

tinyML[®] Summit

Enabling Ultra-low Power Machine Learning at the Edge

tinyML Summit 2021 Proceedings

March 22 – 26, 2021

Virtual Event



www.tinyML.org

Accelerating the next generation of tinyML Applications

Kartik Thakore
Co-Founder



TinyML applications
will be unlike any
other!

Leveraging tinyML in a mesh of
decentralized networks to
develop unique applications



Operating theses for the future of tinyML

Developers will push the boundaries of apps beyond the limitations of devices

- tinyML applications at scale will be able to leverage all domains of computations
- New class of applications that wouldn't be possible with cloud applications
- As we have seen with the cloud, tooling will be key to help developers explore

Service uptime, security and privacy requirements will become crucial as regulations catch up

- Devices in the wild will need to guarantee models cannot be downloaded and reversed
- Signed applications will be needed to prevent unsigned applications from exploiting device resources
- Guarantee service uptimes will be necessary as more people depend on tinyML applications

Applications: Privacy first health sensors

- Private Passive Fall detection
 - Vision sensors
 - Audio sensors
- Topology of several sensors
- AI models working together to develop consensus of fall
- Current Solutions:
 - Apple Watch
 - Pendants
- Data collected by sensors are still sent to the cloud!



Applications: Multi Domain Smart Cities

- Sensors clustered over LoraWAN to collect insights on
 - Traffic
 - Public health signals
 - Cough detection
 - Preventive maintenance on services
- Data is never aggregated!
- Insights are openly shared with local residents only
 - Smart city app
- Local residents can help improve models given incentives



Exploring the tools and technical concepts

Taking the learnings from cloud computing and re-imagining them for tinyML

- Security
- Reliability
- Service Uptime
- Privacy

Let's Build a Container

```
FROM runicos/base

CAPABILITY<_,f32[384]> accelerometer ACCEL -n 128

PROC_BLOCK<f32[384],f32[384]> normalize
hotg-ai/rune#proc_blocks/normalize

MODEL<f32[384],f32[4]> gesture ./model.tflite

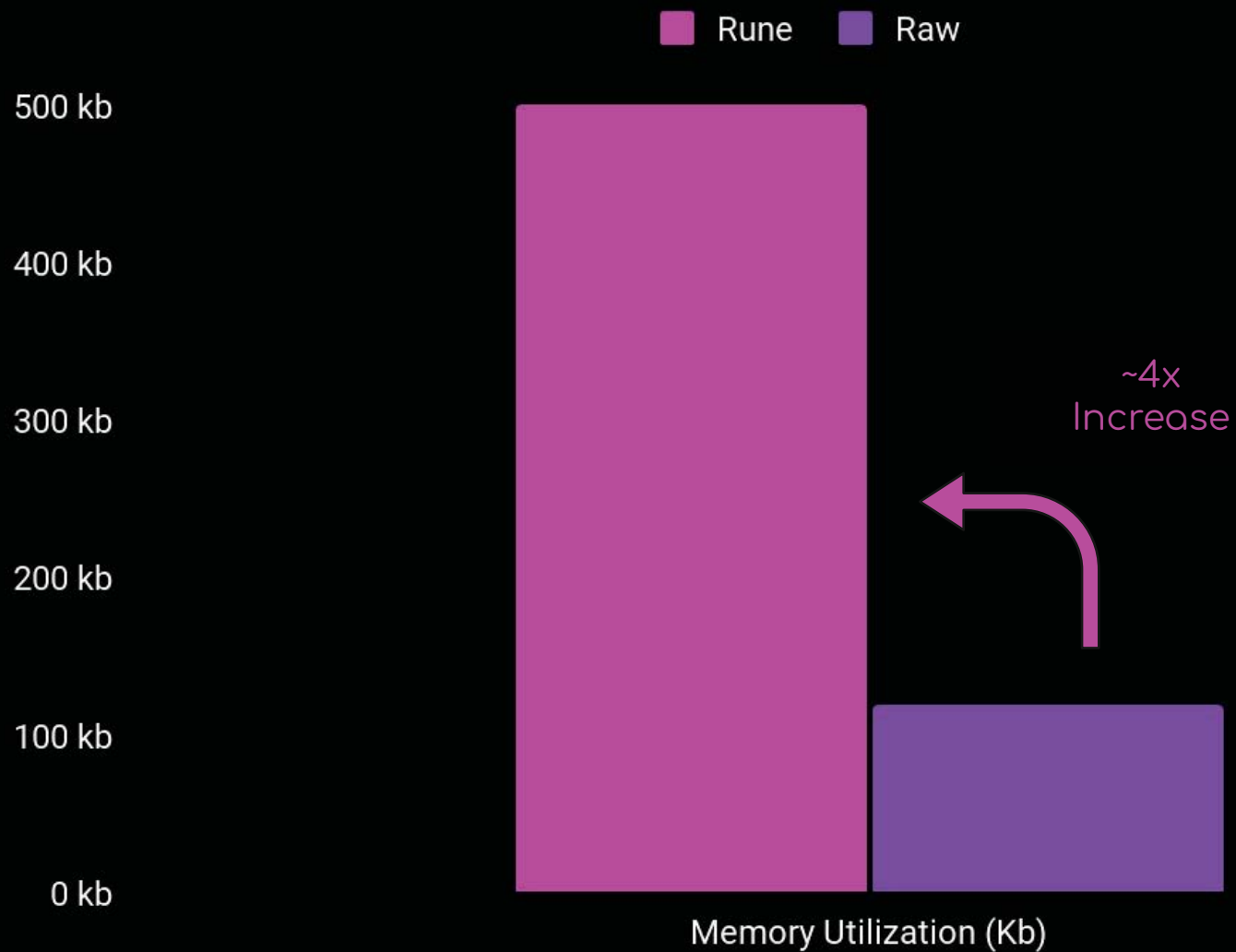
PROC_BLOCK<f32[4], UTF8> label
hotg-ai/rune#proc_blocks/ohv_label \
  --labels=Wing, Ring, Slope, Unknown

RUN accelerometer normalize gesture label

OUT SERIAL
```

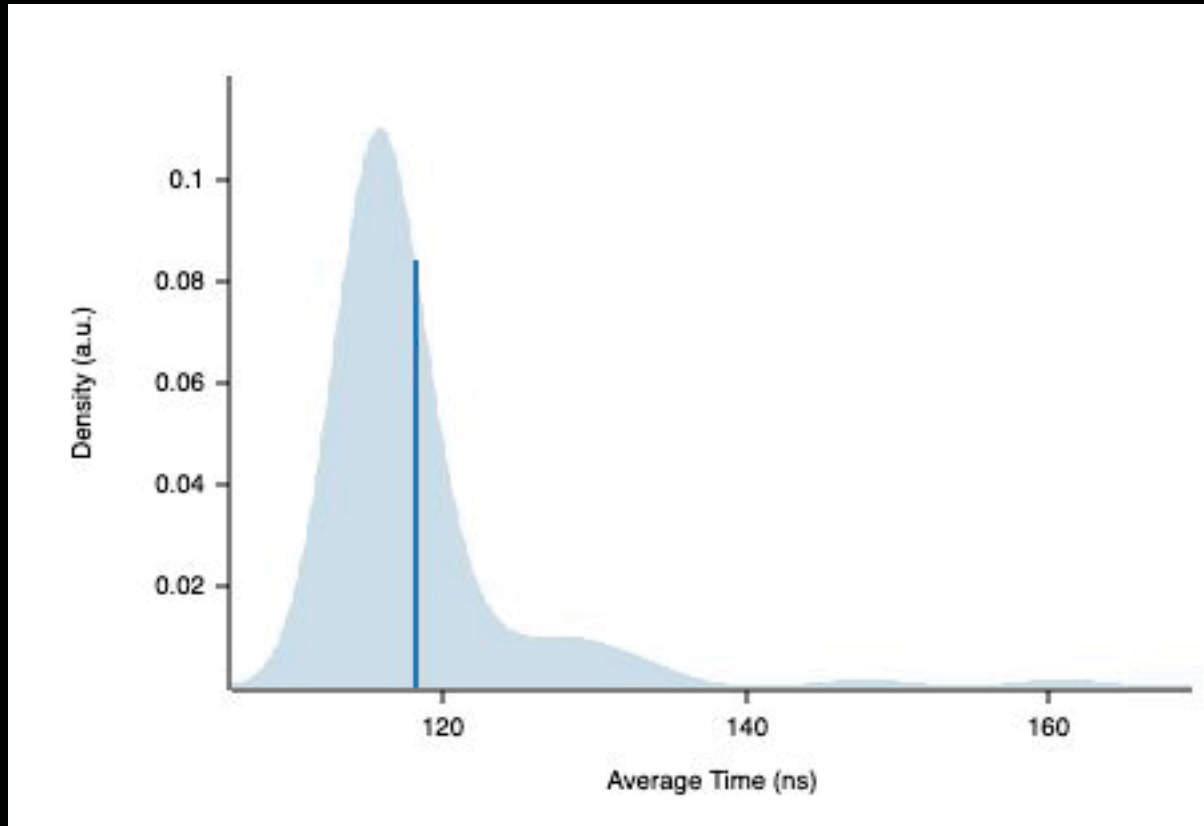
- Declarative file interface called Runefile
- Tutorial!
- DEMO!

Tradeoff: Memory Benchmark

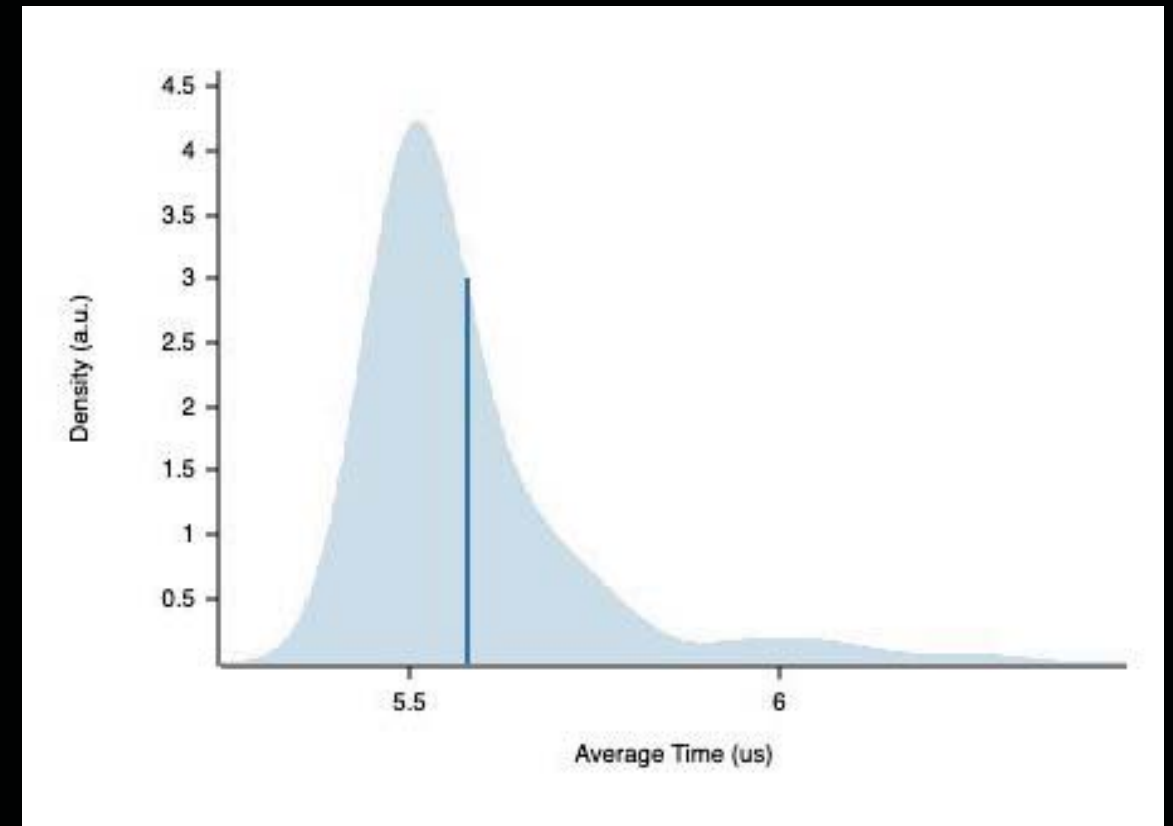


CPU Benchmarks

Native

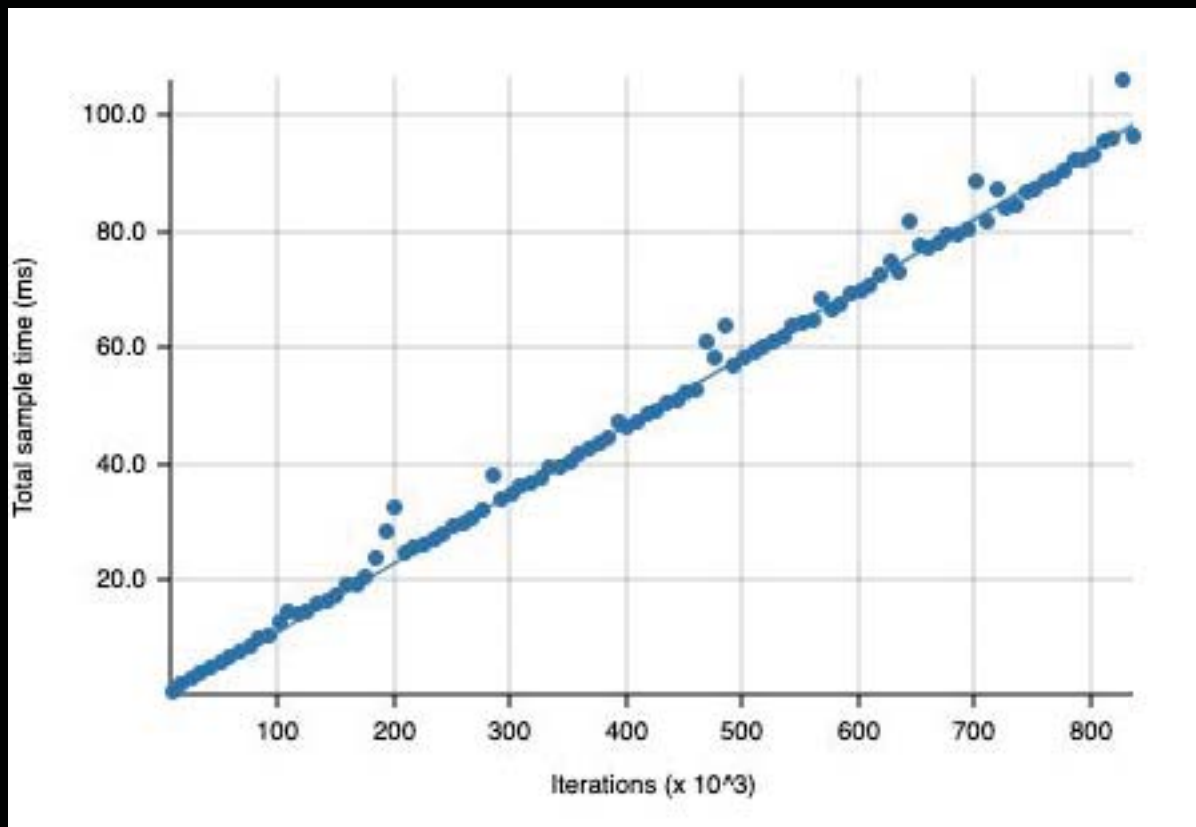


Containerized

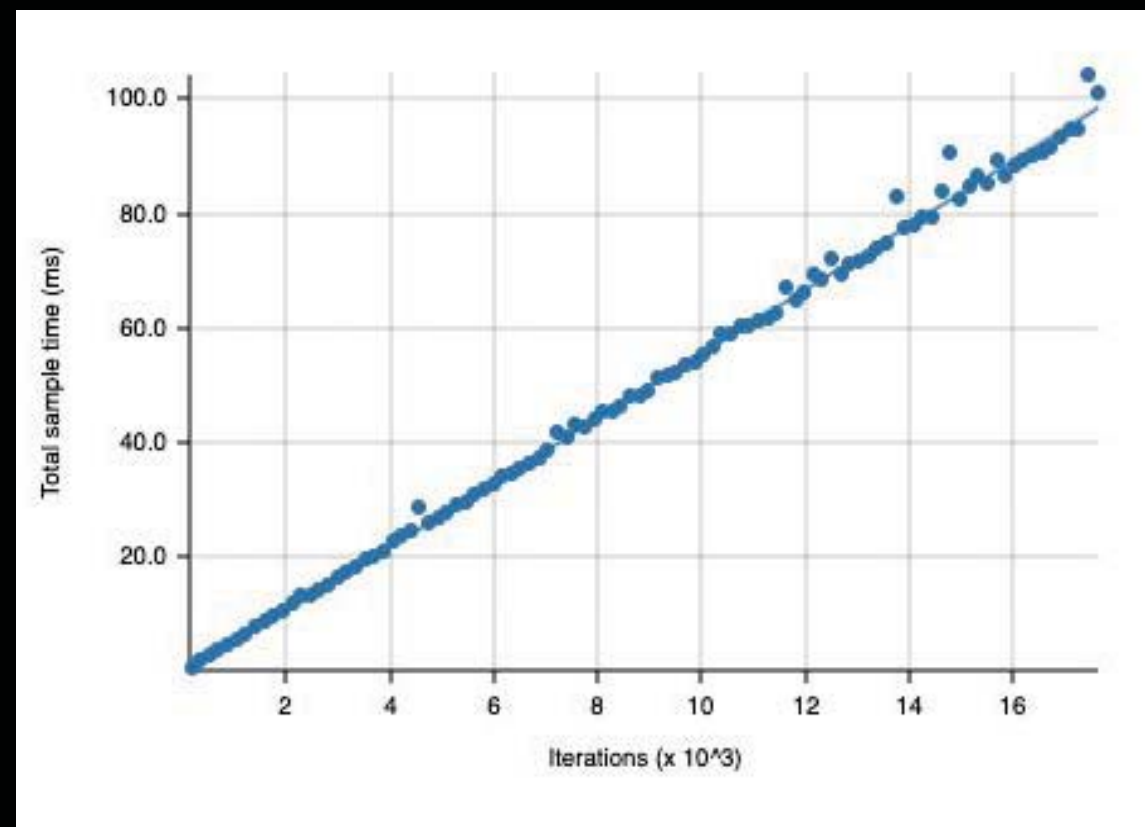


Compute Stability

Native



Containerized



Next Steps

Caching bytecode compilation to host memory to reduce CPU loads

More benchmarking on several devices

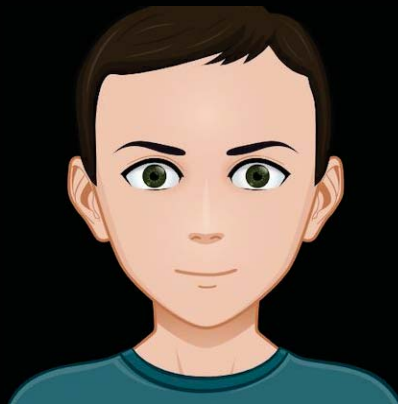
Deployment tools

Open source release!!!

Summary

- Discussion of next gen applications for tinyML
- Developer tooling that would be needed
- Containerization demo and tutorial
- Benchmark results
- Areas for improvement
- Next steps!

Acknowledgements

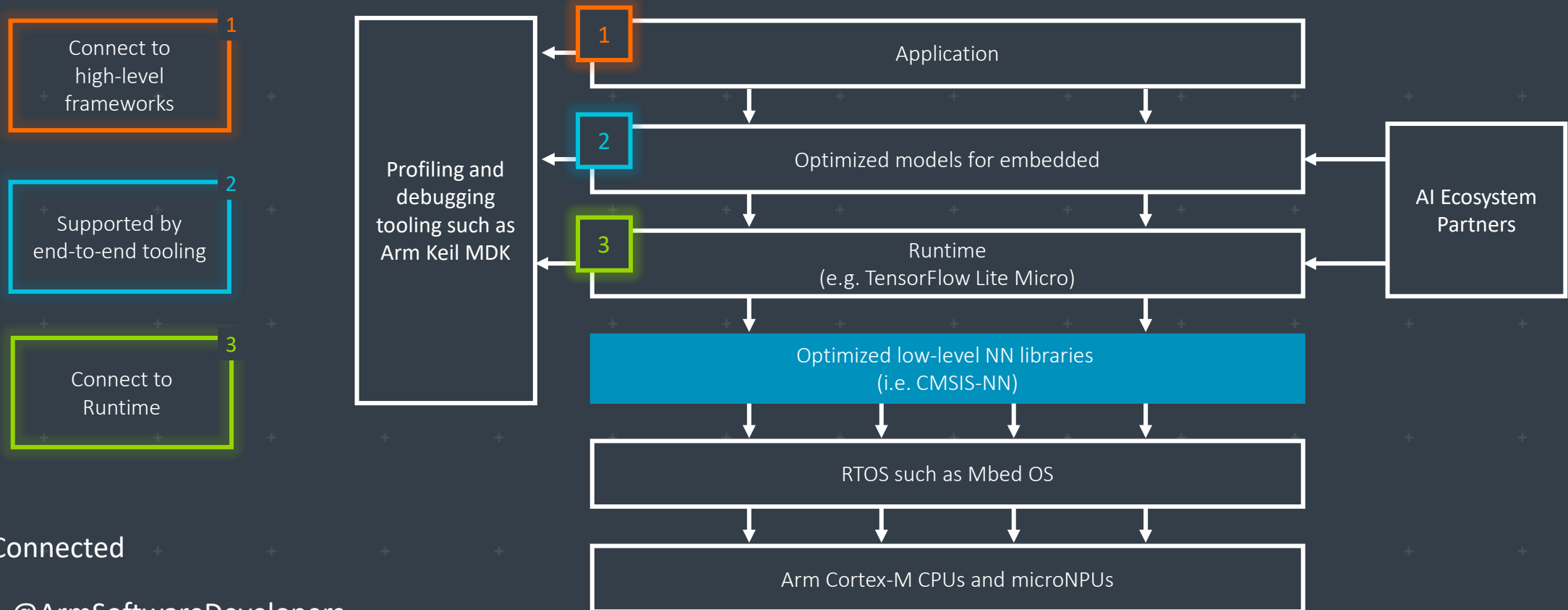


We thank the authors for their presentations and everyone who participated in the tinyML Summit 2021.

Along with a special thank you to the sponsors who made this event possible!

Executive Sponsors

Arm: The Software and Hardware Foundation for tinyML



Stay Connected

 @ArmSoftwareDevelopers

 @ArmSoftwareDev

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

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



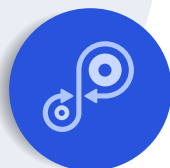
Perception

Object detection, speech recognition, contextual fusion



Reasoning

Scene understanding, language understanding, behavior prediction



Action

Reinforcement learning for decision making



Edge cloud



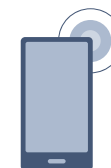
Cloud



IoT/IIoT



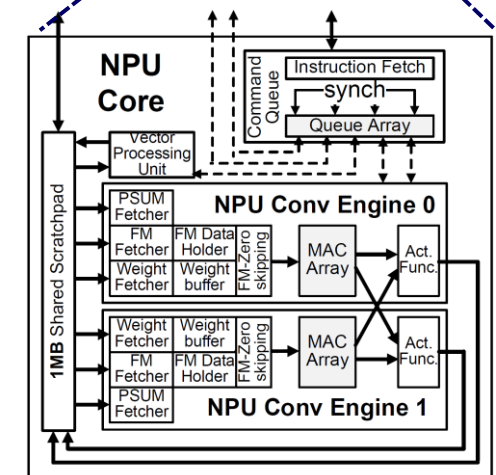
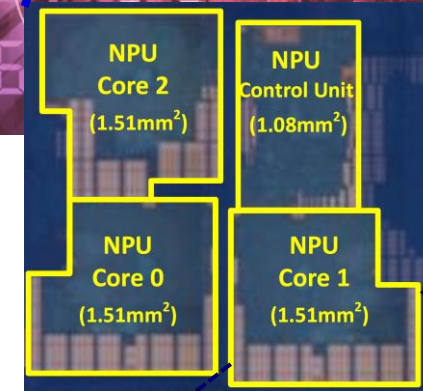
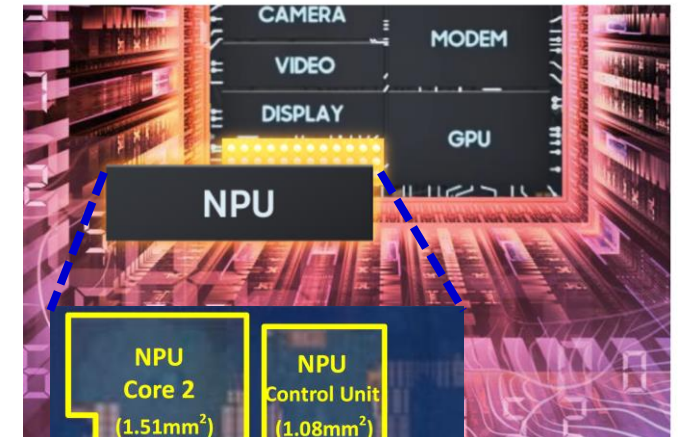
Automotive



Mobile

NEURAL PROCESSING

- Samsung brings AI in the hands of everyone, with >300M Galaxy phones per year. Fingerprint ID, speech recognition, voice assistant, machine translation, face recognition, AI camera; the application list goes on and on.
- In the heart of AI applications is the NPU, the neural processor that efficiently calculates AI workloads. Samsung NPU is a home grown IP that was employed since 2018 inside Samsung Exynos SoC.
- Samsung NPU is brought by global R&D ecosystem that encompasses US, Korea, Russia, India, and China. In US, we are the fore-runner to guide the future directions of Samsung NPU, by identifying major AI workloads that Samsung's NPU needs to accelerate in 3-5 years. For this, we collaborate with world-renowned academia research groups in AI and NPU.



Platinum Sponsors



Eta Compute

DISRUPTION AT THE EDGE

Eta Compute creates energy-efficient AI endpoint solutions that enable sensing devices to make autonomous decisions in energy-constrained environments in smart infrastructure and buildings, consumer, medical, retail, and a diverse range of IoT applications.

www.etacompute.com



THE LOW POWER LEADER

Lattice Semiconductor (NASDAQ: LSCC) is the low power programmable leader. We solve customer problems across the network, from the Edge to the Cloud, in the growing communications, computing, industrial, automotive and consumer markets. Our technology, relationships, and commitment to support lets our customers unleash their innovation to create a smart, secure and connected world.
www.Latticesemi.com.

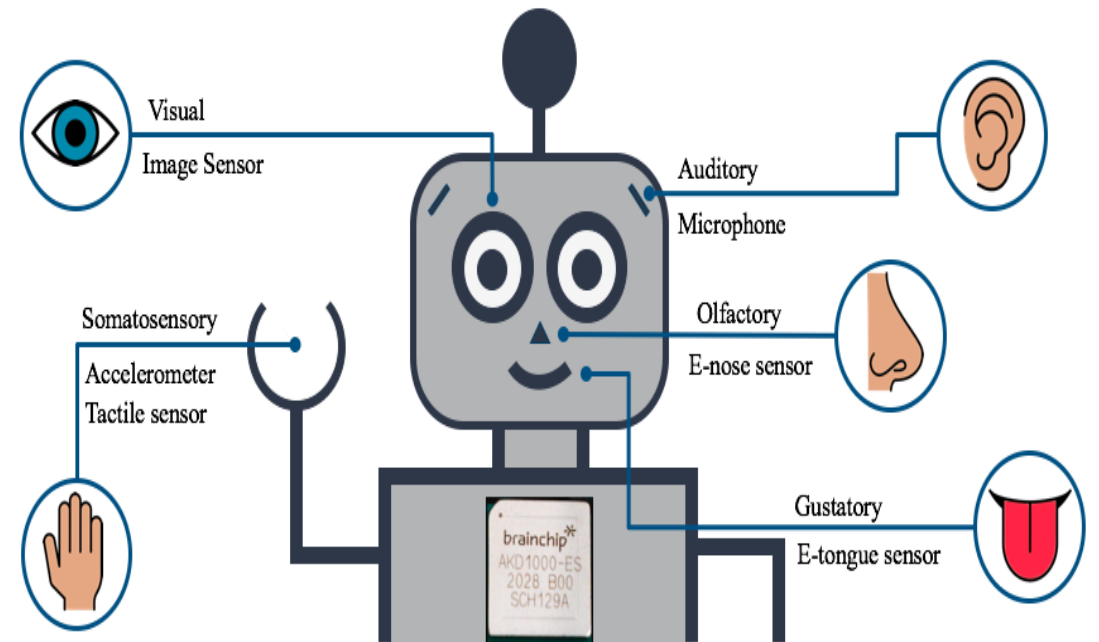
Gold Sponsors



AKIDA™ Neuromorphic Technology: Inspired by the Spiking Nature of the Human Brain

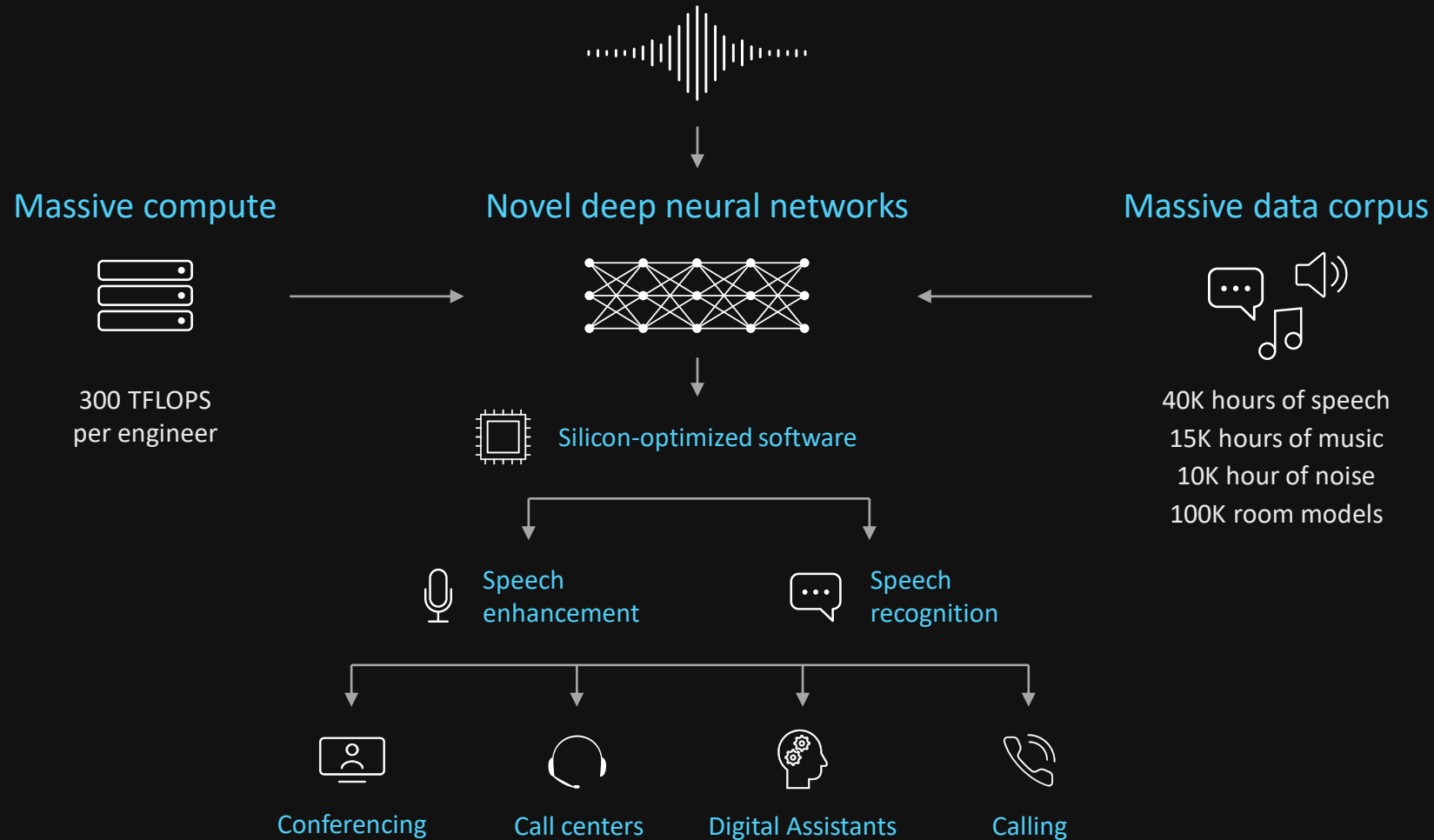
- Supports ultra-low power applications (microwatts to milliwatts)
- Edge capabilities: on-chip training, learning, and inference
- Designed for AI Edge applications: vision, audio, olfactory, and smart transducer applications
- Licensed as IP to be designed into SoC or as silicon
- Sensor inputs are analyzed at the point of acquisition rather than through transmission via the cloud to the data center. Enables real time response for power-efficient systems
- Software Development Platform

AKIDA™ Enables Processing of All Sensor Modalities



BabbleLabs AI speech wizardry in Cisco Webex

AI meets speech - deep experience in speech science, AI/ML, embedded systems



The VOICE of AI

DSP Group, Inc. develops wireless communications and voice processing chipsets, algorithms, and software solutions for converged communications and smart-enabled devices. Core competencies include, but are not limited to, voice processing. Its technology supports the development and integration of voice user interfaces (VUIs) for applications ranging from smartphones to the smart home. Its Ultra-Low Energy (ULE, per the ULE Alliance) wireless solutions enable low-power, long-range, secure communication applications for the IoT and are distinguished by their native support of two-way voice communication. On-going development efforts include the application of machine learning (ML) and artificial intelligence (AI) hardware and algorithms to address the need for accurate AI solutions at the edge for applications such as sound detection, proximity detection, and acoustic beacons.



Unified Communications



Smart Home & Security



Mobile Computing

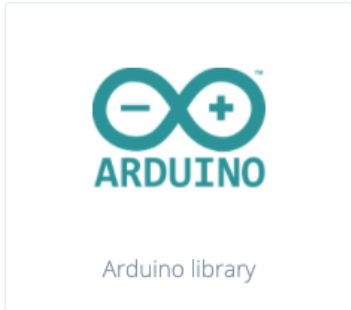


Hearables

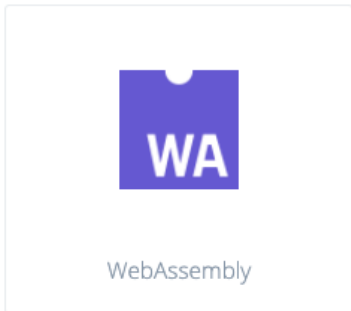
TinyML for all developers



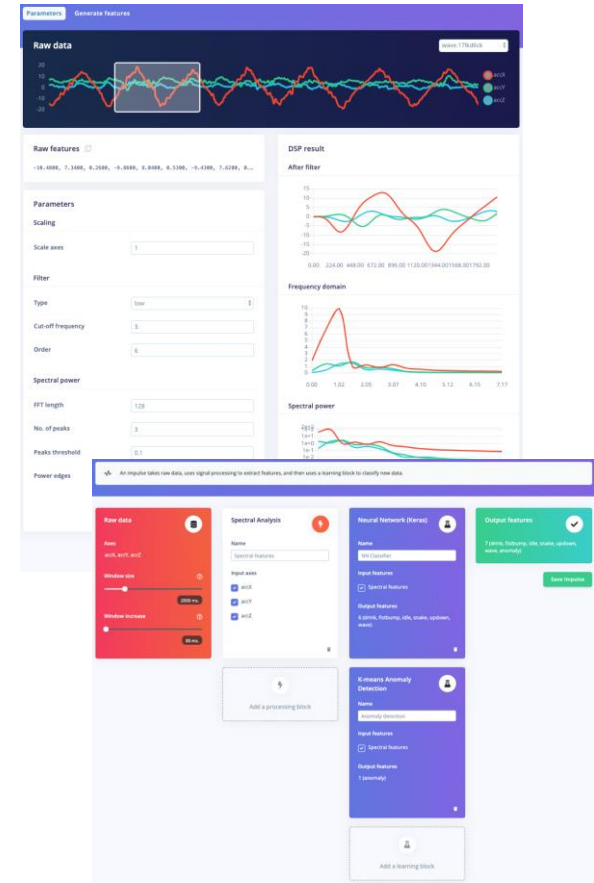
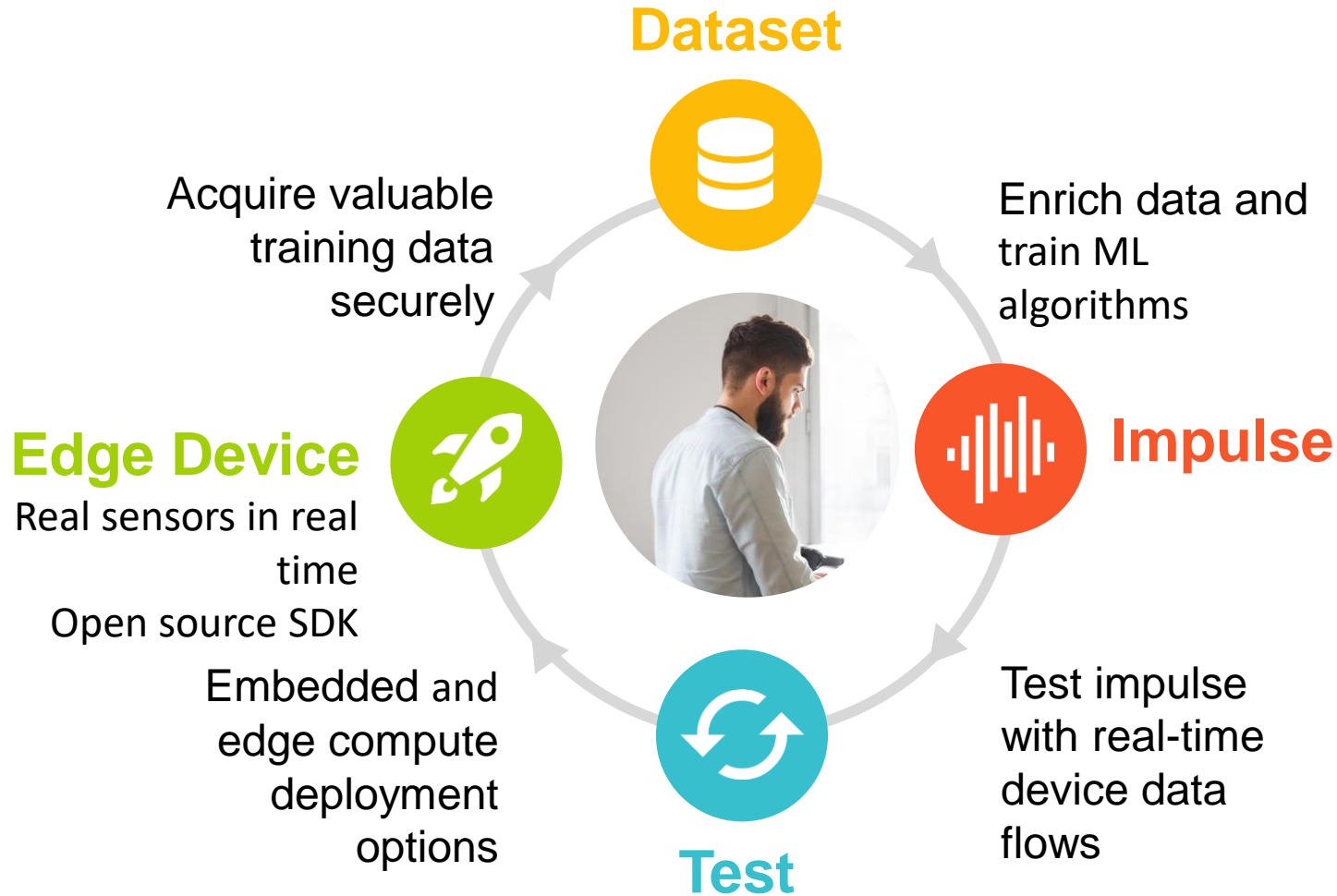
C++ library



Arduino library



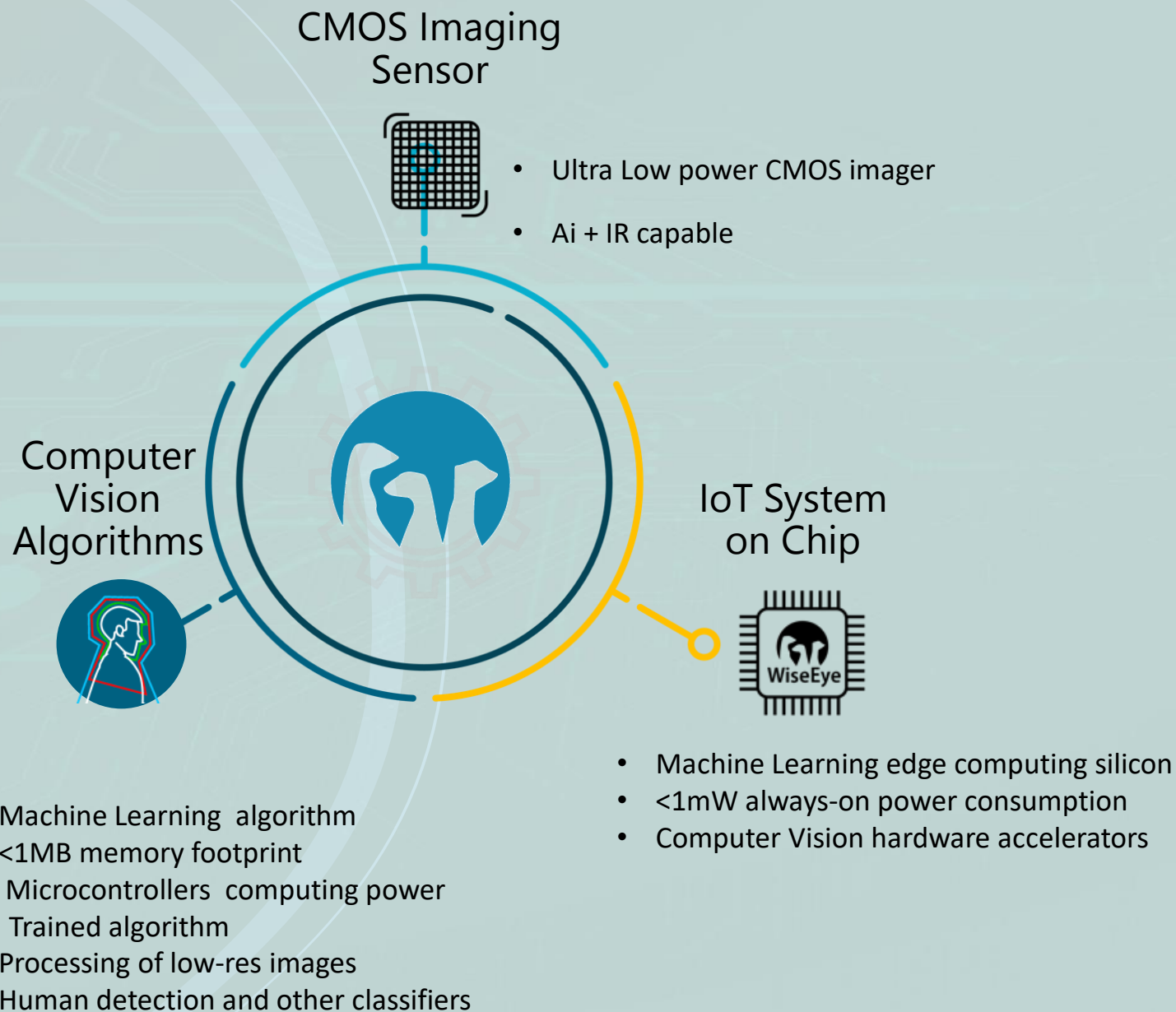
WebAssembly





The Eye in IoT

Edge AI Visual Sensors



info@emza-vs.com





GrAI Matter Labs

GrAI Matter Labs

has created an AI Processor for use in edge devices like drones, robots, surveillance cameras, and more that require real-time intelligent response at low power. Inspired by the biological brain, its computing architecture utilizes sparsity to enable a design which scales from tiny to large-scale machine learning applications.



www.graimatterlabs.ai

Enabling the next generation of **Sensor and Hearable** products to process rich data with energy efficiency

Visible Image



Sound



IR Image



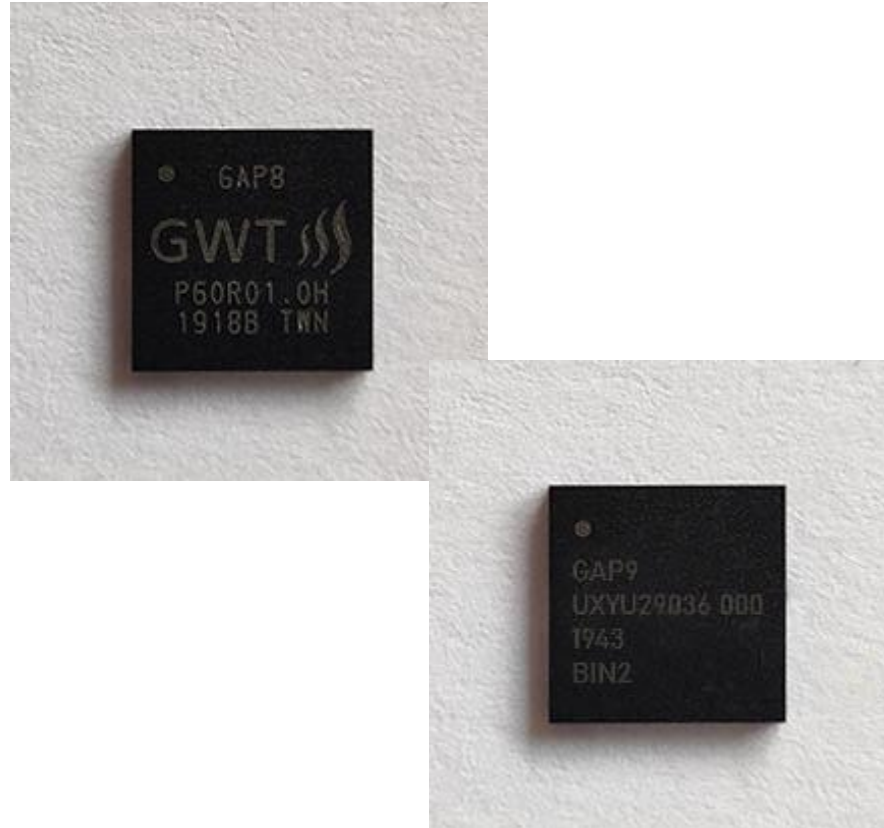
Radar



Bio-sensor



Gyro/Accel



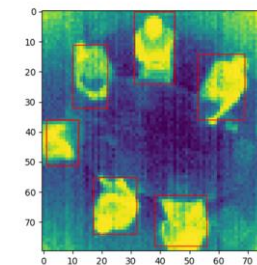
Wearables / Hearables



Battery-powered consumer electronics



IoT Sensors



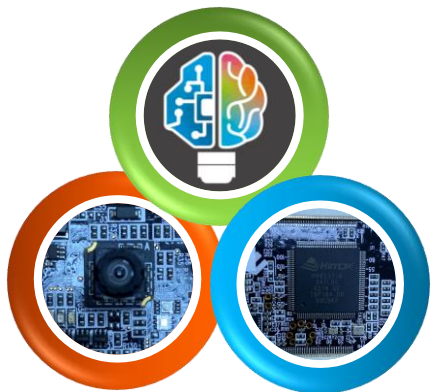


Always-On Ultra Low Power Edge AI



Himax Technologies, Inc. provides semiconductor solutions specialized in computer vision. Himax's WE-I Plus, an AI accelerator-embedded ASIC platform for ultra-low power applications, is designed to deploy CNN-based machine learning (ML) models on battery-powered AIoT devices. These end-point AI platforms can be always watching, always sensing, and always listening with on-device event recognition.

<https://www.himax.com.tw/products/intelligent-sensing/>



Imagimob AI SaaS



- End-to-end development of tinyML applications
- Guides and empowers users through the process
- Support for high accuracy applications requiring low power and small memory
- Imagimob AI have been used in 25+ tinyML customer projects
- Gesture control





Latent AI

Adaptive AI for the Intelligent Edge

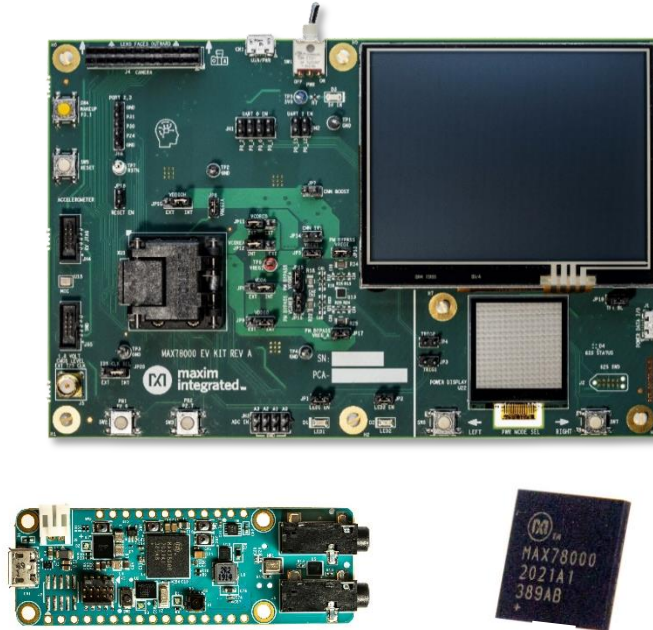
[Latentai.com](https://latent.ai)

Maxim



maxim
integrated™ ge Intelligence

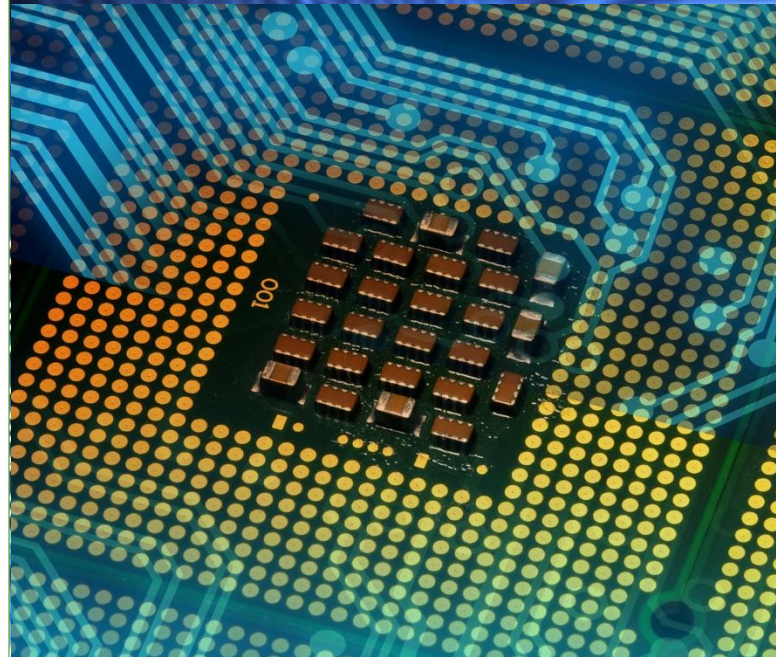
Advanced AI Acceleration IC



The new MAX78000 implements AI inferences at low energy levels, enabling complex audio and video inferencing to run on small batteries. Now the edge can see and hear like never before.

www.maximintegrated.com/MAX78000

Low Power Cortex M4 Micros



Large (3MB flash + 1MB SRAM) and small (256KB flash + 96KB SRAM, 1.6mm x 1.6mm) Cortex M4 microcontrollers enable algorithms and neural networks to run at wearable power levels.

www.maximintegrated.com/microcontrollers

Sensors and Signal Conditioning



Health sensors measure PPG and ECG signals critical to understanding vital signs. Signal chain products enable measuring even the most sensitive signals.

www.maximintegrated.com/sensors

Qeexo AutoML

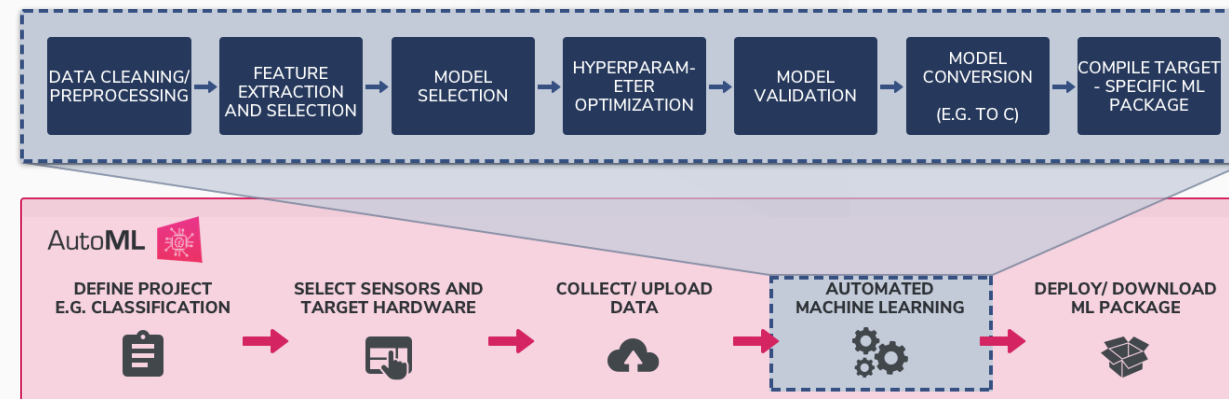


Automated Machine Learning Platform that builds tinyML solutions for the Edge using sensor data

Key Features

- Supports 17 ML methods:
 - Multi-class algorithms: GBM, XGBoost, Random Forest, Logistic Regression, Gaussian Naive Bayes, Decision Tree, Polynomial SVM, RBF SVM, SVM, CNN, RNN, CRNN, ANN
 - Single-class algorithms: Local Outlier Factor, One Class SVM, One Class Random Forest, Isolation Forest
- Labels, records, validates, and visualizes time-series sensor data
- On-device inference optimized for low latency, low power consumption, and small memory footprint applications
- Supports Arm® Cortex™- M0 to M4 class MCUs

End-to-End Machine Learning Platform



For more information, visit: www.qeexo.com

Target Markets/Applications

- Industrial Predictive Maintenance
- Smart Home
- Wearables
- Automotive
- Mobile
- IoT



Reality AI[®]

Add Advanced Sensing to your Product with Edge AI / TinyML

<https://reality.ai>



info@reality.ai



[@SensorAI](https://twitter.com/SensorAI)



[Reality AI](https://www.linkedin.com/company/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

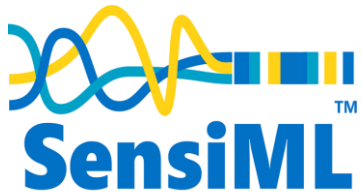
Pre-built automotive solution that lets cars
“see with sound”

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



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



Silicon Labs (NASDAQ: SLAB) provides silicon, software and solutions for a smarter, more connected world. Our technologies are shaping the future of the Internet of Things, Internet infrastructure, industrial automation, consumer and automotive markets. Our engineering team creates products focused on performance, energy savings, connectivity, and simplicity. silabs.com

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



TensorFlow

TensorFlow is an end-to-end open source platform for machine learning. Our ecosystem of tools, libraries, and community resources help users push the state-of-the-art in building and deploying ML powered applications.

[tensorflow.org](https://www.tensorflow.org)



Bringing technology to life



A DEEP TECH COMPANY AT THE LEADING EDGE OF THE AIOT

JOIN OUR SESSIONS DURING THE TINYML SUMMIT

Performing inference on BNNs with xcore.ai
Tuesday, March 23 at 12pm (PST)

TinyML: The power/cost conundrum
Thursday, March 25 at 12pm (PST)

VISIT [XMOS.AI](https://www.xmos.ai) TO FIND OUT MORE

Silver Sponsors



Copyright Notice

The presentation(s) in this publication comprise the proceedings of tinyML[®] Summit 2021. The content reflects the opinion of the authors and their respective companies. This version of the presentation may differ from the version that was presented at the tinyML Summit. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

www.tinyML.org