tinyML® Talks

Enabling Ultra-low Power Machine Learning at the Edge

"Empowering the Edge: Practical Applications of Embedded Machine Learning on MCUs"

Jongmin Lee – Machine Learning Architecture Engineer, NXP

May 25, 2023

www.tinyML.org

TINY ML TALKS webcast

Thank you, **tinyML Strategic Partners**, for committing to take tinyML to the next Level, together

TINY ML TALKS webcast

Executive Strategic Partners

TINY ML TALKS webcast

EDGE IMPULSE

The Leading Development Platform for Edge ML

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

Accelerate Your Edge Compute

SYNTIANT

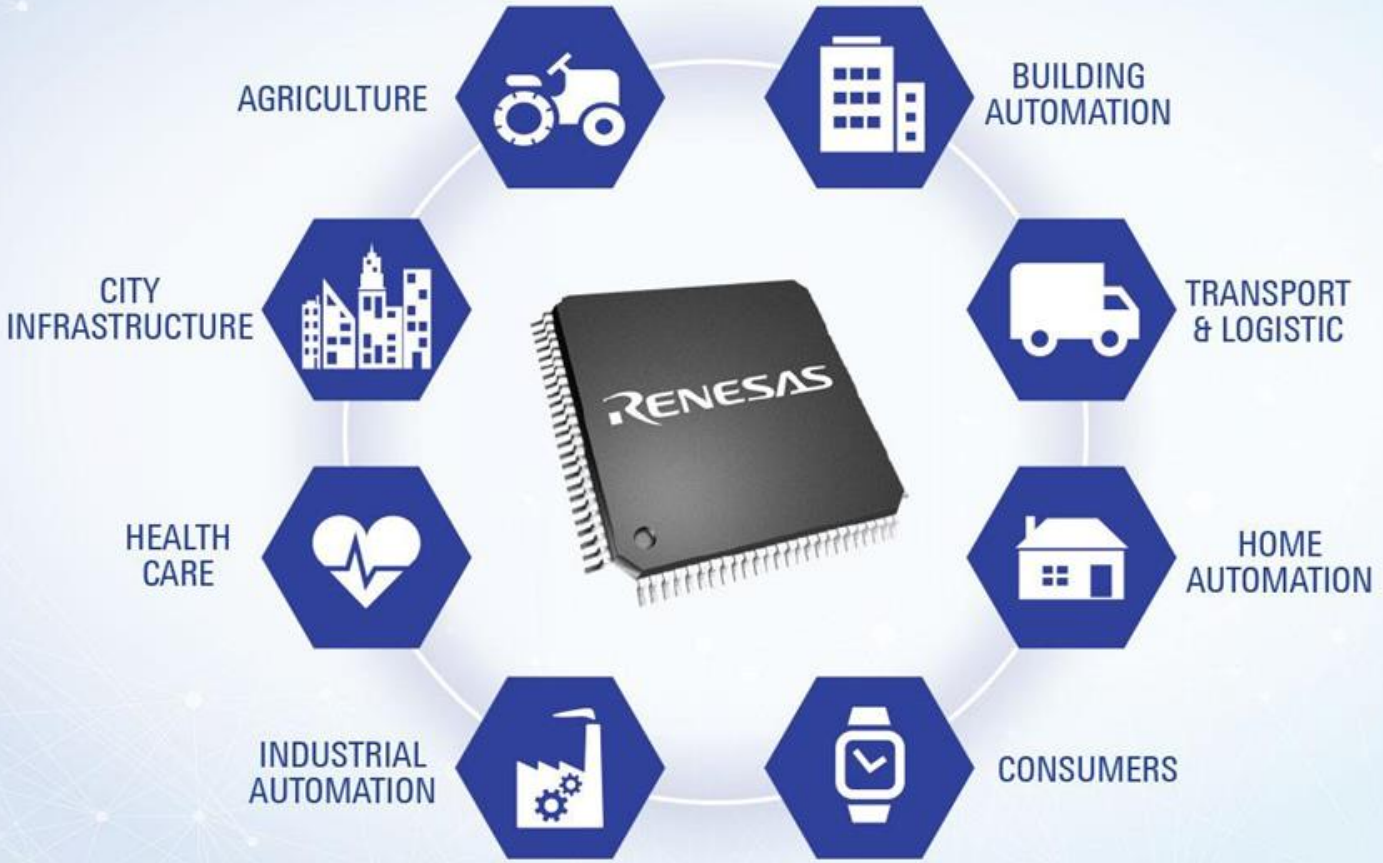
Making Edge AI A Reality

www.syntiant.com

TINY ML TALKS webcast

Platinum Strategic Partners

Renesas is enabling the next generation of AI-powered solutions that will revolutionize every industry sector.



renesas.com

AITRIOS

DEPLOY VISION AI AT THE EDGE AT SCALE

SONY

TINY ML TALKS webcast

Gold Strategic Partners

ANALOG DEVICES

AHEAD OF WHAT'S POSSIBLE™

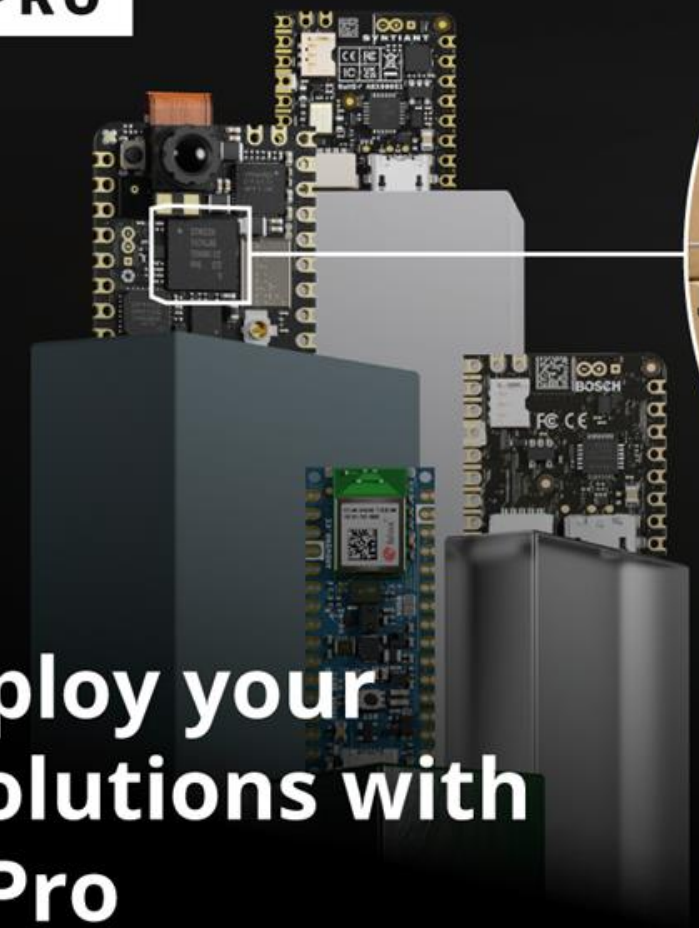
ANALOG DEVICES

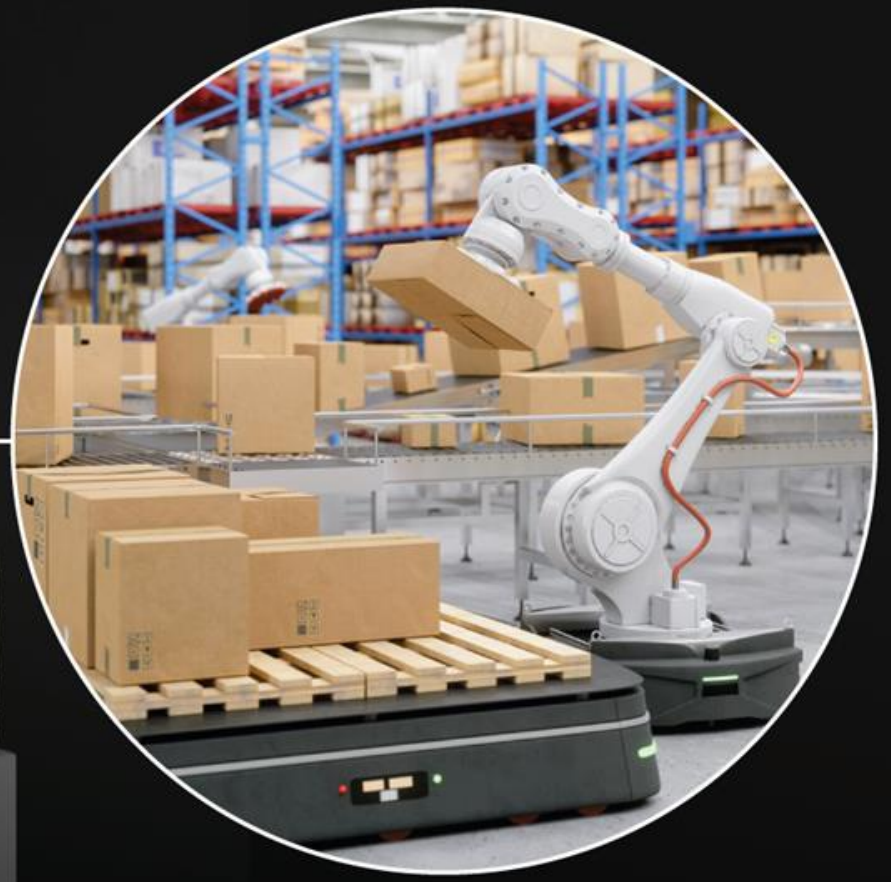
AHEAD OF WHAT'S POSSIBLE™

Where what if becomes what is.

Witness potential made possible at analog.com.

ARDUINO PRO™





Easily deploy your tinyML solutions with Arduino Pro

arduino.cc/pro

Made In Italy

arm AI

Powering tinyML Innovation

Arm AI Virtual Tech Talks

The latest in AI trends, technologies & best practices from Arm and our Ecosystem Partners.

Demos, code examples, workshops, panel sessions and much more!

Fortnightly Tuesday @ 4pm GMT/8am PT

Find out more:
www.arm.com/techtalks

Driving decarbonization and digitalization. Together.

Infineon serving all target markets as
Leader in Power Systems and IoT

www.infineon.com

innatera

NEUROMORPHIC INTELLIGENCE FOR THE SENSOR-EDGE

www.innatera.com


Microsoft

The Right Edge AI Tools Can Make or Break Your Next Smart IoT Product

SensiML

Analytics Toolkit Suite

sensiml.com/tinyML

STMicroelectronics provides extensive solutions to make tiny Machine Learning easy

life.augmented

www.st.com/ai

synaptics®

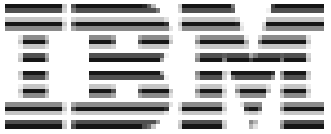
ENGINEERING EXCEPTIONAL EXPERIENCES

We engineer exceptional experiences for consumers in the home, at work, in the car, or on the go.

www.synaptics.com

TINY ML TALKS webcast

Silver Strategic Partners



Join Growing tinyML Communities:

14.8k members in
47 Groups in 39 Countries

tinyML - Enabling ultra-low Power ML at the Edge

https://www.meetup.com/tinyML-Enabling-ultra-low-Power-ML-at-the-Edge/

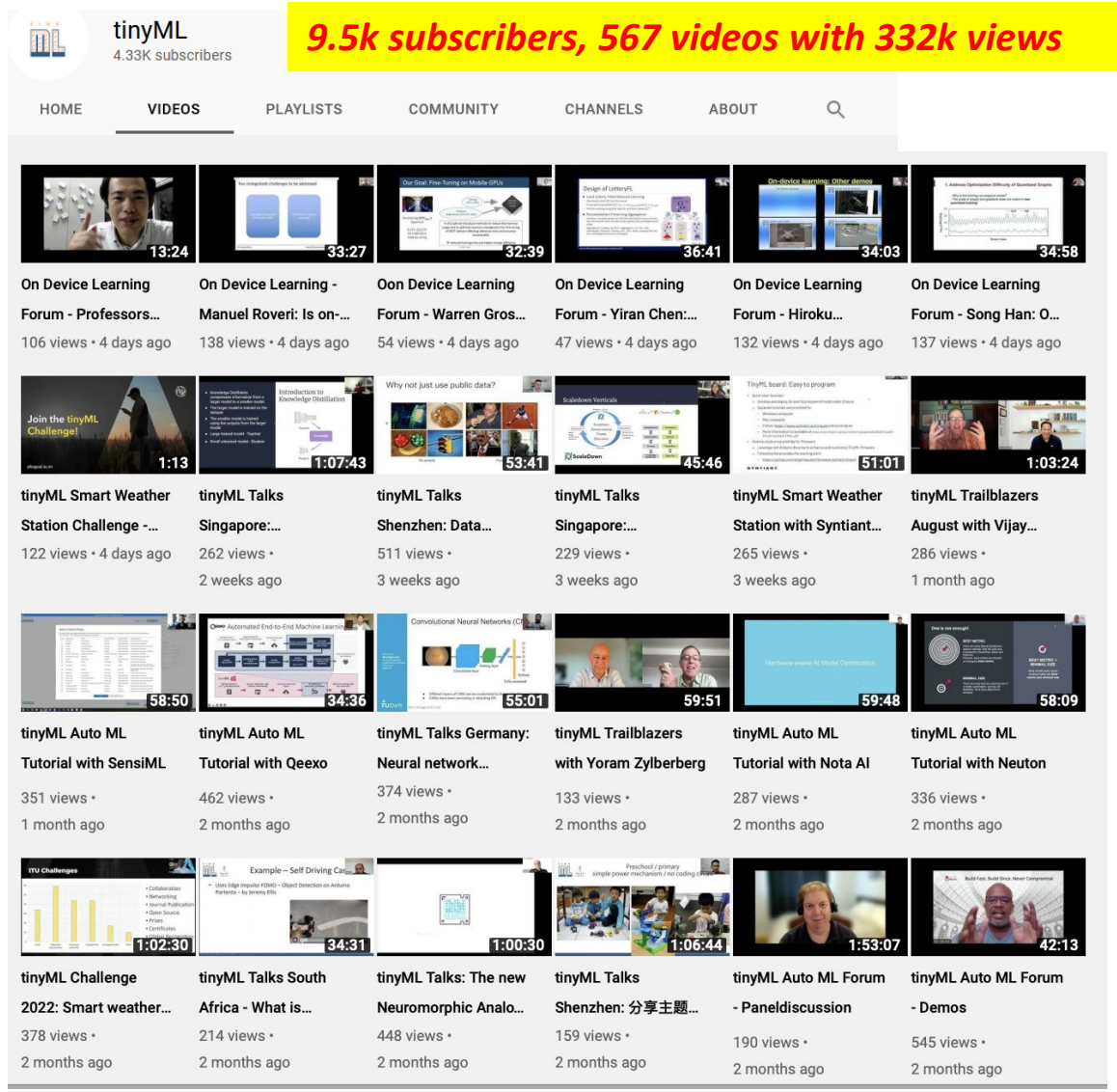
4k members
&
11.6k followers

The tinyML Community

https://www.linkedin.com/groups/13694488/

TINY ML TALKS webcast

YouTube



Subscribe to tinyML YouTube Channel for updates and notifications

(including this video)

www.youtube.com/tinyML

TINY ML TALKS webcast

TINY ML FOUNDATION

EMEA 2023

https://www.tinyml.org/event/emea-2023

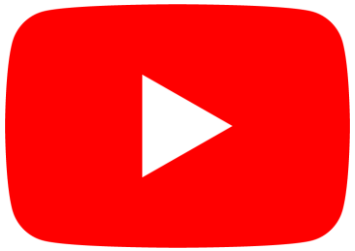
More sponsorships are available: sponsorships@tinyML.org

TINY ML TALKS webcast

Reminders

Slides & Videos will be posted tomorrow

tinyml.org/forums



youtube.com/tinyml

Please use the Q&A window for your questions

TINY ML TALKS webcast

Jongmin Lee

Jongmin Lee is a Machine Learning Architecture Engineer at NXP semiconductors. Throughout his career at NXP, he has primarily focused on developing MCU-based machine learning solutions, and currently, his area of concentration is on advancing the architecture of neural processing unit. He earned his Ph.D. degree in Electrical Engineering from Arizona State University, Tempe, in 2017. He was a research assistant at the Sensor, Signal, and Information Processing (SenSIP) center at Arizona State University.

Empowering the Edge: Practical Applications of Embedded Machine Learning on MCUs

Jongmin Lee, PhD
ML Architecture Engineer

TinyML Talks – Phoenix
May 25, 2023

with contributions from:
Michael Wang, Jianfeng Qin, Anthony Huereca

SECURE CONNECTIONS FOR A SMARTER WORLD

CONTENTS

NXP BROAD-BASED MACHINE LEARNING SOLUTIONS AND SUPPORT

eIQ Machine Learning SW

eIQ® ML SW Development Environment

eIQ Toolkit with eIQ Portal GUI to:

- Import/create, convert, optimize , validate and deploy ML models
- Dataset curation tools to create new, augment, label/annotate datasets

eIQ inference with:

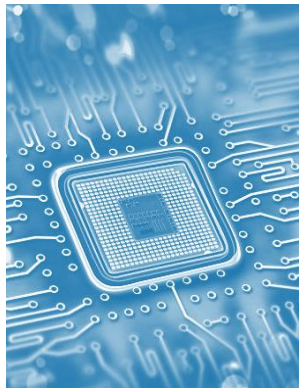
- TensorFlow Lite, TensorFlow Lite Micro and DeepViewRT

eIQ Marketplace:

- Add-on wares available from eco-system partners and NXP for ML applications, optimized models, optimization tools, datasets and sensor solutions

Support for i.MX 8M, i.MX 9, i.MX RT, MCX family of devices

Integrated with NXP dev environments (MCUXpresso, Yocto/Linux)



NXP eIQ Neutron NPU

NXP eIQ Neutron NPU

- Highly scalable ML acceleration cores
- Unified architecture and software support
- Optimized for edge performance and power dissipation

Turnkey Solutions and Voice Technology

Smart HMI solution

- i.MX RT117H (kit - SLN-TLHMI-IOT-RD)

Face & emotion recognition solution with Anti-Spoofing

- i.MX RT106F (kit – SLN-VIZN-IOT)

Local voice control solution

- i.MX RT106L (kit – SLN-LOCAL-IOT)
- i.MX RT106S (kit – SLN-LOCAL2-IOT)

Conversa, VoiceSeeker, VoiceSpot, and Voice Intelligent Technology

VISION KIT MICRO

Third Party SW and HW

- Cloud ML lifecycle and analytics services integration
- Video applications and HW acceleration support with NXP eIQ Toolkit

…. And more

EMBEDDED MACHINE LEARNING AT THE EDGE

- Deploying ML on embedded devices can
  - Reduce data traffic between the edge devices and the cloud.
  - Reduce latency and can respond in nearly real-time to input data.
  - Protect privacy as raw data is not transmitted to the cloud.

- When it comes to MCU-based embedded ML
  - The performance is often limited by HW resources.
    - Memory – typically less than 1MB
    - Clock speed – a few hundreds MHz, less than 1GHz
  - Still, it's possible to deploy lightweight ML models and run them with good performance!

ML Applications Running on MCU

EXAMPLE APPLICATIONS OF MCU-BASED ML

Anomaly Detection

- ✓ Sensor data analytics
- ✓ Classify states or behavior with sensors
- ✓ Predict potential failure behavior early
- ✓ Smartwatch, predictive maintenance, etc

Voice

- ✓ Keyword spotting – wake word
- ✓ Voice commands

Vision

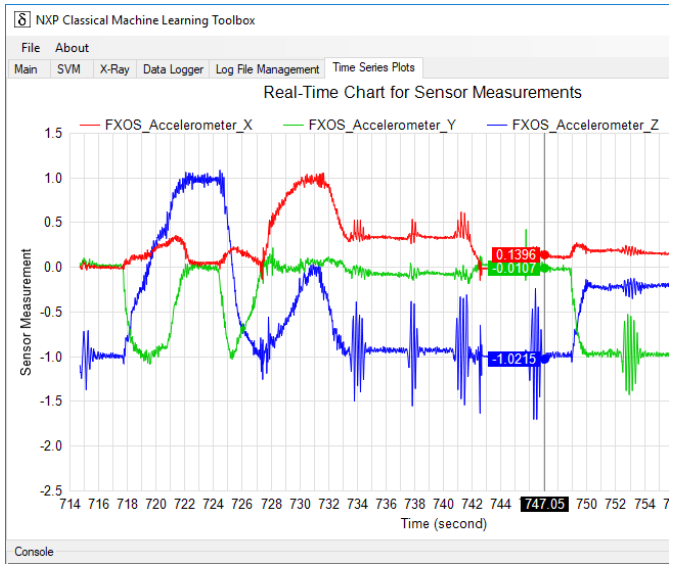
- ✓ Image classification
- ✓ Face recognition
- ✓ Object detection
- ✓ Personalization based on face recognition

ANOMALY DETECTION

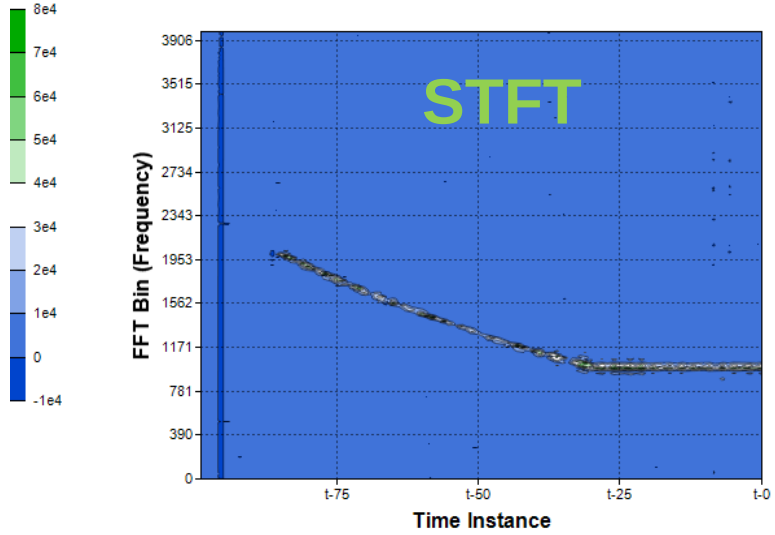
- Sensor data in various forms
  - Motion: accelerometer, magnetometer, gyroscope, etc
  - Environmental: pressure, humidity, temperature, etc
  - Audio: microphone
  - and more

- Combine sensor data, advanced MCUs, and ML to enable
  - Machine condition monitoring
  - Environmental awareness
  - Structural monitoring
  - Audio event detection

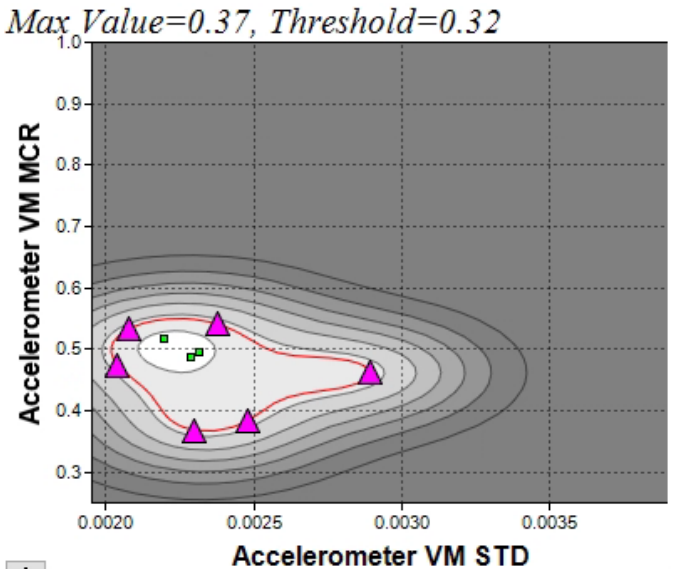
SENSOR DATA LOGGING, FEATURE EXTRACTION, INFERENCING, AND …



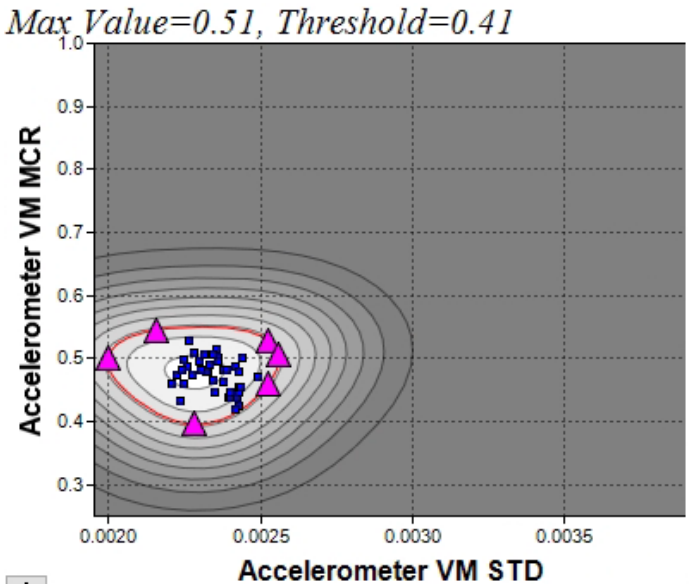
- ✓ Real time sensor data streams



- ✓ Feature extractions (eg, FFT)

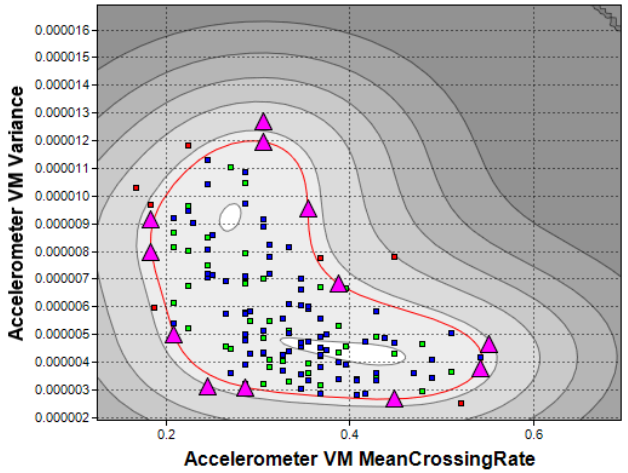


- ✓ SVM Inferencing for anomaly detection



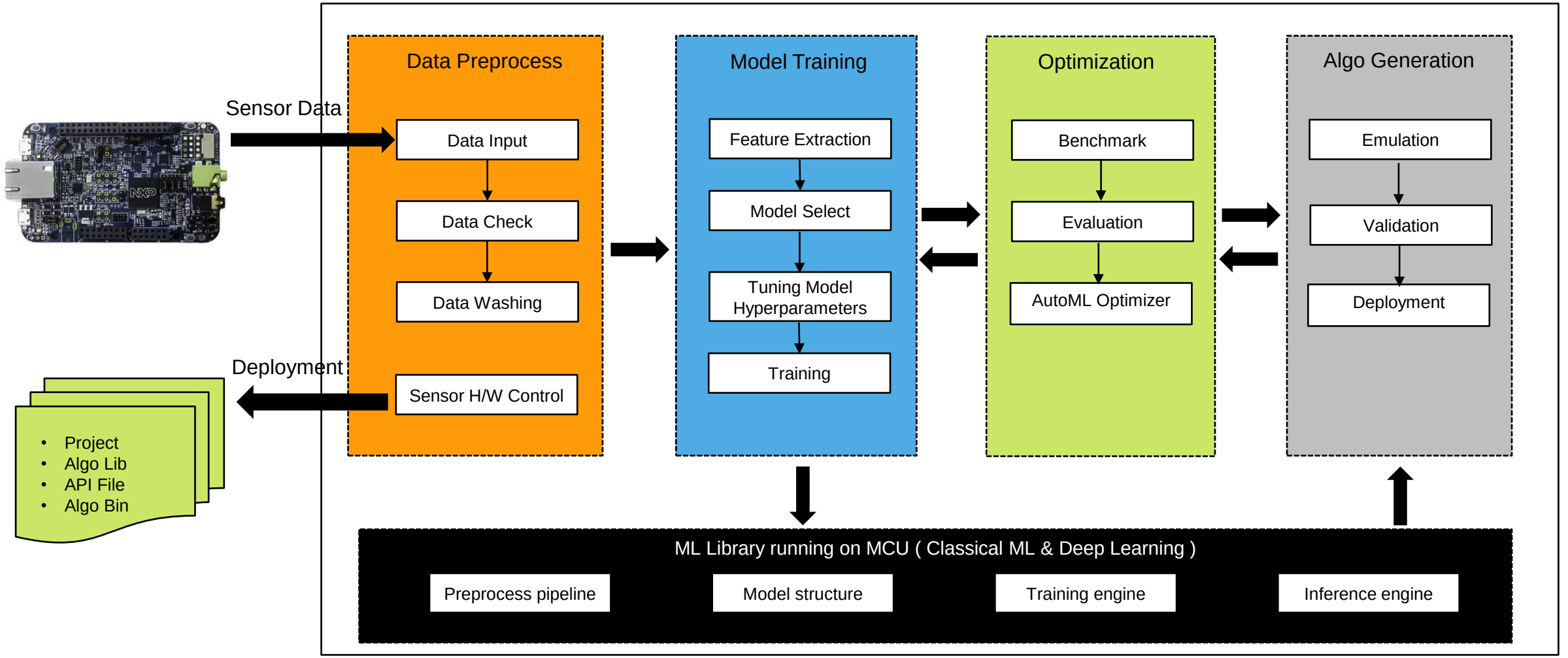
- ✓ Training an SVM model in-situ

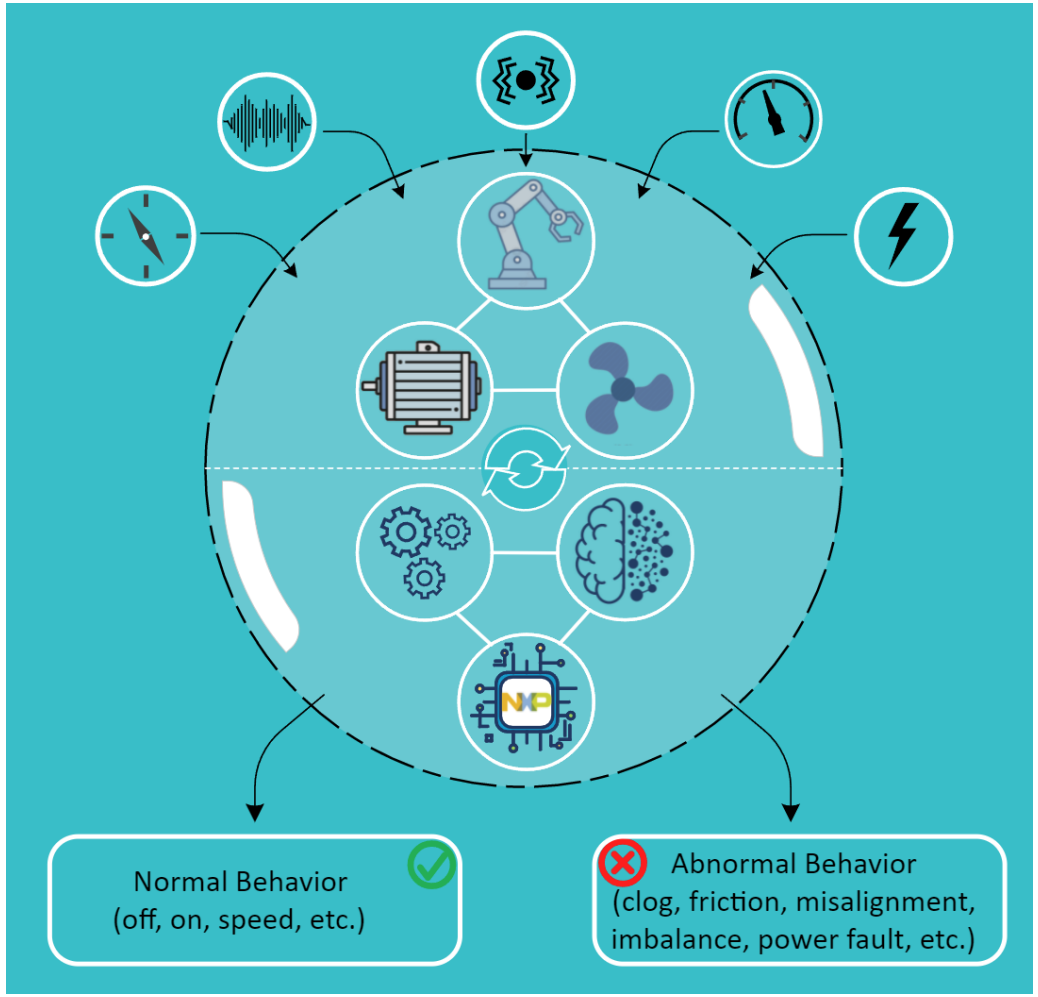
KEY TECHNICAL FEATURES



- ML based Anomaly Detection running in MCUs
  - FRDM K64F (or K66F)
    - Arm Cortex-M4 120 MHz or greater
    - 256KB SRAM
    - Motion sensors
  - i.MX RT1062 custom board
    - Arm Cortex-M7 528 MHz (up to 600 MHz)
    - 1 MB SRAM
    - extended set of sensors and cloud connectivity
- Enough resources to process
  - Sensor data logging
  - Feature extraction
  - SVM inferencing
- Needed more automated way of model optimization and evaluation pipeline

TIME SERIES ML S/W BLOCK DIAGRAM & PROCESS





ML STATE MONITOR APP SW PACK

Goal: Build ML-based Smart Sensing Appliances

How: Relies on Deep Learning and enables developers to build and deploy Neural Networks on NXP's MCU-based systems for developing Smart Sensing Appliances.

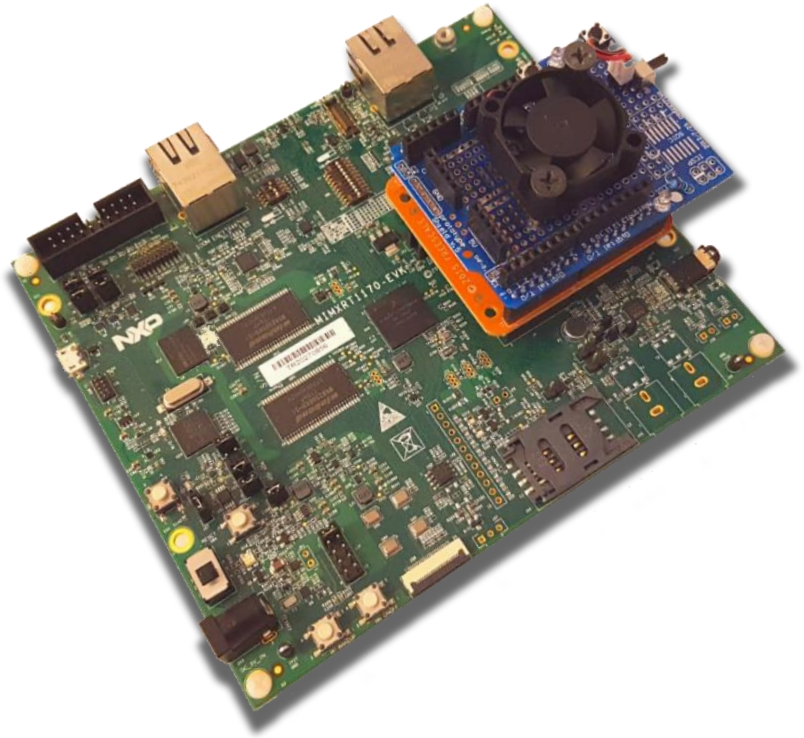
Also provides details on how to validate and evaluate the performance of a model by running it through different inference engines on an embedded sensing device.

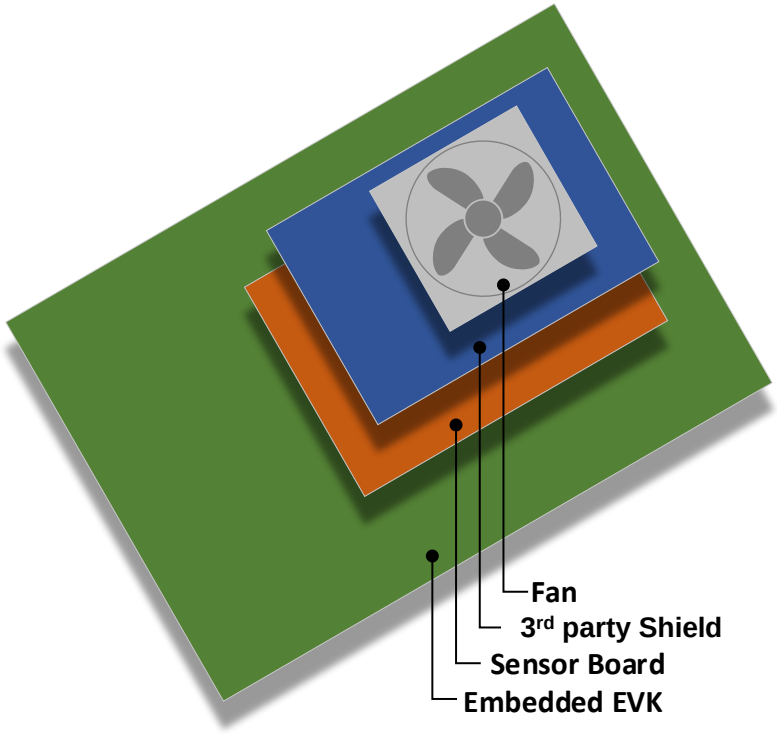
Sensing Device: A device capable of measuring various environment parameters through different sensors (acceleration, magnitude, orientation, temperature, pressure, sound, electric current, etc.)

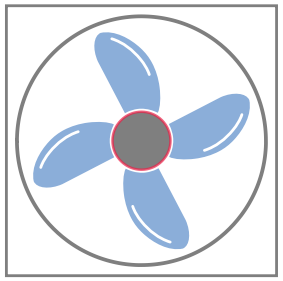
Related application spaces: System state monitoring, Activity recognition, Machine health (preventive maintenance, anomaly detection, failure identification)

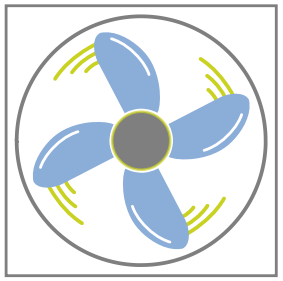
ML STATE MONITOR APP SW PACK – USE CASE

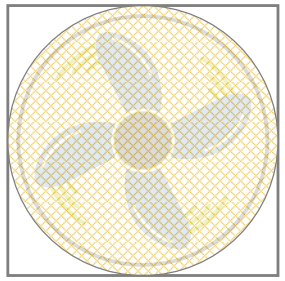
FAN STATE MONITORING AND FAILURE IDENTIFICATION

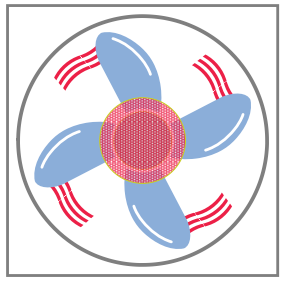





OFF


ON


Clogged


Friction

ML STATE MONITOR APP SW PACK – ONE PAGER

APPLICATION	Fan State Classifier (off, on, clogged, friction) - Fan State Monitoring and Failure Identification using Deep Learning and eIQ Tools **Related application spaces:** • System state monitoring, Activity recognition, Machine health (preventive maintenance, anomaly detection, failure identification)
DATA	Vibration 10h20m of recording at 200Hz (7.44 mega of samples * 6B) **Applicable to:** Any time series data
INPUT SOURCES	3-axis accelerometer **Applicable to:** Other sensors that measure – magnitude, orientation, pressure, temperature, electric current, sound, etc.
EXPANSION HW	Vibration source (fan/motor), Sensors, SD card
MODEL	Custom Convolutional Neural Network Layers: Conv, Dense, Dropout, Pool, Flatten
TOOLS	MCUXpresso IDE, MCUXpresso SDK, Jupyter Notebook eIQ Toolkit, Python, TensorFlow, Keras
DEPLOYMENT	eIQ Portal and command line TFLite Micro, DeepViewRT, Glow
PLATFORMS SUPPORTED	Arm Cortex-M7 (MIMXRT1170-EVK) Arm Cortex-M4 (FRDM-K66F) Arm Cortex-M33 (LPC55xx) Arm Cortex-M33 (MCX-N, coming)
PERFORMANCE METRICS	Inference time: down to 150µs Accuracy: up to 99% Model Size: 10K (total params) / 45KB (float) / 15KB (quant)
COLLATERAL	**App note** **Lab Guide and Video** **Training Dataset and Dataset Creation Guide** **Software Package**
LEARN MORE:	https://nxp.com/appswpack

LOCAL VOICE CONTROL

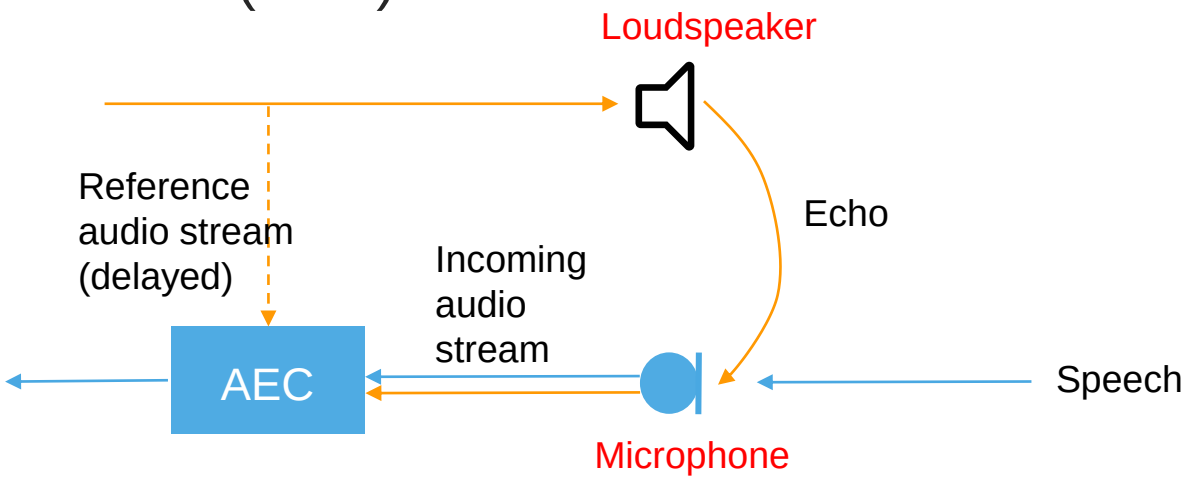
- Anywhere that needs hands-free, private voice control without cloud connectivity
  - Smart Home
  - Smart appliances
  - Smart buildings and industrial

- Audio front end
  - Echo cancellation
    - While speaker is playing, need to recognize wake word(s).
  - Far-field voice recognition (up to 3~5m)

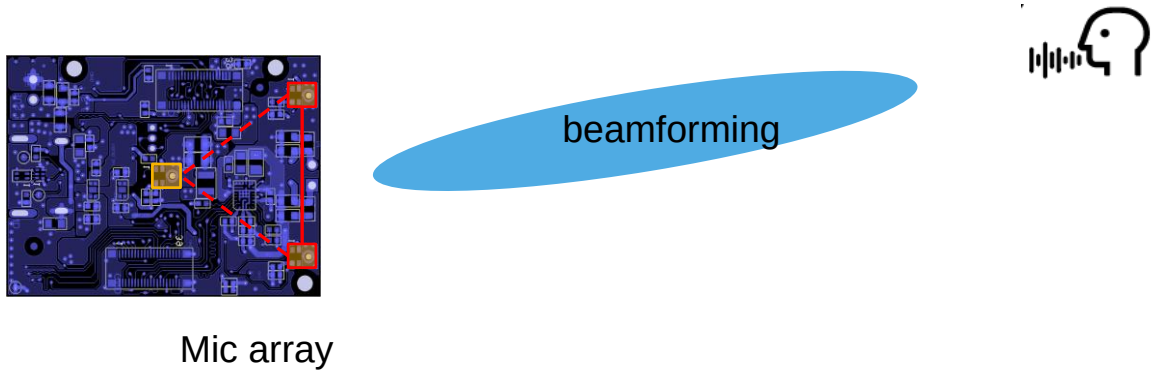
- Voice engine
  - Keyword spotting on wake word(s) and commands

AUDIO PROCESSING

- Acoustic echo cancellation (AEC)



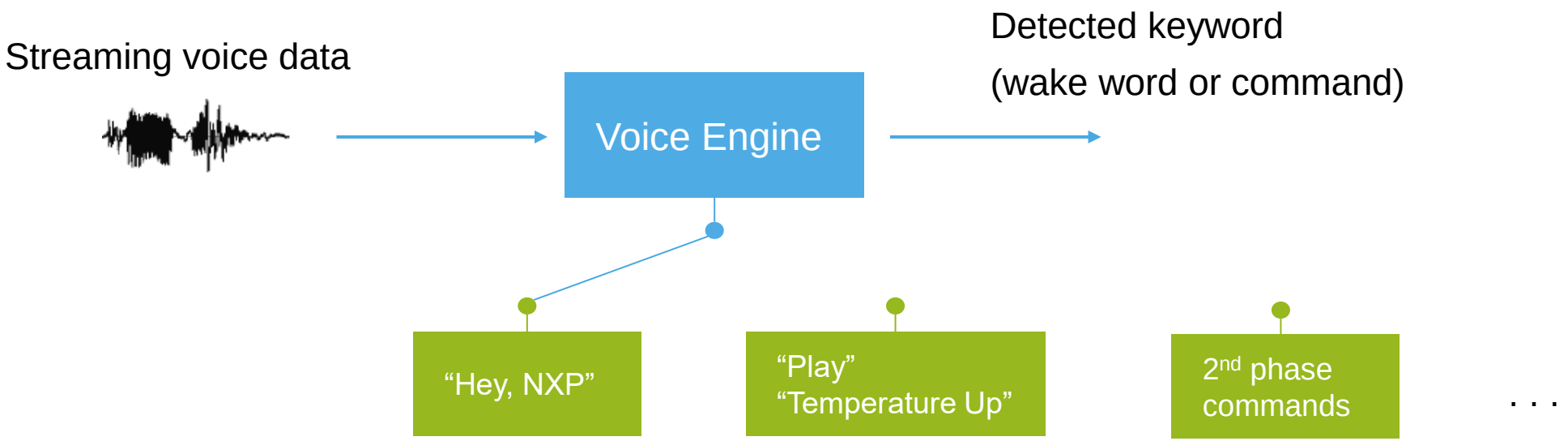
- Far-field voice enhancement



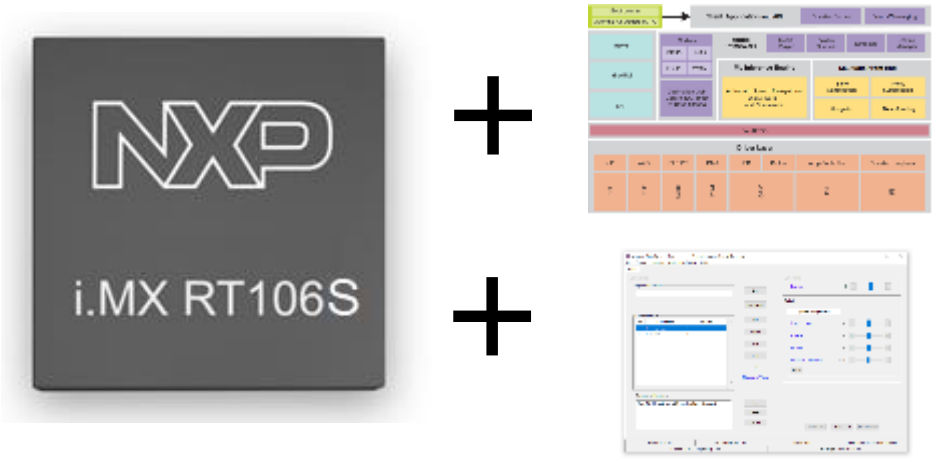
VOICE ENGINE: KEYWORD SPOTTING

- Always on for **wake word(s)**

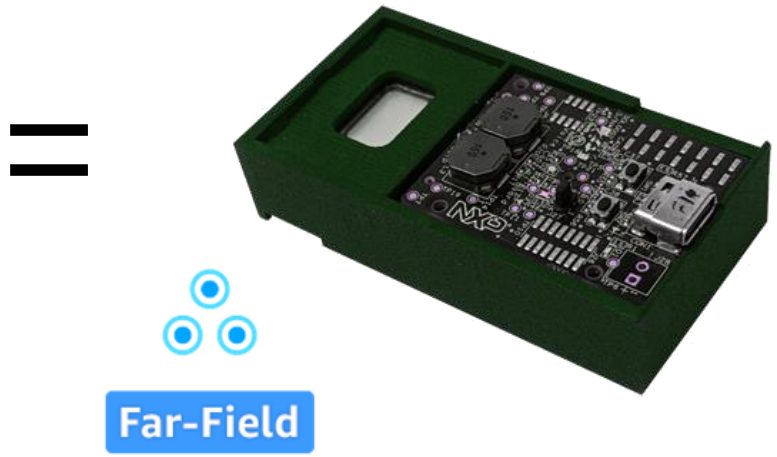
- After being triggered by a wake word, detect a **command**.



MCU BASED FAR-FIELD VOICE CONTROL TURNKEY SOLUTION



SLN-LOCAL2-IOT

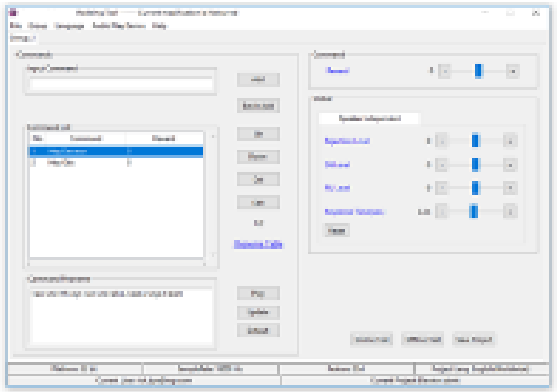


- The MCU - i.MX RT106S (or i.MX RT105S)
  - Arm Cortex-M7 528 MHz (up to 600 MHz)
  - 1 MB SRAM (512 KB on i.MX RT105S)
- Audio Processing
  - PDM to PCM conversion
  - DNN based audio front end to support far-field
  - Two (up to three) microphones array
- Keyword Spotting
  - Multiple custom wake word(s), >100 commands
    - “Hey NXP”, “Temperature Up”, etc
  - Up to four languages simultaneously
  - PC based speech model creation tool
    - Model generation in a few seconds
- Turnkey Solution – one stop shop – includes all SW/HW
  - www.nxp.com/mcu-local2

EXAMPLE WAKE WORDS AND COMMANDS

- Wake words up to four languages
  - English, Mandarin, German & French
    - English – Hey NXP
    - Mandarin – 你好恩智浦 (Nǐ hǎo NXP)
    - German – Hallo NXP
    - French – Salut NXP

- Developers can use speech model tool to quickly create their own commands and wake words



Commands (English, Mandarin, German & French)

- Audio Device Control
  - Turn On
  - Turn Off
  - Play
  - Pause
  - Start
  - Stop
  - Next track
  - Previous track
  - Volume Up
  - Volume Down
- Washing Machine
  - Wash Delicate
  - Wash Normal
  - Wash Heavy Duty
  - Wash Whites
  - Cancel
- LED Control (English only)
  - L, E, D, Red
  - L, E, D, Green
  - L, E, D, Blue
  - Cycle Fast
  - Cycle Slow
- Elevator
  - Floor one
  - Floor two
  - Floor three
  - Floor four
  - Floor five
  - Main lobby
  - Going up
  - Going down
  - Open door
  - Close door
- Smart Home (IoT)
  - Turn On
  - Turn Off
  - Brighter
  - Darker
  - Temperature Up
  - Temperature Down
  - Window Up
  - Window Down

FAR-FIELD VOICE CONTROL DEMO

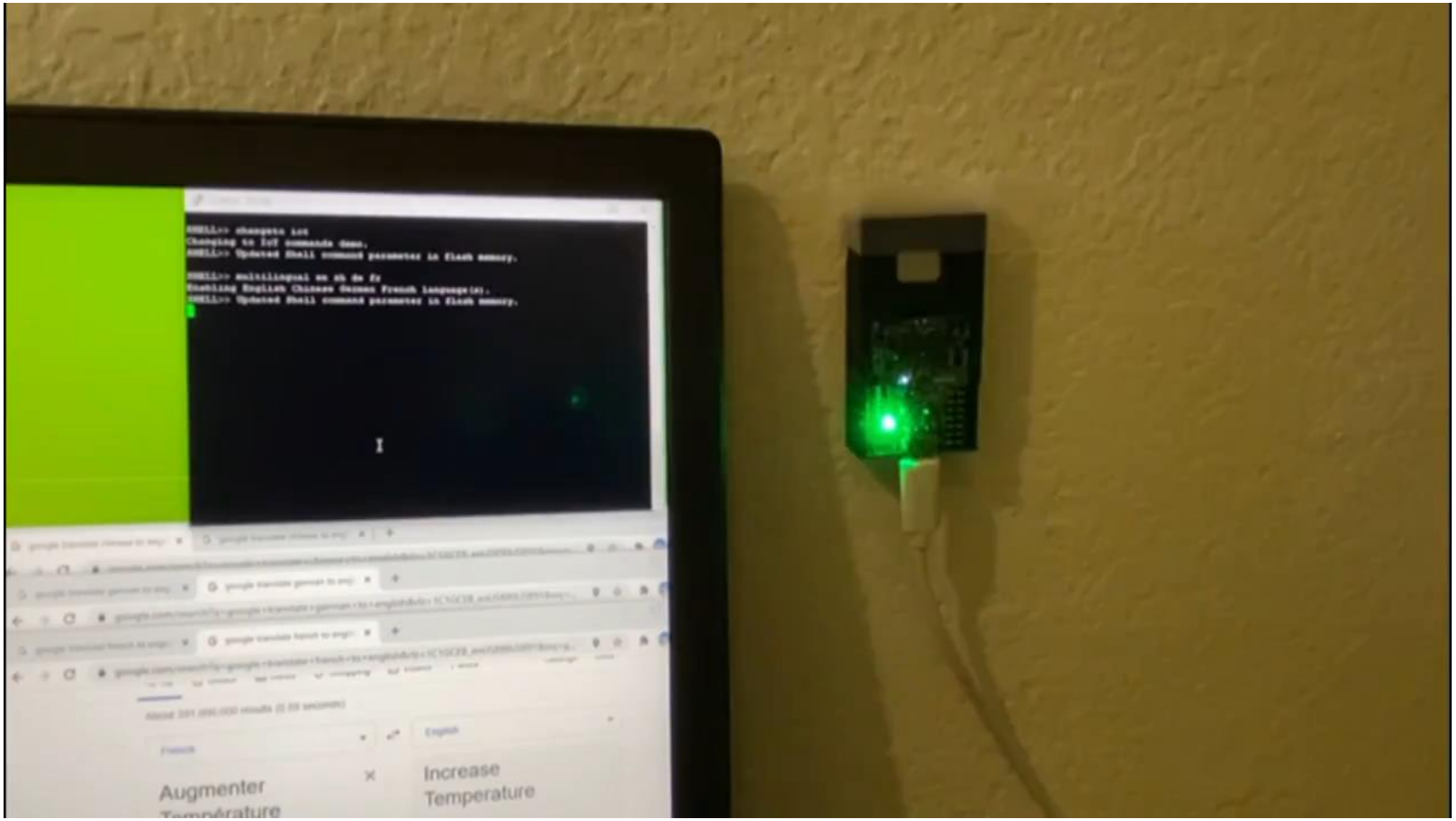
Wake word:
“hey NXP”

Command 1:
“LED Red”

Command 2:
“LED Green”

Command 3:
(LED) “cycle fast”

MULTILINGUAL VOICE CONTROL DEMO



MCU RESOURCE USAGE WITH 2 MICROPHONES

- RAM usage
  - Overall approx. 480 KB for one language (wake word + commands) application
  - Plus approx. 50 KB each additional language
    - Approx. 90 KB for Chinese with tone recognition
    - Actual usage varies with size of voice commands.

- CPU usage
  - One language: approx. up to 65% used @ 528 MHz (QSPI flash)
  - Two languages: approx. up to 80% used @ 528 MHz (QSPI flash)
  - Three languages: approx. up to 95% used @ 528 MHz (QSPI flash)
  - Four languages: approx. up to 70% used @ 528 MHz (HyperFlash)

- Possible to customize for smaller RAM/CPU usage
  - Removal of echo cancellation, PDM to PCM conversion, size of voice commands, etc.
  - Achievable as low as approx. 35% used @ 528 MHz (QSPI flash), approx. 25% used @ 528 MHz (HyperFlash)

The exact numbers of the resource usage can vary with custom solutions.

VISION – PERSON DETECTOR

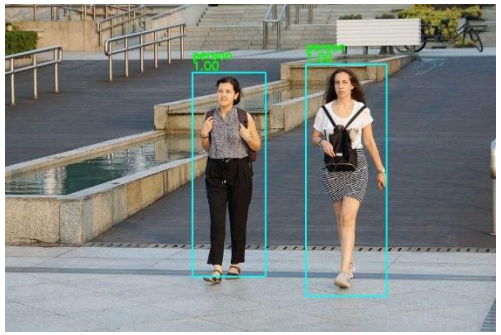
Overview

- Lightweight neural network (NN) model for multi-person detection developed with an open-source NN structure **shufflenetv2** [1] and trained with **COCO** and **PASCAL-VOC** data sets.
- Converts ML model to executable codes through NXP eIQ® Glow NN to increase performance and smaller footprint on i.MX RT MCU.
- Leverages **Micro**controller based **V**ision **I**ntelligence **A**lgorithms (**uVITA**) system to build examples of deploying the vision NN model on NXP's i.MX RT1170evk and RT1060evk.

Example

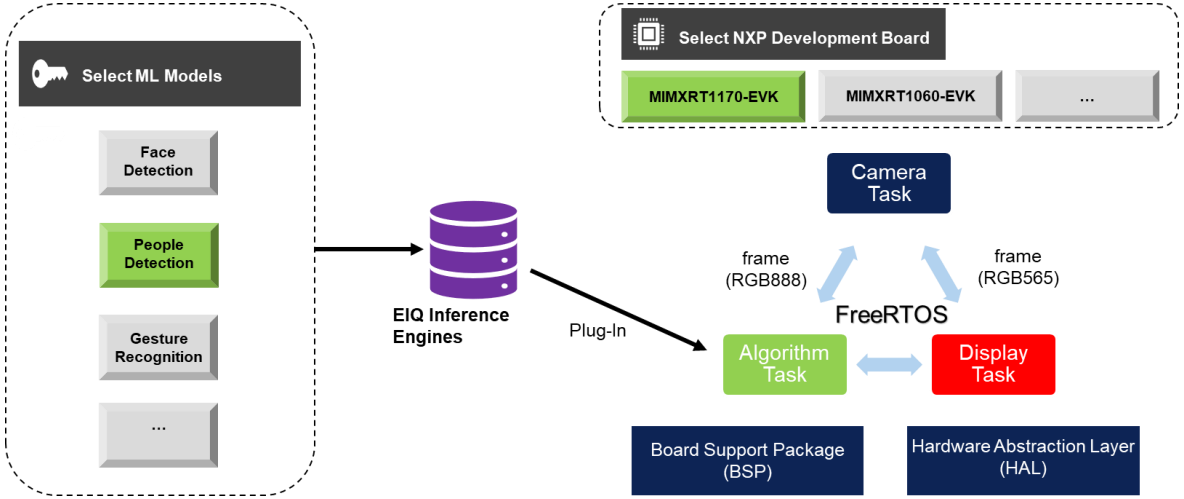
Multiple person detection with i.MX RT1170 EVK and i.MX RT1060 EVK using eIQ tools

- Model Binary size: 241 KB
- Model Peak RAM Usage: 630 KB
- Latency:
  - i.MX RT1060 : 252ms
  - i.MX RT1170: 161ms



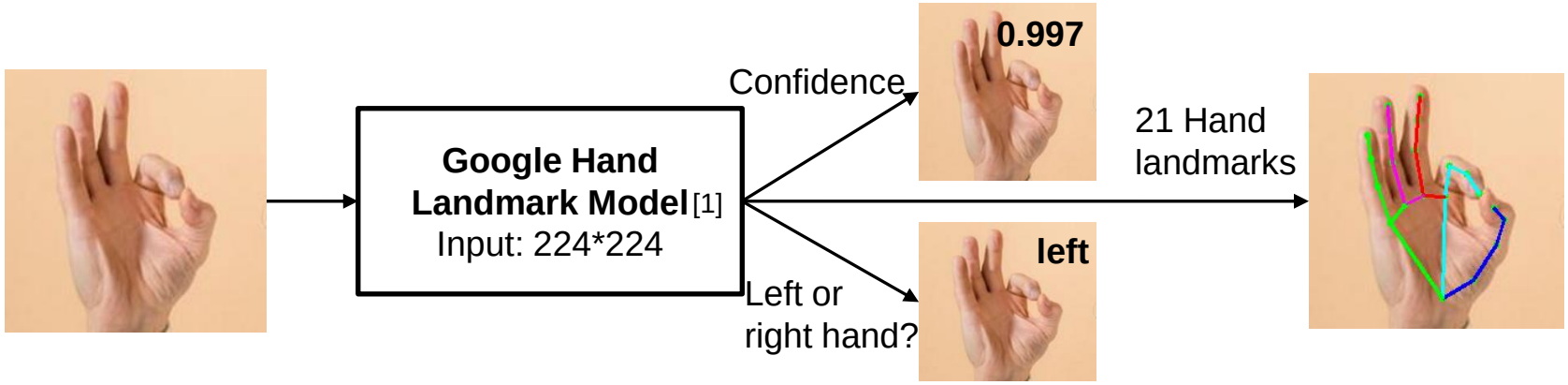
[1] Ma, Ningning, et al. "Shufflenet v2: Practical guidelines for efficient cnn architecture design. " *Proceedings of the European conference on computer vision (ECCV)*. 2018.

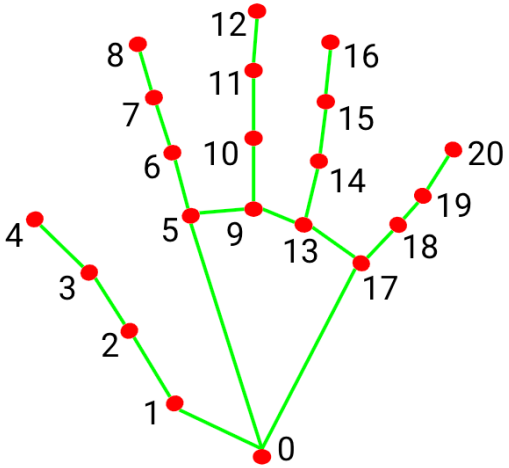
System Architecture



VISION – PERSON DETECTOR

VISION – GESTURE RECOGNITION





Model Type

• Convolutional Neural Network

Model Architecture

• Convolutional Neural Network: MobileNetV2-like with customized blocks for real-time performance.

Input(s)

• Image of proportionally cropped hand area with a flexible margin on each side and size 224x224

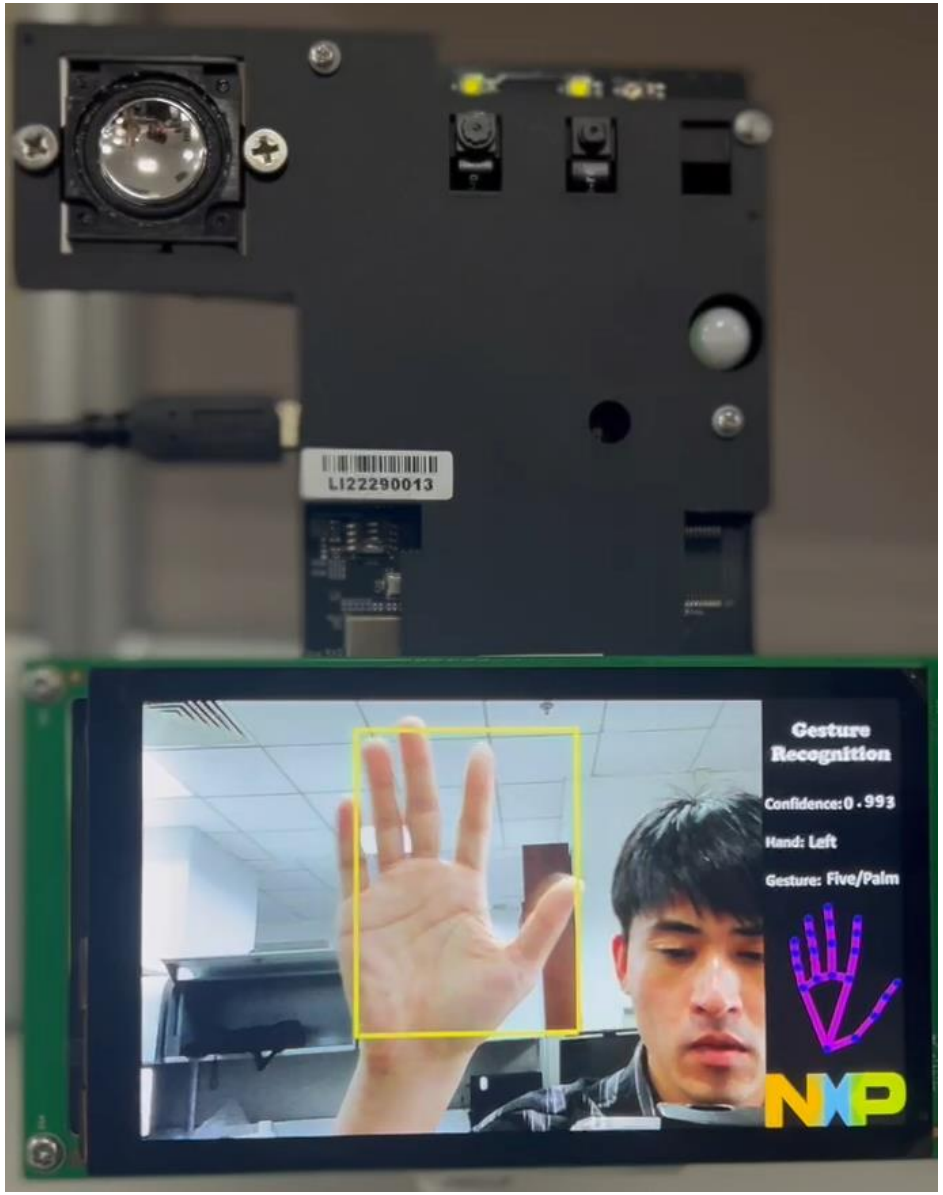
Output(s)

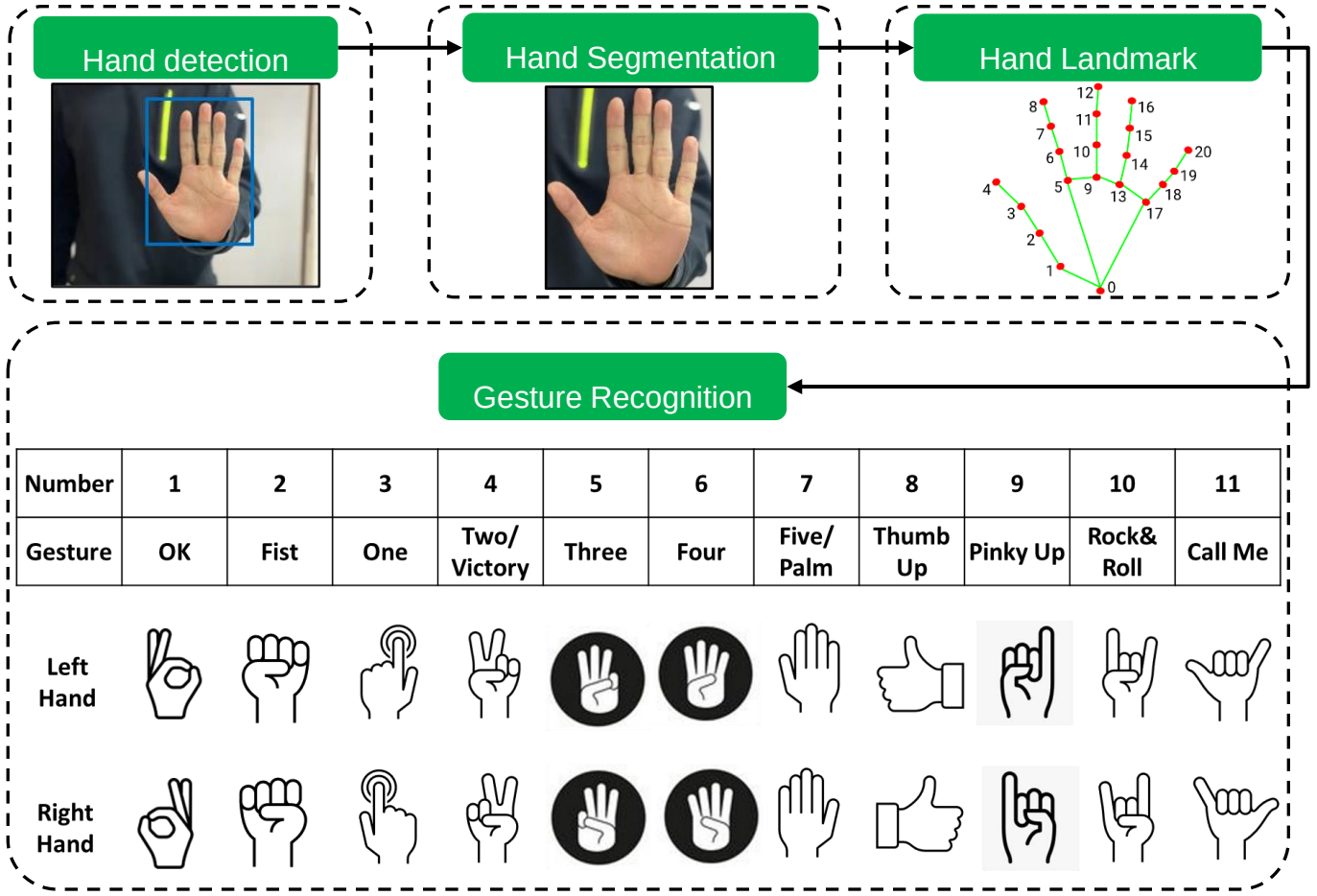
• Confidence of whether it is a hand;
• Confidence of whether it is the left hand or right hand;
• Hand region surface represented as 21 2D landmarks flattened into a 1D tensor: (x1,y1), (x2, y2), ...; x- and y-coordinates follow the image pixel coordinates.

Edge Deployment	
Quantization	8-bits
Weights size (Flash or RAM)	1074 KB
Input/Output + Activations size (RAM)	1459 KB
Time cost on RT117H	530ms

[1] https://developers.google.com/mediapipe/solutions/vision/hand_landmarker#models

VISION – GESTURE RECOGNITION



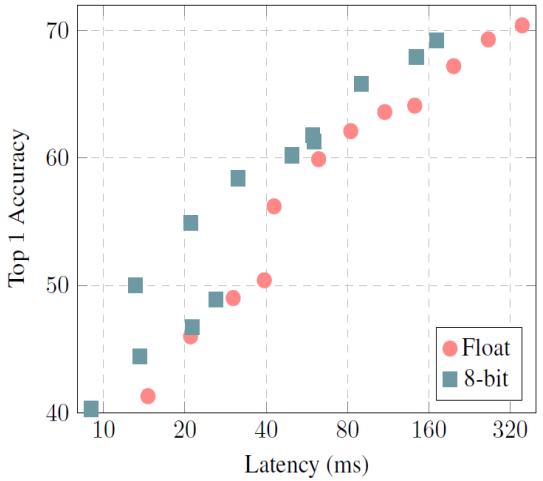


Light-weighted MCU optimized pipeline and model provides low latency and reduced footprint.

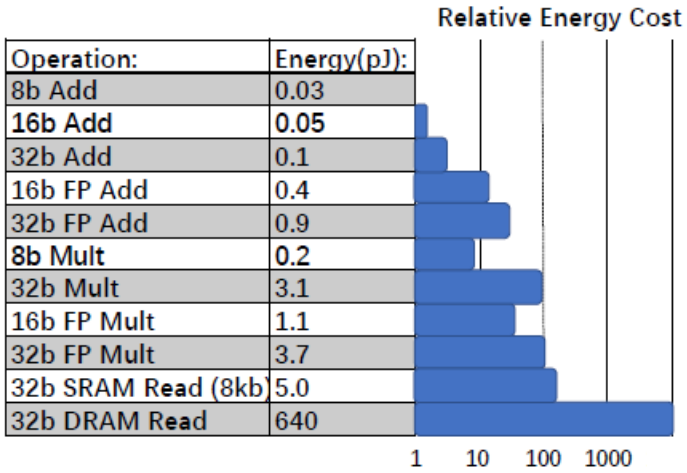
Enabling NN Accelerator to MCU

LATENCY, ENERGY EFFICIENCY, AND ACCURACY

- Quantization and Pruning: common approaches to improve the performance.
  - Lower precision (via quantization) provides lower energy consumption as well as lower latency.



ImageNet latency-vs-accuracy tradeoff [1]



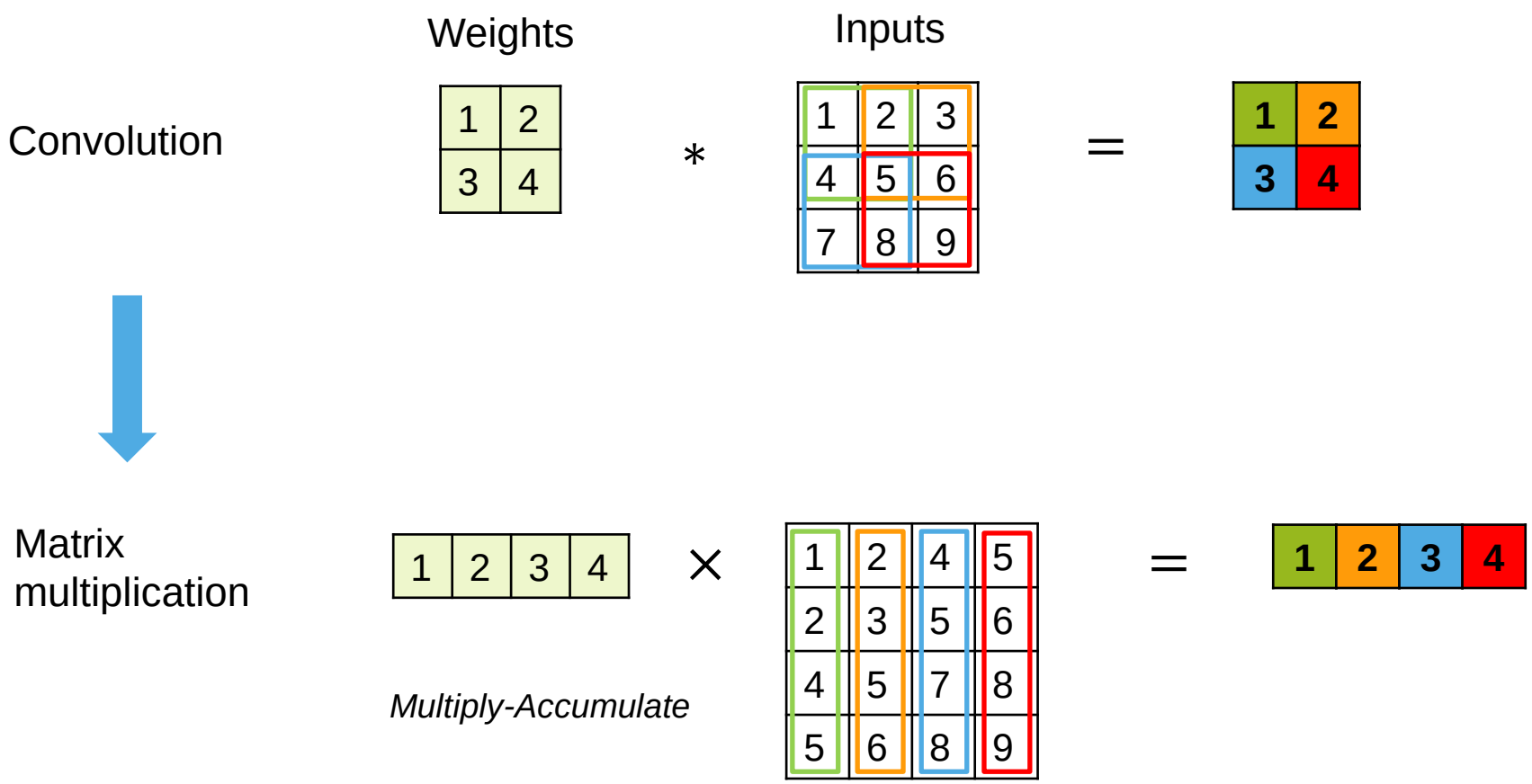
Energy cost for different precisions [2] [3]

- Enabling HW NN accelerator to MCU can provide a new opportunity for further improvement!
  - e.g. leveraging the reusability of 2D convolutional kernel in CNN.

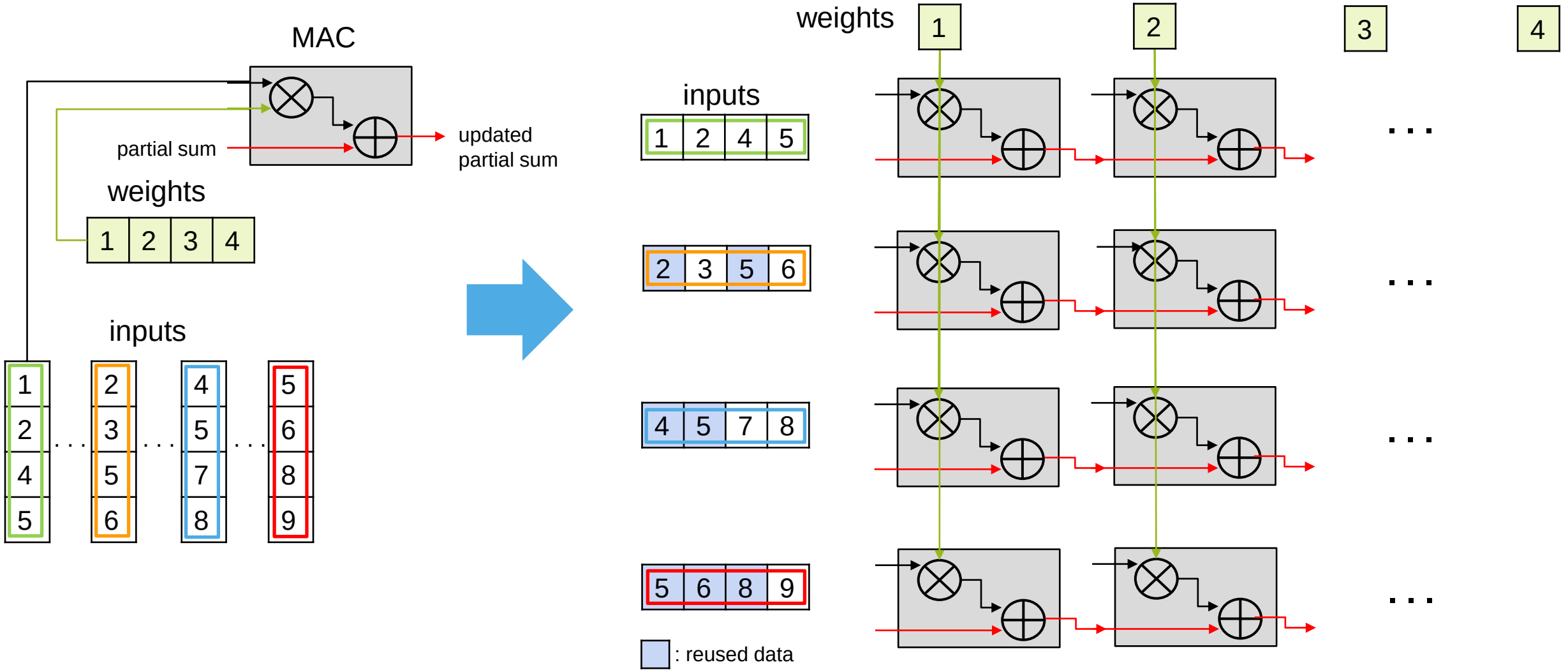
[1] Benoit Jacob et al., “Quantization and Training of Neural Networks for efficient Integer-Arithmetic-Only Inference,” IEEE CVPR, 2018
[2] Amir Gholami et al., “A Survey of Quantization Methods for Efficient Neural Network Inference,” arXiv:2103.13630, 2021
[3] Mark Horowitz, “1.1 Computing’s Energy Problem (and what we can do about it),” IEEE ISSCC, 2014

CONVOLUTION TO MATRIX MULTIPLICATION

- Convolution can be treated as matrix multiplication.



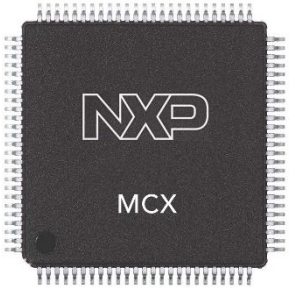
PARALLELISM & REUSABILITY TO ACCELERATE THE CONVOLUTION OPERATION



Inputs are divided and processed in parallel.

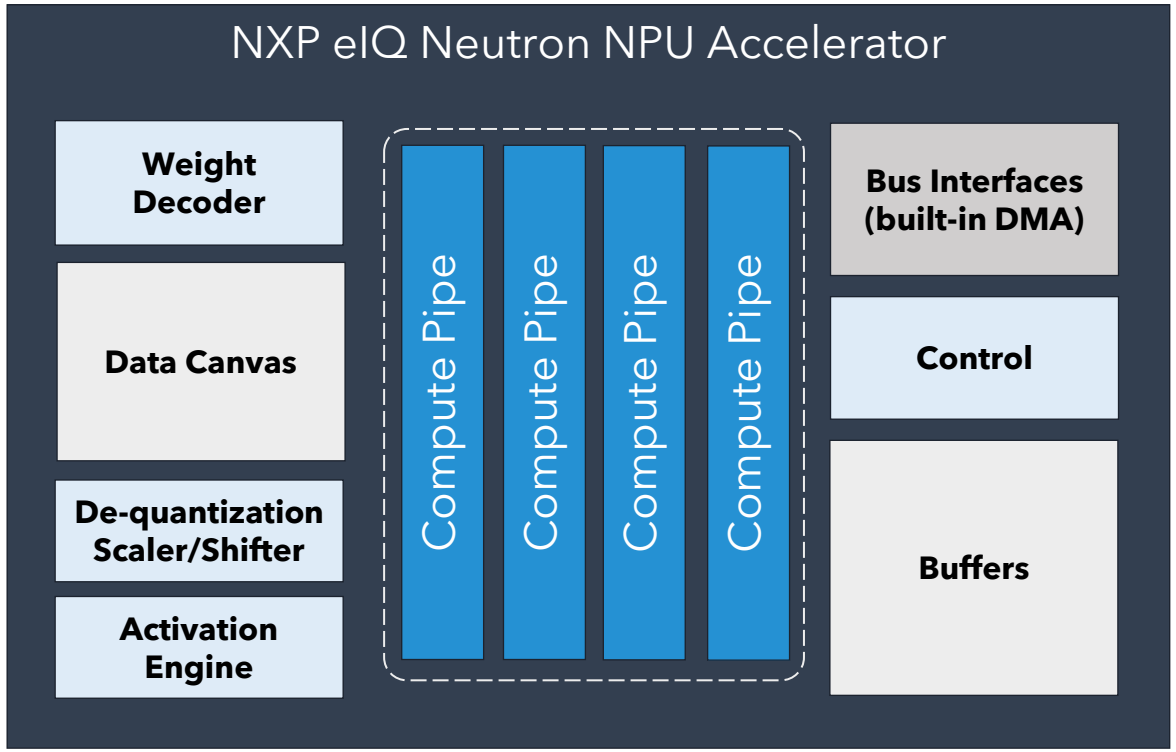
Weights and inputs are reused.

MCX N SERIES - HW ACCELERATOR ENABLED MCU



- TinyML Summit 2023 Best ML Processing MCU Award.
- Arm Cortex-M33 + NPU + DSP accelerator + eIQ SW + more features.

HW ACCELERATOR FOR NEURAL NETWORKS RUNNING ON MCU



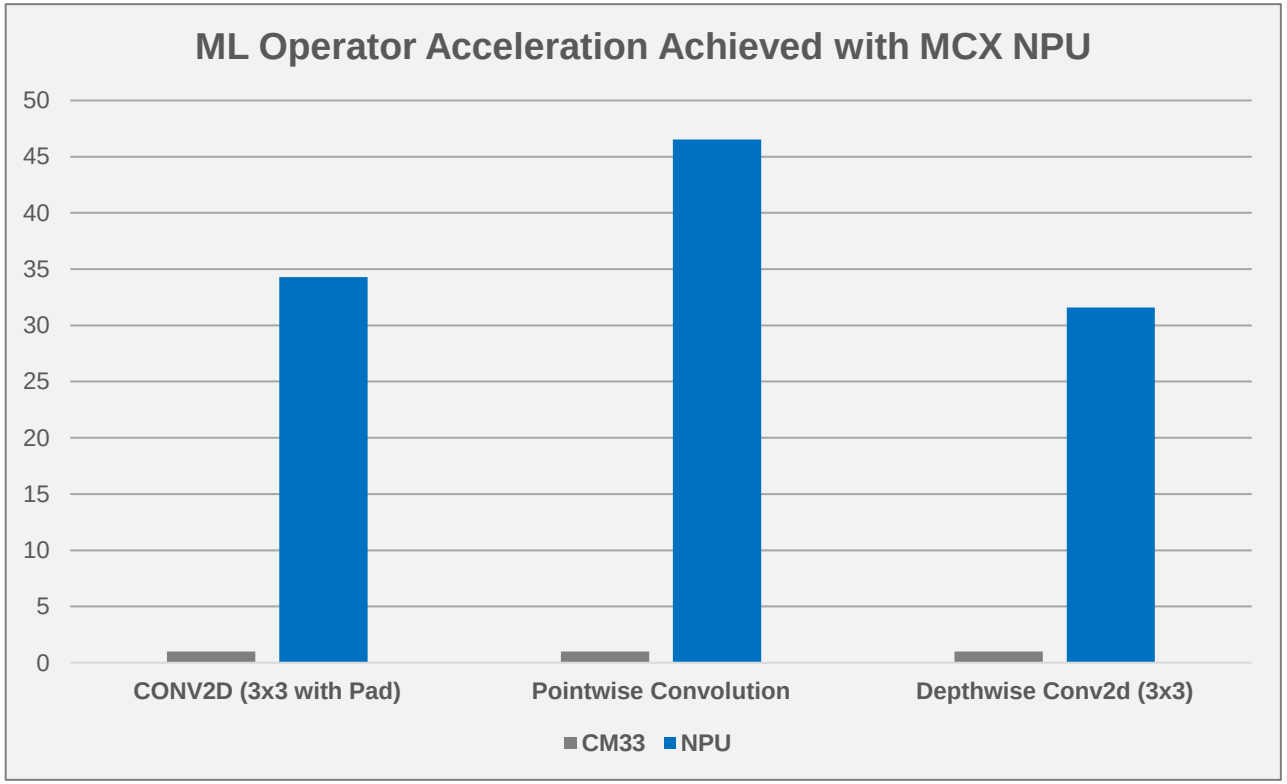
- Optimized for performance and power efficiency
- The neural processing unit (NPU) is designed to accelerate matrix computation to reduce inference time.
- ML solution development support with eIQ® ML SW Development Environment
  - Supports major Neural Networks types (CNN, RNN, LSTM, TCN, and more)

For MCX N Series MCU:

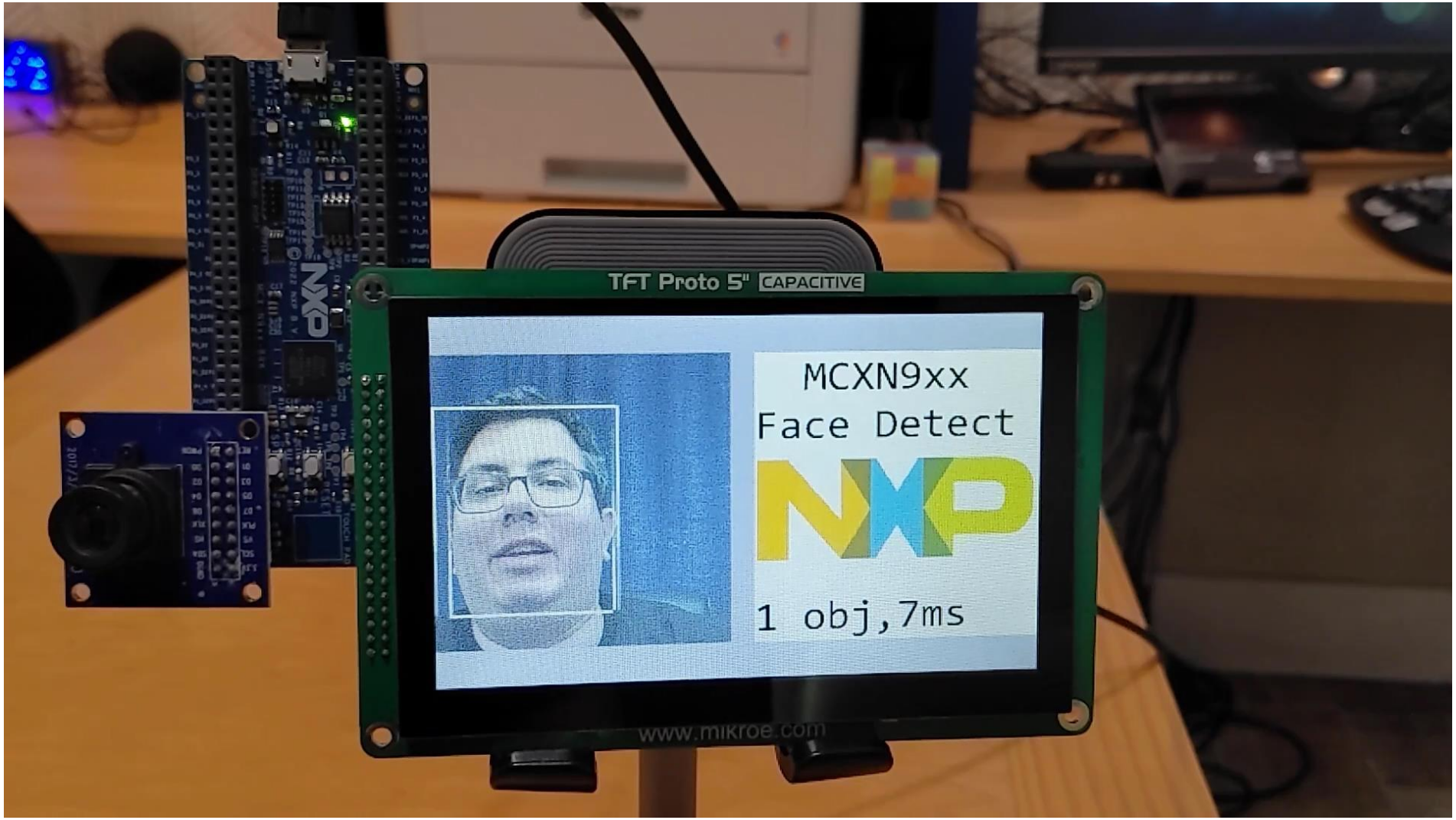
(16 MAC/cycle * 2 op/MAC + 4 activation/cycle * 1 op/activation + 4 rescale/cycle * 1 op/rescale) * 150 MHz = 6 GOPS

eIQ NEUTRON NPU ACCELERATION

- Over 30x relative acceleration.
- 37x less energy consumption.



FACE DETECTION ON MCX-N eIQ NEUTRON NPU



Face Detection demo available by request for MCX-N Breakout Board.

SUMMARY

- Deploying lightweight ML models on MCU-based edge devices have shown good performance in various application domains:
  - Anomaly detection
  - Voice control
  - Vision

- Recent advancement such as enabling neural processing unit to MCU can significantly accelerate ML performance at the Edge.

NXP eIQ RESOURCES

- eIQ for iMX RT on MCUXpresso SDK Builder
  - DeepViewRT, Glow, and TFLite eIQ User Guides located in SDK documents package
- Porting guides for Glow and TensorFlow Lite available.
- Save camera data to an SD card for training data
- Complete eIQ FAQ

- Hands-on Labs
  - DeepViewRT Getting Started Labs
  - Glow Getting Started Lab
  - TFLite for Microcontrollers Getting Started Lab

- App Notes
  - Anomaly Detection App Note (AN12766)
  - Handwritten Digit Recognition (AN12603)
  - Datasets and Transfer Learning App Note (AN12892)
  - Glow Memory Usage App Note (AN13001)

ML/AI TRAINING SERIES

- 20+ training modules
- Available at www.nxp.com/mltraining

1 **MACHINE LEARNING CONCEPTS**
- Concepts and Introduction

2 **eIQ™ SOFTWARE DEVELOPMENT ENVIRONMENT**
- eIQ Overview
- Transfer Learning Intro & Lab
- Handwritten Digit Recognition Example

3 **eIQ TOOLKIT**
- eIQ Toolkit: How to BYOD
- eIQ Toolkit: How to BYOM
- eIQ Toolkit: Command Line interface

4 **MACHINE LEARNING ON MCUS**
- Machine Learning with i.MX RT
- Get started with eIQ on i.MX RT
- Create your own Model using Glow
- eIQ inference with TensorFlow Lite for MCUs – Overview & Lab
- eIQ inference with Glow NN Compiler – Overview & Lab
- eIQ inference with DeepViewRT

5 **MACHINE LEARNING ON MPUS**
- eIQ inference with ONNX
- eIQ inference with Arm NN
- eIQ inference with TensorFlow Lite (for MPUs)
- TensorFlow Lite support for Android ML

6 **PARTNER ML SOLUTIONS**
- Solutions and topics contributed by NXP eIQ partners

More information on our products:

www.nxp.com

Our distribution partners:

https://www.nxp.com/support/sample-and-buy/distributor-network:DISTRIBUTORS

Q & A

SECURE CONNECTIONS
FOR A SMARTER WORLD

SECURE CONNECTIONS
FOR A SMARTER WORLD

Copyright Notice

www.tinyml.org

Copyright Notice

www.tinyml.org