# 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

# tinyML EMEA Technical Forum 2021
# June 7-10, 2021

## Runtime DNN Performance Scaling though Resource Management on Heterogeneous Embedded Platforms

Lei Xun, PhD Candidate, University of Southampton, UK

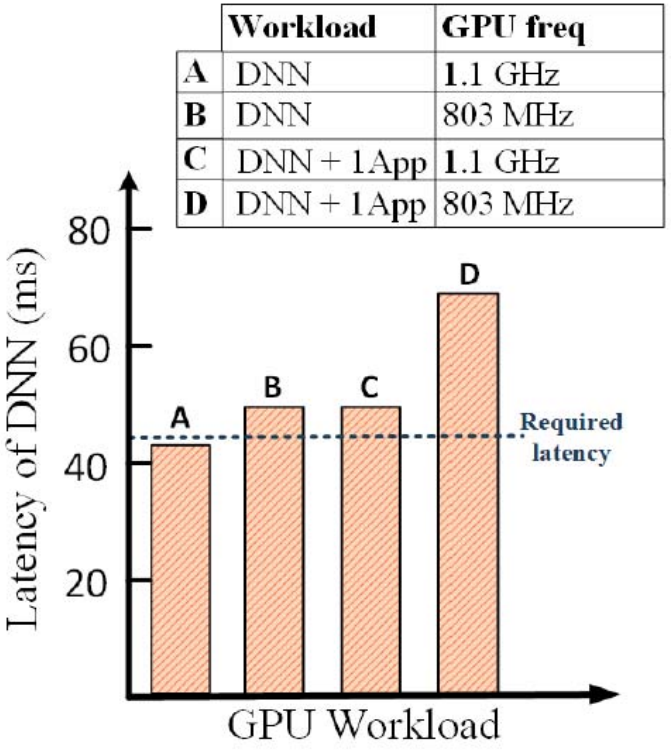
Advisor: Geoff V. Merrett, Jonathon Hare, Bashir M Al-Hashimi

International Centre for Spatial Computational Learning (EPSRC EP/S030069/1)

June 10, 2021

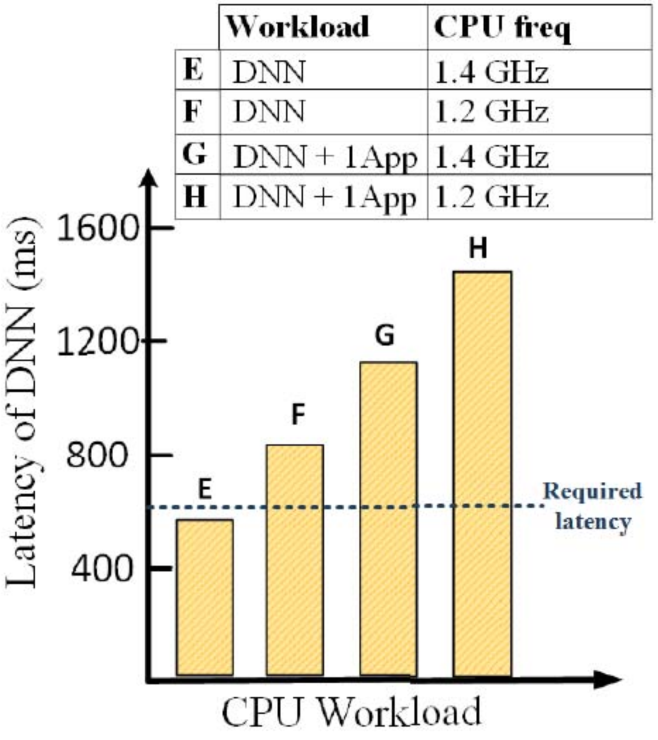
# Motivation for dynamic DNNs

- DNNs are typically compressed before deployed on embedded platform
- However, the assumed hardware resources may not be available at runtime



| | Workload | GPU freq |
|---|---|---|
| A | DNN | 1.1 GHz |
| B | DNN | 803 MHz |
| C | DNN + 1App | 1.1 GHz |
| D | DNN + 1App | 803 MHz |

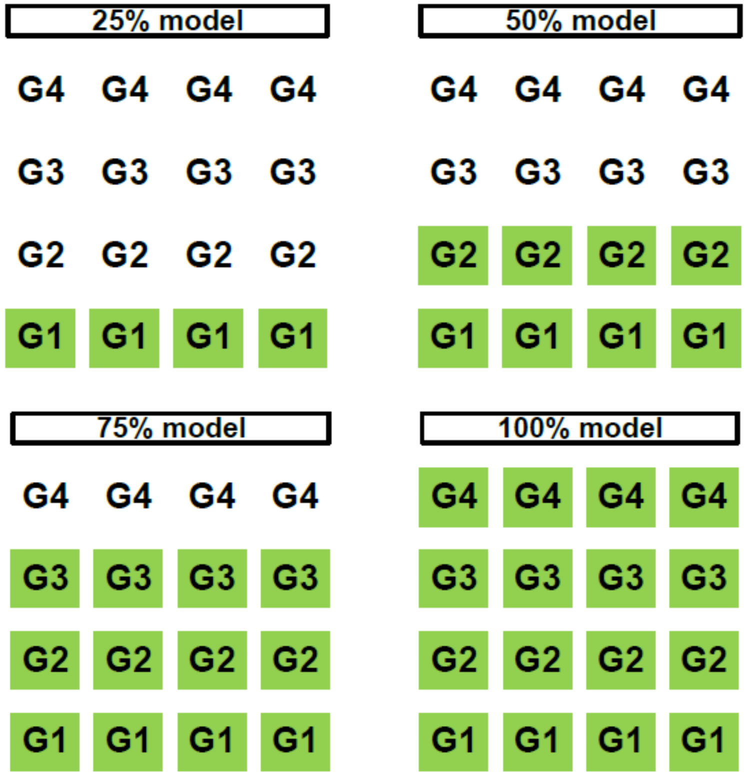
(a)



| | Workload | CPU freq |
|---|---|---|
| E | DNN | 1.4 GHz |
| F | DNN | 1.2 GHz |
| G | DNN + 1App | 1.4 GHz |
| H | DNN + 1App | 1.2 GHz |

(b)

- Dynamic DNNs can be executed partially to trade-off accuracy for latency/power/energy reduction



25% model

| | | | |
|---|---|---|---|
| G4 | G4 | G4 | G4 |
| G3 | G3 | G3 | G3 |
| G2 | G2 | G2 | G2 |
| 🟩 G1 | 🟩 G1 | 🟩 G1 | 🟩 G1 |

50% model

| | | | |
|---|---|---|---|
| G4 | G4 | G4 | G4 |
| G3 | G3 | G3 | G3 |
| 🟩 G2 | 🟩 G2 | 🟩 G2 | 🟩 G2 |
| 🟩 G1 | 🟩 G1 | 🟩 G1 | 🟩 G1 |

75% model

| | | | |
|---|---|---|---|
| G4 | G4 | G4 | G4 |
| 🟩 G3 | 🟩 G3 | 🟩 G3 | 🟩 G3 |
| 🟩 G2 | 🟩 G2 | 🟩 G2 | 🟩 G2 |
| 🟩 G1 | 🟩 G1 | 🟩 G1 | 🟩 G1 |

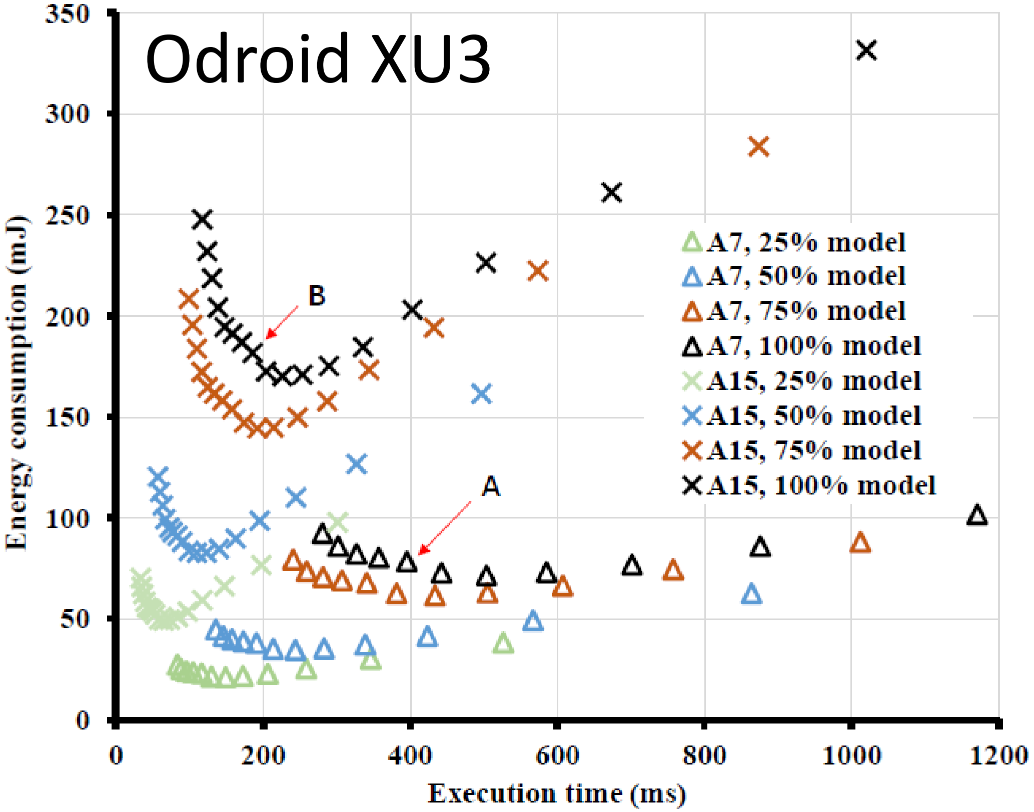
100% model

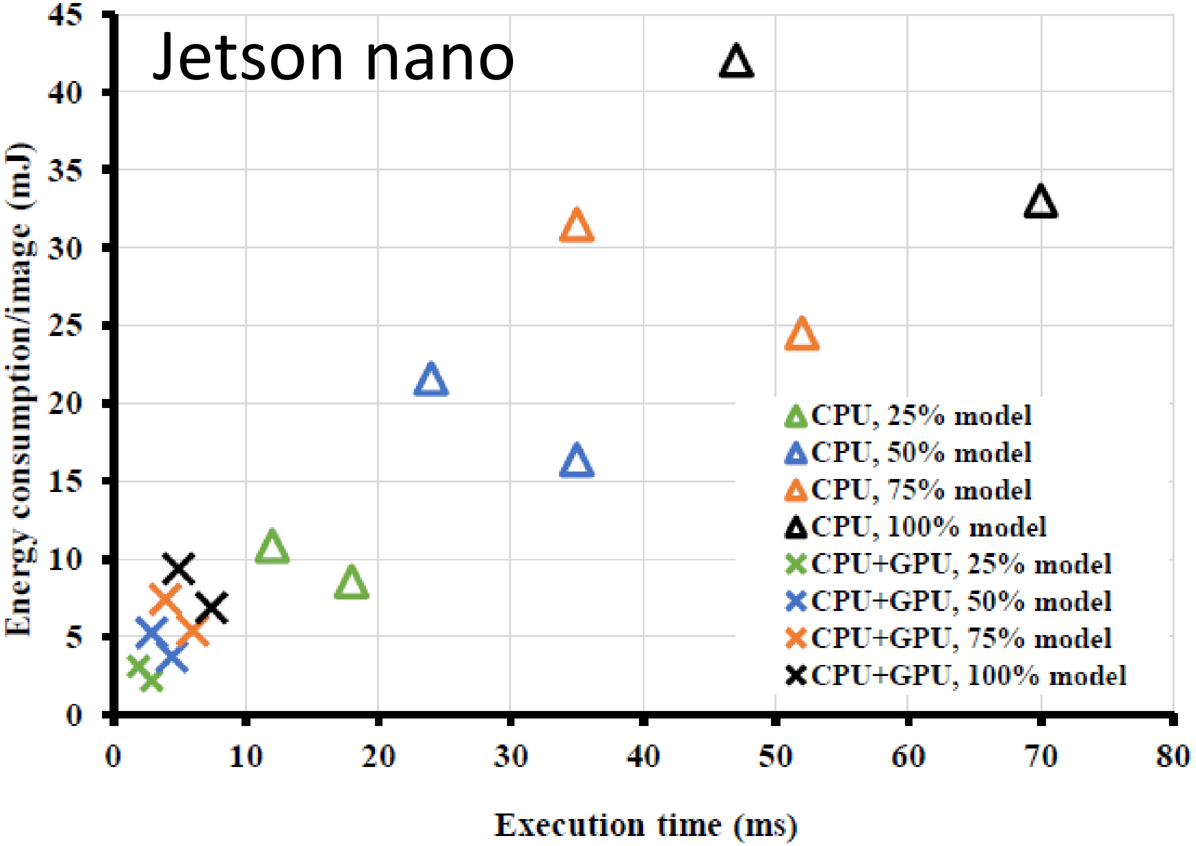
| | | | |
|---|---|---|---|
| 🟩 G4 | 🟩 G4 | 🟩 G4 | 🟩 G4 |
| 🟩 G3 | 🟩 G3 | 🟩 G3 | 🟩 G3 |
| 🟩 G2 | 🟩 G2 | 🟩 G2 | 🟩 G2 |
| 🟩 G1 | 🟩 G1 | 🟩 G1 | 🟩 G1 |

Lei Xun, Long Tran-Thanh, Bashir M Al-Hashimi, and Geoff V Merrett (2019) *Incremental training and group convolution pruning for runtime DNN performance scaling on heterogeneous embedded platforms*. in **Workshop on Machine Learning for CAD (MLCAD).**

# Accuracy/latency/energy trade-offs

- Subnetworks are shown in different colors, computing elements are shown in different symbols) and frequency scaling are shown in points
- Operating points example: On Odroid XU3, A has the best trade-off under 100mJ and 400ms requirements, B is the best for 200mJ and 200ms





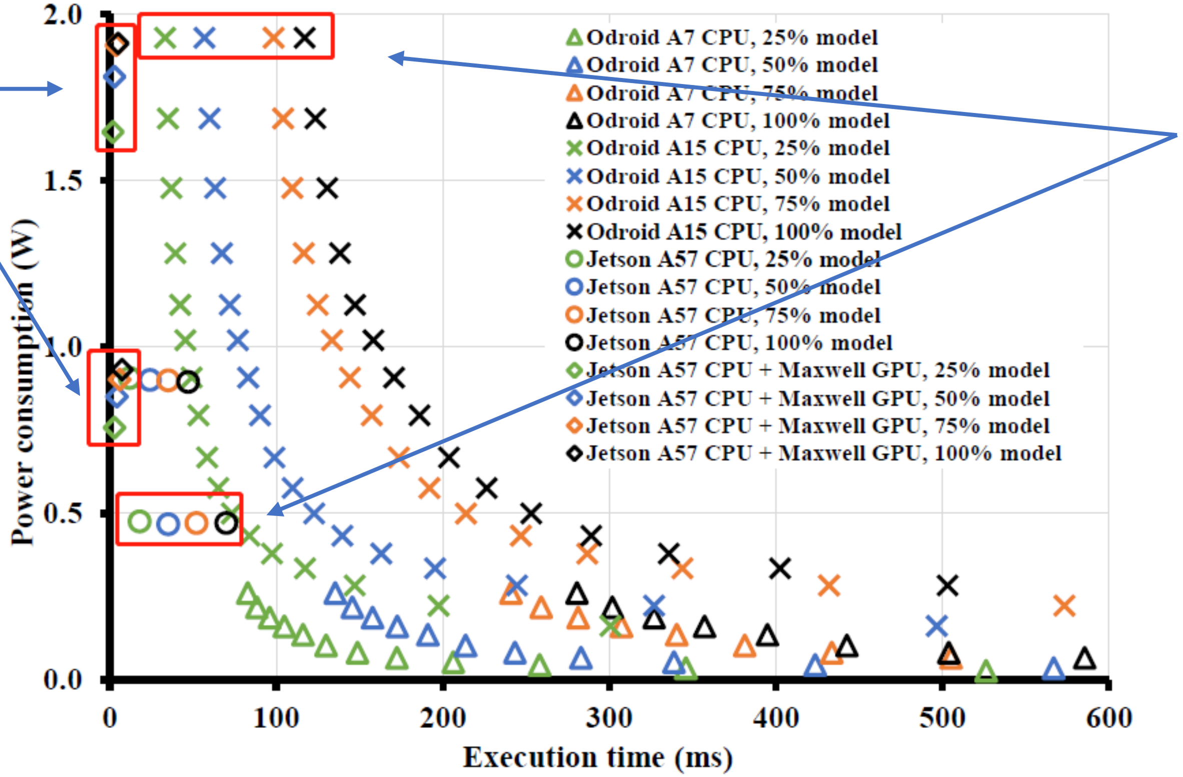
Lei Xun, Long Tran-Thanh, Bashir M Al-Hashimi, and Geoff V Merrett (2020) *Optimising resource management for embedded machine learning*. In **Design, Automation and Test in Europe Conference (DATE).**

# Accuracy/latency/power trade-offs

Now we know that dynamic DNNs can trade-off accuracy for latency and energy, what about power?

The power of GPU

- Scales with frequency scaling
- Scales with dynamic DNN, this provide us new opportunities to meet power target



The power of a single-core CPU:

- Scales with frequency scaling
- Does not scale with dynamic DNN, since the computation intensity does not change, only the latency changes

To know more about dynamic DNNs, please check out our SOTA dynamic DNN paper:

Wei Lou*, Lei Xun*, Mohammadamin Sabetsarvestani, Jia Bi, Jonathon Hare, Geoff V Merrett (2021) *Dynamic-OFA: Runtime DNN architecture switching for performance scaling on heterogeneous embedded platforms*. In **Conference on Computer Vision and Pattern Recognition Workshops (CVPR'W).** https://arxiv.org/abs/2105.03596

# Premier Sponsor

Intelligent Agent
Neuton

# Automated TinyML

Zero-code SaaS solution

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

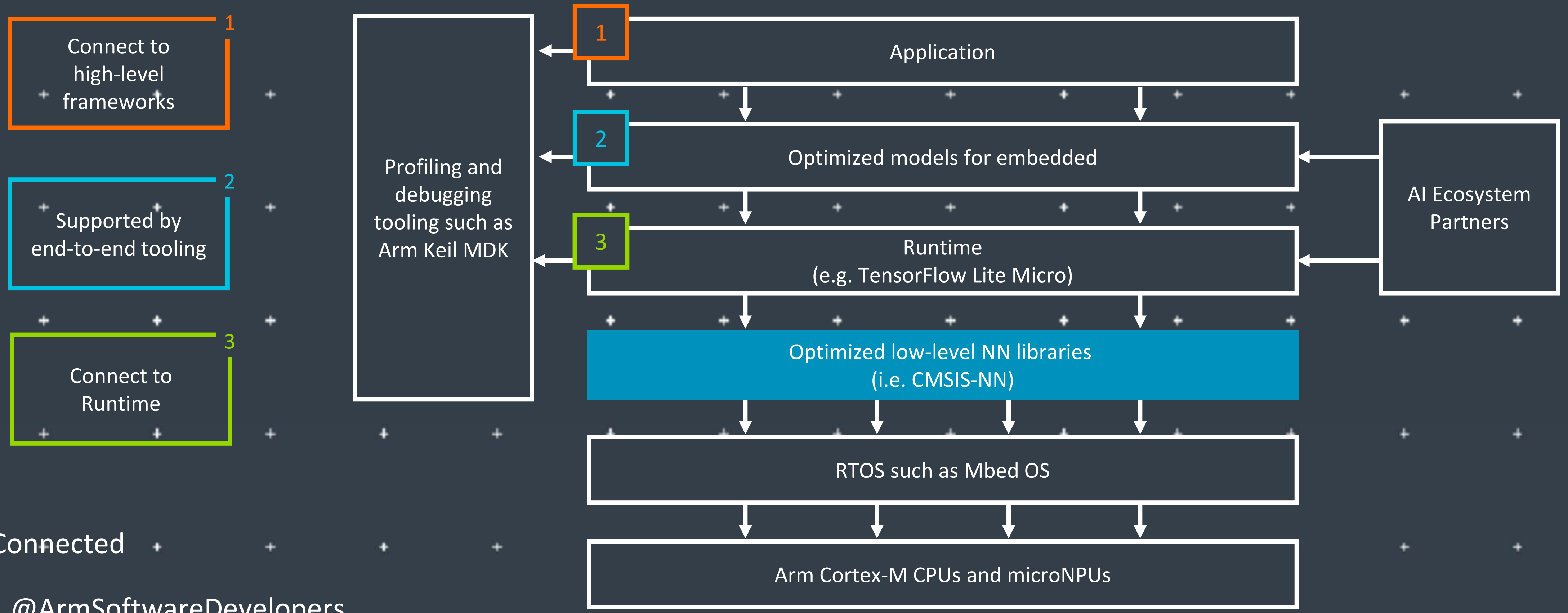
Compare the benchmarks of our compact models to those of TensorFlow and other leading neural network frameworks.

***Build Fast. Build Once. Never Compromise.***

TINY ML EMEA
Virtual
JUNE 7-10,2021

# Executive Sponsors

# Arm: The Software and Hardware Foundation for tinyML



1
Connect to high-level frameworks
2
Supported by end-to-end tooling
3
Connect to Runtime
Profiling and debugging tooling such as Arm Keil MDK
1
Application
2
Optimized models for embedded
3
Runtime (e.g. TensorFlow Lite Micro)
AI Ecosystem Partners
Optimized low-level NN libraries (i.e. CMSIS-NN)
RTOS such as Mbed OS
Arm Cortex-M CPUs and microNPUs

Stay Connected

@ArmSoftwareDevelopers

@ArmSoftwareDev

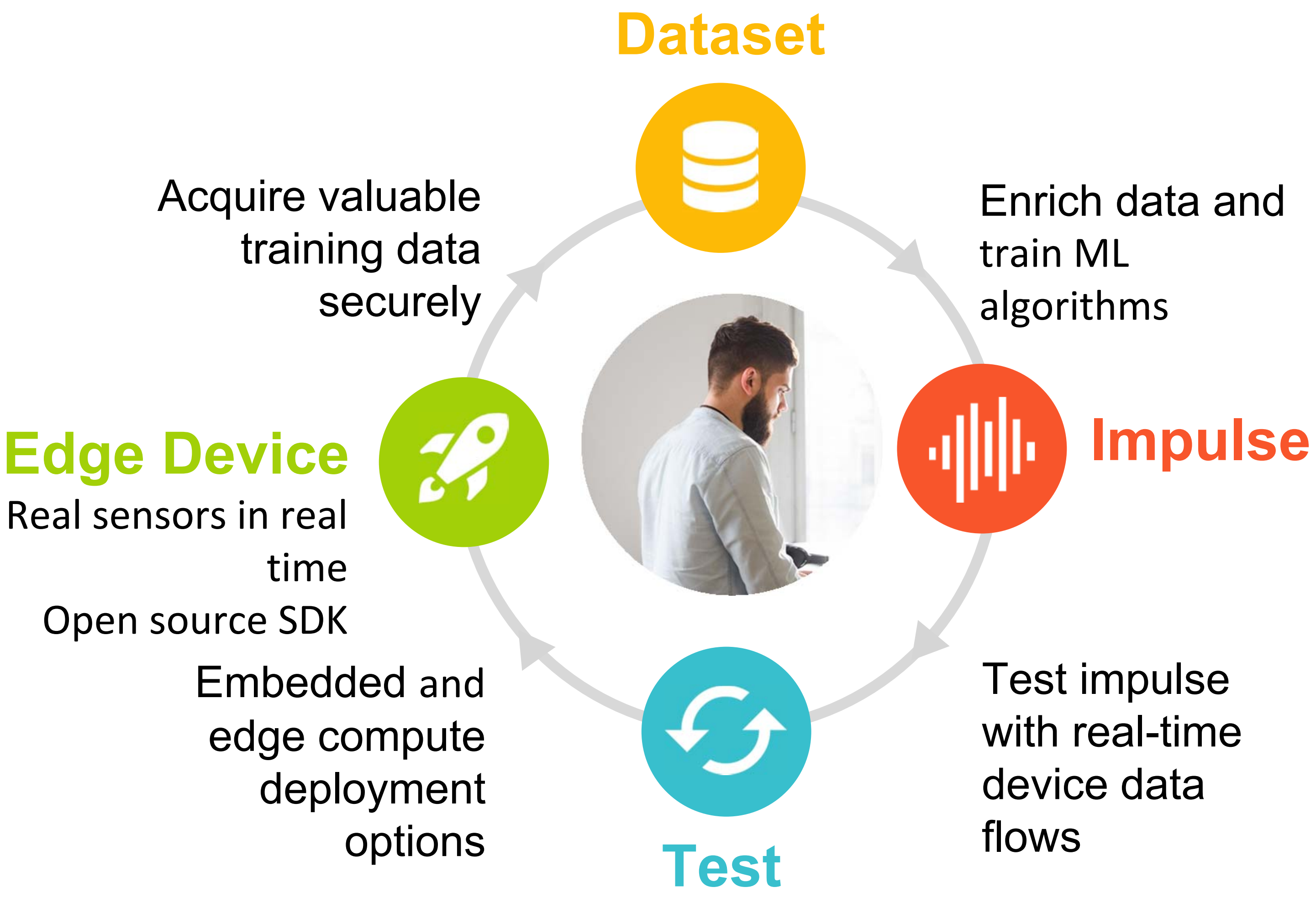
Resources: developer.arm.com/solutions/machine-learning-on-arm

# TinyML for all developers

C++ library

Arduino library

WebAssembly



**Dataset**

Acquire valuable training data securely

Enrich data and train ML algorithms

**Impulse**

Test impulse with real-time device data flows

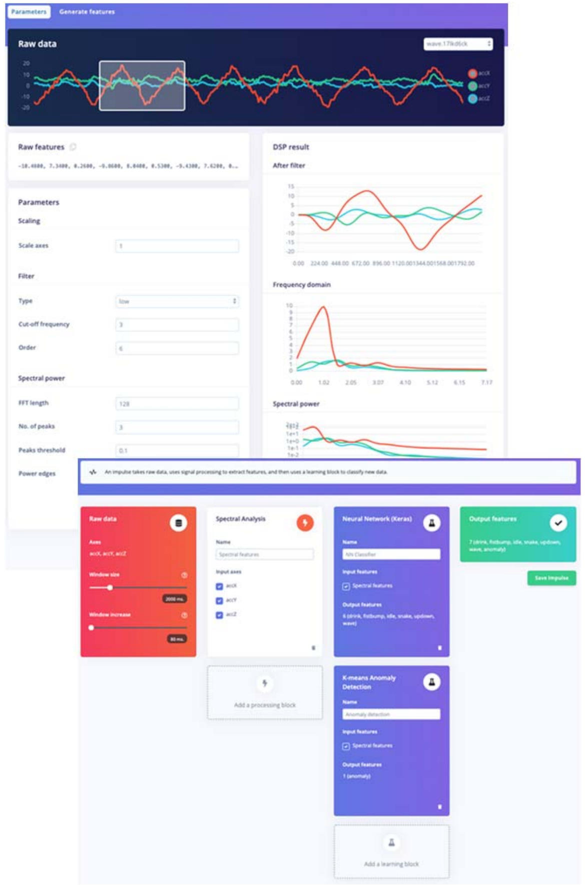
**Test**

Embedded and edge compute deployment options

**Edge Device**

Real sensors in real time

Open source SDK



Parameters | Generate features

Raw data — wave.17lkd6ck

Raw features

-10.4000, 7.3400, 0.2600, -0.0600, 0.0400, 0.5300, -5.4300, 7.6200, 0...

Parameters

Scaling

Scale axes: 1

Filter

Type: low

Cut-off frequency: 3

Order: 6

Spectral power

FFT length: 128

No. of peaks: 3

Peaks threshold: 0.1

Power edges

DSP result

After filter

Frequency domain

Spectral power

An impulse takes raw data, uses signal processing to extract features, and then uses a learning block to classify new data.

Raw data

Axes: accX, accY, accZ

Window size: 2000 ms.

Window increase: 80 ms.

Spectral Analysis

Name: Spectral features

Input axes: ☑ accX ☑ accY ☑ accZ

Add a processing block

Neural Network (Keras)

Name: NN Classifier

Input features: ☑ Spectral features

Output features: 6 (drink, fistbump, idle, snake, updown, wave)

K-means Anomaly Detection

Name: Anomaly detection

Input features: ☑ Spectral features

Output features: 1 (anomaly)

Add a learning block

Output features

7 (drink, fistbump, idle, snake, updown, wave, anomaly)

Save Impulse

[www.edgeimpulse.com](http://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

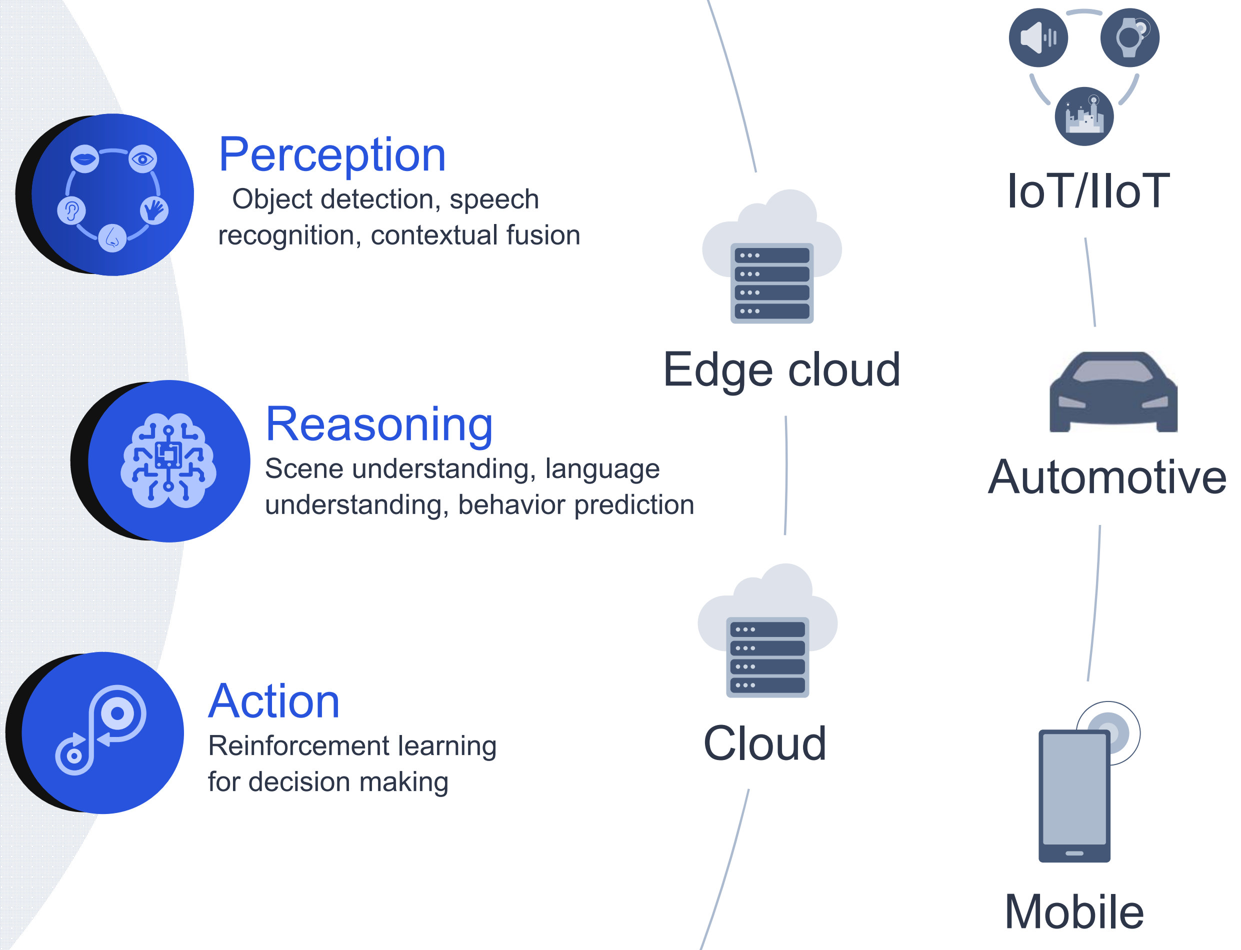
## Personalization

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

# 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$^{TM}$ 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

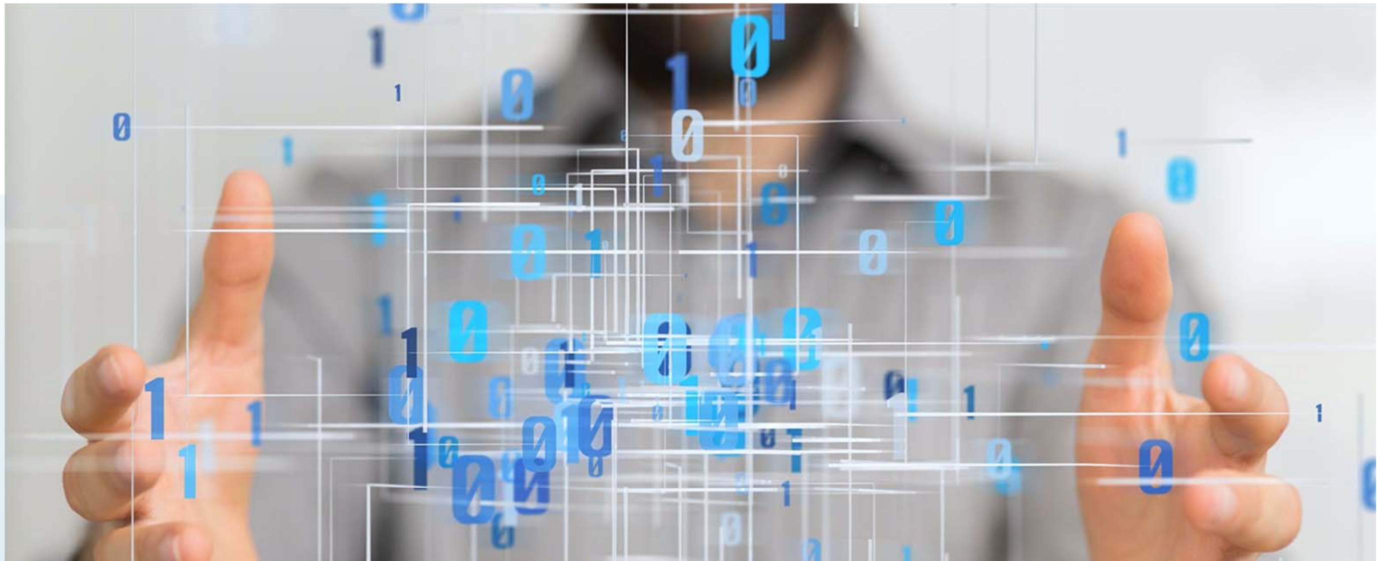
SensiML™

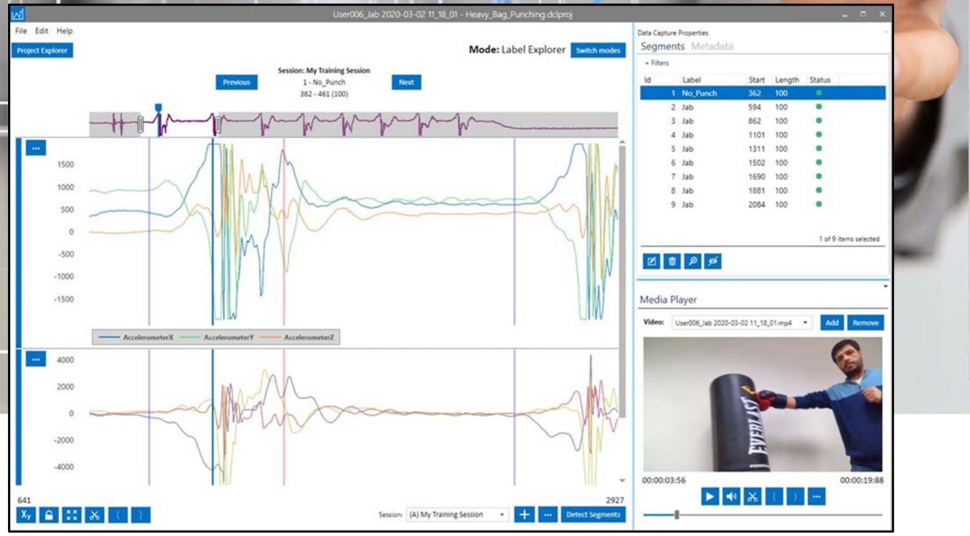
# Build Smart IoT Sensor Devices From Data

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.





Session: My Training Session
1 - No_Punch
362 - 461 (100)

Session: (A) My Training Session

| Id | Label | Start | Length | Status |
|---|---|---|---|---|
| 1 | No_Punch | 362 | 100 | ● |
| 2 | Jab | 594 | 100 | ● |
| 3 | Jab | 862 | 100 | ● |
| 4 | Jab | 1101 | 100 | ● |
| 5 | Jab | 1311 | 100 | ● |
| 6 | Jab | 1502 | 100 | ● |
| 7 | Jab | 1690 | 100 | ● |
| 8 | Jab | 1881 | 100 | ● |
| 9 | Jab | 2084 | 100 | ● |

1 of 9 items selected

Video: User006_Jab 2020-03-02 11_18_01.mp4

sensiml.com

# Silver Sponsors

emza
visual sense

GREENWAVES
TECHNOLOGIES

HOTG

imagimob

Qeexo

seeed studio
The IoT Hardware Enabler

ST
life.augmented

# Copyright Notice

www.tinyML.org