tinyML® Summit

Enabling Ultra-low Power Machine Learning at the Edge

tinyML Summit 2021 Proceedings

March 22 – 26, 2021
Virtual Event

www.tinyML.org

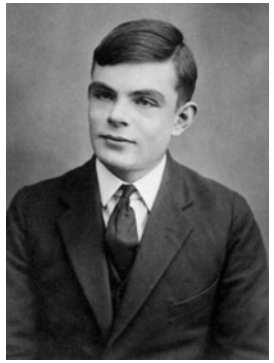
Ultra-low Power and Scalable Compute-In-Memory AI Accelerator for Next Generation Edge Inference

Behdad Youssefi – Founder and CEO

AREANNA

TINY ML

VON NEUMANN - HISTORY



- Blueprint of all modern computing platforms
- Proposed by Alan Turing in 1930s and developed by J. Presper Eckert and John Mauchly in 1940s
- A breakthrough from fixed-program computing
- Bifurcated computer engineering into software and hardware engineering for the first time

Source: www.semiengineering.com

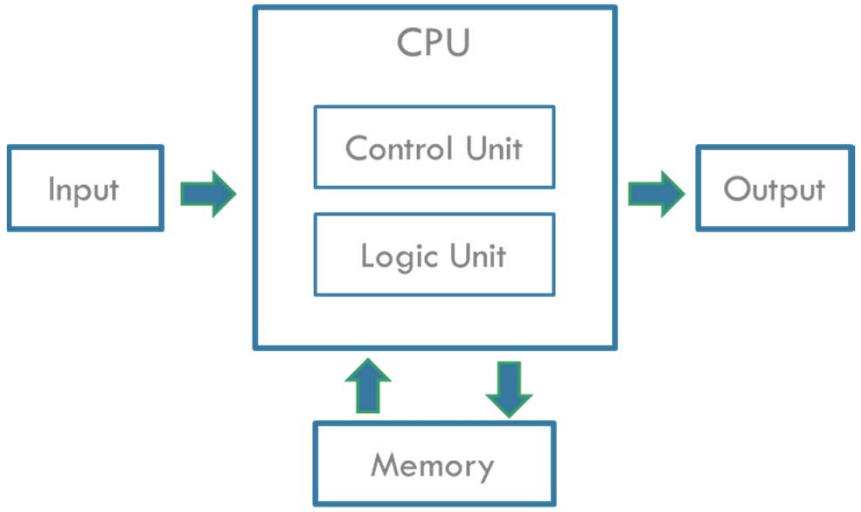
AREANNA

VON NEUMANN – WHY SO SUCCESSFUL?

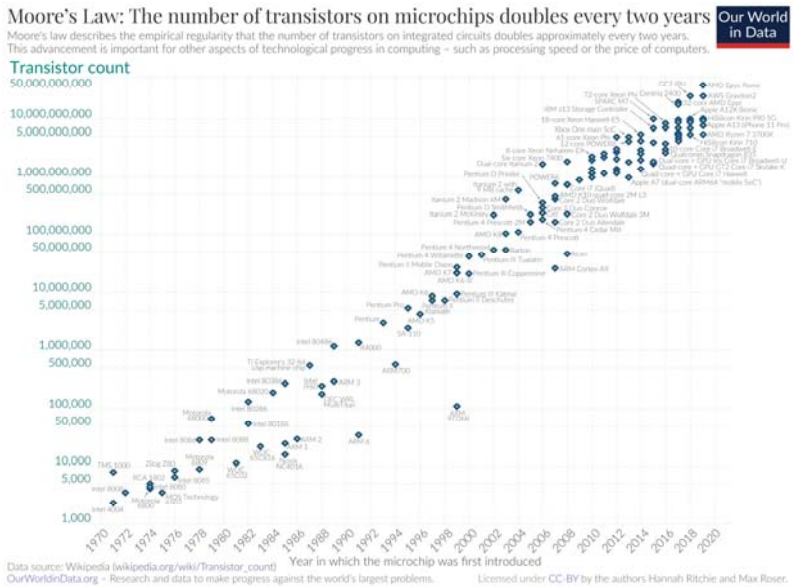
Two golden rules of von Neumann architecture:

1. **Flexibility**
   - Given enough time and memory it can complete any mathematical task

2. **Scalability**
   - Memory can be expanded
   - Width of data can be enlarged
   - Speed of computation can increase
   - Fueled by Moore's law



von Neumann architecture

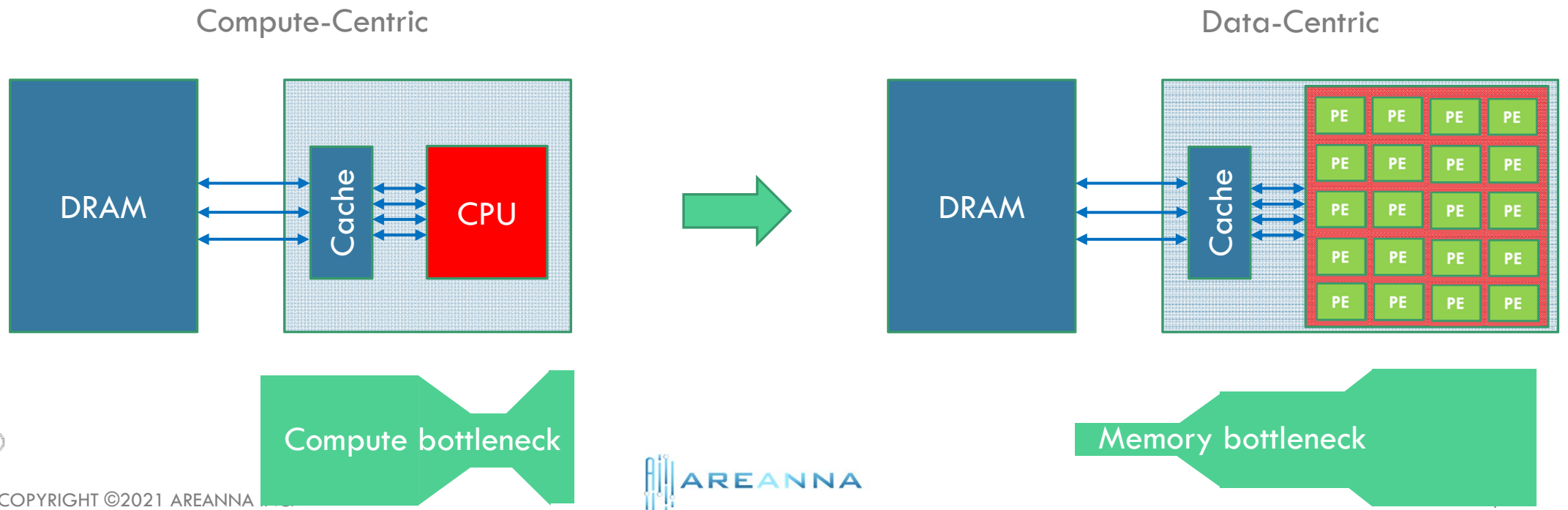


Source: www.semiengineering.com & Wikipedia

AREANNA

VON NEUMANN – SHORTCOMINGS

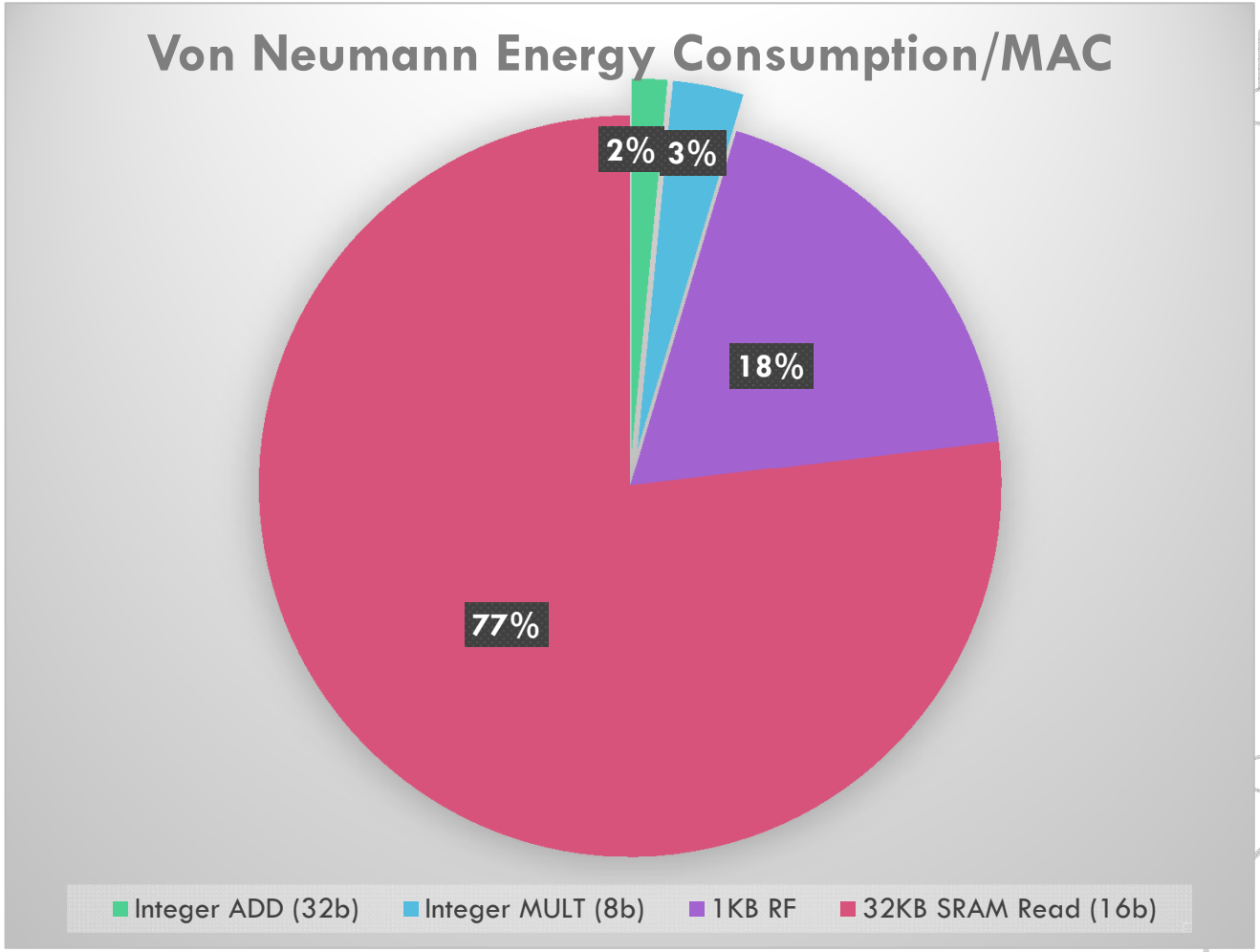
- Moving away from Compute-Centric towards Data-Centric processing
- Flexibility comes at the cost of power and performance
- Memory bottleneck



DATA MOVEMENT DOMINATES POWER CONSUMPTION

- **95% of energy consumption is in data movement**

- 8-bit MAC operation
- 32-bit accumulator
- 32KB on-chip SRAM



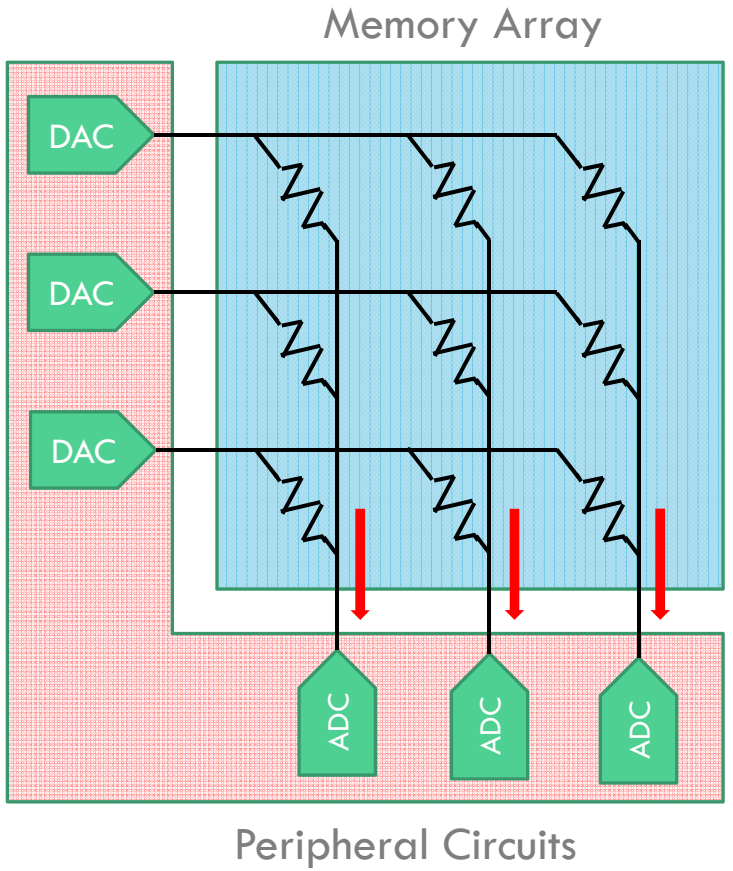
Source: "Computing Energy Problem", M. Horowitz & "Using Dataflow to Optimize Energy Efficiency…", V. Sze

AREANNA

EMBED COMPUTATION WITHIN MEMORY

Compute-In-Memory (CIM)

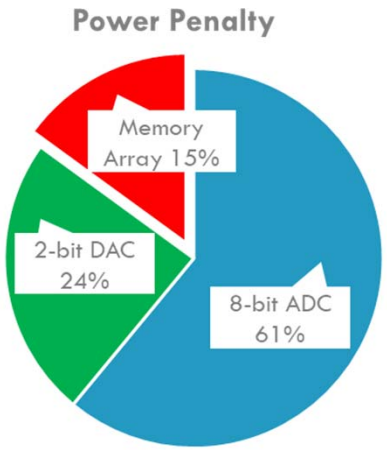
- Computation takes place inside memory array
  - Weights are stored in NVM
  - DACs generate inputs as voltages
  - Input voltages are multiplied by bit-cells' conductance
  - Outputs are collected as currents
  - ADCs quantize outputs back to digital

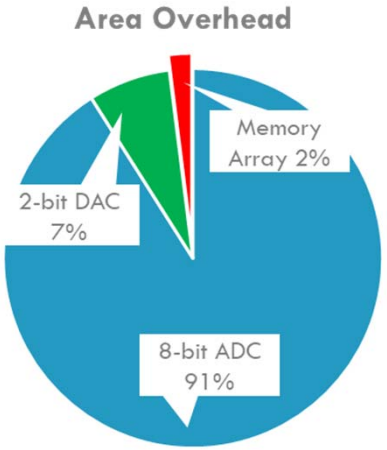


AREANNA

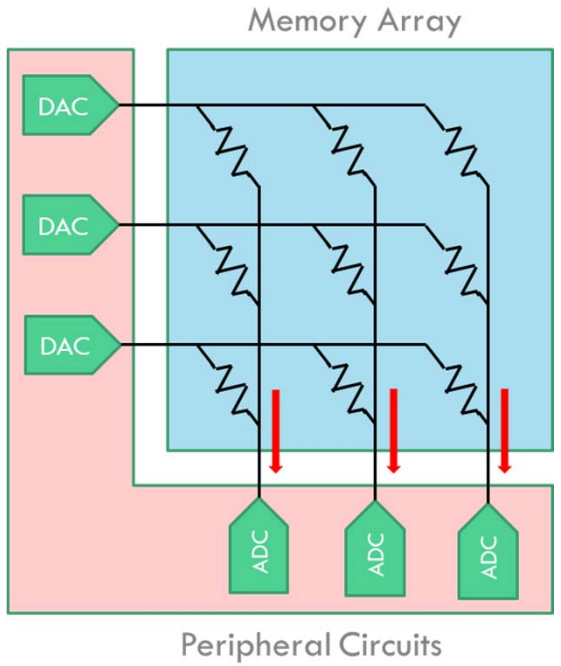
COMPUTE-IN-MEMORY – LIMITATIONS

- Data Conversion Bottleneck
  - 85% of overall power
  - 98% of overall area
- Lack of Scalability
  - Logic technology optimized for power and performance
  - Memory technology optimized for density
  - Putting them together gives worst of both worlds
- High Power Write
  - All weight coefficients must be on-chip
  - Stuck with Weight Stationary (WS) dataflow





Source: ISSCC 2020

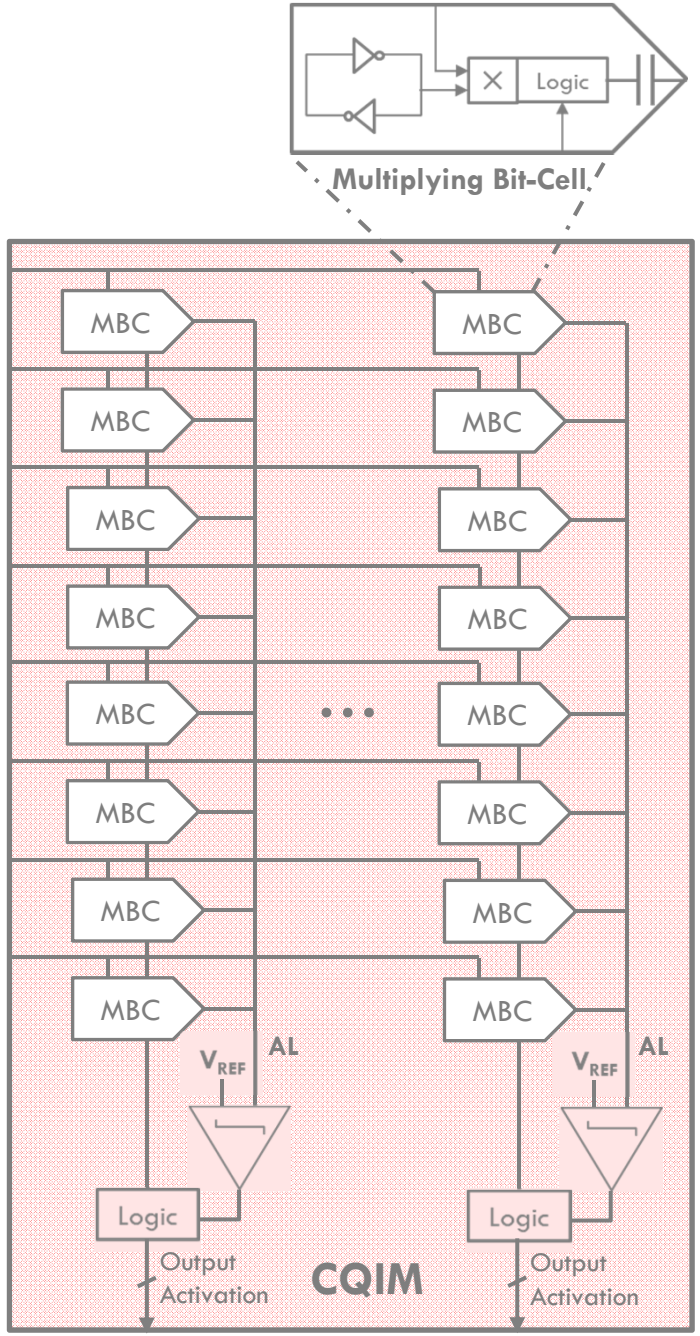


AREANNA

EMBED DATA CONVERSION IN MEMORY

Compute-and-Quantize-In-Memory (CQIM)

- DACs and ADCs are realized by the same circuit structures within memory array
  - Obviates the need for sampling circuit
  - Significant power and area saving
- In-Memory Computation
  - Weight parameters are stored in SRAM bit-cells
  - Product of weights and input activations converted to charge by unit capacitors
  - MAC result accumulates on Accumulation Line (AL)
- Only one quantization per dot product
  - 8x reduction in data conversion energy
- Fully programmable resolution comes for free
  - 1-8 bit computation while maintaining hardware utilization rate



AREANNA

CQIM: RECORD BREAKING ARCHITECTURE

Highest power efficiency

- 40 TOPS/W baseline power efficiency for 8-bit compute

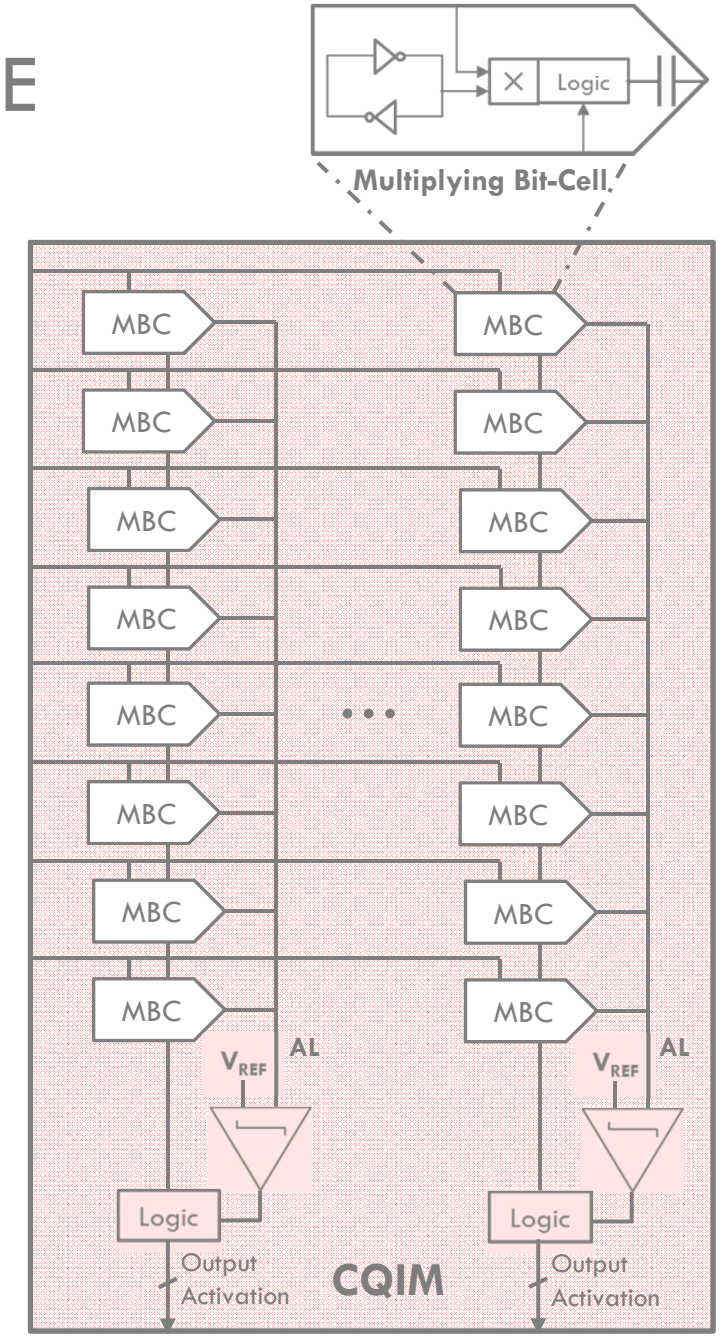
Highest computational density

- 2 TOPS/mm^2 for 8-bit compute

Highest memory bandwidth

- 2TB/s per core

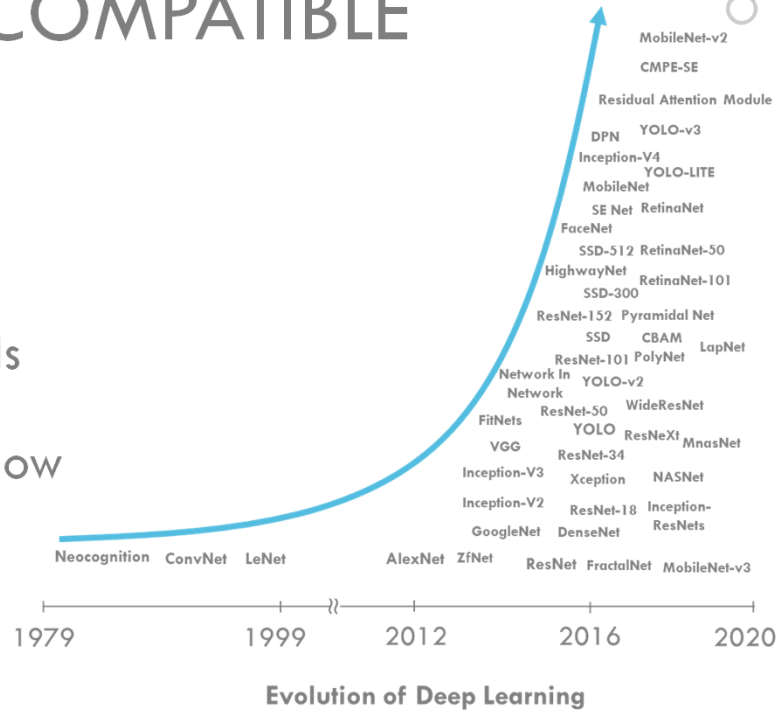
Highest dynamic range



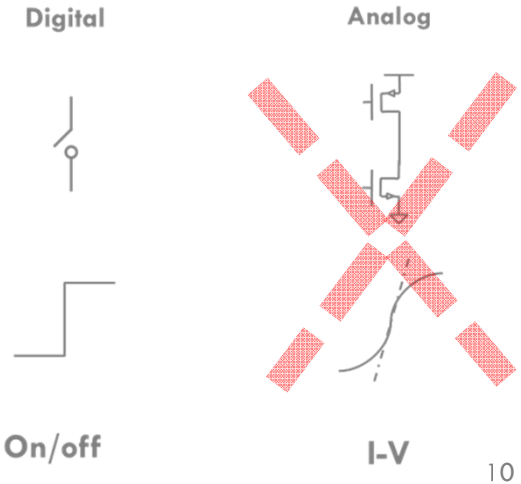
AREANNA

CQIM: VON NEUMANN GOLDEN RULES COMPATIBLE

1. Flexibility
   - Efficiency is maintained across wide range of DNN workloads
   - More efficient data reuse with Double Stationary (DS) dataflow
   - Variable precision across different layers & data types
   - Sparsity-aware
2. Scalability
   - Fabricated in CMOS process and tracks with Moore's law
   - Built almost entirely on digital blocks
   - Increased performance by increasing number of tiles



Evolution of Deep Learning



AREANNA

AREANNA STATUS

- NSF backed
- Started in late 2019
- Located in Silicon Valley
- Taped out first test chip in 2020
- Issued two patents
- Small and rapidly growing

THANK YOU!!!

TINY ML SUMMIT

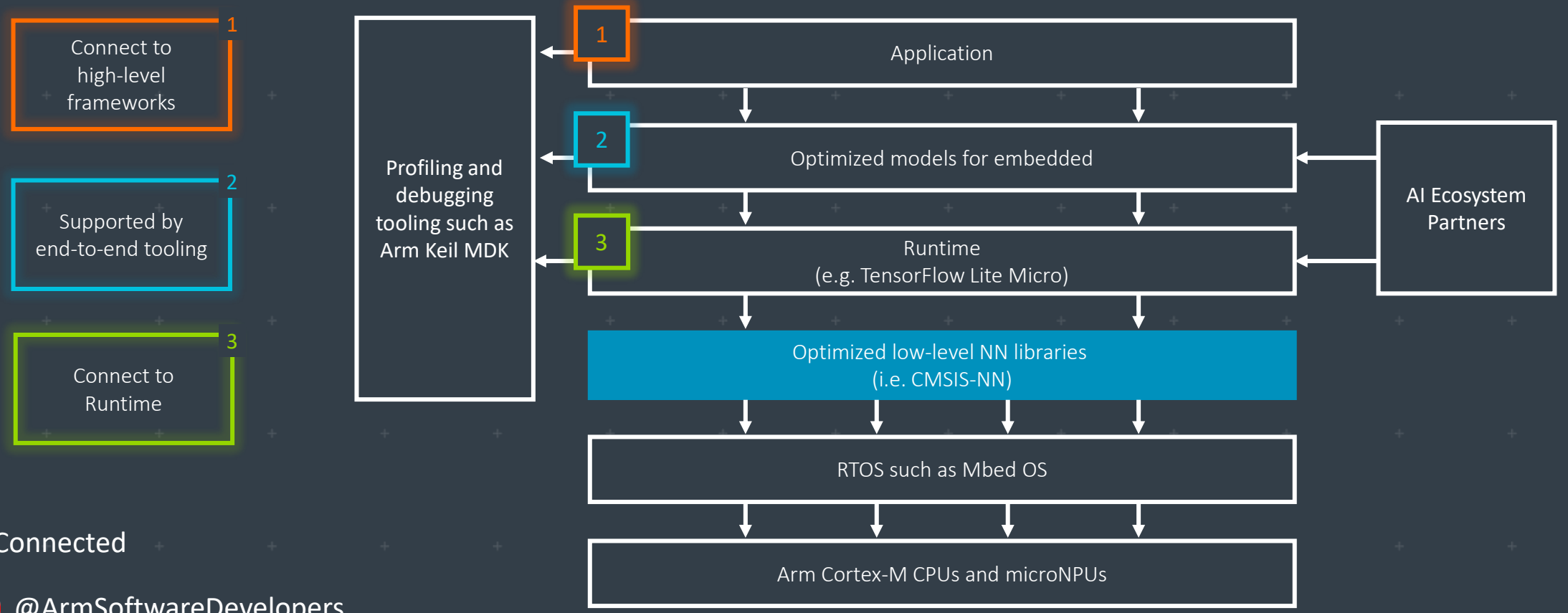
We thank the authors for their presentations and everyone who participated in the tinyML Summit 2021.

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

TINY ML SUMMIT

Executive Sponsors

Arm: The Software and Hardware Foundation for tinyML



Stay Connected

@ArmSoftwareDevelopers

@ArmSoftwareDev

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

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

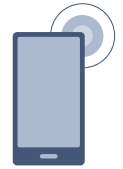
Reinforcement learning for decision making

Edge cloud

Cloud

IoT/IIoT

Automotive



Mobile

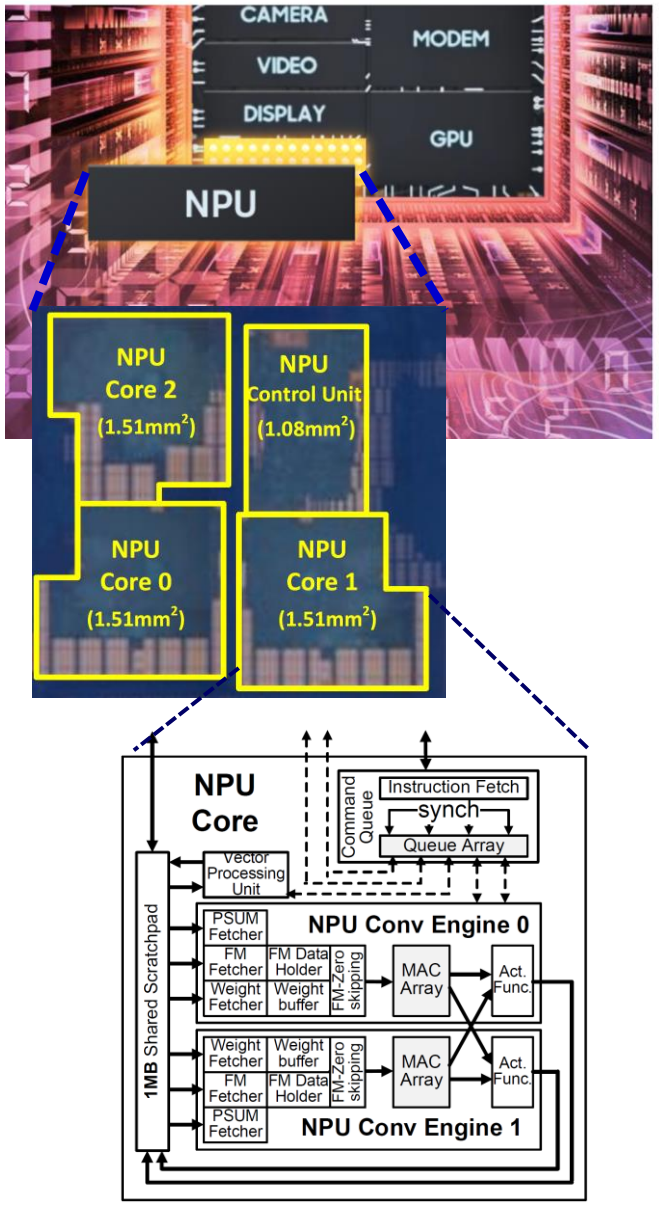
NEURAL PROCESSING

SAMSUNG

- 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.



TINY ML SUMMIT

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.

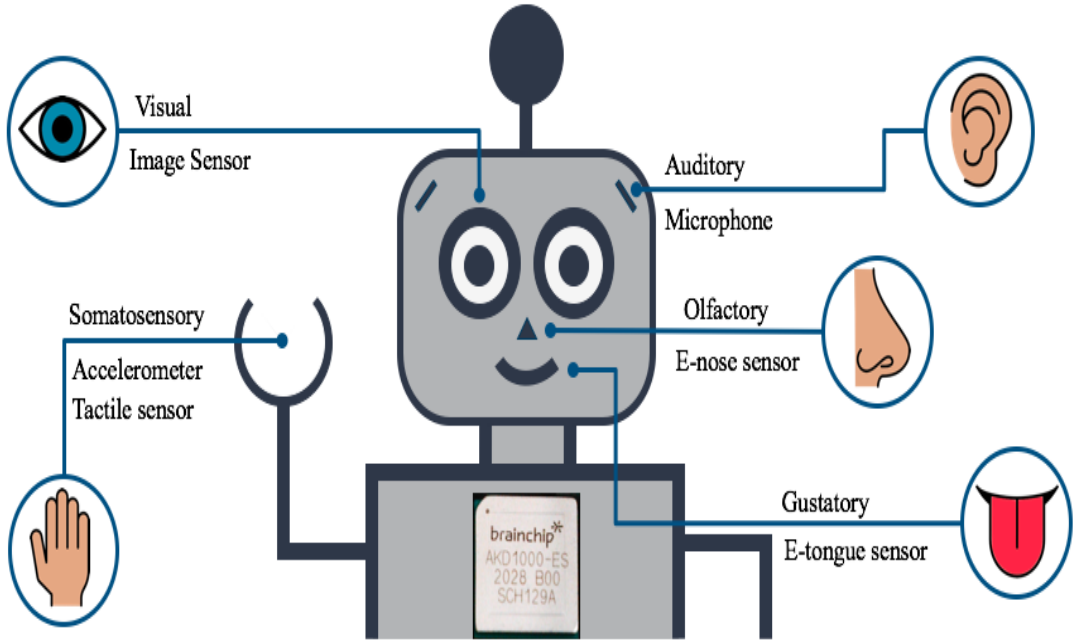
TINY ML SUMMIT

Gold Sponsors

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

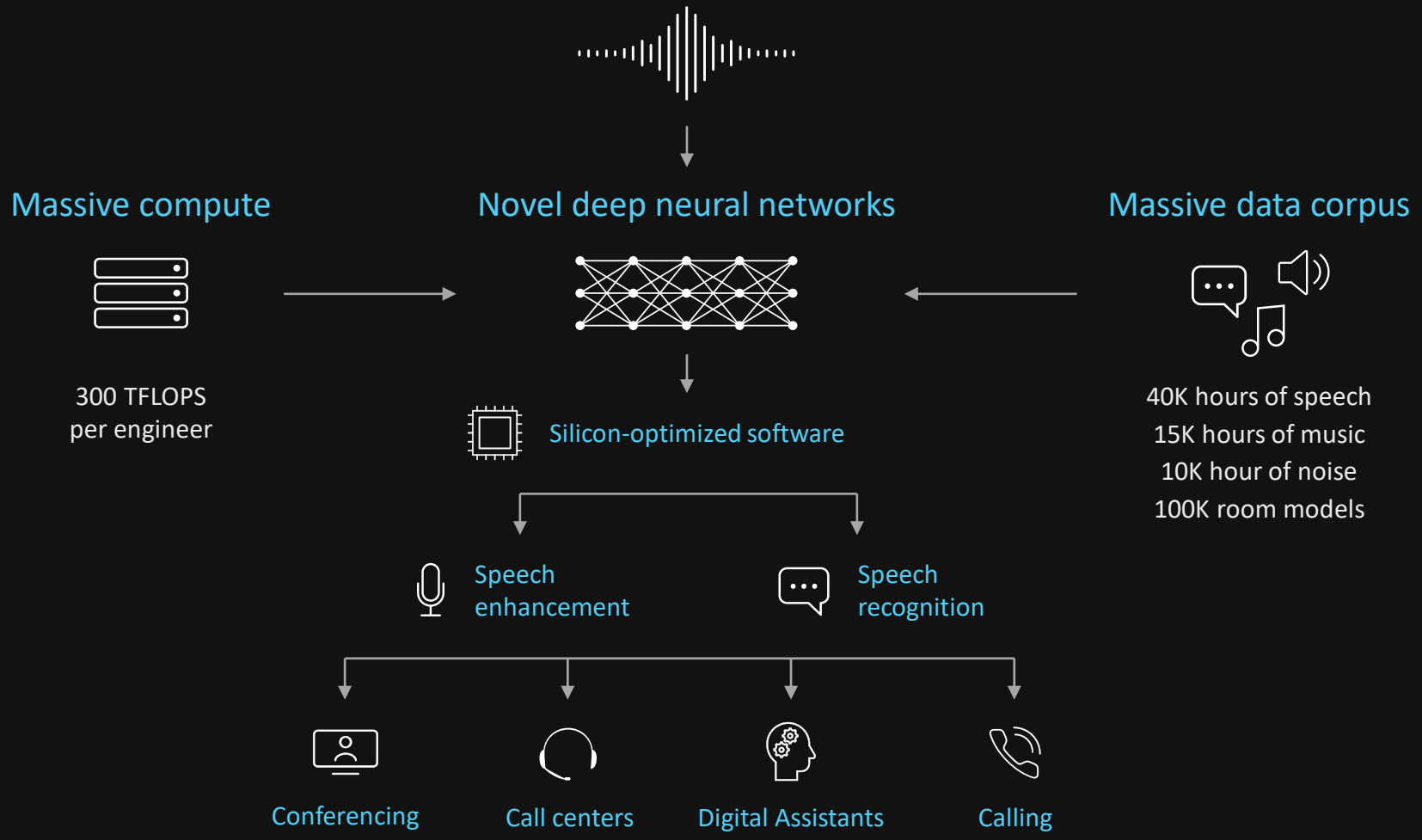
- Supports ultra-low power applications (microwatts to milliwatts)
- Edge capabilities: on-chip training, learning, and inference
- Designed for AI Edge applications: vision, audio, olfactory, and smart transducer applications
- Licensed as IP to be designed into SoC or as silicon
- Sensor inputs are 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



Cisco Webex

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.

TinyML for all developers

C++ library

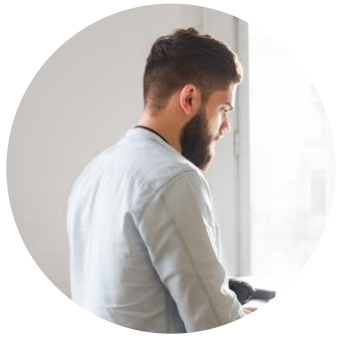
Arduino library

WebAssembly

Dataset

Acquire valuable training data securely

Enrich data and train ML algorithms



Edge Device

Real sensors in real time

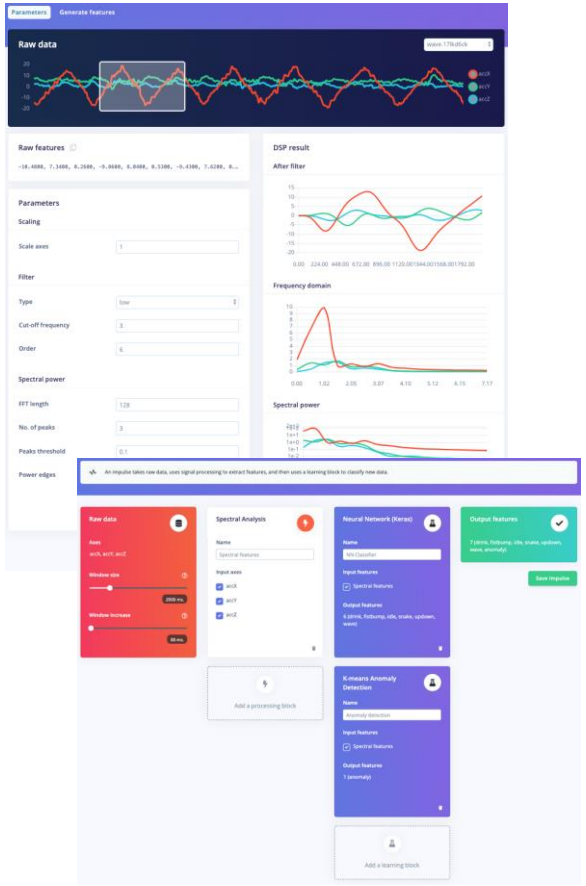
Open source SDK

Impulse

Embedded and edge compute deployment options

Test impulse with real-time device data flows

Test

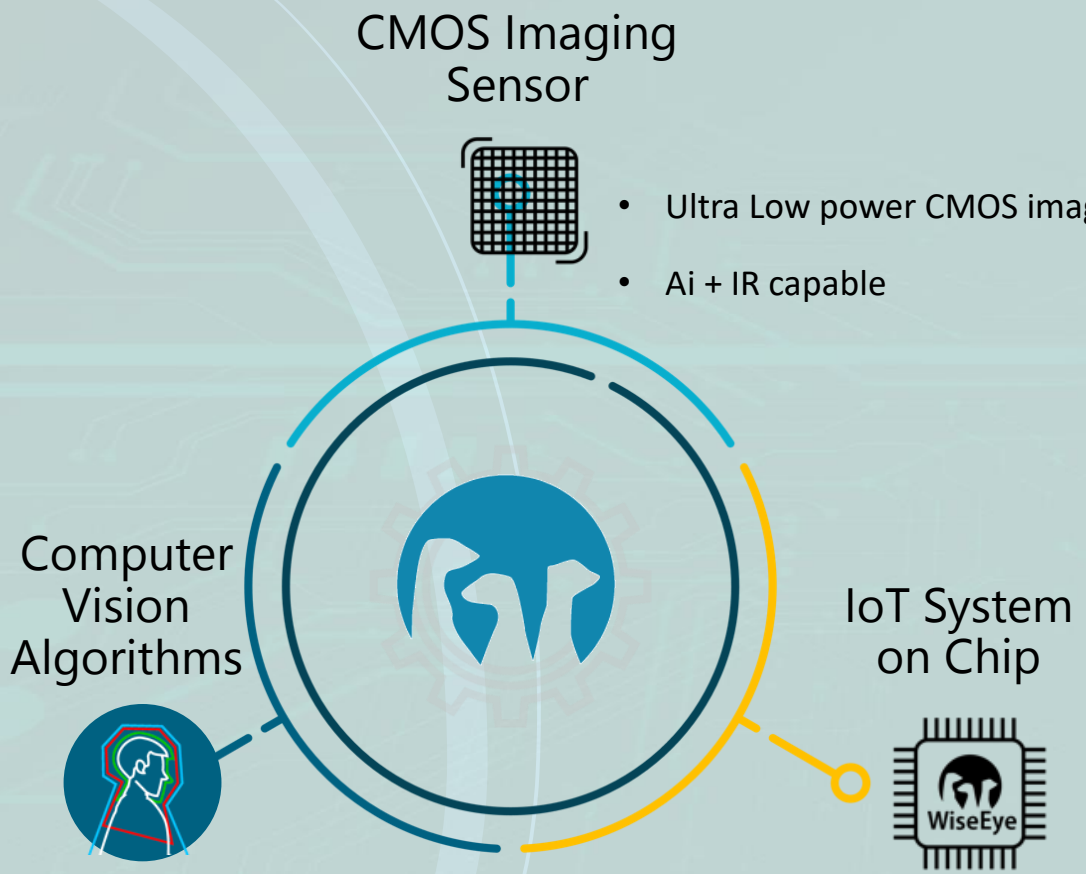


www.edgeimpulse.com

emza
visual sense

The Eye in IoT

Edge AI Visual Sensors



- Ultra Low power CMOS imager
- Ai + IR capable

- Machine Learning algorithm
- <1MB memory footprint
- Microcontrollers computing power
- Trained algorithm
- Processing of low-res images
- Human detection and other classifiers

- Machine Learning edge computing silicon
- <1mW always-on power consumption
- Computer Vision hardware accelerators

info@emza-vs.com

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

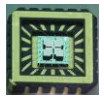
Sound

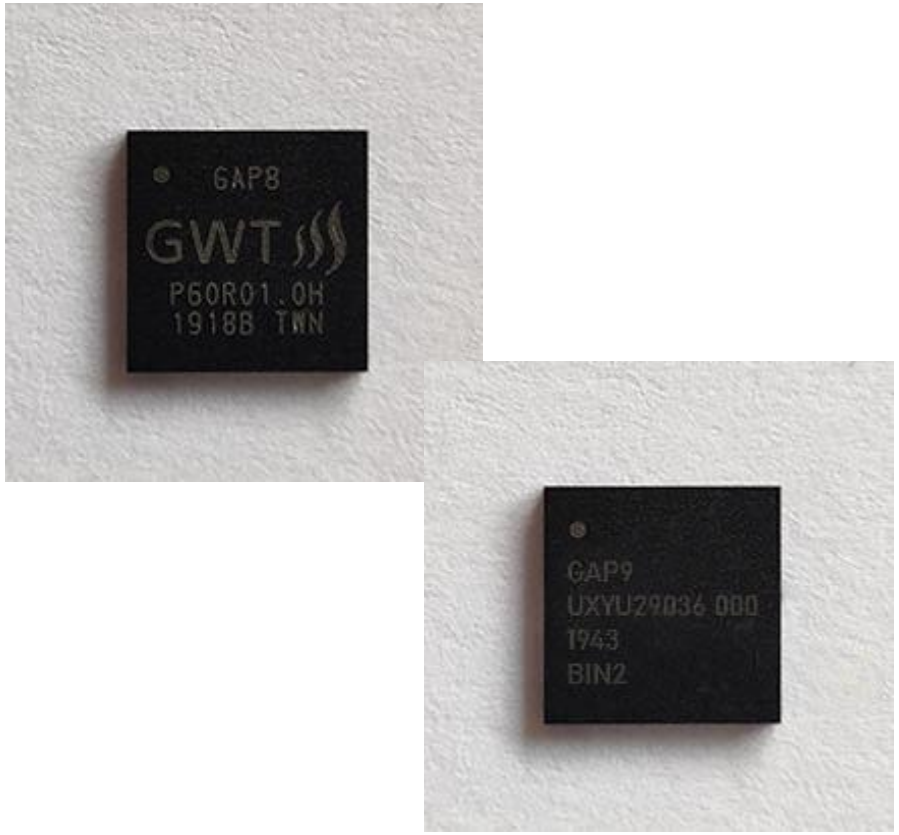
IR Image

Radar

Bio-sensor

Gyro/Accel

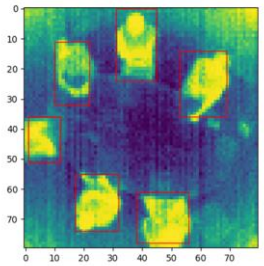




Wearables / Hearables

Battery-powered consumer electronics

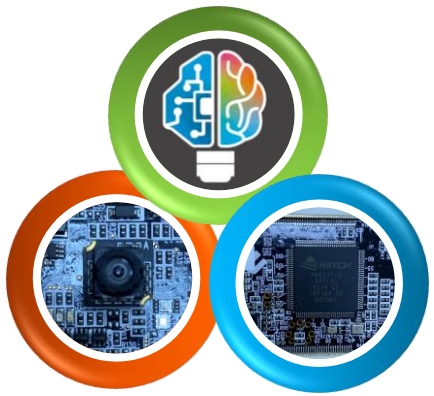
IoT Sensors



GREENWAVES TECHNOLOGIES

Himax

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

imagimob.com

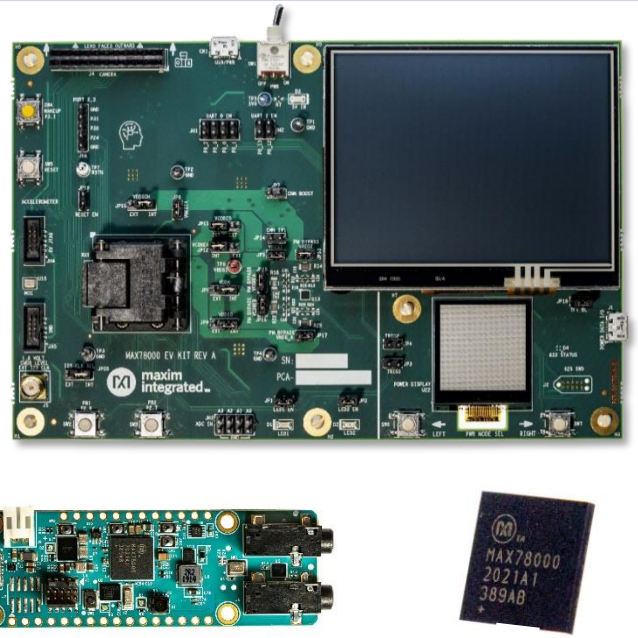
LatentAI

Adaptive AI for the Intelligent Edge

Latentai.com

Maxim maxim integrated™ ge Intelligence

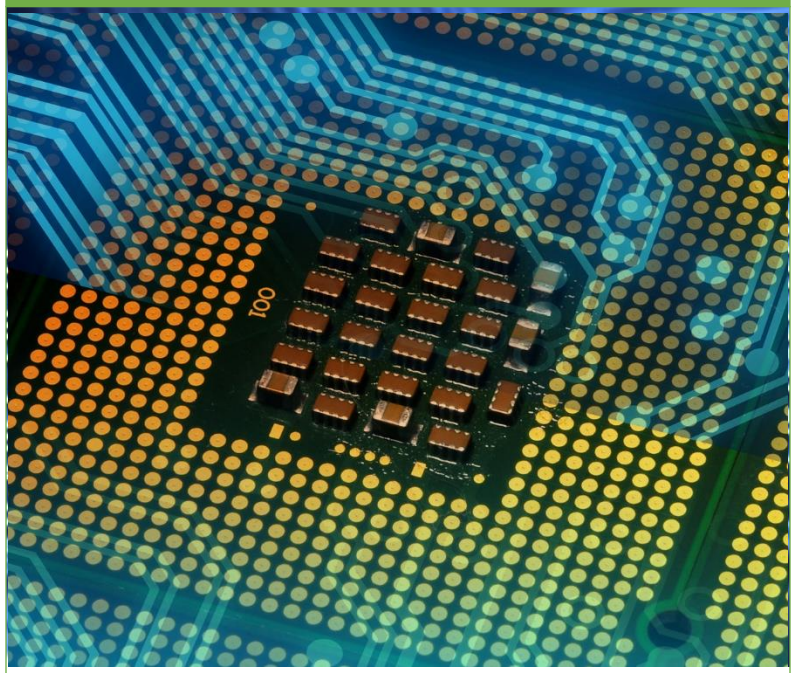
Advanced AI Acceleration IC



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

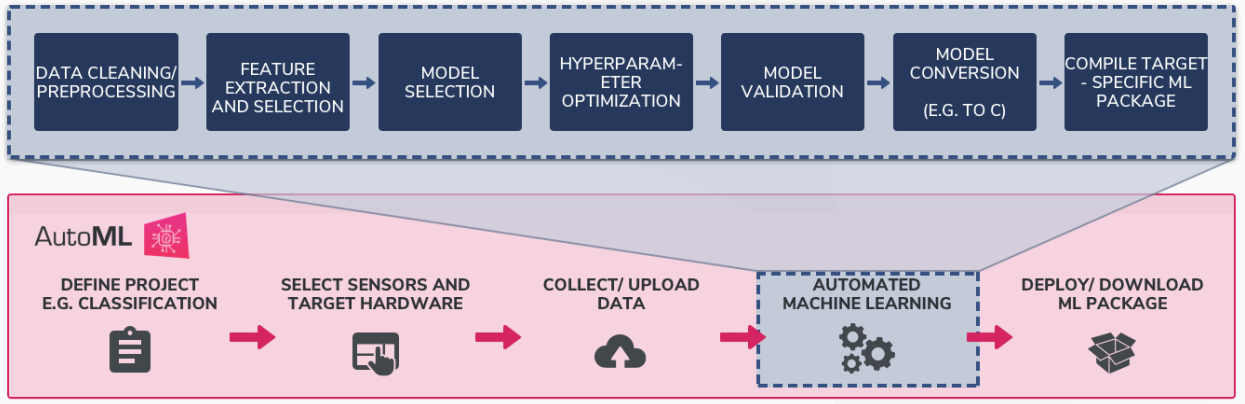
Qeexo AutoML

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 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

SensiML™

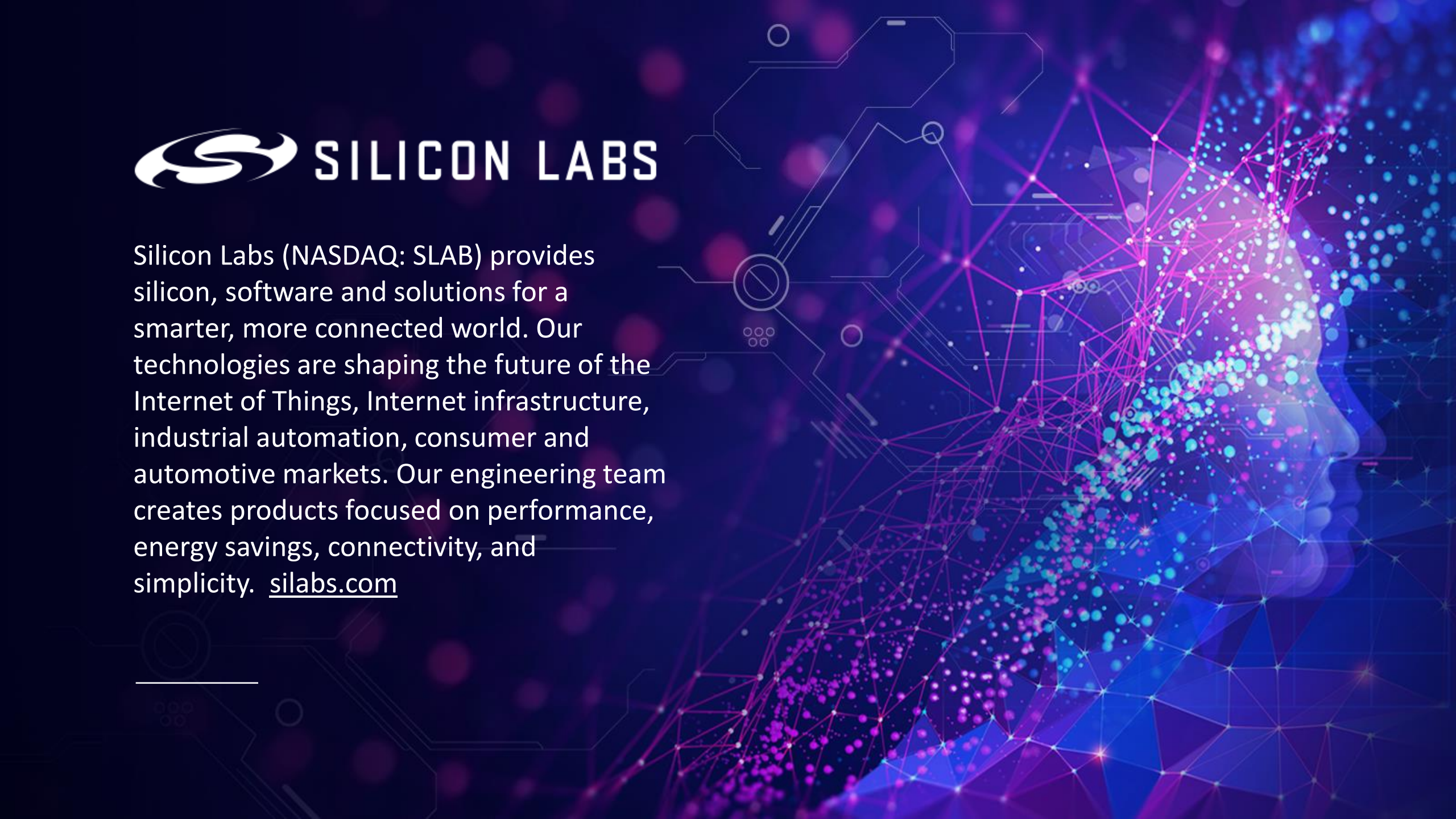
Build Smart IoT Sensor Devices From Data

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



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

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 TO FIND OUT MORE

TINY ML SUMMIT

Silver Sponsors

Copyright Notice