tinyML® EMEA

Enabling Ultra-low Power Machine Learning at the Edge

tinyML EMEA Technical Forum 2021 Proceedings

June 7 – 10, 2021

Virtual Event

TINY ML

www.tinyML.org

PULP PLATFORM
Open Source Hardware, the way it should be!

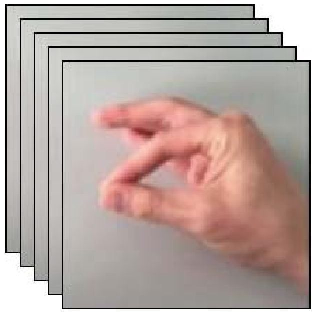
Energy-efficient TCN-Extensions for a TNN accelerator

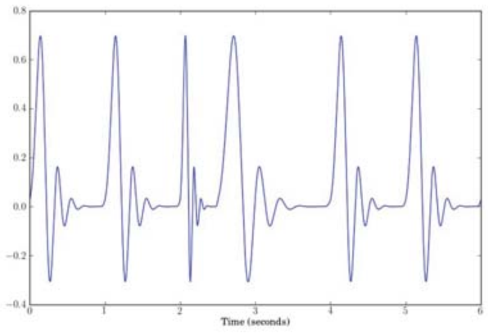
Tim Fischer
fischeti@iis.ee.ethz.ch

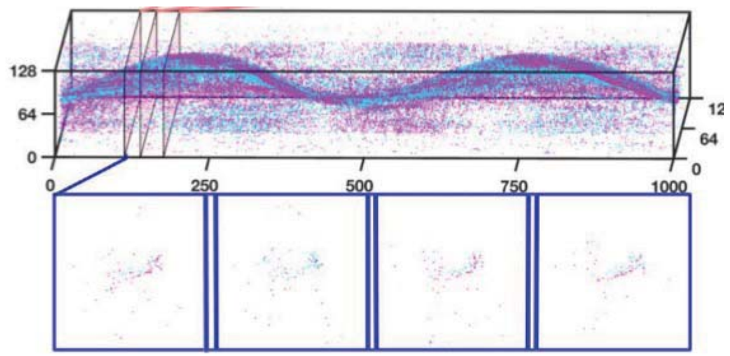
Special Thanks to:
Moritz Scherer, Georg Rutishauser, Matteo Spallanzani, Luca Benini

Processing temporal data on the edge

- **Temporal data is very prevalent in edge applications**
  - Audio (speech recognition)
  - Biomedical signals (EEG, ECG)
  - Video (gesture recognition)
- **Edge resources are scarce**
  - Battery powered: ~mWs vs. MFLOPS
  - Small memory: 100s KBs vs. Mparams
- **Processing time-series data is complex**
  - Additional dimension
  - Capture temporal dependencies







CUTIE – Ternary NN accelerator

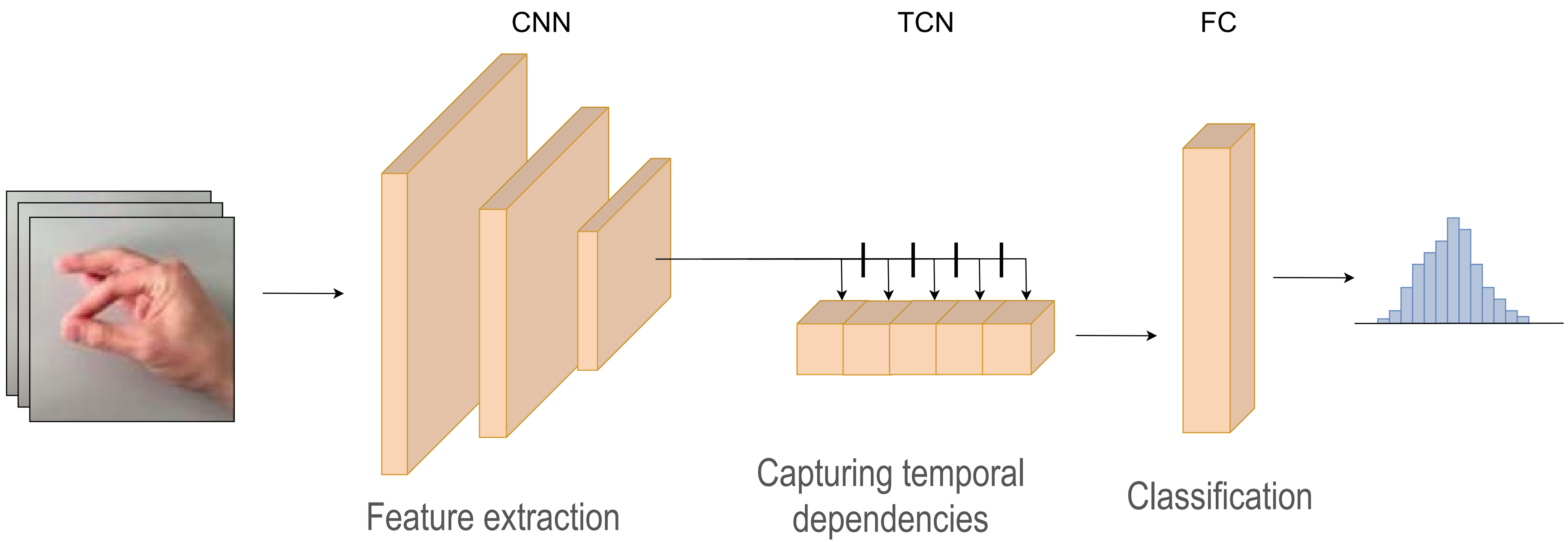
What is the SoA for energy-efficient processing on the edge?

- **Ternary quantization**
  - Encodes 3 values {-1, 0, 1}
  - Achieves reasonable accuracy for some applications
  - Better than binary accuracy
  - Higher efficiency due to sparsity
- **CUTIE – Completely Unrolled Ternary Inference Engine[1]**
  - Hardware accelerator for ternary CNNs
  - Peak energy efficiency: 3.1 [7nm] resp. 0.6 [22nm] **POp/s/W** @ 0.65V

[1] Scherer et al.: CUTIE – Beyond PetaOp/s/W Ternary DNN Acceleration

Time-series processing on the edge

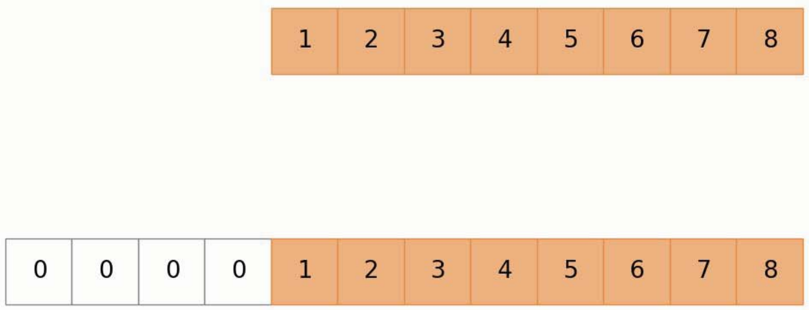
- **Steps required for time-series processing**
  - Feature extraction to generate embeddings from frames (i.e. CNN)
  - Capturing temporal dependencies (i.e. RNN)
  - Classify output (i.e. FC)



Contribution – Ternary TCNs

How can we bring time-series data processing to the edge?

- **Temporal Convolutional Neural Networks (TCNs)**
  - Causal 1D Convolution
  - Flexible receptive field
  - High parallelism and low memory requirements
  - Competitive accuracy w.r.t. RNNs

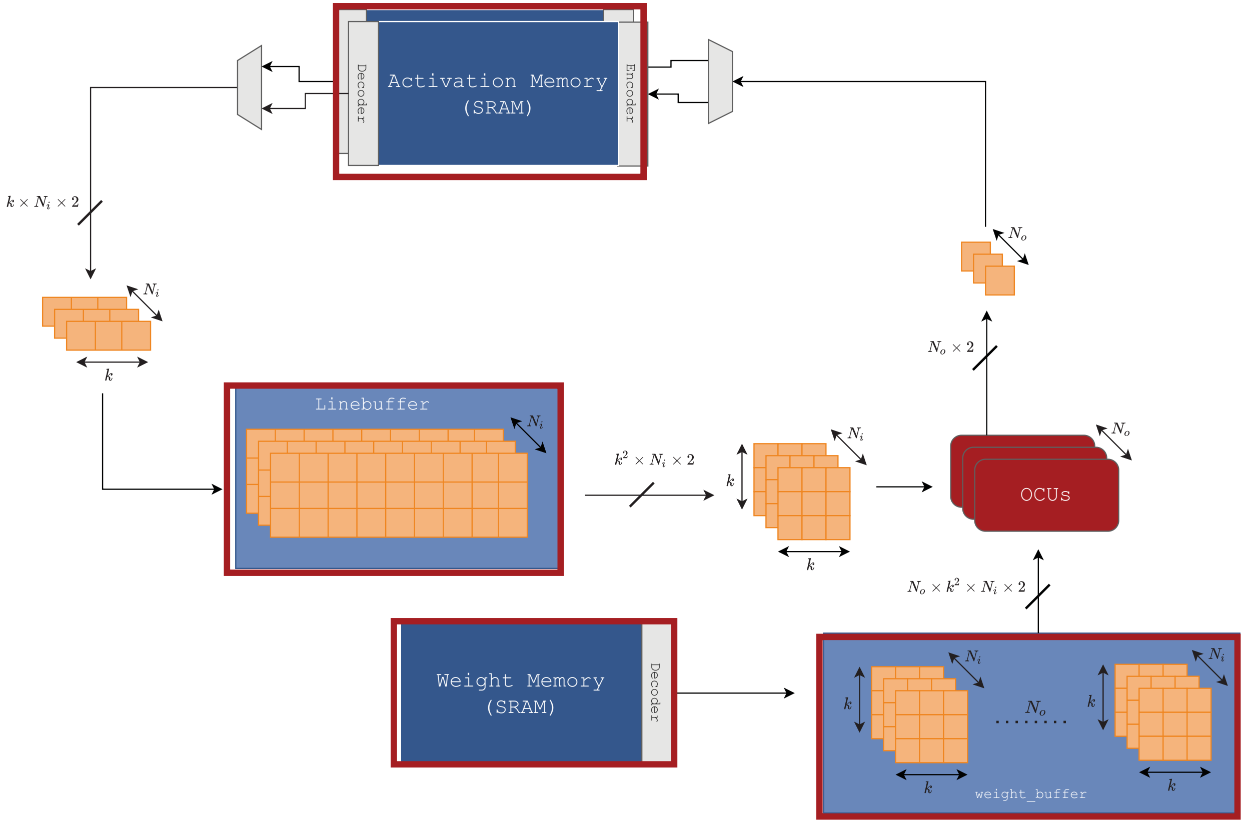


- **Contributions**
  - Ternarization of TCNs
  - Mapping of 1D TCN layers to 2D CNN layers
  - Reuse existing hardware of CUTIE
  - Leverage high parallelism of CNN layers

TNN-Accelerator - CUTIE

Completely Unrolled Ternary Inference Engine

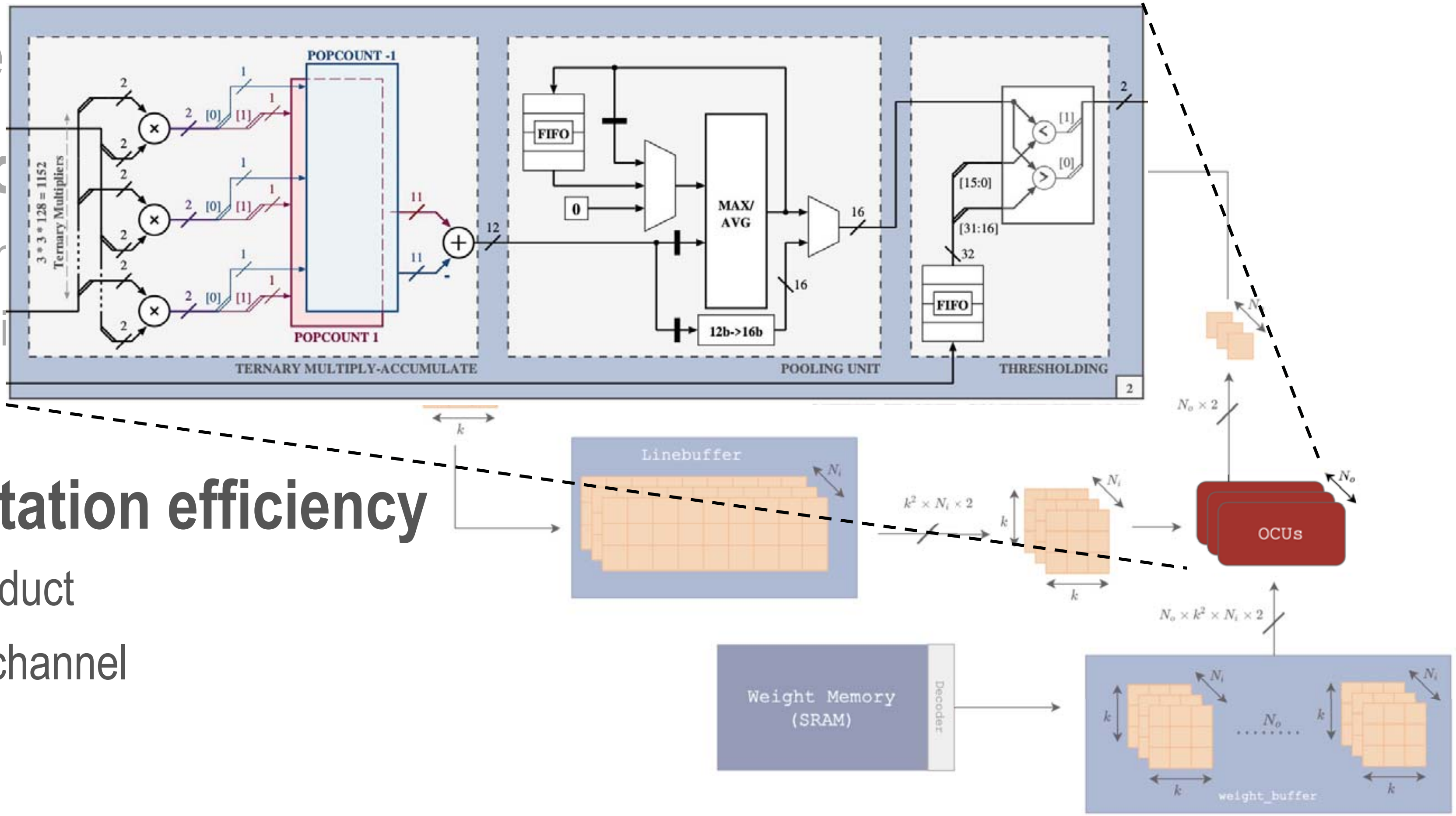
- Minimize Data movement
  - Local memory for weights and activations
  - Maximize data reusability



TNN-Accelerator - CUTIE

Completely Unrolle

- Minimize Data mo
  - Local memory for weigh
  - Maximize data reusabili

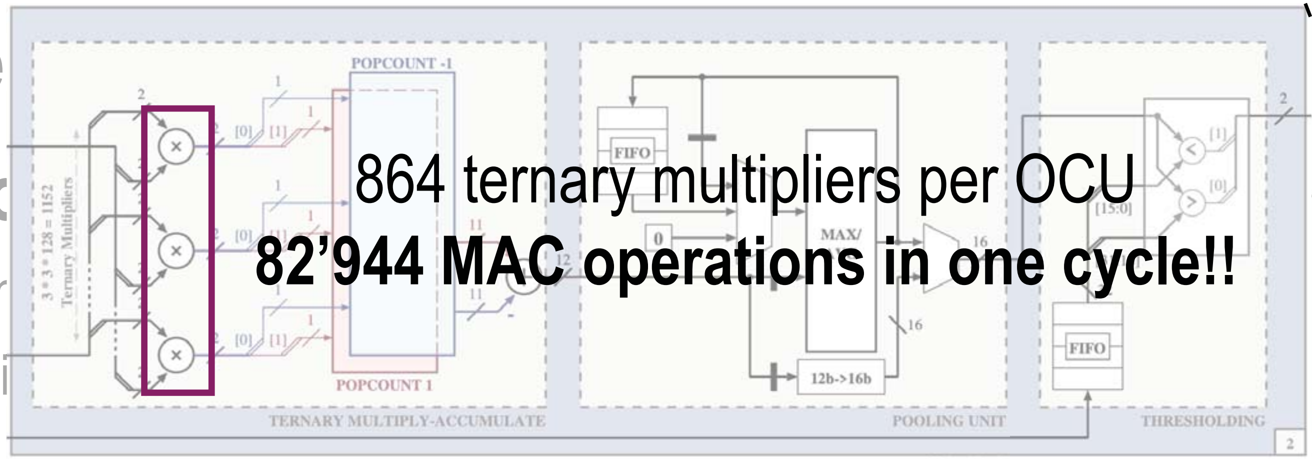


- Maximize computation efficiency
  - Fully unrolled inner product
  - One compute unit per channel

TNN-Accelerator - CUTIE

Completely Unrolle

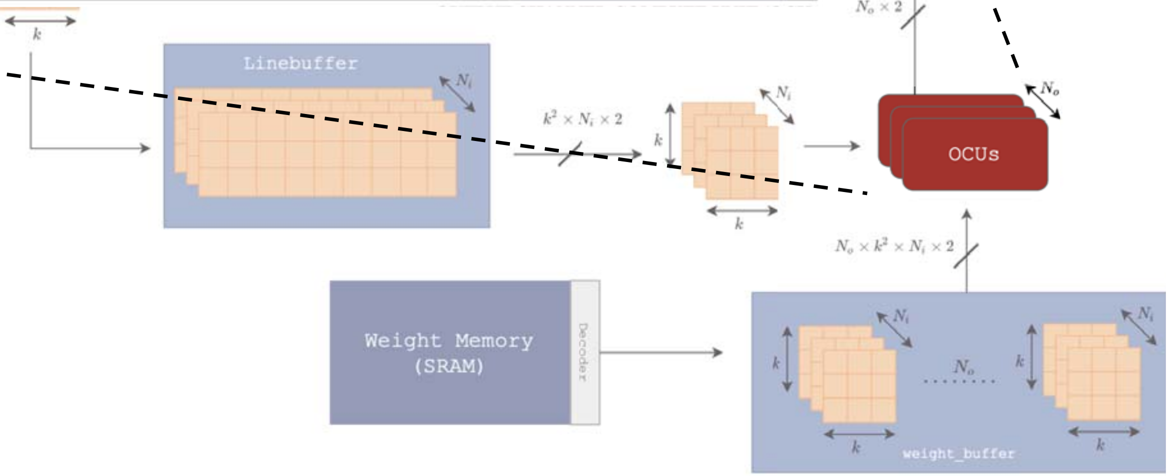
- **Minimize Data mo**
  - Local memory for weigh
  - Maximize data reusabili



864 ternary multipliers per OCU

82’944 MAC operations in one cycle!!

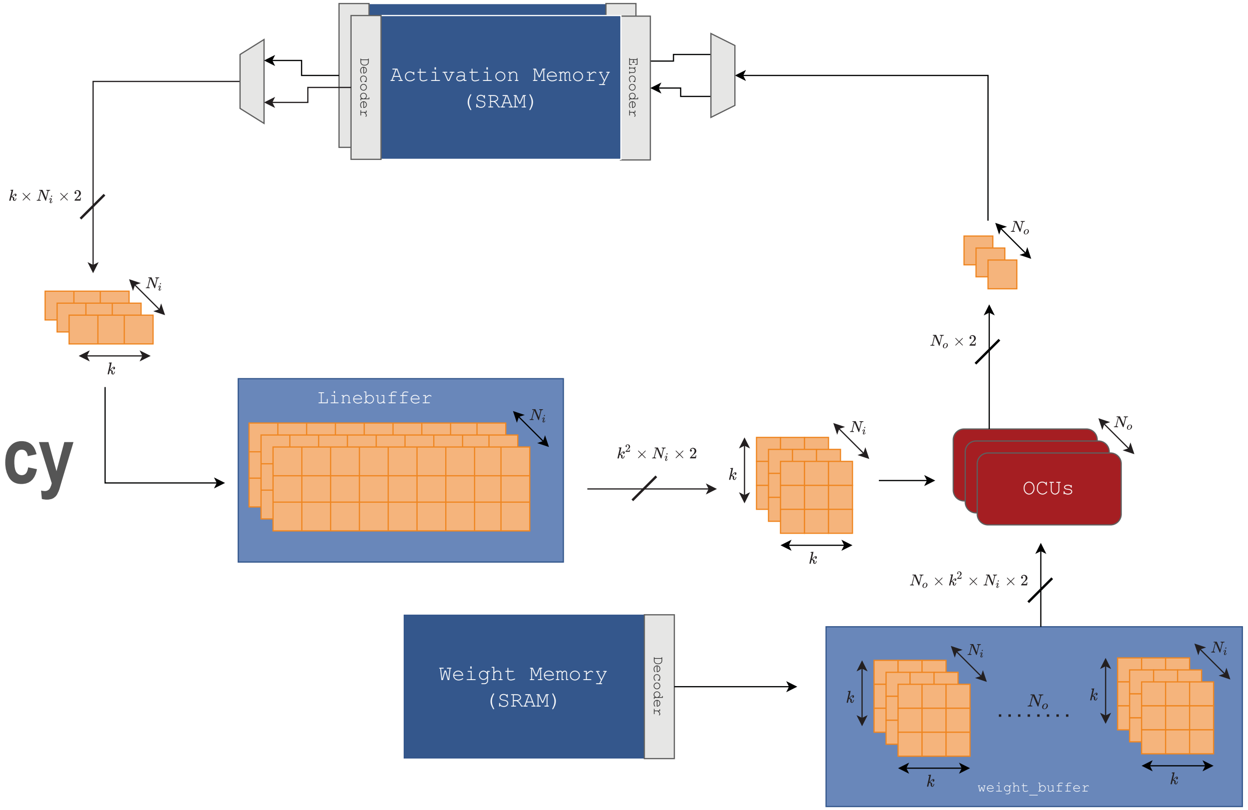
- **Maximize computation efficiency**
  - Fully unrolled inner product
  - One compute unit per channel



CUTIE –TNN Accelerator

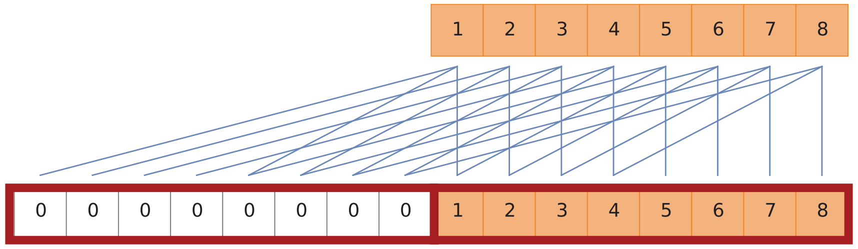
Completely Unrolled Ternary Inference Engine

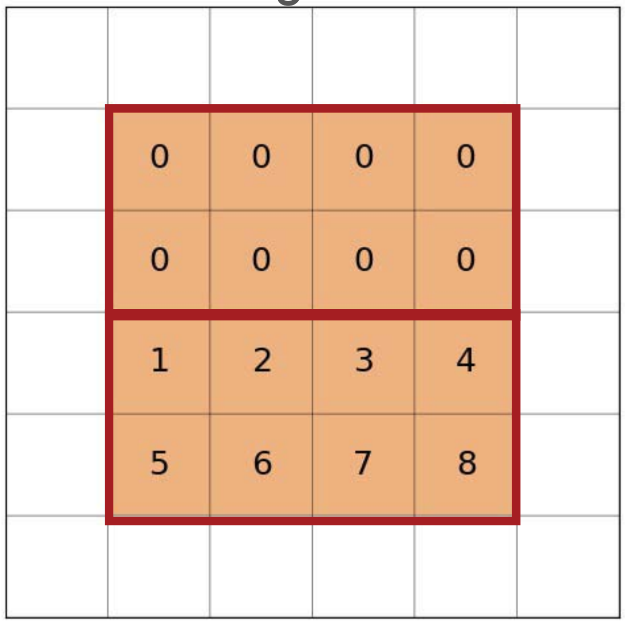
- **Minimize Data movement**
  - Local memory for weights and activations
  - Maximize data reusability
- **Maximize computation efficiency**
  - Fully unrolled inner product
  - One compute unit per channel
- **Minimize switching activity**
  - Exploit sparsity of values



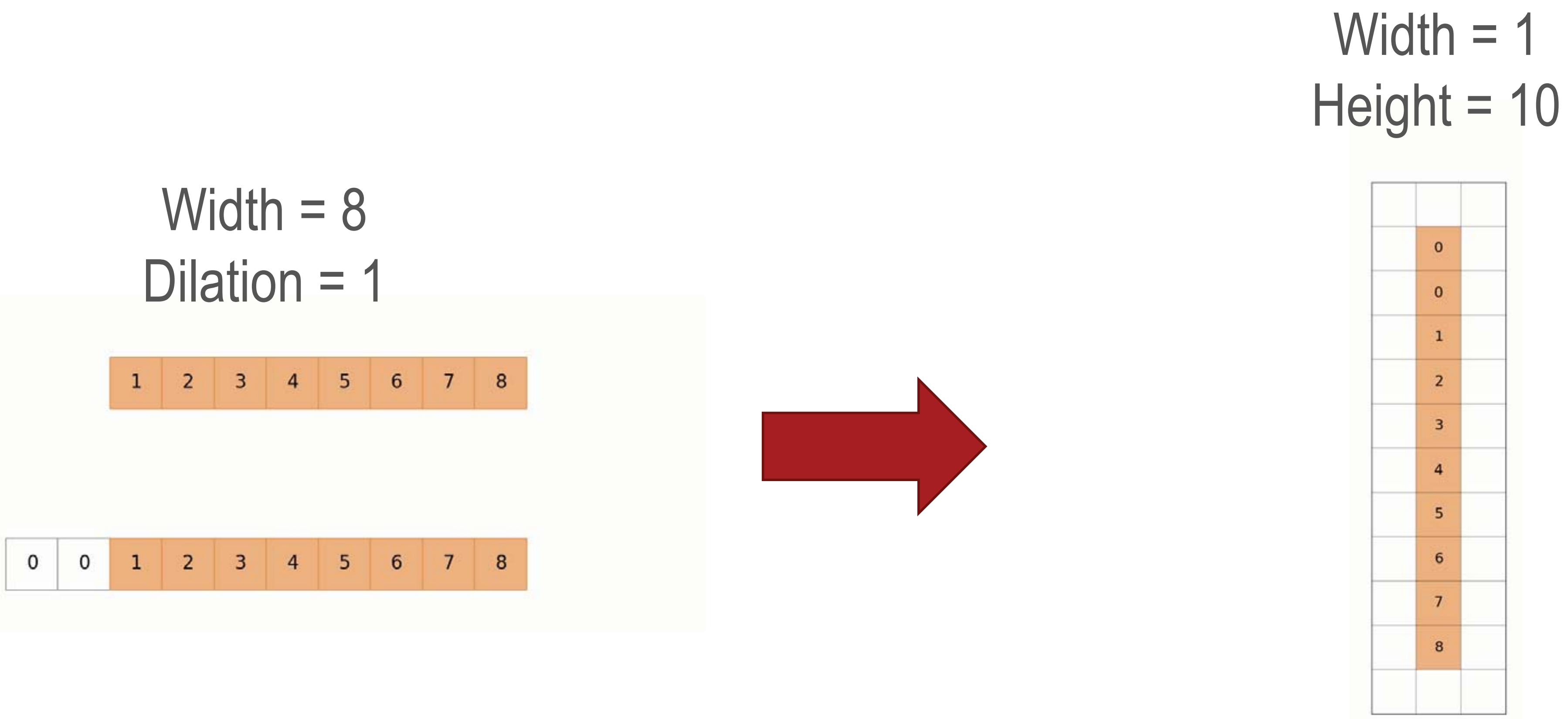
Mapping - TCN to CNN

How can we map 1D TCNs to 2D CNNs?

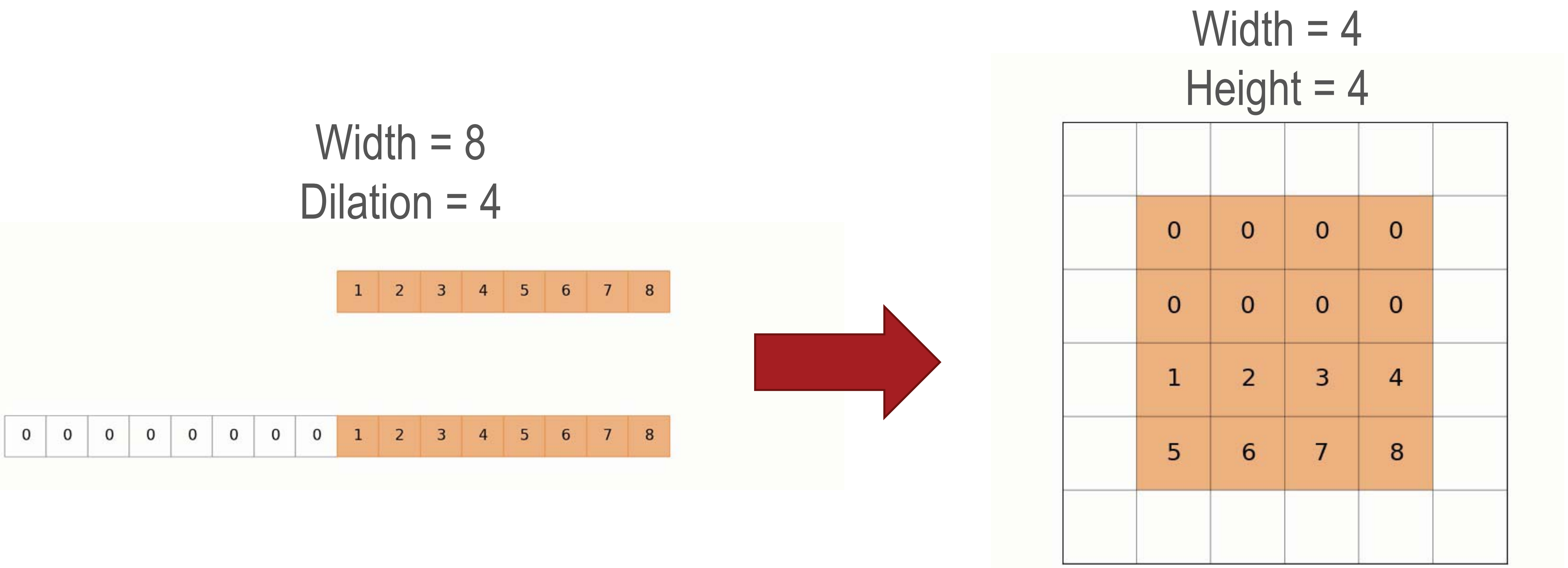




Mapping – TCN to CNN

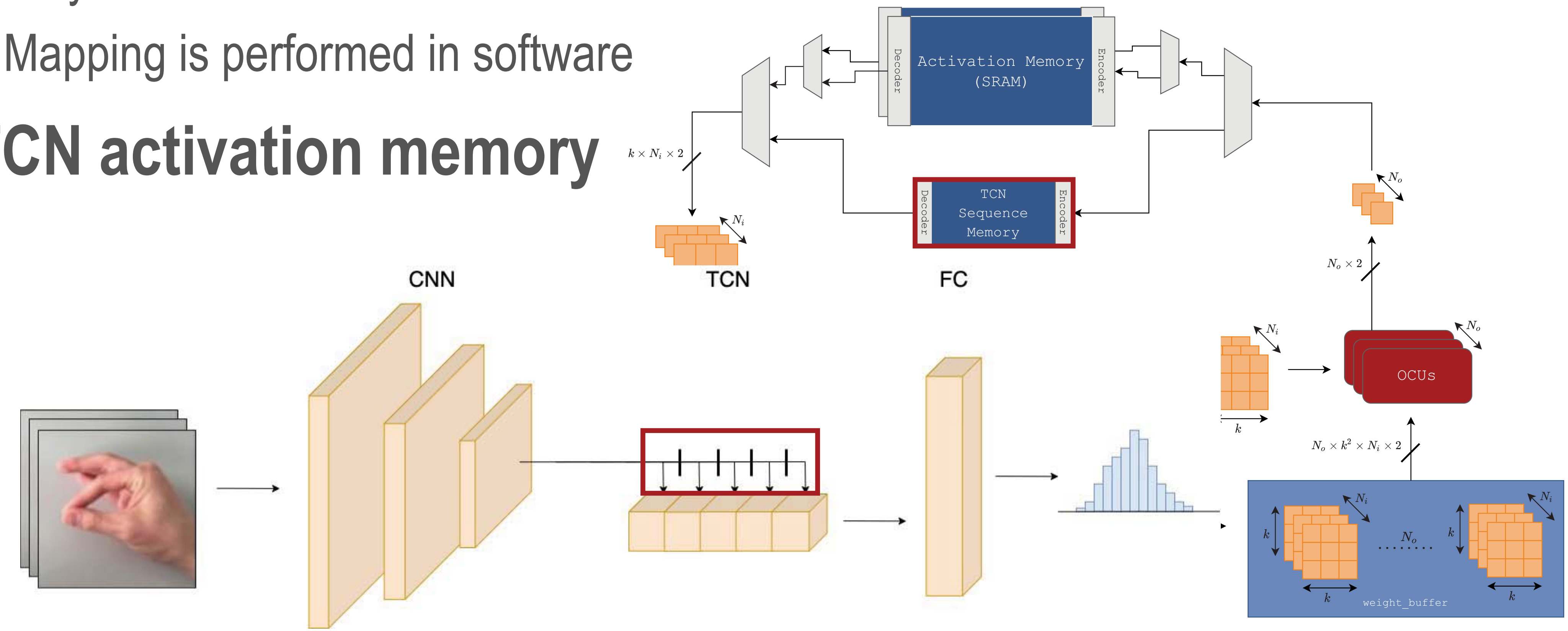


Mapping – TCN to CNN



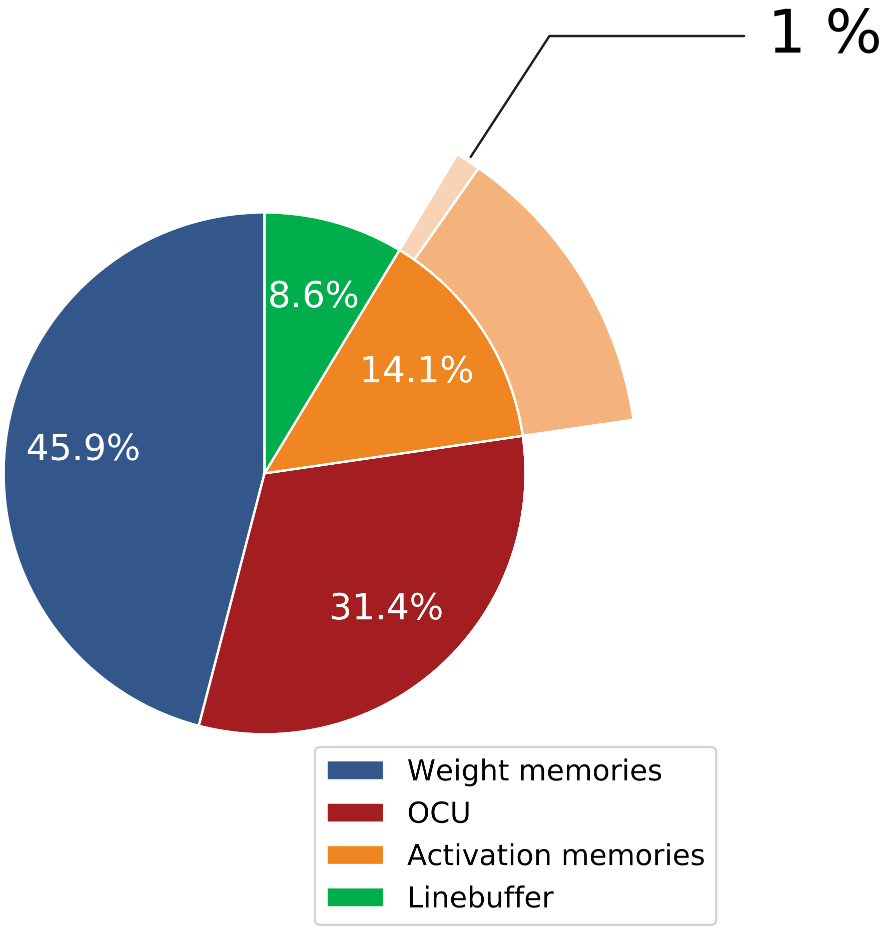
Implementation - TCN

- Control is extended to support TCNs
  - Very minimal hardware modifications
  - Mapping is performed in software
- TCN activation memory



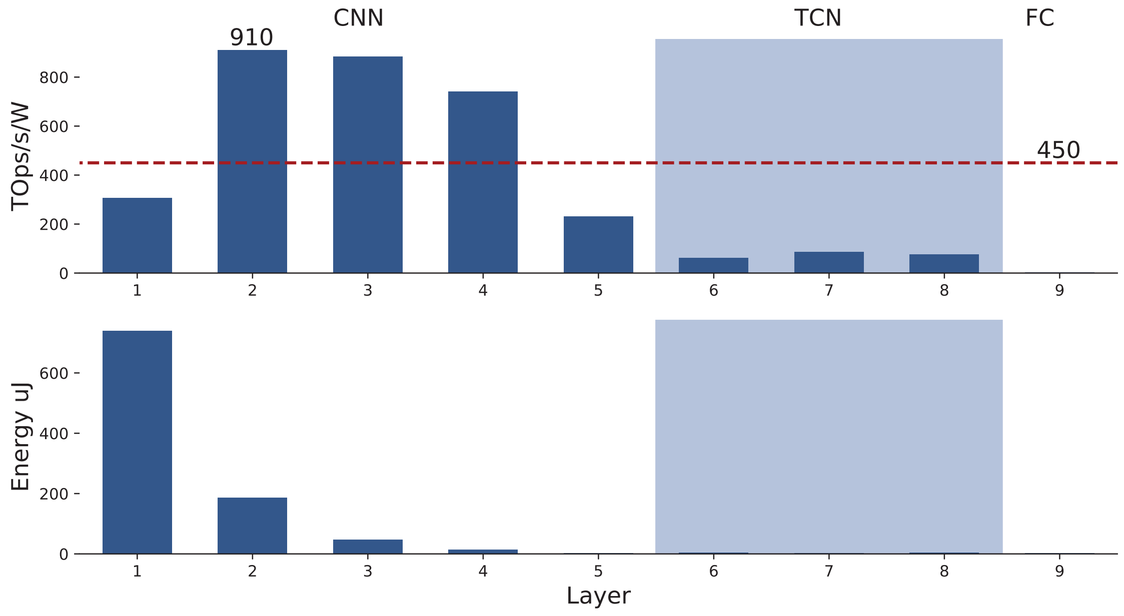
What is the cost for TCN support?

- **Area overhead**
  - Control modifications
  - Additional activation memory
- **GF 22 nm technology**
  - 96 input/output channels, 64x64 feature maps
  - 8-layer, 3x3 kernels
  - 1.86 mm^2 / 9.3MGE, post-synthesis
  - **Only 1% increase for TCN support**



Results – Power and Energy

- **GF 22 nm technology**
  - 0.65 V @ 66 MHz, post-synthesis
  - **Peak energy efficiency: 910 TOp/s/W**
  - Avg. energy efficiency: 450 TOp/s/W



- **Inference on DVSGesture**
  - 92.6% accuracy
  - Core energy per inference: 1 µJ
  - Core throughput: 5.48 TOp/s
  - Core inference Time: 78 µs

Conclusion

- **Processing temporal data on the edge is possible**
  - Ternary TCNs are well-suited for edge processing
- **Smart Mapping of 1D TCN to 2D CNN**
  - Mapping is primarily done in software
  - only 1% hardware area increase
- **CUTIE achieves exceptional energy-efficiency**
  - Peak performance of **910 TOp/s/W** in GF 22 nm

TINY ML EMEA
Virtual
JUNE 7-10, 2021

Premier Sponsor

Intelligent Agent Neuton

Automated TinyML

Zero-code SaaS solution

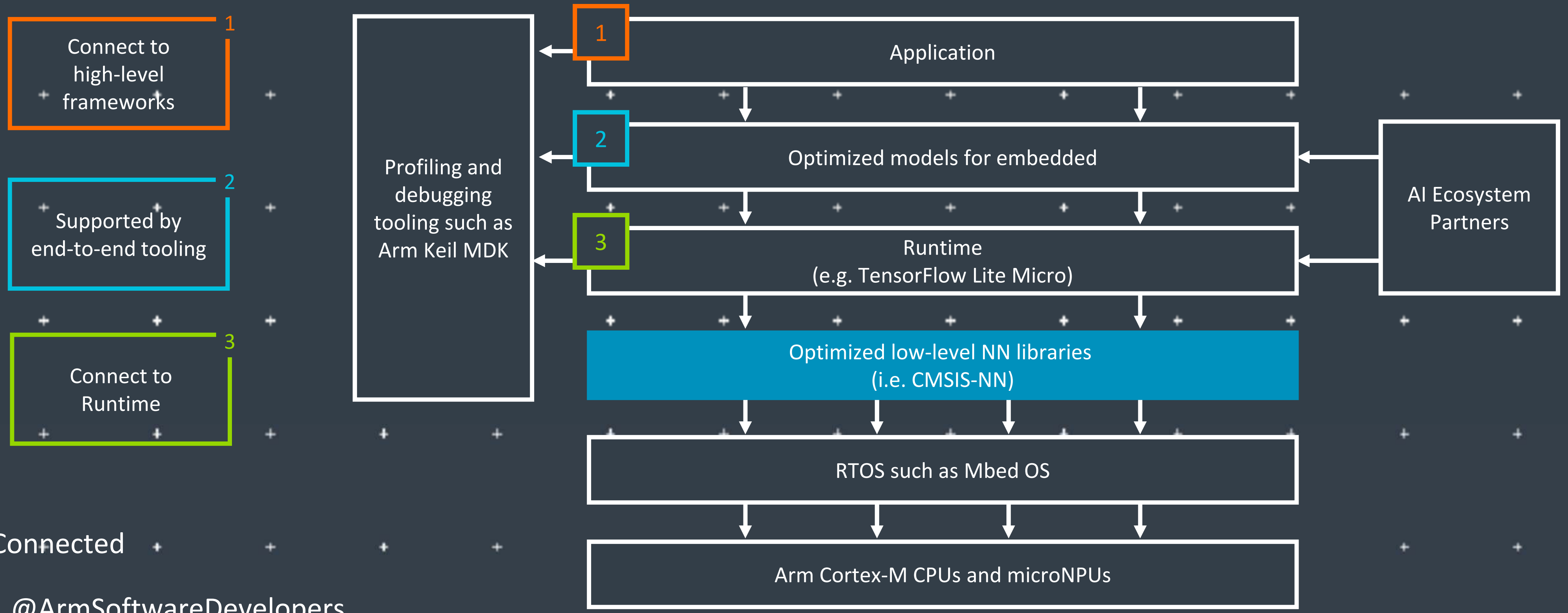
Create tiny models, ready for embedding, in just a few clicks!

Compare the benchmarks of our compact models to those of TensorFlow and other leading neural network frameworks.

Build Fast. Build Once. Never Compromise.

Executive Sponsors

Arm: The Software and Hardware Foundation for tinyML



Stay Connected

@ArmSoftwareDevelopers

@ArmSoftwareDev

Resources: developer.arm.com/solutions/machine-learning-on-arm

arm

TinyML for all developers

C++ library

Arduino library

WebAssembly

Dataset

Acquire valuable training data securely

Enrich data and train ML algorithms

Edge Device

Real sensors in real time

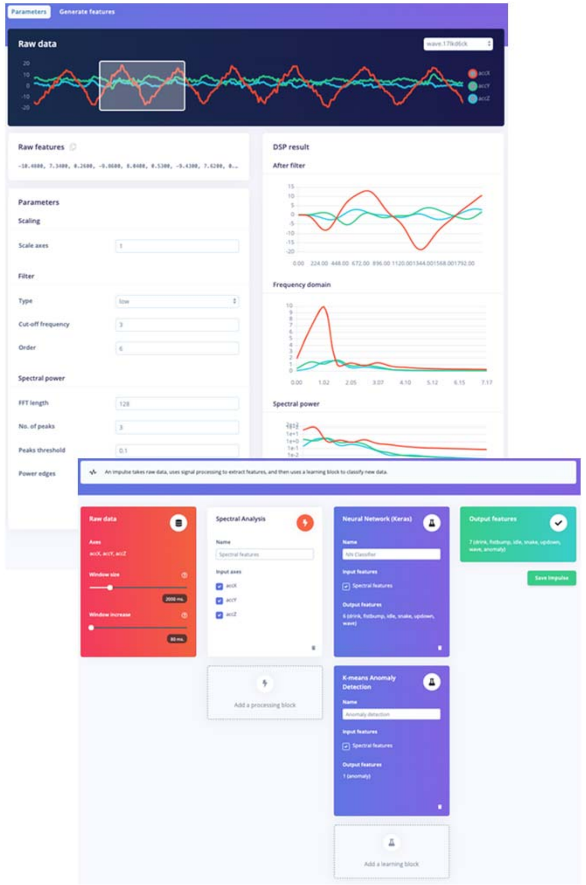
Open source SDK

Impulse

Embedded and edge compute deployment options

Test impulse with real-time device data flows

Test



www.edgeimpulse.com

Qualcomm
AI research

Advancing AI research to make efficient AI ubiquitous

Power efficiency

Model design, compression, quantization, algorithms, efficient hardware, software tool

Personalization

Continuous learning, contextual, always-on, privacy-preserved, distributed learning

Efficient learning

Robust learning through minimal data, unsupervised learning, on-device learning

A platform to scale AI across the industry

Perception

Object detection, speech recognition, contextual fusion

Reasoning

Scene understanding, language understanding, behavior prediction

Action

Reinforcement learning for decision making

Edge cloud

Cloud

IoT/IIoT

Automotive

Mobile

SYNTIANT

Syntiant Corp. is moving artificial intelligence and machine learning from the cloud to edge devices. Syntiant's chip solutions merge deep learning with semiconductor design to produce ultra-low-power, high performance, deep neural network processors. These network processors enable always-on applications in battery-powered devices, such as smartphones, smart speakers, earbuds, hearing aids, and laptops. Syntiant's Neural Decision Processors™ offer wake word, command word, and event detection in a chip for always-on voice and sensor applications.

Founded in 2017 and headquartered in Irvine, California, the company is backed by Amazon, Applied Materials, Atlantic Bridge Capital, Bosch, Intel Capital, Microsoft, Motorola, and others. Syntiant was recently named a CES® 2021 Best of Innovation Awards Honoree, shipped over 10M units worldwide, and unveiled the NDP120 part of the NDP10x family of inference engines for low-power applications.

www.syntiant.com

@Syntiantcorp

Platinum Sponsors

Part of your life. Part of tomorrow.

www.infineon.com

Reality AI®

Add Advanced Sensing to your Product with Edge AI / TinyML

https://reality.ai   info@reality.ai   @SensorAI   Reality AI

Pre-built Edge AI sensing modules, plus tools to build your own

Reality AI solutions

- Prebuilt sound recognition models for indoor and outdoor use cases
- Solution for industrial anomaly detection
- Pre-built automotive solution that lets cars "see with sound"

Reality AI Tools® software

- Build prototypes, then turn them into real products
- Explain ML models and relate the function to the physics
- Optimize the hardware, including sensor selection and placement

Gold Sponsors

LatentAI

Adaptive AI for the Intelligent Edge

Latentai.com

SensiML™

Build Smart IoT Sensor Devices From Data

SensiML pioneered TinyML software tools that auto generate AI code for the intelligent edge.

- End-to-end AI workflow
- Multi-user auto-labeling of time-series data
- Code transparency and customization at each step in the pipeline

We enable the creation of production-grade smart sensor devices.

sensiml.com

TINY ML EMEA
Virtual
JUNE 7-10, 2021

Silver Sponsors

emza visual sense

GREENWAVES TECHNOLOGIES

HOTG

imagimob

Qeexo

seeed studio
The IoT Hardware Enabler

life.augmented

Copyright Notice

www.tinyML.org