TinyML Solution Power without Tiny Models: the NDP120
NDP120 is the TinyML Target for Modelers

Run your architecture on the NDP120

Codesign without compromise
Can we update Confirming Model 2?
Can we update Confirming Model 1?
Can we update the Front End Model?
"Alexa"
Cascade Models

Front End Model Update Checklist

1. 3 jointly calibrated and tested models
   a. Latency
   b. Memory
   c. Task performance

2. Simultaneous deployment
   a. Microcontroller update
   b. Application Firmware
   c. Cloud model updates

3. Coordinate
   a. Model vendors, OEM, OS updates, Etc.
Solve Model Entanglement Dependency Hell

Move the Cascade into the NDP120 and You Can Update Every 10 Minutes

"Alexa"

NDP120
Front End Model
Confirming Model 1
Confirming Model 2

Application Processor
No-Op

Voice Agent

GPU
No-Op
Solve Model Entanglement Dependency Hell

Move the Cascade into the NDP120 and You Can Update Every 10 Minutes

This is not possible: needs to be tiny. :(

"Alexa"

NDP120
Front End Model
Confirming Model 1
Confirming Model 2

Consumer Device

Application Processor

Voice Agent

GPU
No-Op

No-Op
Solve Model Entanglement Dependency Hell

Move the Cascade into the NDP120 and You Can Update Every 10 Minutes

"Alexa"

This is not possible: needs to be tiny. :(

But what is "Tiny"?

"Alexa"
Ok Google Porting Story

Google has their model on billions of devices and wants to keep it

- No retraining: "here is the model"
- Must be low latency
- Must be low power

Can highly-specialized edge silicon handle this without changing the hardware or the model?
Setting 768kHz PDM audio
Okay-G Hotword version 0x14b8c880
Audio Buffer [0x300850e0:0x300968e0]
Monitoring wakeword for 60.0 sec

Monitoring NDP120 VDDD: 305.0μA, 274.5μW
Monitoring NDP120 VDDD: 307.5μA, 276.8μW
Monitoring NDP120 VDDD: 310.0μA, 287.9μW
Monitoring NDP120 VDDD: 315.0μA, 283.5μW

-----
OKAY GOOGLE! - match 0 at frame 975

Monitoring NDP120 VDDD: 305.0μA, 274.5μW
Monitoring NDP120 VDDD: 302.5μA, 272.2μW
Monitoring NDP120 VDDD: 310.0μA, 287.9μW
Monitoring NDP120 VDDD: 312.5μA, 281.2μW
Monitoring NDP120 VDDD: 312.5μA, 281.2μW
Monitoring NDP120 VDDD: 312.5μA, 281.2μW
Monitoring NDP120 VDDD: 312.5μA, 281.2μW
Monitoring NDP120 VDDD: 310.0μA, 279.0μW
OK Google Implementation

Total power is **280 uW**

- Lowest power implementation of “Okay Google” in the world
- 97% of the NDP120 DNN engine is still available for speaker verification, song identification, ensemble modeling, staged architectures, and more.
Ok Google Porting Story

Key Lessons

1. We learned from shipping 10+ million prior generation chips
2. We have solid hardware+software co-design practices
3. We have excellent tooling and support
4. Flexibility in neural accelerators is a reality
Key Lessons

1. We learned from shipping 10+ million prior generation chips
2. We have solid hardware+software co-design practices
3. We have excellent tooling and support
4. Flexibility in neural accelerators is a reality
NDP120 Multi-modal, Multi-feature Neural Decision Processor

Neural Accelerator
- Highly Parallel
- Highly efficient
- Specialized

General Compute
- Highly flexible
- Connective
Ok Google Porting Story

Key Lessons

1. We learned from shipping 10+ million prior generation chips
2. We have solid **hardware+software** co-design practices
3. We have excellent tooling and support
4. Flexibility in neural accelerators is a reality
Software Bridging the Gap

Syntiant chip sim is implemented in Tensorflow
Ok Google Porting Story

Key Lessons

1. We learned from shipping 10+ million prior generation chips
2. We have solid hardware+software co-design practices
3. We have excellent tooling and support
4. Flexibility in neural accelerators is a reality
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
Key Lessons

1. We learned from shipping 10+ million prior generation chips
2. We have solid hardware+software co-design practices
3. We have excellent tooling and support
4. **Flexibility** in neural accelerators is a reality
Supporting multiple models without joint training

ALEXA & GLASS BREAKING
CONCURRENT MODELS
SYNTIANT NDP120
FAR FIELD DEMONSTRATION WITH ECHO CANCELLATION

SYNTIANT NDP120

CONTACT: INFO@SYNTIANT.COM
WWW.SYNTIANT.COM/NDP120
NDP120: Tiny Power without Compromises

- Empowers modelers
- Highly efficient
- Highly flexible
- Tiny

<table>
<thead>
<tr>
<th>Key Specifications</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Solution Power</td>
<td>&lt; 1mW</td>
</tr>
<tr>
<td>Audio format</td>
<td>16-bit, 16-48Ksps</td>
</tr>
<tr>
<td>Parameter precision</td>
<td>1,2,4,8-bit</td>
</tr>
<tr>
<td>Maximum layers</td>
<td>256</td>
</tr>
<tr>
<td>Maximum layer width</td>
<td>4096 neurons</td>
</tr>
<tr>
<td>Layer types</td>
<td>FC, Conv, DSConv, LSTM, GRU, Pooling, Deconv, ...</td>
</tr>
<tr>
<td>Dimensions</td>
<td>3.1mm x 2.5mm WLBGA 5mm x 5mm QFN</td>
</tr>
</tbody>
</table>
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

Connect to high-level frameworks
Profiling and debugging tooling such as Arm Keil MDK
Supported by end-to-end tooling
Connect to Runtime

Application
Optimized models for embedded
Runtime (e.g. TensorFlow Lite Micro)
Optimized low-level NN libraries (i.e. CMSIS-NN)
RTOS such as Mbed OS
Arm Cortex-M CPUs and microNPUs

AI Ecosystem Partners

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

- Novel deep neural networks
- Silicon-optimized software
- Speech enhancement
- Speech recognition

Massive data corpus

- 300 TFLOPS per engineer
- 40K hours of speech
- 15K hours of music
- 10K hour of noise
- 100K room models

Applications:

- 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

**Dataset**
- Acquire valuable training data securely
- Enrich data and train ML algorithms

**Edge Device**
- Real sensors in real time
- Open source SDK
- Embedded and edge compute deployment options

**Impulse**
- Test impulse with real-time device data flows

**Test**

[www.edgeimpulse.com](http://www.edgeimpulse.com)
The Eye in IoT

Edge AI Visual Sensors

- CMOS Imaging Sensor
  - 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
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
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
Adaptive AI for the Intelligent Edge

LatentAI

Latentai.com
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

Maxim Integrated: Enabling Edge 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
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
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

Add Advanced Sensing to your Product with Edge AI / TinyML

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