“Get Ahead of the Curve: Develop Software in the Cloud for the Ethos-U55 and Cortex-M55 Processors”

Stefano Cadario – Director Product Management, IoT Group, Arm

February 8, 2022
tinyML Talks Strategic Partners

Additional Sponsorships available – contact Olga@tinyML.org for info
Executive Strategic Partners
Arm: The Software and Hardware Foundation for tinyML

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

Application

1. Optimized models for embedded

Runtime (e.g. TensorFlow Lite Micro)

2. Optimized low-level NN libraries (i.e. CMSIS-NN)

RTOS such as Mbed OS

3. Arm Cortex-M CPUs and microNPUs

Stay Connected

@ArmSoftwareDevelopers
@ArmSoftwareDev

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

Device

Impulse

Test

The leading edge ML platform

www.edgeimpulse.com
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
SYNTIANT

Neural Decision Processors
- At-Memory Compute
- Sustained High MAC Utilization
- Native Neural Network Processing

ML Training Pipeline
- Enables Production Quality Deep Learning Deployments

Data Platform
- Reduces Data Collection Time and Cost
- Increases Model Performance

End-to-End Deep Learning Solutions for TinyML & Edge AI

partners@syntiant.com
www.syntiant.com
Platinum Strategic Partners
WE USE AI TO MAKE OTHER AI FASTER, SMALLER AND MORE POWER EFFICIENT

**Automatically compress** SOTA models like MobileNet to <200KB with little to no drop in accuracy for inference on resource-limited MCUs

**Reduce** model optimization trial & error from weeks to days using Deeplite’s **design space exploration**

**Deploy more** models to your device without sacrificing performance or battery life with our **easy-to-use software**

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
BROAD AND SCALABLE EDGE COMPUTING PORTFOLIO

Microcontrollers & Microprocessors

Arm® Core
- Arm® Cortex®-M 32-bit MCUs
  - Arm ecosystem, Advanced security, Intelligent IoT
- Arm®-based High-end 32 & 64-bit MPUs
  - High-resolution HMI, Industrial network & real-time control
- Arm® Cortex®-M0+ Ultra-low Power 32-bit MCUs
  - Innovative process tech (SOTB), Energy harvesting

Renesas Core
- Ultra-low Energy 8 & 16-bit MCUs
  - Bluetooth® Low Energy, SubGHz, LoRa®-based Solutions
- High Power Efficiently 32-bit MCUs
  - Motor control, Capacitive touch, Functional safety, GUI
- 40nm/28nm process Automotive 32-bit MCUs
  - Rich functional safety and embedded security features

Renesas Synergy™
- Arm®-based 32-bit MCUs for Qualified Platform
- Qualified software and tools

Core technologies

AI
- A broad set of high-power and energy-efficient embedded processors

Security & Safety
- Comprehensive technology and support that meet the industry’s stringent standards

Digital & Analog & Power Solution
- Winning Combinations that combine our complementary product portfolios

Cloud Native
- Cross-platforms working with partners in different verticals and organizations

© 2021 Renesas Electronics Corporation. All rights reserved.
Gold Strategic Partners
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](http://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](http://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](http://www.maximintegrated.com/sensors)
LatentAI

Adaptive AI for the Intelligent Edge

Latentai.com
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
SynSense builds sensing and inference hardware for ultra-low-power (sub-mW) embedded, mobile and edge devices. We design systems for real-time always-on smart sensing, for audio, vision, IMUs, bio-signals and more.

https://SynSense.ai
Silver Strategic Partners
tinyML Summit 2022
Miniature dreams can come true…
March 28-30, 2022
Hyatt Regency San Francisco Airport
https://www.tinyml.org/event/summit-2022/

The Best Product of the Year and the Best Innovation of the Year awards are open for nominations between November 15 and February 28.

tinyML Research Symposium 2022
March 28, 2022
https://www.tinyml.org/event/research-symposium-2022

More sponsorships are available: sponsorships@tinyML.org
tinyML Trailblazers Series
Success Stories with Marian Verhelst
(Professor, EE Department of KU Leuven)

LIVE ONLINE March 2nd, 2022 at 8 am PST

Register now!
Join Growing tinyML Communities:

tinyML - Enabling ultra-low Power ML at the Edge

The tinyML Community
https://www.linkedin.com/groups/13694488/

8k members in 42 Groups in 33 Countries

2.6k members & 4.6k followers
Subscribe to tinyML YouTube Channel for updates and notifications (including this video)
www.youtube.com/tinyML

6k subscribers, 347 videos with 177k views
Next tinyML Talks

<table>
<thead>
<tr>
<th>Date</th>
<th>Presenter</th>
<th>Topic / Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tuesday, February 15</td>
<td>Jenny Plunkett, Senior Developer Relations Engineer, Edge Impulse</td>
<td>Advanced Anomaly Detection Made Easy</td>
</tr>
</tbody>
</table>

Webcast start time is 8:00 am Pacific time

Please contact talks@tinyml.org if you are interested in presenting
Reminders

Slides & Videos will be posted tomorrow

tinyml.org/forums  youtube.com/tinyml

Please use the Q&A window for your questions
Stefano Cadario

Stefano is Director Product Management in the Arm IoT group, focusing on driving new growth initiatives, as well as engaging in strategic relationships with key cloud partners. He obtained his Bachelor's and Master's degree in Computer Engineering at Politecnico di Milano and, more recently, an MBA at Warwick Business School. Stefano joined Arm in 2012 as a Software Engineer in Cambridge (UK) and he has taken on several roles including Technical Specialist, Product Manager and Sr. Strategy Manager in San Jose (US).
Get Ahead of the Curve: Develop Software in the Cloud for the Ethos-U55 and Cortex-M55

TinyML Talks

Stefano Cadario
February 8th 2022
Agenda

- Software development for Intelligent Edge
- Why Arm Virtual Hardware?
- A new generation of TinyML applications with Ethos-U55
- A practical example of Arm Virtual Hardware use for ML
- Q&A
IoT and ML Challenges

A steep increase in complexity

- Connectivity and integration with cloud services
- Security updates and over-the-air (OTA)
- Machine Learning
Traditional Embedded Development

Software development for embedded has not changed considerably in the past 30 years

- **Deploy-and-forget**
  - Same firmware for entire life of the device
  - Updates are rare and require manual intervention

- **Software team in control of the entire development flow**
  - Local development/build environment
  - Testing is limited and done before deployment

- **Security is an afterthought**
  - Most devices are not connected, and security threat managed by limiting physical access
An Intelligent Edge Software Development Flow

Connectivity is enabling intelligence at the Edge

Training data → NN Design & Training → Trained NN → Model Optimization → Integration → Binary Image → Deployment → Devices fleet

Software development

Cloud services → Monitor & data collection
Cloud-Native Development

A new paradigm to develop modern cloud enabled applications

- **Software updates are deployed regularly**
  - Small changes made very often
  - Continuous deployment is considered from day one

- **DevOps and MLOps**
  - Development environment is code
  - Automated tests are extremely important
  - Data scientists can independently update ML models

- **Security addressed throughout the full development lifecycle**
What’s Hard with Hardware

Hardware constraints limit the use in a cloud native environment

- **No scalability**
  - Adding more hardware is not a linear effort
  - Infostructure costs can be considerably high
  - Bottleneck for developers

- **Complexity and high maintenance costs**
  - Board farms are expensive and not standardized
  - High engineering costs

- **Limited ability to stimulate I/O**
  - Additional infrastructure complexity needed
  - Corner cases cannot be reproduced consistently

- **Not available in public clouds**
What’s Arm Virtual Hardware?
Virtual Hardware Targets are the IoT equivalent of Virtual Machines

- An Arm Virtual Hardware Target is a functionally accurate representation of a physical SoC, simulating its software-visible behavior
- Runs as a simple application in a Linux environment for easy scalability in the cloud
- Remove dependency hardware or silicon availability
Arm Virtual Hardware advantages
Simplifying the developer experience and accelerating IoT

- **Cloud development**
  - Virtual targets can be encapsulated and managed through cloud infrastructure

- **Maximum scalability**
  - Can easily create 10-100-1000s different instances of the same device thanks to cloud infrastructure

- **Accelerate and simplify CI/CD flow**
  - Remove dependency from physical hardware
  - Pre-silicon availability development

![Virtual Hardware instances diagram](image)
An Intelligent Edge Software Development Flow

- Training data
- NN Design & Training
- Trained NN
- MLOps
  - NN Optimization for Edge
  - Virtual Hardware
  - Performance Metrics

- Software development
  - Virtual Hardware
  - Software development
  - Integration
    - Binary Image
    - Deployment
    - Devices fleet

- Monitor & data collection
- Cloud services
What Can You Already do Today With a Cortex-M4?

Simple speech recognition

Gesture recognition using accelerometer

Person detection
https://github.com/tensorflow/tensorflow/tree/master/tensorflow/lite/micro/examples/person_detection_experimental

Colored object recognition
Ethos-U55 Performance

Enabling a new generation of TinyML applications which weren’t possible before

- DS-CNN-L: Keyword Spotting
- MobilNet: Image Classification
- Wav2letter: Speech Recognition
- Model is stored in Flash.
- 4GB/s SRAM Bandwidth | 500MB/s Flash Bandwidth
- Host CPU for Ethos-U55 is Cortex-M55
- CMSIS_NN Optimized kernels used for Cortex-M55 performance

110x
DS-CNN-L

73.7x
MobilNet v1_0.5_128

253.7x
Wav2letter
Arm Virtual Hardware - Corstone-300

- Arm reference design for an AI MCU based on Cortex-M55 CPU and Ethos-U55 NPU
- Include some virtual peripherals such as UART and Ethernet
- Functionally accurate – behaves like real hardware
- Provide instruction counts for CPU and cycle-approximate performance for NPU
- Available on AWS Marketplace today!
Weight Clustering

Objective: determine the performance impact of weight clustering

+ Optimization algorithm aimed at reducing the storage and network transfer size of models
+ Limited impact on accuracy
+ Replace similar weights in a layer with the centroids of clusters
+ Ethos-U55 runtime / Vela compiler leverage this optimization

End-to-end Flow with Ethos-U55

Objective: determine the performance impact of weight clustering
End-to-end Flow with Ethos-U55

Objective: determine the performance impact of weight clustering

With Weight clustering enabled

-36% NPU cycles

Network topology → Training network → Weight clustering → Quantization → Vela Optimizer → Integrate in C++ example → Build Image → Run

Not clustered vs Clustered

NPU Total cycles (lower is better)

1,508,097 → 964,097

With Weight clustering disabled

INFO - NPU AX10 RD DATA BEAT RECEIVED beats: 57004
INFO - NPU AX10 WR DATA BEAT WRITTEN beats: 25396
INFO - NPU AX11 RD DATA BEAT RECEIVED beats: 173168
INFO - NPU ACTIVE cycles: 1507110
INFO - NPU IDLE cycles: 987
INFO - NPU TOTAL cycles: 1508097

INFO - Profile for Inference:
INFO - NPU AX10 RD DATA BEAT RECEIVED beats: 56732
INFO - NPU AX10 WR DATA BEAT WRITTEN beats: 25372
INFO - NPU AX11 RD DATA BEAT RECEIVED beats: 96984
INFO - NPU ACTIVE cycles: 963134
INFO - NPU IDLE cycles: 963
INFO - NPU TOTAL cycles: 964097
Summary

- Challenges of IoT/ML software development and how Arm Virtual Hardware enables scale through cloud

- A new generation of rich TinyML applications thanks to the Ethos-U55 NPU delivering 100x more performance than traditional MCUs

- A practical example of ML flow where Arm Virtual Hardware is used to optimize models to run best on the target device
More Resources

Learn more about Arm Virtual Hardware
bit.ly/arm-virtual-hardware

Join our upcoming Arm Virtual Hardware Lab Series, starting 22 Feb
bit.ly/arm-virtual-hardware-series

Get in touch
Stefano.cadario@arm.com

Try for yourself and run an inference on the Ethos-U55 microNPU
bit.ly/arm-virtual-hardware-workshop
Thank You
Danke
Gracias
Grazie
谢谢
ありがとう
Asante
Merci
감사합니다
धन्यवाद
شكراً
ধন্যবাদ
תודה
Copyright Notice

This multimedia file is copyright © 2022 by tinyML Foundation. All rights reserved. It may not be duplicated or distributed in any form without prior written approval.

tinyML® is a registered trademark of the tinyML Foundation.

www.tinyml.org
Copyright Notice

This presentation in this publication was presented as a tinyML® Talks webcast. The content reflects the opinion of the author(s) and their respective companies. 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