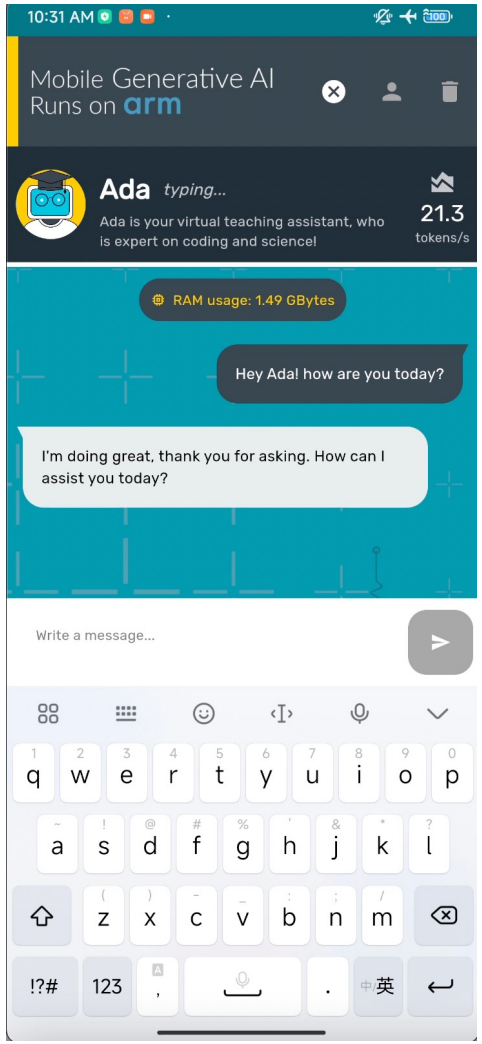
What’s so unique about tinyML?

Keynote

Gian Marco Iodice, GenAI lead @arm
06/05/2024

Generative AI



LLM: Phi2 2.7B

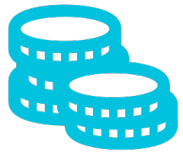
CPU: 4x Arm Cortex-A715 CPUs

Why generative AI? Isn't this talk about tinyML?

https://community.arm.com/arm-community-blogs/b/ai-and-ml-blog/posts/generative-ai-on-mobile-on-arm-cpu

tinyML

Bring intelligence to objects around us with a focus on power consumption, data privacy, and cost



But…..why?

To easily scale applications powered by ML

- For example:
  - It does not need cables to operate
  - It does not need an internet connection.

Ingredients

TinyML is the set of technologies in ML and embedded systems to make use of smart applications on low-power devices. Generally, these devices have limited memory and computational capabilities, but they can sense the physical environment through sensors and act based on the decisions taken by ML algorithms*.

Ingredients

Characteristics of low-power devices

System input

Level of computing on top of sensors that allows smartness in a minimally intrusive way

*TinyML Cookbook, Gian Marco Iodice – April 2022

arm

Applications

TinyML finds its natural home wherever a power supply from the mains is impossible or complex to have, and the application must operate with a battery for as long as possible

Applications

Battery-powered solutions are not limited to consumer electronics only…

There are scenarios where we might need devices to monitor environments. For example, we may consider deploying battery-powered devices running ML in a forest to detect fires and prevent fires from spreading over a large area

Microcontrollers

Microcontrollers are suitable devices for tinyML

- Designed to be power-efficient
- Already popular in many industries (e.g., automotive, consumer electronics, healthcare,..).
  - 31.2 billion devices sold worldwide in 2021!
- Low-cost
- Easy to program
- …

My first encounter with tinyML

“If walls could talk!”

First lesson of Ambient Intelligence, University of Pisa - 2013

2013

Software development

- Microcontrollers were difficult to program
  - Assembly and C were the main programming languages
- IDEs did not offer a plug-and-play experience.
- We could not port the code easily among different platforms

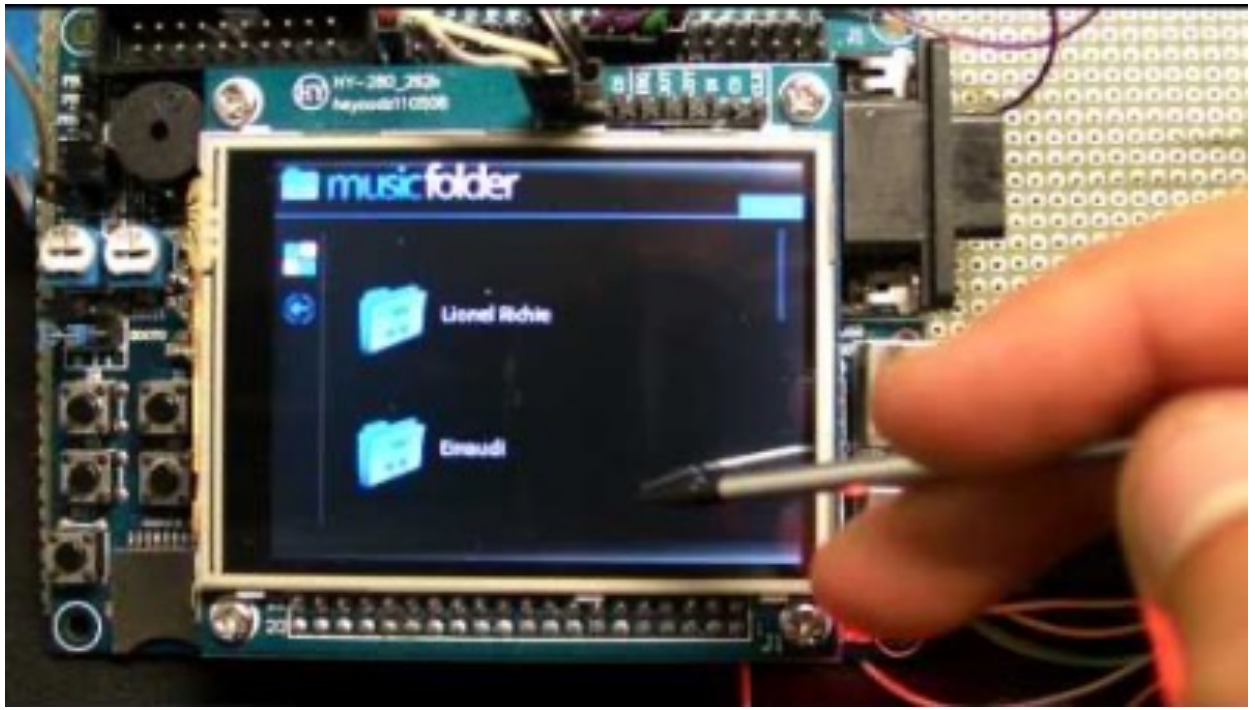
ML

- Very few free and open-source frameworks to develop ML models
- No end-to-end tools that can make ML design straightforward

But…what about the hardware capabilities?

arm

Personal project in 2013



https://www.youtube.com/watch?v=LXm6-LuMmUU

Personal project to build an MP3 player **from scratch in C** on an Arm Cortex-M3 microcontroller (STM32F1x)

- Program memory: 512 KBytes
- SRAM: 64 Kbytes
- Songs stored on SD card
- Responsive touchscreen user interface (UI)
- Everything runs on the Arm Cortex-M CPU without O.S.

At that time, there weren't SW libraries for the UI!

Today: From few to many

Software and open-source contributions were the keys to scaling tinyML

Software development

- Microcontrollers are now easy to program
  - Also, with Python!
- IDEs offer a plug-and-play experience.
- Code can be easily ported among different devices

ML

- Many free, open-source, and easy-to-use frameworks to develop ML models
- No end-to-end tools that can make ML design straightforward

arm

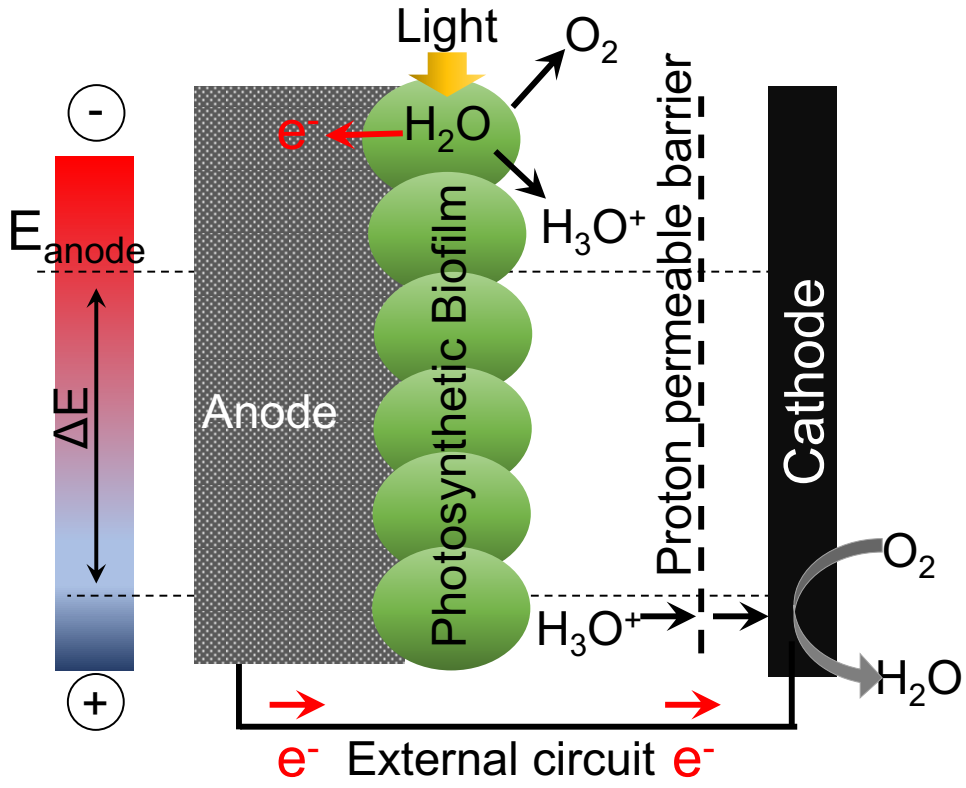
When the technology scales, there are often huge opportunities to solve real-world big problems

Part 1:

Collaboration with Dr. Paolo Bombelli, University of Cambridge

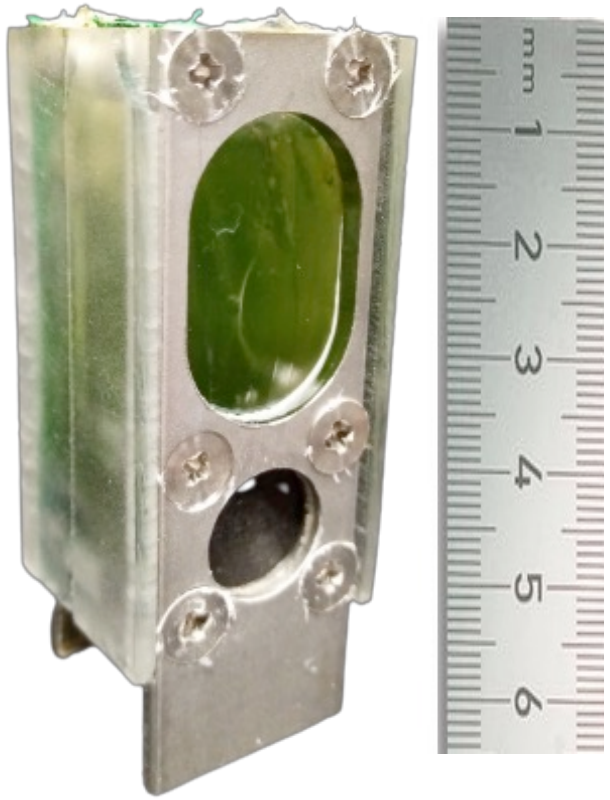
Can we generate electricity from photosynthesis?

Photosynthetic microorganisms (e.g., micro-algae) are able to generate electrons that can be harvested by a suitable electrochemical setup and be used as a source of electrical current. This concept forms the basis of Bio Photo Voltaic (BPV) devices[1,2]. The main components forming a BPV are shown in the diagram below.

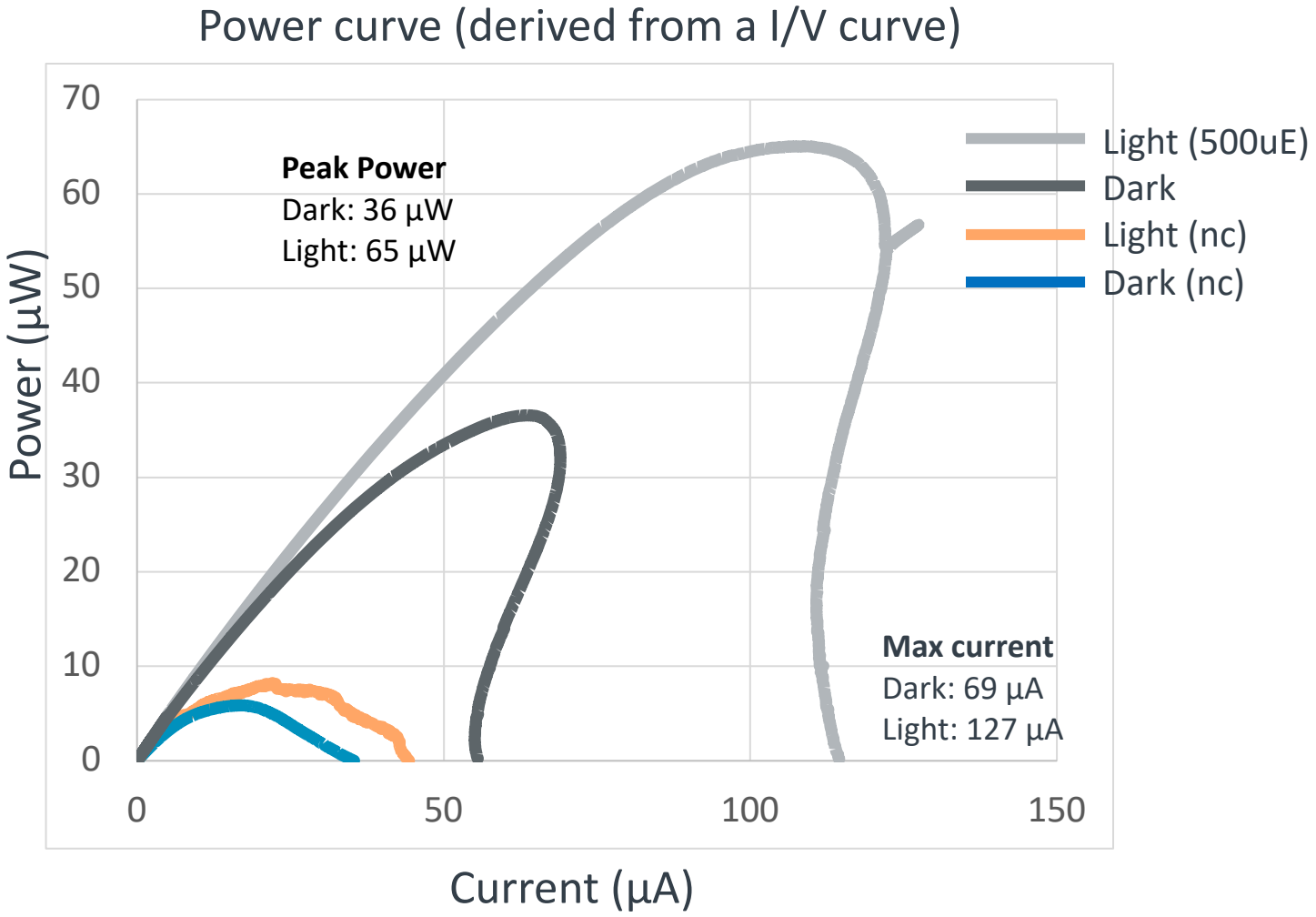


[1] McCormick *et al*., (2015), *Energy & Environmental Science* 8 (4), 1092-1109.
[2] Howe and Bombelli (2020) Joule 4 (10), 2065-2069

How much current can we generate?



Active volume (anode): 16.4 cm^3



Volumetric power density: dark: ~2 µW / cm^3 ; light: ~4 µW / cm^3

Bombelli *et al*. (2022) *Energy & Envtl Sci*.

Domain of application

- Power consumption +

application today

application tomorrow

Micro and small electronic devices

Small domestic appliances

Large domestic appliances and industrial electric devices

µWs **mWs** **Ws** **kW** **kWs**

Powered by disposable and/or removable batteries

Powered by built-in butteries charged through the national grid

Powered by national grid

Powering tinyML with algae

- Monitoring environments through many sensors
- Rural areas where a small amount of power might be beneficial to power environmental sensors.
- Provide energy for longer than traditional batteries
  - It can last years!

2023 Project

Monitoring weather conditions and water quality with tinyML in the river Thames (London) using native algae of the river.

How

By powering the first ML application with algae on an Arm-based microcontroller

Goal

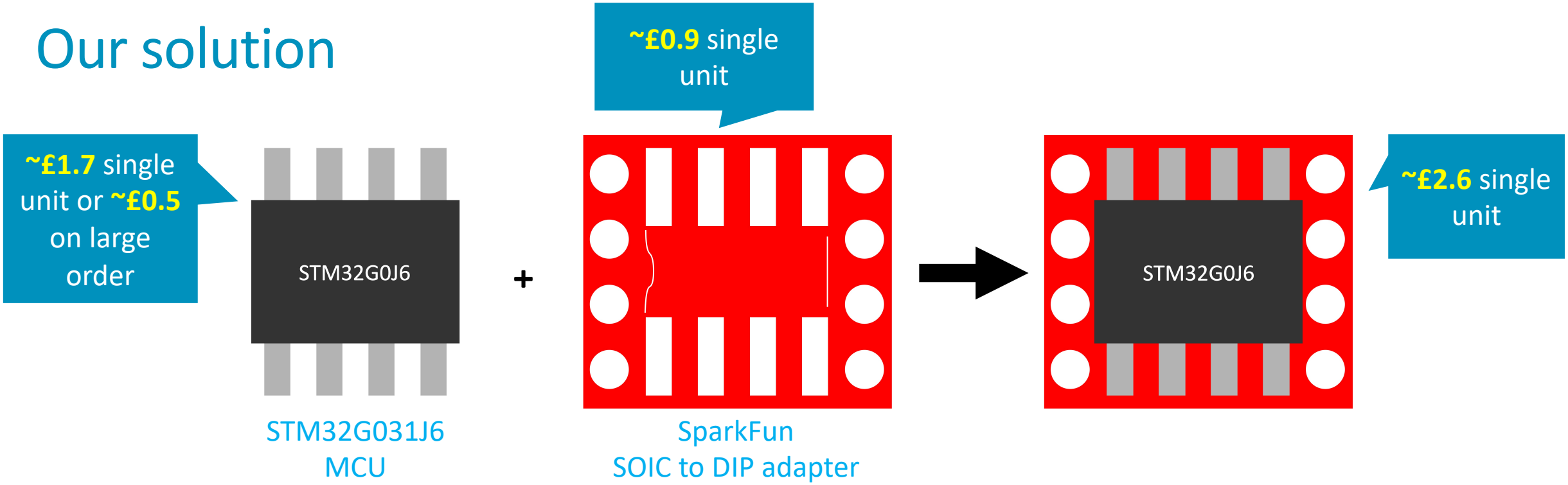
Demonstrating that people, regardless of their expertise, have the power to address environmental challenges using today's technology in a sustainable and affordable manner.

ML deployment framework/library selection considerations

- It depends on:
  - The target device (microcontroller)
  - The features of the target device (on-board memory)

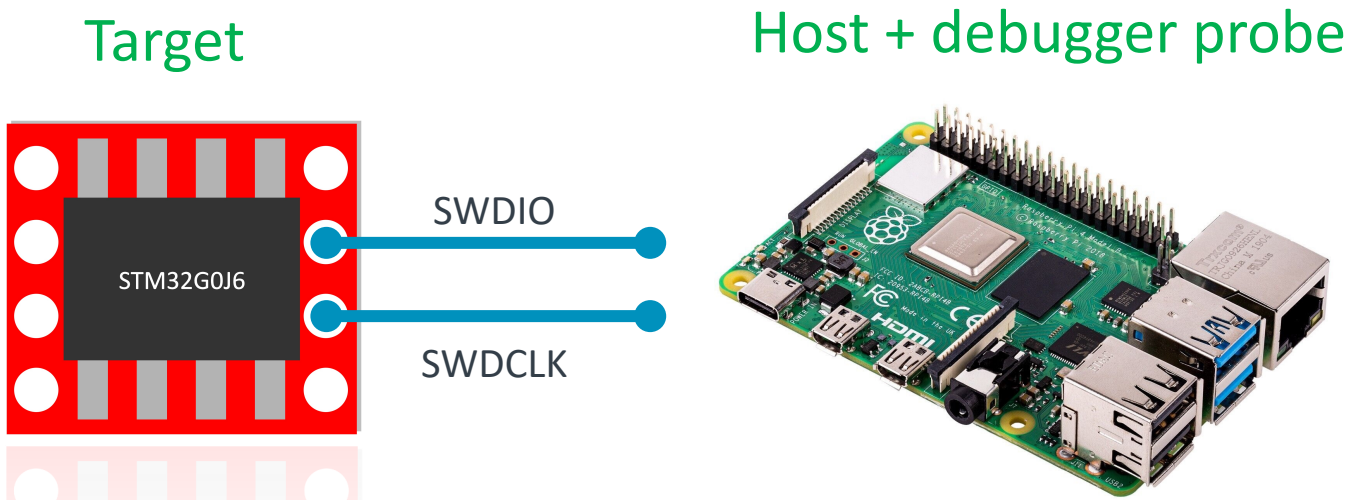
- There are many great tools (and often free and/or open source) for microcontrollers.
  - However, we should consider the feasibility of fitting the ML model onto the chosen microcontroller.
  - Frameworks providing ahead-of-time (AOT) capabilities may be preferred to reduce program memory utilization.
    - The model should NOT be loaded at runtime

Our solution



Feature	
CPU	Arm Cortex-M0+ @16 MHz
Operating voltage - Power	2 - 3.6V - 93uA/MHz
FLASH - SRAM	32 Kbytes – 8 KBytes
Peripherals	I2C, SPI, UART, USART, Timers, DMA, RTC, ADC

Programming and debugging with Raspberry Pi



- We can use the SWD IO pins on Raspberry Pi to communicate with the microcontroller
- We need to install OpenOCD and GDB to use the Raspberry Pi as debugger probe
- Raspberry Pi can be used as a standard computer for programming the microcontroller.

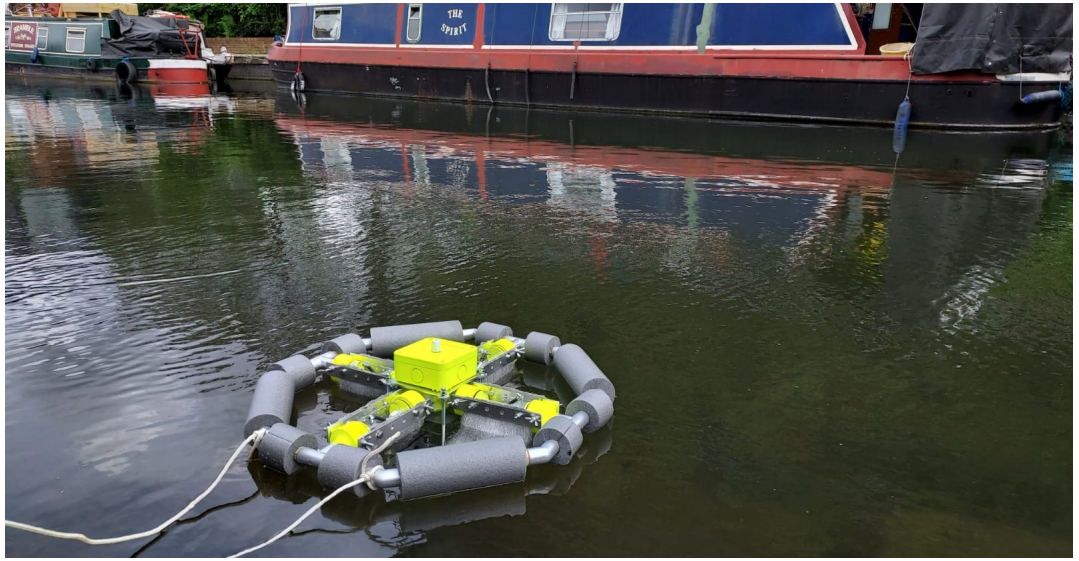
Designing and deploying the first ML model powered by algae!

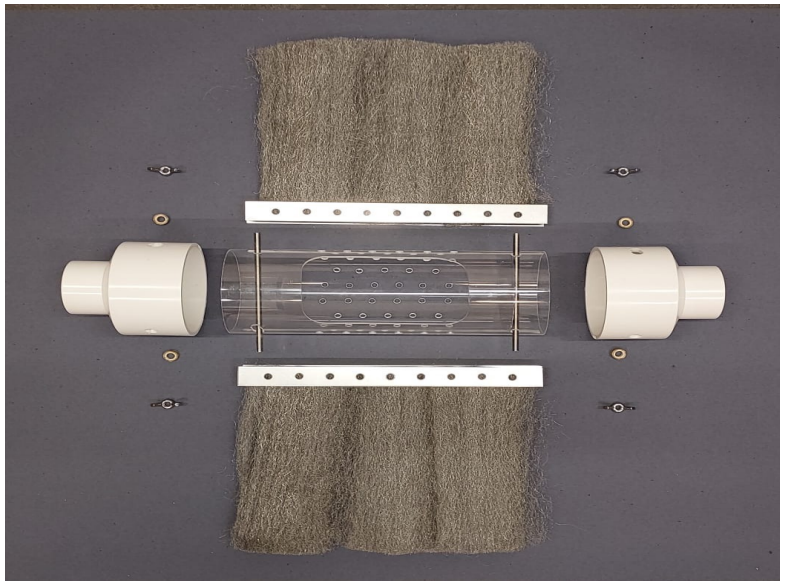
- As first model powered by algae, we have considered the sine wave model.
  - This model represents the “Hello, World” for the tinyML community.
  - We express our deepest gratitude to Pete Warden and Daniel Situnayake for their contribution.
- The model was trained and quantized (8-bit) directly in Google Colaboratory
- The tinyML application was written in C++
  - The application has been compiled and uploaded from the Raspberry Pi
- The model has been accelerated on the microcontroller using CMSIS-NN

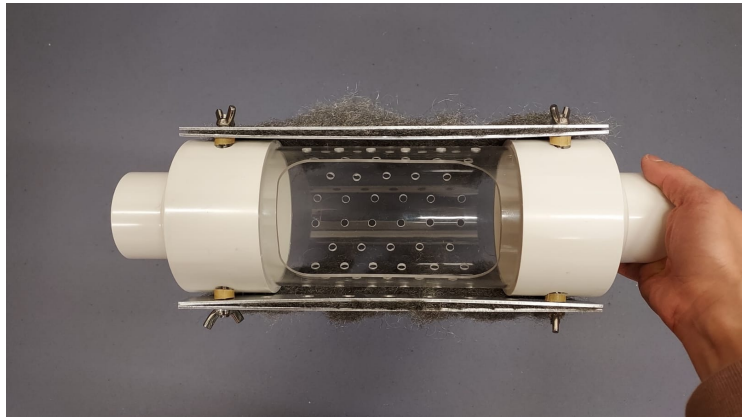
Stats	Values
Model params	3 Fully Connected layers with a total of 321 params
Program memory	10 Kbytes
SRAM	TBD
Current consumption	1.8 mA

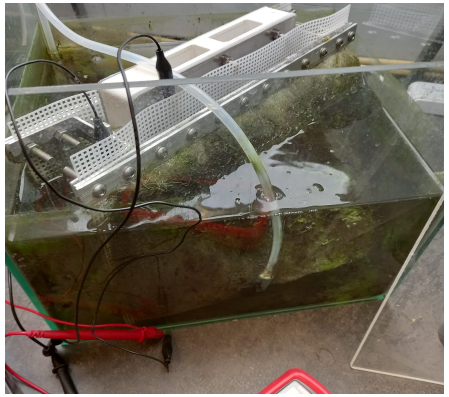
arm

Powering ML with algae









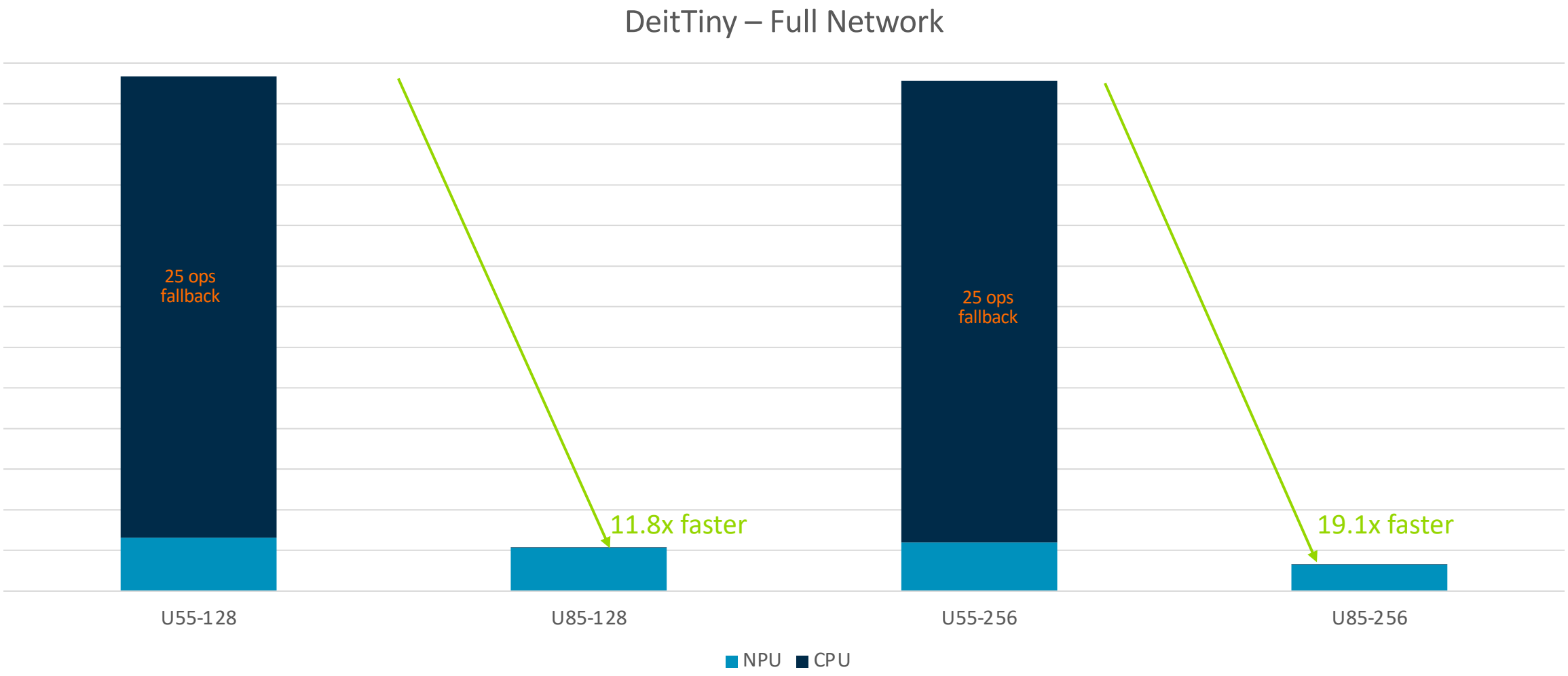
Unlocking new use cases

Microcontrollers are more compute capable and more power efficient than 10 years

- Now, they can be equipped with Neural Processing Units (NPUs) for efficient AI solutions:
  - For example, Arm Ethos-U65*

*https://www.arm.com/products/silicon-ip-cpu/ethos/ethos-u65

Performance Report



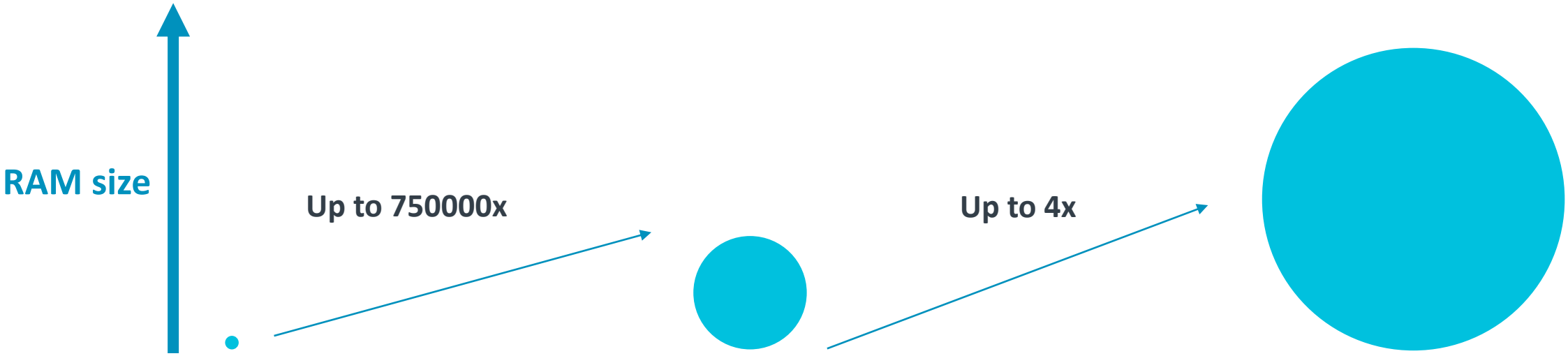
All numbers measured on FPGA with pre-EAC RTL and early software support

Unlocking new use cases

However, the pace of memory storage development has been steady for the last 10 years.

- Therefore, you might find microcontrollers with the same Flash and SRAM storage size as 10 years ago, although more power-efficient

RAM size (Typical range)



Microcontrollers
[16-256KBytes]

Mobile phones
[8-12GBytes]

Laptops
[16-32GBytes]

From large to tiny

Models tailored for memory-constrained devices

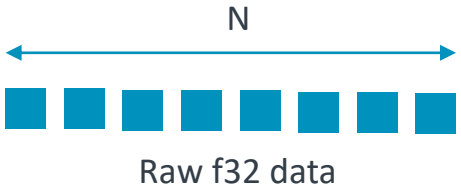
- MobileNet/EfficientNet for image image classification
- Faster Objects More Objects (FOMO) for image detection

Compression techniques and lower bit precision were the keys to bringing sophisticated ML models to microcontrollers.

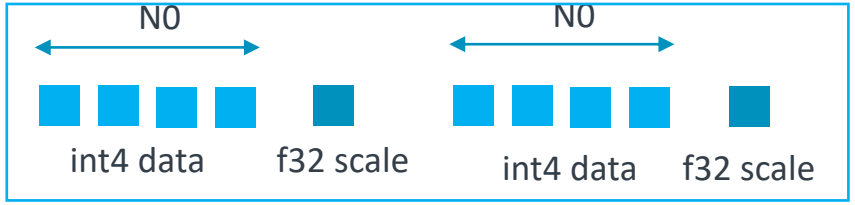
- Clustering and pruning
- Knowledge distillation
- 8-bit quantization and lower-bit precision

4-bit quantization

- Many formats. For example:



Per-channel



Per-block

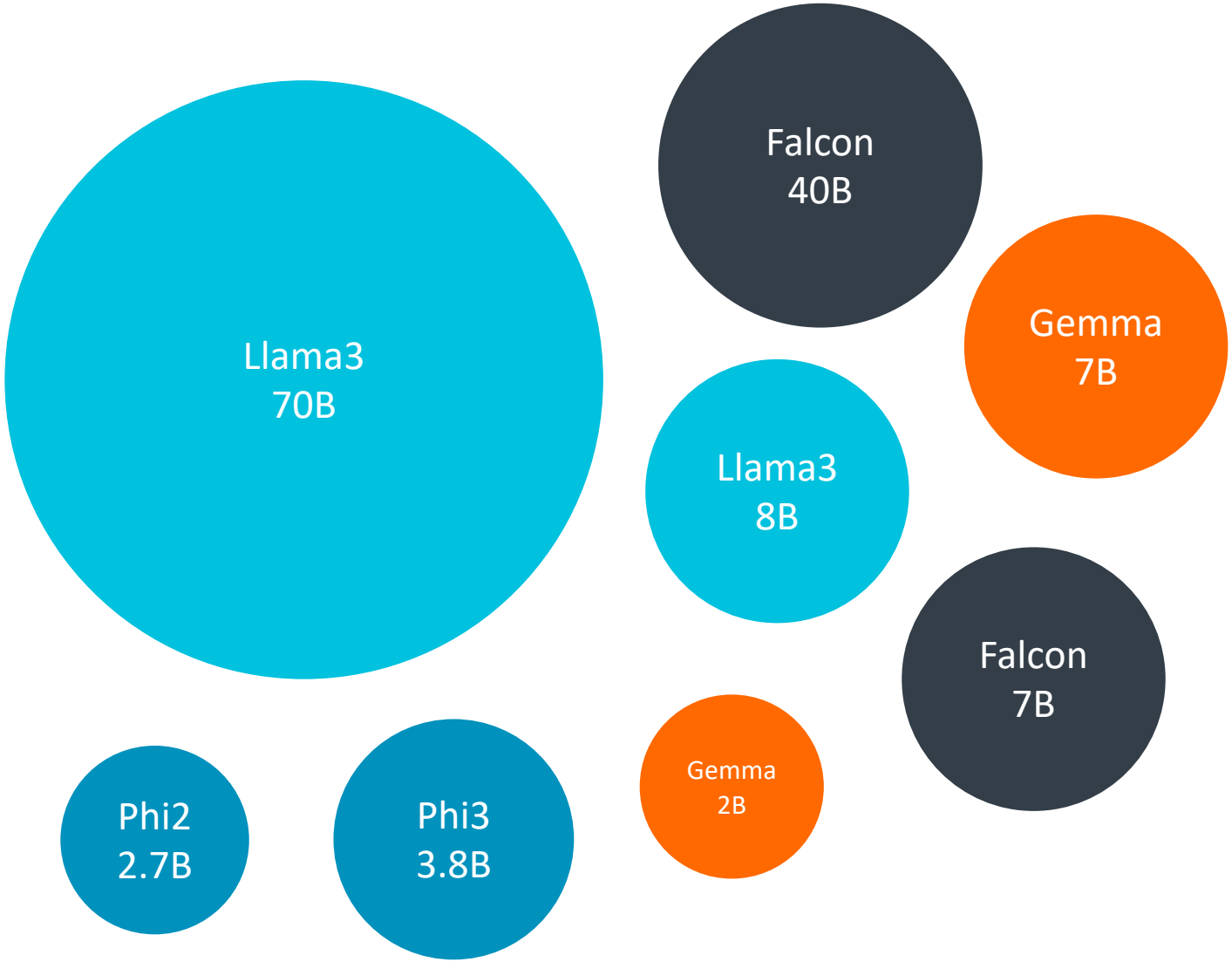
The generative AI era

OpenAI launched ChatGPT in November 2022

Models are very sophisticated but extremely large.

At the moment, LLMs are not for microcontrollers.

Some open LLMs



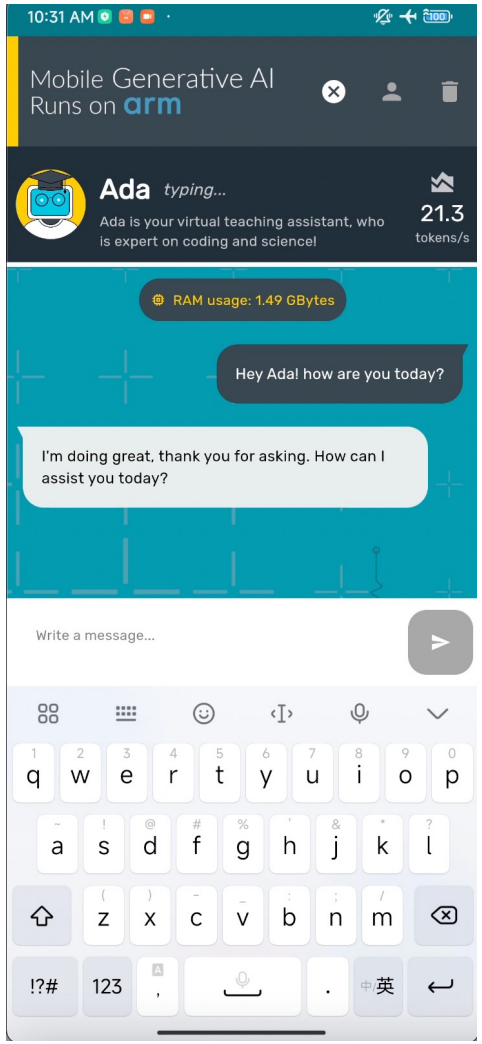
Estimated memory usage with Fp16:

- 70B -> 140 Gbytes
- 40B -> 80 Gbytes
- 7B -> 14 Gbytes
- 3.8B -> 7.6 GBytes
- 2.7B -> 5.4 Gbytes
- 2B -> 4 GBytes

arm

The LLM model deployment has the same type of constraint as tinyML. The memory.

Exploiting tinyML techniques to accelerate LLMs at the edge



LLM: Phi2 2.7B

CPU: 4x Arm Cortex-A715 CPUs

Estimated memory usage with int4:

- 70B -> 35 Gbytes
- 40B -> 20 Gbytes
- 7B -> 3.5 Gbytes
- 3.8B -> 1.9 GBytes
- 2.7B -> 1.4 Gbytes
- 2B -> 1 GByte

Suitable models for mobile and edge deployment

But…..why?

To easily scale applications powered by LLM

arm

When the technology scales, there are often huge opportunities to solve real-world big problems

Part 2

Collaboration with Unicef

No one is left behind

Definition

*The Global Goals are an agreed-upon set of goals, targets, and indicators that all member states of the UN agreed to in 2015 with the goal of ending poverty, fighting inequality, and urgently addressing the climate crisis

How are designed

*…designed to provide **lifelong education for all**.

This is the story we are going to tell you today

*https://www.unicef.ca/en/global-goals-sustainable-development-for-every-child%E2%80%99s-future

Ed Tech for Good Curation Framework

- We are seeing today a plethora of educational technology (EdTech) tools emerging:
  - Adaptive learning programs for math, gamified language learning apps, and robotics kits powered by **tinyML**!

- However, how can governments, organizations, schools, teachers, and parents know which tool better suits their needs?

- Unicef Learning Innovation Hub-Office of Innovation with the Asian Development Bank (ADB) has been developing, along with partners such as Arm, the EdTech for Good Curation Framework.

- In August 2023, we met in Helsinki to shape the framework's pillars together.

A scalable learning experience

How can we help kids learn new technologies or programming languages in a scalable way in the global south?

A scalable learning experience

- Consider the following example:

Let’s assume we provide a fancy, cool robot in the Global South to teach the foundations of technology.

- Would it be enough to teach technology?

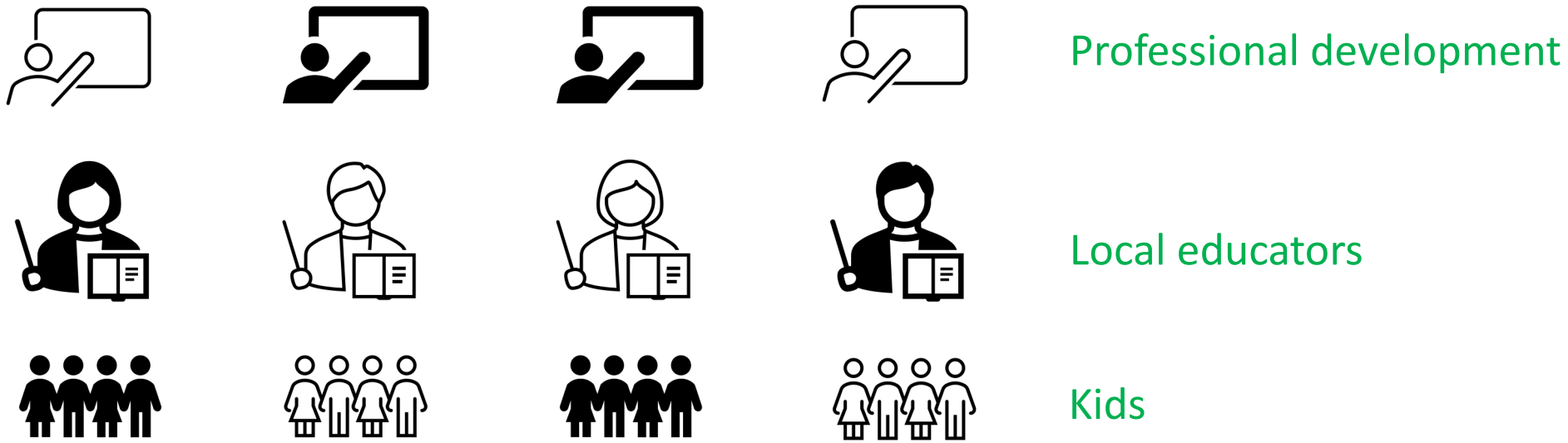
The answer depends upon whether educators/teachers know the technology and the robot.

If not, the robot will be just a toy.

A scalable learning experience

- **Educators/teachers are indispensable contributors to children's education.**

- However, they need continuous professional development to help them broaden their knowledge, which can be very tricky in some countries.



GenAI to scale professional development and learning experience

- GenAI at the Edge is a MUST in some countries (you will know why in a few slides)
- However, is the edge technology ready for genAI?
- The answer is…YES!

LLM performance on Raspberry Pi 5

	TinyLLaMA (1.1B)	Phi-2 (2.7)
Text generation speed (tokens/second)	10.05	5.90
RAM use	606 MBytes	1526 MBytes

- 2 threads on Arm Cortex-A76 @ 2.4GHz
  - SDOT instruction available to speed up the matrix multiplication
- Out-of-the-box performance from the llama.cpp framework
- Int4 blockwise weights compression

What’s so unique about tinyML?

- It holds the answer to AI everywhere.
  - Many opportunities from a business and sustainable perspective
- We can bring to life real projects with just a few dollars.
- It is an excellent field for learning
  - It thrives on the contributions of the open-source community

arm

Thank You
Danke
Gracias
Grazie
谢谢
ありがとう
Asante
Merci
감사합니다
धन्यवाद
Kiitos
شكرًا
ধন্যবাদ
תודה
ధన్యవాదములు