TinyML Enables The Intelligence of Things

The New “IoT”

Processing data where it is generated and used in battery powered devices

Mobile  Consumer  Health  Industrial
What Is Driving TinyML?

TinyML Powers Rich & Robust Human Machine Interfaces
SUCCESS depends on processing and analyzing data at the source.

Requirements For Human Machine Interfaces:
- Privacy
- Reliability
- Latency
- Energy Efficiency
TinyML Relies On Three Pillars

1. **Silicon**
   - Chip architecture is defined for optimal efficiency of deep learning workloads

2. **Training Pipeline**
   - End to end software tools to create production ready deep learning models

3. **Data**
   - Data flywheel allows for accuracy and improvement over time
Architecture Opportunities For TinyML Silicon

1. **Memory Dominated**
   Performance and power are driven by memory bandwidth and data movement costs

2. **Parallel Processing**
   Identical operations on multiple blocks of data

3. **Modest Precision**
   ≤8-bits often sufficient

SYNTIANT CORE®
A New Kind of Processor
Semiconductor Design
Deep Learning
Three Pillars Of SYNTIANT® Products

Silicon
~100x the efficiency, ~10x the performance
~½ the die size

Training Pipeline
End to end software tooling to create production deep learning models

Data
Syntiant owned data flywheel allows continuous performance improvements
SYNTIANT® NDP100

- Negligible impact to battery life
- Quickest time-to-market
- Smallest footprint
- Most cost effective
- Adding Always-on voice interface to anything
- Adding Always-on sensing to anything

Private  Reliable  Responsive  Ultra-low-power
NDP100: Ultra-low Power Neural Decision Processor

Features
- 2 PDM mics, I2S or PCM-over-SPI input
- Frequency, time-domain & batch input models
- 3-sec 16-bit input audio holding tank

Applications
- Keyword speech interface
- Wake word detection
- Speaker identification
- Event detection
- Other 'low-speed' tasks

Specifications
- Fully connected Deep Neural Network (1600 • 256 • 256 • 256 • 64)
- Output classes: 64
- Parameters: 560K
- Always-on Power: 140uW
NDP120: Multi-feature, Multi-modal Neural Decision Processor

Ideally Suited for:
- Mid Range to Premium Smartphones
- High End Hearables & Wearables
- Smart Speakers
- Laptops
- Home Entertainment (STBs, media streamers, sound bars, remotes)
- Health & Industrial IoT devices

Applications
- Far-, Near-, Close-field Audio
- Key word speech interface
- Echo Cancellation
- Noise Suppression
- Speaker identification
- Audio Event Classification
- Sensor hub & 'low-speed' tasks

Key Specifications
- Far-field audio power: < 1mW
- Audio format: 16-bit, 16-48Ksps
- Parameter precision: 1, 2, 4, 8-bit
- Maximum layers: 256
- Maximum layer width: 4096 neurons
- Layer types: FC, Conv, DSConv, RNN, LSTM

Application:
- HiFi DSP
- I2C
- Quad SPI

Audio Inputs:
- Audio Input #1: 16KHz / 48KHz
- Audio Input #2: 16KHz / 48KHz

Syntiant Core 2®
- DNN Instruction set
- DNN Data Path
- CNN/FC/RNN
- Holding Tank
- 7M+ Neural Network Parameters

Features:
- Far-, Near-, Close-field Audio
- Key word speech interface
- Echo Cancellation
- Noise Suppression
- Speaker identification
- Audio Event Classification

3.1mm x 2.5mm WLBGA
NDP120 is the TinyML Platform for Modelers

Run your architecture on the NDP120

Codesign without compromise
Number of Employees: 75

Median Experience: 20 yrs

Funds Raised: $65M (A:$5M, B:$25M, C:$35M)

Turn-key solution: Silicon, Software, Data

Shipped over 10M units in just over 3 years of founding

Locations: Irvine, CA (HQ); India; China; SG; UK; NL; DN
THANK YOU

www.syntiant.com
Backup
Major Milestones

Series A Financing $5M
First Product Samples
Series B Financing $25M
Series C Financing $35M
Shipped over One Million Units
Shipped over 10 Million Units


funding → first samples: under 1 year
funding → first million units: under 3 years
funding → first 10 million units: just over 3 years
NDP120: Multi-feature, Multi-modal Neural Decision Processor

Neural Accelerator
- Highly Parallel
- Highly efficient
- Specialized

General Compute
- Highly flexible
- Connective
Syntiant Development Flow Bridges The Gap Between Modelling And Silicon

Syntiant chip sim is implemented in Tensorflow
NDP120
Performance Model

- Modelers can negotiate with system properties
  - **Accessible Design Space:** Strong estimates of power, latency, and memory
  - **Hyperparameter Optimization:** Joint optimization of task, power, latency, and memory performance
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

1. Connect to high-level frameworks
   - Profiling and debugging tooling such as Arm Keil MDK

2. Supported by end-to-end tooling
   - Application
   - Optimized models for embedded
   - Optimized low-level NN libraries (i.e. CMSIS-NN)
   - RTOS such as Mbed OS
   - Arm Cortex-M CPUs and microNPUs

3. Connect to Runtime
   - Connect to high-level frameworks
   - Supported by end-to-end tooling
   - Connect to Runtime

Stay Connected

@ArmSoftwareDevelopers
@ArmSoftwareDev

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

Perception
Object detection, speech recognition, contextual fusion

Reasoning
Scene understanding, language understanding, behavior prediction

Action
Reinforcement learning for decision making

A platform to scale AI across the industry
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** 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](http://www.etacompute.com)
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
BabbleLabs AI speech wizardry in Cisco Webex

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

- Massive compute
  - 300 TFLOPS per engineer

- Novel deep neural networks
  - Silicon-optimized software

- Massive data corpus
  - 40K hours of speech
  - 15K hours of music
  - 10K hour of noise
  - 100K room models

- Speech enhancement
- Speech recognition

- Conferencing
- Call centers
- Digital Assistants
- Calling
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

Acquire valuable training data securely

Edge Device
Real sensors in real time
Open source SDK

Embedded and edge compute deployment options

Test

Dataset

Enrich data and train ML algorithms

Impulse

Test impulse with real-time device data flows

www.edgeimpulse.com
The Eye in IoT

Edge AI Visual Sensors

- Ultra Low power CMOS imager
- AI + IR capable

Computer Vision Algorithms

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

IoT System on Chip

- 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.
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.

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
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

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

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
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
Add Advanced Sensing to your Product with Edge AI / TinyML

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

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

EDGE
cortix

HOTG

SynSense
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