Integrate and deploy machine learning at the edge with embedded software containers

Semir Haddad – Chief Product and Strategy Officer, MicroEJ

January 17, 2023
Thank you, tinyML Strategic Partners, for committing to take tinyML to the next Level, together
Executive Strategic Partners
The Leading Development Platform for Edge ML

dgeimpulse.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

Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.
Platinum Strategic Partners
Fastest Video Analytics Solutions on Arm CPUs
Renesas is enabling the next generation of AI-powered solutions that will revolutionize every industry sector.
Sony Semiconductor Solutions Corporation
Where what if becomes what is.

Witness potential made possible at analog.com.
Arm AI Virtual Tech Talks

The latest in AI trends, technologies & best practices from Arm and our Ecosystem Partners.

Demos, code examples, workshops, panel sessions and much more!

Fortnightly Tuesday @ 4pm GMT/8am PT

Find out more: www.arm.com/techtalks
Making Over-the-Air Firmware and ML models Updates Simple and Accessible!

- Securely update your IoT devices regardless of their Hardware Platform (Silicon) Provider and physical location.
- Unlock TinyML business value through OTA Firmware and ML models update.
- Pay-as you-go

www.fotahub.com
contact@fotahub.com
TOGETHER, WE ACCELERATE THE BREAKTHROUGHS THAT ADVANCE OUR WORLD

www.nxp.com/ai
The Right Edge AI Tools Can Make or Break Your Next Smart IoT Product

Analytics Toolkit Suite

AutoML

Data Collection

Test & Validation

Code Generation

Code Editing

Version Control and Model Management

Team Collaboration

sensiml.com/tinyML
STMicroelectronics provides extensive solutions to make tiny Machine Learning easy
ENGINEERING
EXCEPTIONAL EXPERIENCES

We engineer exceptional experiences for consumers in the home, at work, in the car, or on the go.

www.synaptics.com
Silver Strategic Partners

- Arduino
- Azip
- AONdevices
- Greenwaves Technologies
- Grovety Inc.
- IBM
- Imagimob
- Itemis
- Nota AI
- OctoML
- Prophesee
- Qeexo
- Reexen Technology
- SAP
- Schneider Electric
- Silicon Labs
- Stream Analyze
- TDK InvenSense
Join Growing tinyML Communities:

**Meetup**

[ tinyML - Enabling ultra-low Power ML at the Edge

12.7k members in
46 Groups in 37 Countries

---

**LinkedIn**

[ The tinyML Community
https://www.linkedin.com/groups/13694488/](https://www.linkedin.com/groups/13694488/)

3.3k members
&
10.7k followers
Subscribe to tinyML YouTube Channel for updates and notifications (including this video)
www.youtube.com/tinyML
Reminders

Slides & Videos will be posted tomorrow

tinyml.org/forums  youtube.com/tinyml

Please use the Q&A window for your questions
Our next tinyML Trailblazers Series

Success Stories with Luca Verre
(Co-founder and CEO, Prophesee)

LIVE ONLINE January 25th, 2023 at 7:30 am PDT

Register now!
The tinyML Research Symposium 2023 will be held in conjunction with the tinyML Summit. The Research Symposium is the premier annual gathering of senior level technical experts and decision makers representing fast growing global tinyML community. The Call for Papers is now open.
tinyML EMEA Innovation Forum 2023

Connect, Unify, and Grow the tinyML EMEA Community
June 26 - 28, 2023

https://www.tinyml.org/event/

Event will be held in person in Amsterdam, Netherlands

EMEA 2023 Call for Presentations will open in January, 2023.

More sponsorships are available: sponsorships@tinyML.org
Semir Haddad

Semir is the Chief Product and Strategy Officer at MicroEJ, and is in charge of product strategy and ecosystem, for the company that makes everything software defined. Semir has over 20 years of experience working with industry leaders and startups, bringing innovative technologies to industrial and consumer markets. Early in his career, he worked on digital video and audio, creating the first DVD player, and developing one of the earliest PVR software solutions for which he received four patents. At ST and Renesas he drove the creation of award winning lines of microcontrollers and microprocessors, including the STM32, RX and RZ. Recently, Semir was part of the innovators at Eta Compute that created one of the first TinyML implementation. Semir graduated from CentralSupelec with a Master of Science in Electrical and Electronics Engineering and from ESSEC/Mannheim with an MBA.
Integrate and deploy machine learning at the edge with embedded software containers

AGENDA

1. TinyML for product owners
   1. TinyML is cool, how do I use it in my product?
   2. Integration and deployment challenges

2. Devops solutions: Containers
   1. Concept and Origin
   2. Cloud-Native Made Edge-Compliant

3. Containers for Edge ML
   1. Concept and Implementation
   2. AuZone-MicroEJ Demo

4. Q&A

© MICROEJ 2022
Semir Haddad
Chief Product officer
MicroEJ
San Francisco, CA
semir.haddad@microej.com

Semir Haddad
Chief Product officer
MicroEJ
San Francisco, CA
semir.haddad@microej.com
THE TINY CONTAINER FOR THE EMBEDDED WORLD

MICROEJ ENABLES THE DIGITAL TRANSFORMATION

IP provider & runtime licensing
- User Interface
- Edge computing and AI
- Device-to-Cloud
- Device-to-devices

$40M R&D total investment since inception

US, France, Germany, Romania, China, Japan, Korea

+150M units MicroEJ IP sold
THE DEMAND FOR SMART DEVICES IS EXPLODING

2011 - Marc Andreessen  
Software Is Eating the World

2015 - Klaus Schwab  
The Fourth Industrial Revolution

2017 - 3GPP  
Release of the 5G standard

2022 - McKinsey  
The pandemic has sped up the adoption of digital technologies by several years

5G to cover 60% of the world by 2026

75B+ connected devices by 2030

Smart home growth 20% CAGR over '21-'25

~60% electrified vehicles by 2030

---

1. MicroEJ Strategy Office, Omdia, Ericsson Mobility Report, IDTechEx, ABI Research, Strategy Analytics
TINYML AND DISTRIBUTED INTELLIGENCE HAVE MANY ADVANTAGES

1. Reducing the cost of data transfers and data storage as most data is treated locally

1. Reducing the energy consumption and carbon footprint of the system

1. Improving service latency, as the local device is able to react to inputs in real-time

1. Increased resilience and availability as devices are more independent from the cloud and less impacted by network outages, connection losses, etc.
TINYML IS COOL – BUT WHERE DO WE NEED IT?

Let’s talk about devices (or “EDGE NODES”, or “THINGS”)
TINYML NEXT CHALLENGE: INTEGRATION AND DEPLOYMENT

Smart Cooker with various sensors
- Adjust cooking vs food
- Adjust cooking vs user (learn user’s preferences)
- Advanced cooking cycles
- etc… (you name it)

BUT… The cooker must still perform its main task
- Cook (manage heating element, mixing element (motor)
- Display information and user interface (display, button, sound)
- Connect to the cloud (status, recipe)

TinyML must integrate within the existing device
TinyML must be easy to deploy and update
MACHINE LEARNING DEVELOPMENT FLOW

- Data Collection from board or other methods
- Data Collection for training
- Data management Collection for training
- Data analysis/Feature engineering
- Train
- Model Design
- Pruning / Quantize / Compress
- Quantize and optimize
- Validate
- Release
- NNCompilation Acceleration for edge device
- Test on real device
- Results & Measurement
- Test board

At the end of the process:

Trained neural network for the edge

NN network

1/17/2023
WHAT HAPPENS NEXT?

- Mass Storage
- Ethernet Wi-Fi/LTE
- Display
- UART
- Bluetooth

HARDWARE

SOFTWARE

Mass Storage Ethernet Wi-Fi/LTE Display UART Bluetooth

Application

High-level libraries: Web / REST servers | MQTT / LWM2M clients | JSON | CBOR | Crypto | Widgets …

File System Internet Protocol Graphical Engine Sensors I/O Bluetooth

BSP

Drivers

RTOS / Linux / QNX

PROCESSOR CORE

NN network

NN runtime

Drivers

Graphical Engine

RTOS / Linux / QNX

Bluetooth

I/O

Sensors

Drivers

BSP

File System Internet Protocol Graphical Engine Sensors I/O Bluetooth

Drivers

Application

High-level libraries: Web / REST servers | MQTT / LWM2M clients | JSON | CBOR | Crypto | Widgets …
THEN THE REAL HARD WORK STARTS

• **The matrix of pain:** Hundred variants of processors, tools, software components

• **The impossible dilemma:** mixing quality, reliability, real-time with useability, flexibility and connectivity

• **Multivariate constrained optimization:** cost, power, memory, compute, connection bandwidth

• **Talent shortage:** limited talent pool for embedded + application + ML
NEURAL NETWORKS ARE SOFTWARE

Code

```
model = Sequential()
model.add(Conv2D(32, (3, 3), padding='same',
    input_shape=x_train.shape[1:])))
model.add(Activation('relu'))
model.add(Conv2D(32, (3, 3)))
model.add(Activation('relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))
model.add(Dropout(0.25))
model.add(Conv2D(64, (3, 3), padding='same'))
model.add(Activation('relu'))
model.add(Conv2D(64, (3, 3)))
model.add(Activation('relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))
model.add(Dropout(0.25))
model.add(flatten())
model.add(Dense(512))
model.add(Activation('relu'))
model.add(Dropout(0.5))
model.add(Dense(num_classes))
model.add(Activation('softmax'))
```

Data

```
0x00, 0x30, 0x00, 0x30, 0x30, 0x30, 0x30, 0x03, 0x30, 0x30, 0x30, 0x31, 0x30, 0x30, 0x30, 0x31, 0x30, 0x30,
0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30,
0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30,
0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30, 0x30,
```

= Result

```
int x = 20;
int y = 30;
int yOffset = 150;
g.setColor(Colors.LIME);
VectorGraphicsPainter.drawString(g, "Hello MicroE!", font0, 20, x, y);
VectorGraphicsPainter.drawString(g, "Hello MicroE!", font1, 20, x + yOffset, y);
```

```
VectorGraphicsPainter.drawString(g, "Hello MicroE!", font0, 40, x, y);
VectorGraphicsPainter.drawString(g, "Hello MicroE!", font1, 40, x + y, y + yOffset);
```

```
VectorGraphicsPainter.drawString(g, "Hello MicroE!", font0, 60, x, y);
VectorGraphicsPainter.drawString(g, "Hello MicroE!", font1, 60, x + y, y + yOffset);
```

```
g.setColor(Colors.YELLOW);
y == 40;
VectorGraphicsPainter.drawString(g, "Hello MicroE!", font0, 80, x, y);
VectorGraphicsPainter.drawString(g, "Hello MicroE!", font1, 80, x + y, y + yOffset);
display.flush();
```

```
```
VectorFont font0 = VectorFont.loadFont("/fonts/Arial.ttf");
VectorFont font1 = VectorFont.loadFont("/fonts/BAVIE.ttf");
```

= Result
NEURAL NETWORKS ARE SOFTWARE

TENSORFLOW-LITE MICRO INTERPRETER

Trained Models -> TF-Lite model -> TF-Lite executable

TF .pb
Keras .h5
Trained Models

TF-Lite Converter (FlatBuffer)
Post training quantization

TF-Lite model

Convert to C array

TensorFlow library
tensorflow/tensorflow/lite/micro/

Model_data.c

Application Code

tflite::MicroInterpreter interpreter(model, resolver,
tensor_arena, tensor_arena_size,
error_reporter);
interpreter.AllocateTensors();
...
Getinput (interpreter.input, input)
invoke_status = interpreter.Invoke();
output = interpreter.output(0);
...
THEN THE REAL HARD WORK STARTS (THE SEQUEL)

Developers need to maintain and update their neural network along with the device software

- Integration in the DEVOPS process – Configuration management and CI/CD
- Life cycle management with secure and safe update in the field
“Even in the cloud, machine learning projects require you to manage the compatibility and complexities of an ever-evolving software stack, which can be frustrating, time-consuming.” Google cloud

Google Cloud solve the issue with **Deep Learning Containers**
Microsoft Azure with **Azure Kubernetes Service**

“Containers are packages of software that contain all of the necessary elements to run in any environment. In this way, containers virtualize the operating system and run anywhere, from a private data center to the public cloud or even on a developer’s personal laptop.” Google cloud

**Advantages:**
- Portability and consistency to move from on-premises to cloud scale.
- Container encapsulate all the resources needed to run the neural network
- Containers allow to deploy multiple instances and microservices, update and manage models
- Fully automated MLOps
CONTAINER FOR TINYML

WHAT WORKS FOR THE CLOUD

- Cloud-native development
- Microservices
- Docker and Kubernetes
- Agile
- DevOps
- Virtualization

HOW TO DO IT IN A DEVICE?

Each device has Linux, Kubernetes and Docker
Expensive, Wastes lots of energy, Unpractical

Or

Start from the device

Build a standard low-power container technology for the edge

- Containers and containerized apps with app orchestration, and bringing microservices to the devices.

- Virtualization to insure portability and best use of any hardware platform

This endeavor was attempted several times by various actors and was achieved successfully by MicroEJ with the MICROEJ VEE virtual execution environment.
BENEFITS OF CONTAINERS FOR TINYML

1. Portability and consistency to move from workstation to device
   Fast deployment, simulation and validation on workstation

1. Container encapsulate all the resources needed to run the neural network: portability
   Abstraction of physical resources, portable across many processors

1. Containers allow to deploy multiple instances and neural networks, update and manage models
   Evolutive products: maintain agility and flexibility, from product development to post-release upgrades

1. Fully automated DevOps/ MLOps
   ML integrated in CI/CD process
BUILD SOFTWARE ASSETS

STANDARD SOFTWARE CONTAINER

- Software assets for reuse
  - Independent application code
  - Secure by design
  - Virtual device
  - Fast ideas to prototype
  - Device-to-cloud holistic system
  - Fast redesign

Virtual Device on PC

Real Device 1

Real Device 2

Real Device 3

Software components / assets

Cloud

App and Service
**EDGE AI FLOW**

Event, detection
Classification

Cloud

New training data, Anomalies

Updated NN remotely (without rebooting the device)

1. MicroEJ VEE secure Core
2. Containerized NN

**Model Design and Training**

- Updated NN remotely
- New training data, Anomalies
- Event, detection Classification

**HARDWARE**

- MicroEJ
- Trained Neural Network
- NN File

**EDGE AI FLOW**

- Edge AI Flow Diagram
- Model Design and Training
- Updated NN remotely (without rebooting the device)

1/17/2023
EDGE AI INTEGRATION

Trained model

AI runtime integrated in MicroEJ
VIRTUAL DEVICE WITH AI

Better Virtual device slide

PC running Windows
DESIGN NN ONCE, RUN NN EVERY WHERE

ST F746 : https://developer.microej.com/supported-hardware/stmicroelectronics-stm32f7/
NXP RT595 : https://developer.microej.com/supported-hardware/nxp-imx-rt595/
Virtual Device : https://developer.microej.com/virtual-devices/
SAFETY & SECURITY BY DESIGN

Formal Specification
http://www.e-s-r.net/download/specification/ESR-SPE-0020-KF-1.4-F.pdf

- Set of API & resources available to Apps
- Can’t access directly to code, objects, threads
- No (bypassing) native code call allowed
- No memory inter-dependencies between Apps
- Fully managed in binary
  Install, uninstall, start, stop

BINARY COMPATIBILITY CHECK

Statically linked
System built-in

Dynamically linked
Downloaded

Ecosystem API (= the Kernel)

- Set of API & resources available to Apps
- Can’t access directly to code, objects, threads
- No (bypassing) native code call allowed
- No memory inter-dependencies between Apps
- Fully managed in binary
  Install, uninstall, start, stop
Secure Isolation Layer

APP binary code Verifier

Trusted code

Runtime checks

Granted Access

Unrestricted access and editing

User data needed by the App

System resources needed by the App

Sandbox

MicroEJ VEE secure Core

HARDWARE

Malicious Code

Compilation Src → Bin

Binary App

APP Source code

Static validation of the BINARY correctness

Trusted code CANNOT take control of the MICROEJ VEE secure core, nor access C/ASM resources without being granted explicitly (native calls) - only such Proven technology since 1998

The MICROEJ VEE takes advantage of all software and hardware secure mechanism, around keys, unique IDs, random generation, etc. Means to enhance software IP protection, steganography to hide data, etc.
Easily implementable on secure KF semantic
Define the runtime semantic
  - Registry of shared runtime services (=objects) created by the APP, life cycle of APPs
  - One global security manager (=object) owned by the Kernel to grants rights access to Services
DEMO WITH AU-ZONE

MicroEJ VEE on the NXP RT1170 EVK

Face detection algorithm running on DeepView runtime inside MicroEJ VEE
DEMO WITH AU-ZONE

Event, detection Classification

Cloud

New training data, Anomalies

Updated NN remotely (without rebooting the device)

Model Design and Training with DeepView™

Model Design and Training with DeepView™

DeepView™ Bring Your Own Data Workflow

DeepView™ Bring Your Own Model Workflow

PyTorch ONNX

Public or Proprietary Model

Model Conversion, Optimization, Quantization

On Target Profiling and Production

1 2 NN

MicroEJ VEE secure Core NN API

HARDWARE

Containerized NN

MicroEJ

Trained Neural Network

AuZone NN File

1/17/2023
AU-ZONE INTEGRATION

AU-ZONE trained model

Au-Zone Vision Pack and Model Pack
Au-Zone Services

APPLICATIONS for MICROEJ VEE

ADD-ON LIBRARIES
Web / REST servers | MQTT / LWM2M clients | JSON | CBOR | Crypto | Widgets | Components | Eclasspath |

FOUNDATION LIBRARIES
FS | NET/SSL | MicroUI | Android core | ECOM | BLE |

BAF
AI Library
NN API

SANDBOXED APPLICATIONS
Managed Code (Java, JavaScript, ...)

VIRTUALIZATION
BSP
Drivers

HARDWARE
AI Hardware support | Mass Storage | Ethernet Wi-Fi / LTE | Display / GPU | PROCESSOR CORE | UART | Bluetooth |

1/17/2023
DESIGN – INTEGRATE - DEPLOY

SANDBOXED APPLICATIONS

APPLICATIONS for MICROEJ VEE

AI Library
- Web / REST servers
- MQTT / LWM2M clients
- JSON
- CBOR
- Cryptography
- Components
- EclipsePath

FOUNDATIONS LIBRARIES
- NN API
- FS
- NET/SSL
- MicroUI
- Android core
- ECOM
- BLE

ADD-ON LIBRARIES
- Structured - "Add" libraries - Web / REST servers
- Structured - "Add" libraries - MQTT / LWM2M clients
- Structured - "Add" libraries - JSON
- Structured - "Add" libraries - CBOR
- Structured - "Add" libraries - Cryptography
- Structured - "Add" libraries - Components
- Structured - "Add" libraries - EclipsePath

VIRTUALIZATION

ABSTRACTION LAYERS
- MEJ 32
- File System
- Internet Protocol
- Graphical Engine
- I/O
- Bluetooth

HARDWARE
- AI Hardware support
- Mass Storage
- Ethernet / Wi-Fi / LTE
- Display / GPU
- PROCESSOR CORE
- UART
- Bluetooth

1/17/2023
Semir Haddad
semir.haddad@microej.com
www.microej.com
This multimedia file is copyright © 2023 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