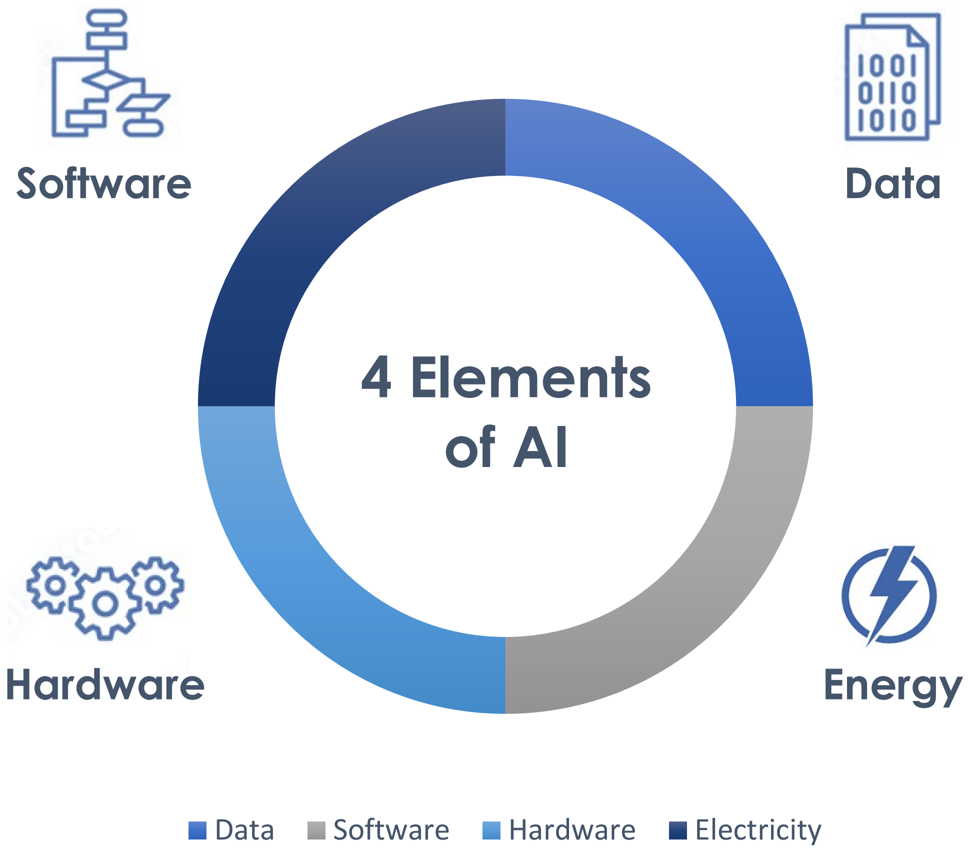
Efficient Models on Efficient Architectures

AI Hardware & Systems
@aiandsystems

Steven Brightfield
Chief Marketing Officer

brainchip

4 Elements of AI

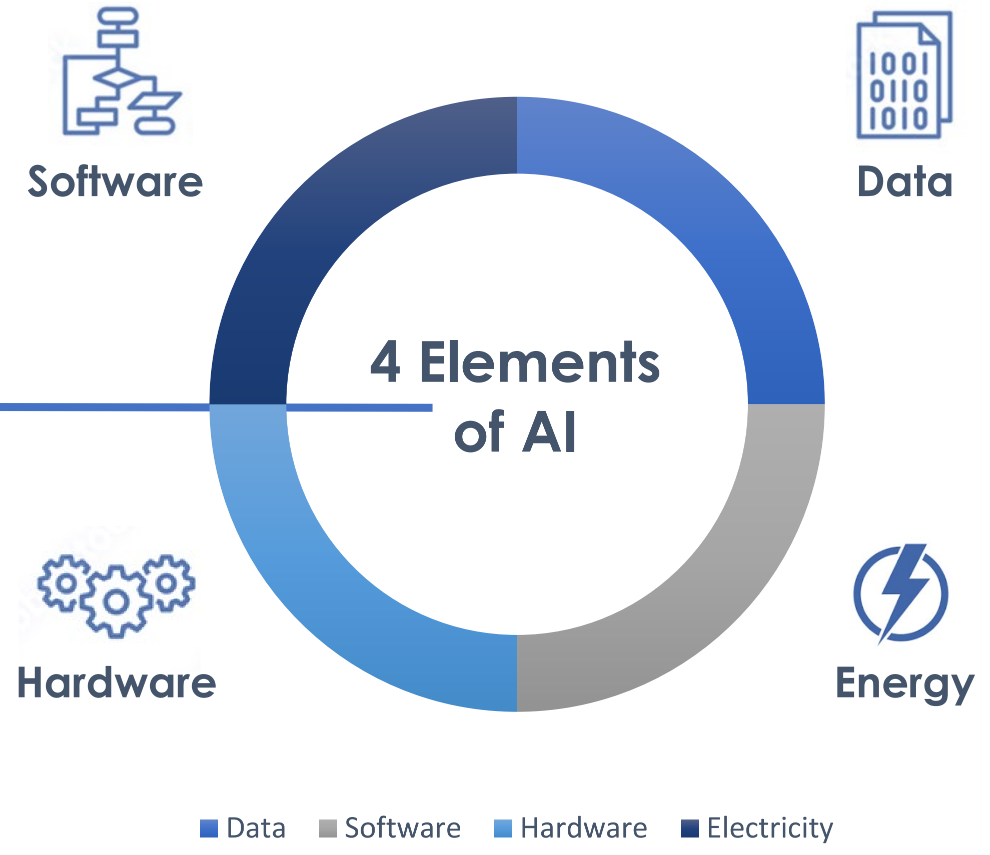


Datasets growing exponentially and the resulting model parameters

Power costs rising in cloud, hard limits on power on the edge

The Model Efficiency Equation

$$\text{Model Execution Power} = \frac{\text{Neural Model Complexity (operations/model)}}{\text{Neural Model Execution (operations/watt)}}$$



Neural Model Complexity

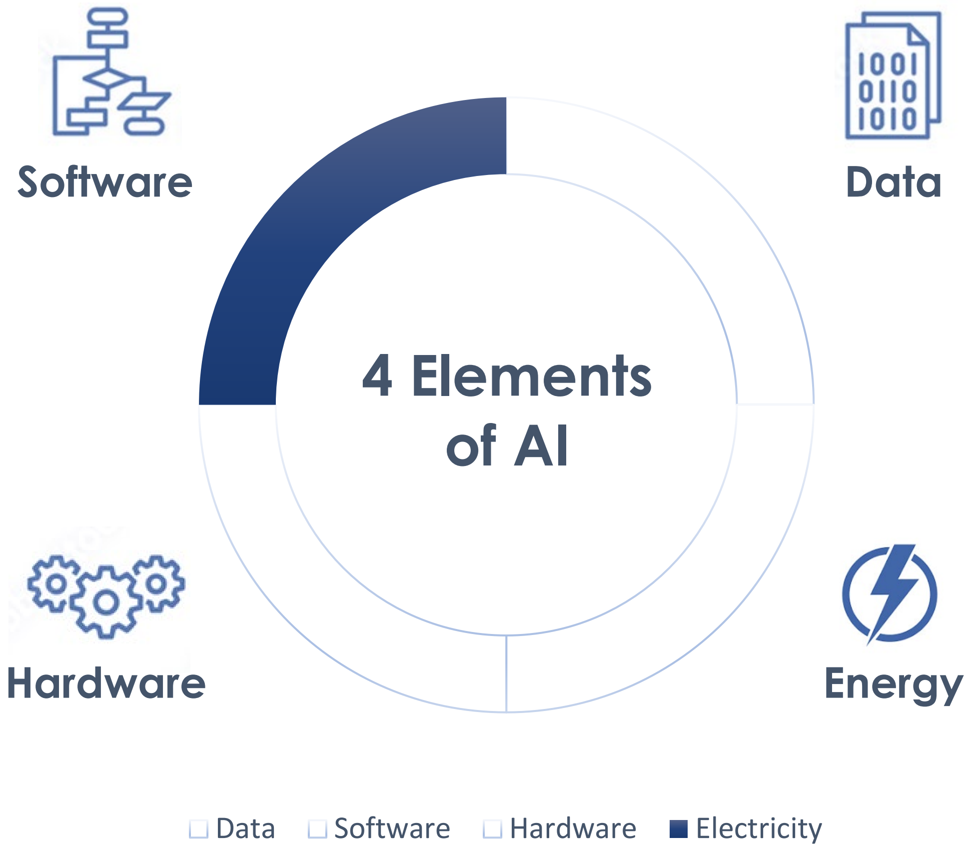
Using Foundation Models

- Pruning and distillation
- Fine tuning
- Trade off quality versus model size
- Use smaller context windows
- RAG Assistance

- More efficient training
  - Incremental training
  - Relevant Subset training

New Foundation Models

- New models suited for edge use cases



The Neural Model Efficiency

$$\textbf{Algorithmic Compute Efficiency} = \frac{\text{Model Metric (PESQ, Perplexity, mAP)}}{\text{MACs/inference (power + area)}}$$

$$\textbf{Algorithmic Memory Efficiency} = \frac{\text{Model Metric}}{\text{Parameters (memory movement)}}$$

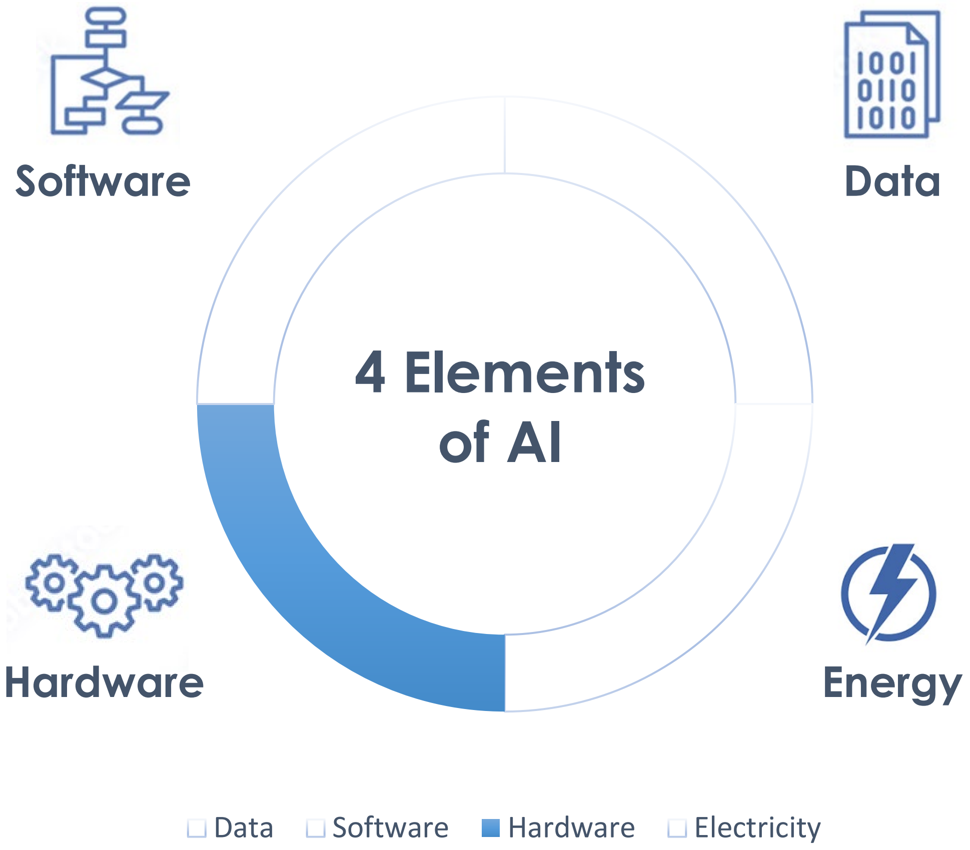
Neural Model Execution

New NPU chip architectures

- Reduced precision
- In-memory compute
- Analog compute
- High sparsity execution
- Efficient scheduling compilers
- Dedicated Transformer accelerators
- Optical
- Quantum

New silicon

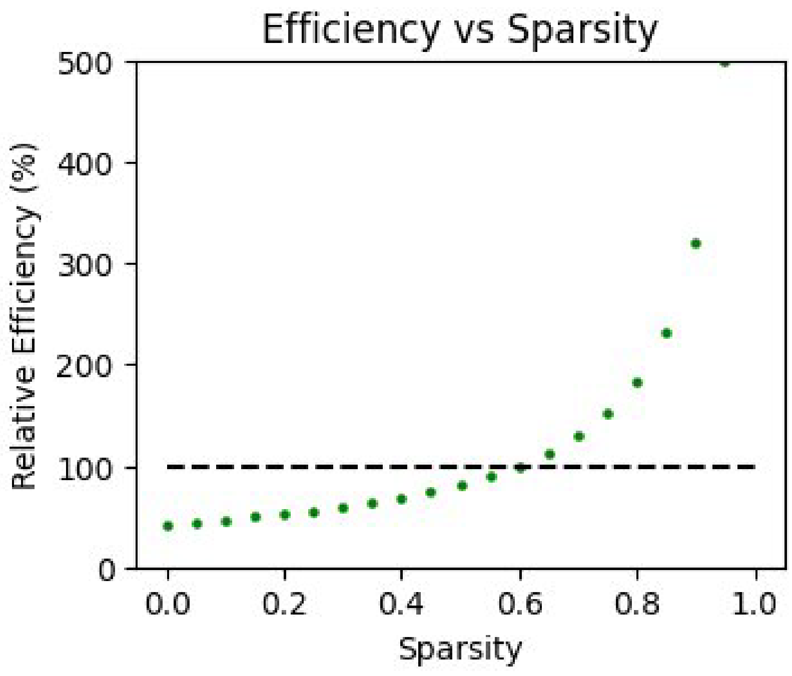
- Smaller process nodes
- Lower voltages
- Better heat dissipation



The Compute Efficiency Equation

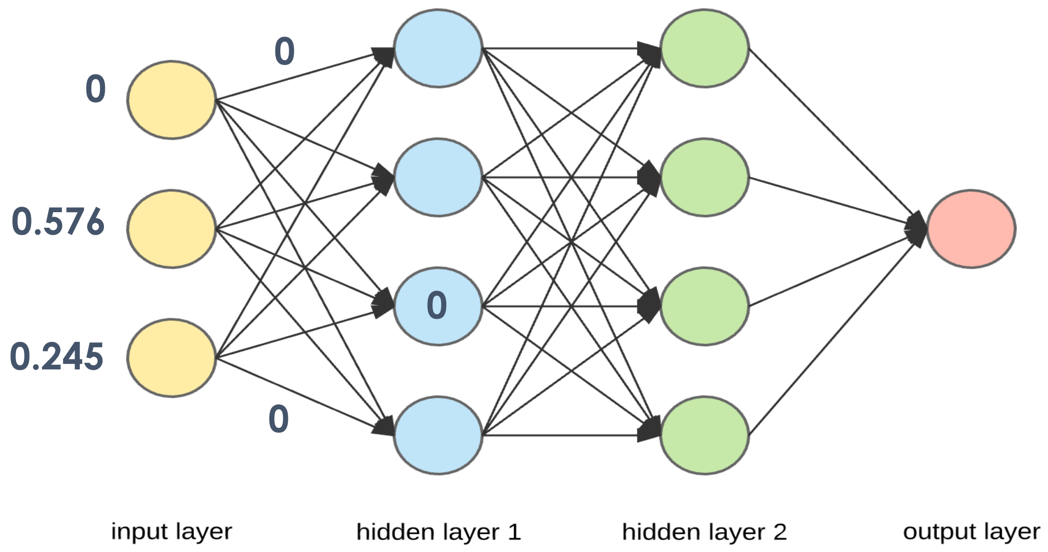
$$\text{Compute Efficiency} = \frac{\text{Actual MACs/sec Computed}}{\text{Total MACs/sec Possible}}$$

- **What percentage of available MACs can be scheduled for a given model**
- Take advantage of sparsity to reduce the number of MACs/sec that need to be computed
- At high-sparsity, >100% efficiency when compared to non event-based accelerators



Sparsity

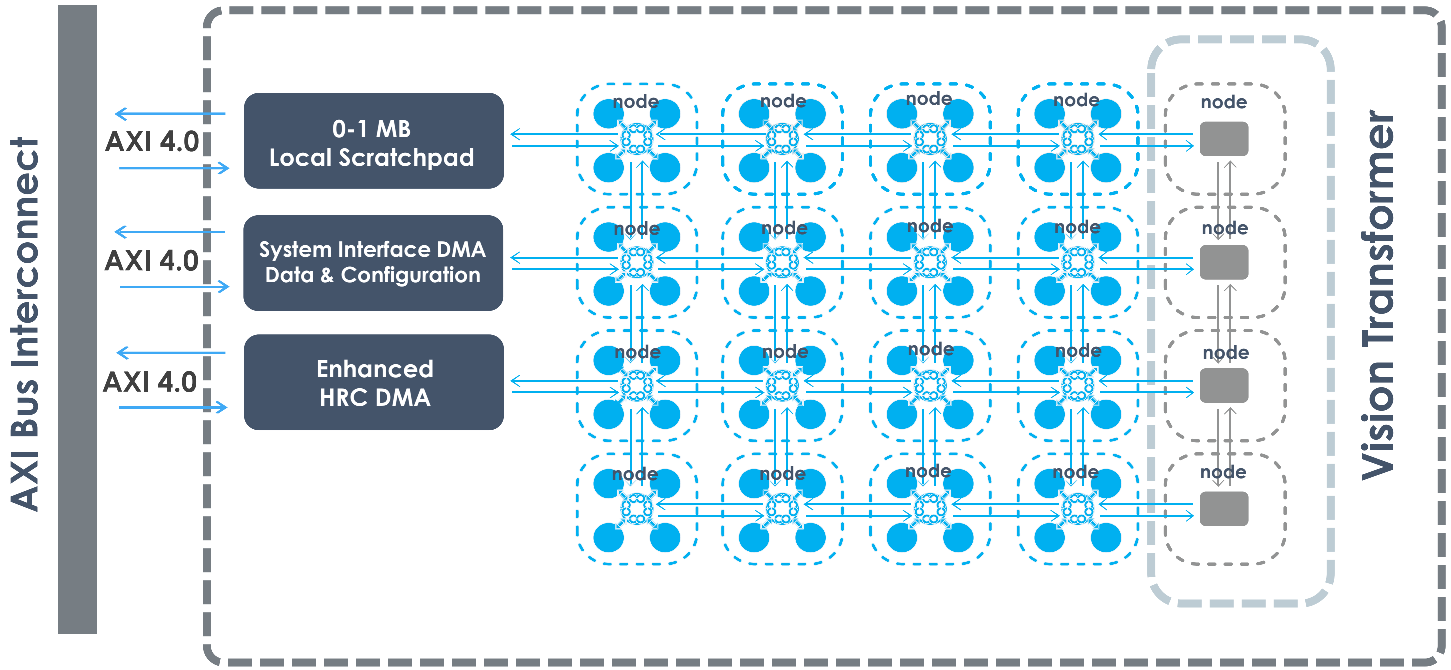
- Weight Sparsity (Model Architecture + Training + HW)
- Activation Sparsity (Model Architecture + Training + HW)
- Input Event Sparsity (Signal)



Akida Event-Based Computing Platform

Akida2 Key Attributes

- **Event-based processing** only processes and communicates on events.
- **At-memory compute:** Dedicated SRAM for each Neural Processing Engine (NPE) in a mesh-connected array,
- **Quantized parameters and activations:** Supports 8, 4, 2-bit parameters and activations
- Scalable, configurable inference platform
- Multi-layer model execution without host
- CNN/RCNN/ViT/SNN/SSM/TENN support
- Digital, event-based, at memory compute



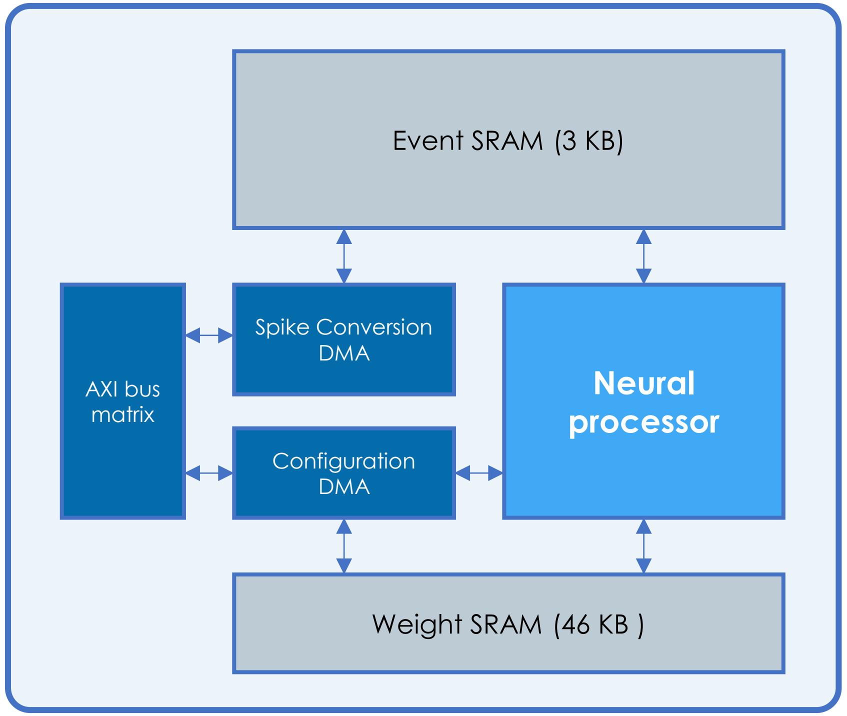
*ViT specialized nodes
**TENN integrated in all nodes

Akida leverages sparsity in weights and activations to reduce computational complexity

Akida 2 Ultra low power configuration

Key Attributes

- 1<mW operation[1]
- 100 % self managed execution from flash
- Total core area[2] **= 0.18 mm2** in GF22nm
- Can use in power island for always on/wake up



1 Power dependent on use case and silicon implementation
2 Total core shown with 21KB SRAM, configurable
3 Event & Weight SRAM sized for Key Word Spotting

Akida leverages sparsity in weights and activations to reduce computational complexity

The Trajectory of Models to the Edge

Global AI Trends and Predictions 2010 - 2030

Technology transitions in AI Roadmap

Convolutional
Neural Networks
(CNNs, e.g. AlexNet)

2012 Paper → 2014 Cloud → 2016 Edge

Transformer
Neural Networks
(Transformers, e.g. ViTs)

2017 Paper → 2019 Cloud → 2023 Edge (high compute complexity)

State Space
Neural Networks
(SSMs, e.g. S4, Mamba, TENNs)

2021 Paper → 2023 Cloud → 2024 Edge (linear compute complexity)

Structured State Space Models

Mamba
is the most well known State Space Model
(SSM)

Mamba supports LLMs

- Demonstrating much faster inferencing than transformers
- Demonstrating lower latency than transformers
- Improves with longer context windows
- Quality versus Transformers on benchmarks ongoing, see below

Several new versions released

- Mamba-2 – a faster version of Mamba
- Falcon Mamba 7B – Technology Innovation Institute (TII) in Abu Dhabi
- ML-Mamba - A new multi-modal Model supporting images and text

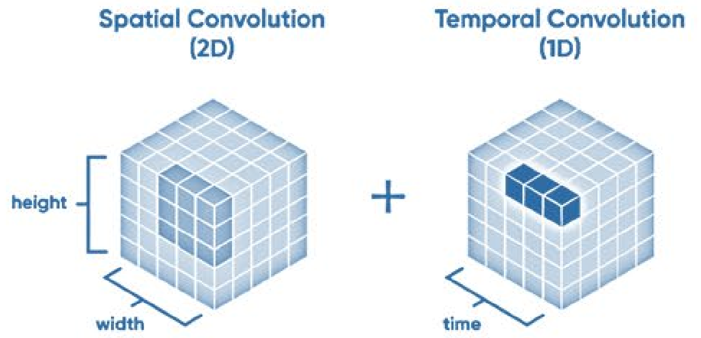
Is Attention All You Need?

[2312.00752] Mamba: Linear-Time Sequence Modeling with Selective State Spaces (arxiv.org)

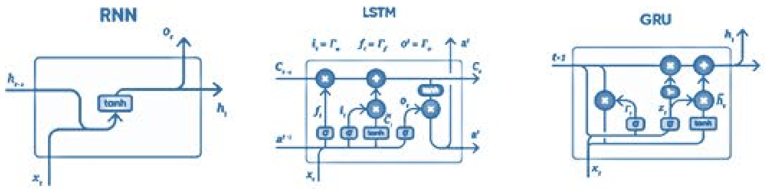
A More Efficient Network for the Edge

Temporal Event Based Neural Nets (TENN)

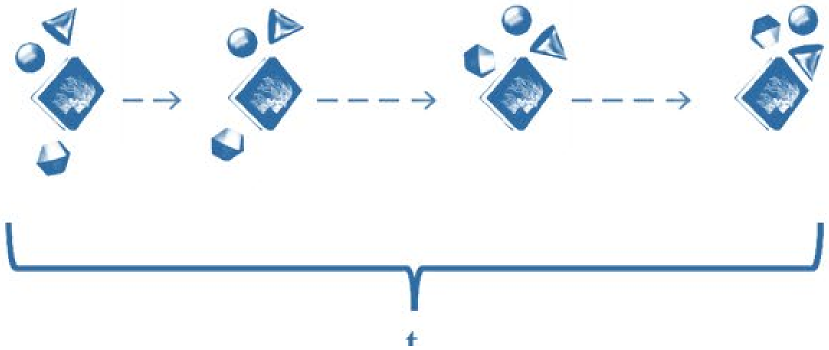
Extremely efficient 3D convolutions



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



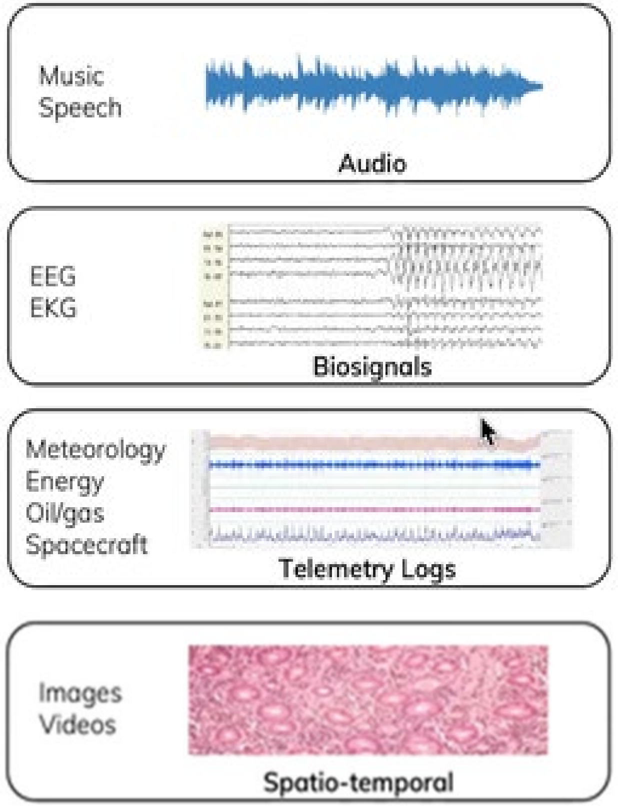
3D Time Series



- Simplifies solution to complex problems
- Reduces model size and footprint without loss in accuracy
- Easy to train (CNN-like pipeline)

Edge Applications for TENNs

- Sequence classification and generation in time:
  - **Raw audio classification:** keyword spotting
  - **Audio denoising:** single mic noise suppression
  - **ASR and GenAI**: compressing LLMs
- Sequence prediction algorithms
  - **Healthcare: v**ital signs estimation
  - **Industrial:** vibration prediction
  - **Robotics:** Path prediction
  - Any time-series/sequence prediction problem
- Multi-dimensional streaming video
  - **Video object detection** – frames are correlated in time.
  - **Action recognition** – classifying across many frames
  - **Video frame prediction** – path prediction & planning



Key Word Spotting on Akida

Key Word Spotting Model Power

TENNs
Vs.
DSP/CNN

=

1/7X
Less MACs

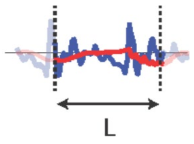
1/3.5X
Memory

Akida
Event-based Ops

Model	Accuracy	Total Memory (KB)	MACs (M/sec)
DS-CNN	**92.43%**	**93.61**	**128**
TENNs Akida	**97.02%**	**26**	**19**
Comparison	**+5%**	**3.5x**	**7x**

TENNs Model Approach

Mixture waveform



→ **TENNs** → Keyword

4.6% > accuracy improvement

Audio Denoising on Akida

Denoising Model Power

TENNs
Vs
Deep Filter Net
V3

=

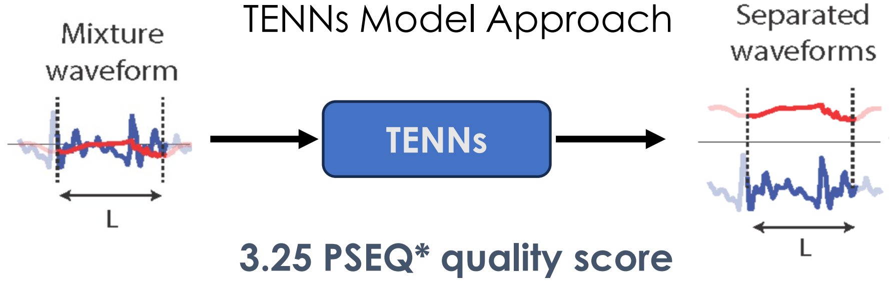
1/11.5X
Less MACs

1/3.6X
Memory

Akida Pico
Event-based Ops

- **Audio denoising isolates a voice signal from background noise**
- Traditional approach employs computationally intensive time domain to frequency domain transform and the inverse transform
- TENNs approach avoids expensive FFT transformations

TENNs Model Approach



3.25 PSEQ* quality score

Note: PESQ score is for a 32fp version of the model

Efficient Models on Efficient Architectures

Goals:

- As few MACs/model inference,
- As little power per effective MAC
- Minimize memory size and movement

Utilize:

- Event-based compute architectures in hardware
- New model algorithms in software
- Model size fits in-memory compute

Visit Us @ Booth #58

Akida 2
https://brainchip.com/wp-content/uploads/2023/03/BrainChip_second_generation_Platform_Brief.pdf

TENNs White Paper
Introducing TENN: Revolutionizing Computing with an Energy Efficient Transformer Replacement - BrainChip

BrainChip Enablement Platforms
https://brainchip.com/akida-enablement-platforms/

Akida Technology Foundations

Fundamentally different. Extremely efficient.

Silicon-Proven, Fully Digital Neuromorphic Implementation

Cost-effective, predictable design and implementation

Event-based Hardware Acceleration

Minimized compute and communication - Minimizes host CPU usage

At-Memory-Compute

Maximum throughput, Lowers latency and system bandwidth usage

On-chip Learning

One-shot/few-shot learning. Minimizes sensitive data sent. Improves security and privacy

Configurable And Scalable

Extremely configurable and post-silicon flexibility

Complex Models, High Accuracy

Unique spatial-temporal capabilities, accelerates Vision Transformers.