CS249r: Sustainable AI

Nov 20

Course Logistics

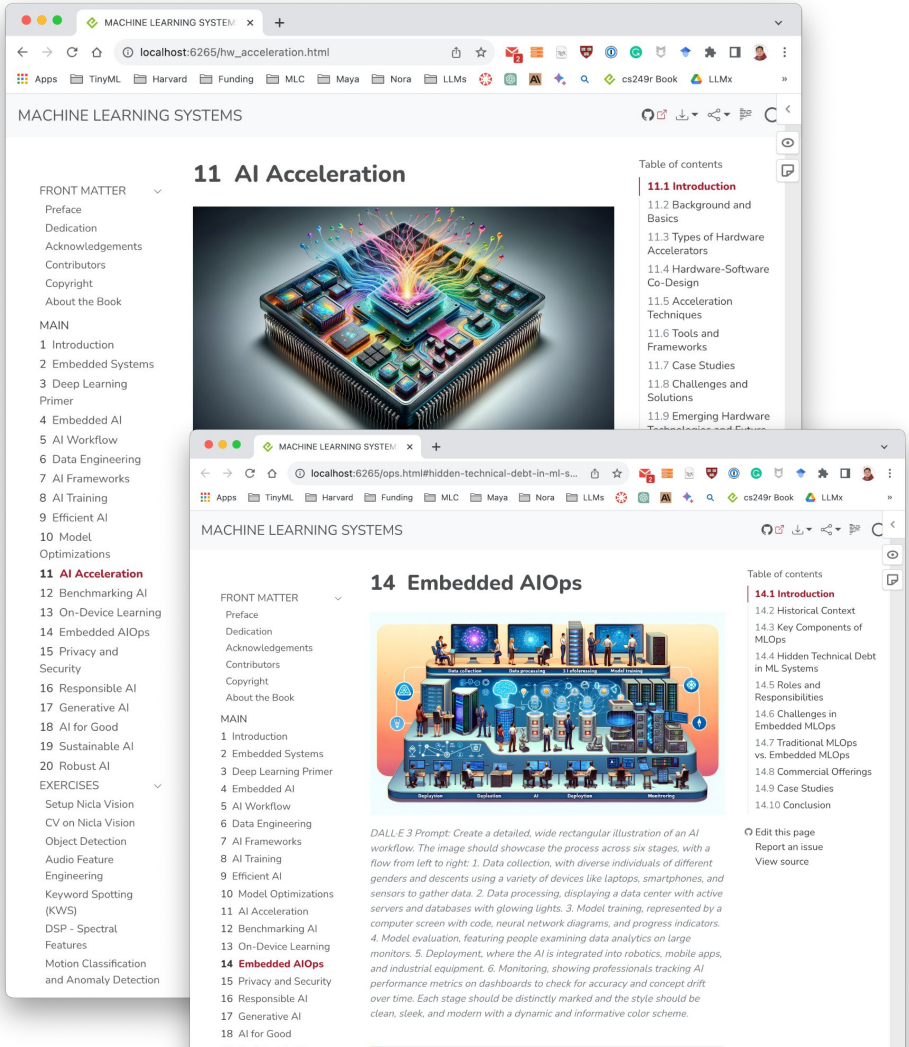
Assignment Schedule Updates

- Assignment 2
  - Due: October 23rd (Monday)
- Mid-Project Review
  - Due: October 30th (Monday)
- Assignment 3
  - Due: November 6th (Monday)
- ~~Assignment 4 Part 1~~
  - ~~Due: November 20th (Monday)~~
- ~~Assignment 4 Part 2~~
  - ~~Due: November 27th (Monday)~~
- Project Presentations
  - Due: December 4th (Monday)
- Final Report
  - Due: December 11th (Monday)

Scribing

- Peer review - Generally, 1 detailed review
  - Security and Privacy PR by EOD
  - Only 1 detailed review
  - Next week Responsible AI
  - Last chapter - Sustainable AI today

- ES91r
  - Opportunity to work with me and my students in the Edge Computing Lab
  - Deepen your knowledge of TinyML and more broadly ML systems.



Project Check-Ins

Check-ins

- Slack check-in messages
- Not mandatory, but **strongly encourage** you to meet with the TAs to discuss the status of the project
- Check-in this week and/or next week

Who to check-in with

- TA: Meet with TAs for informal updates (nothing to prepare)
- VJ: Extended office hours on Monday from 3:30pm to 5pm

Project Rubric

- Final Project Rubric - **45 points**
  - **December 4th - Presentation (~5-7 mins, ~3 min QnA): 10 points**
    - Reuse prior presentation and add in your new approach content
  - **December 11th - Write-up (4-5 pages max): 10 points**
    - Template for write-up here
    - Structure
      - Introduction
      - Background/Related work
      - Approach/Method
      - Insights/Findings
      - Contributions - who did what? **5 points**
  - **December 11th - Video (~2 mins max): 5 points**
    - Demo
  - **December 11th - Technical Deliverables - Code/Hardware: 15 points**

Check List

Scribing

- 24 people completed Peer Reviewing so far
- Check canvas Peer Reviewing assignment

Paper Group Leads

- Check canvas for your grades and adjust if necessary

Assignments

- 1 and 2 (done)
- 3 (done)

Final Projects

- Grading breakdown (slide 5)

Participation/Attendance - check canvas

Course Topics

1. Overview and Introduction to Embedded Machine Learning
2. Data Engineering
3. Embedded Machine Learning Frameworks
4. Efficient Model Representation and Compression
5. ~~Performance Metrics and Benchmarking of ML Systems~~
6. Learning on the Edge
7. Hardware Acceleration for Edge ML: GPUs, TPUs and FPGAs
8. Embedded MLOps
9. Secure and Privacy-Preserving On-Device ML
10. Responsible AI
11. **Sustainable AI**
12. Generative AI at the Edge

Today's Schedule

12:45pm to 1pm

- Course logistics and setting up the context

1pm to 2pm

- Prof. Udit Gupta from Cornell Tech on Sustainability

2:05pm to 2:35pm

- Paper discussion

2:40pm to 3:30pm

- TinyML footprint calculator

Swag!


TINYML


tiny ML
CS249R


Tiny
ML


tiny ML
CS249R


CS249R: TINYML

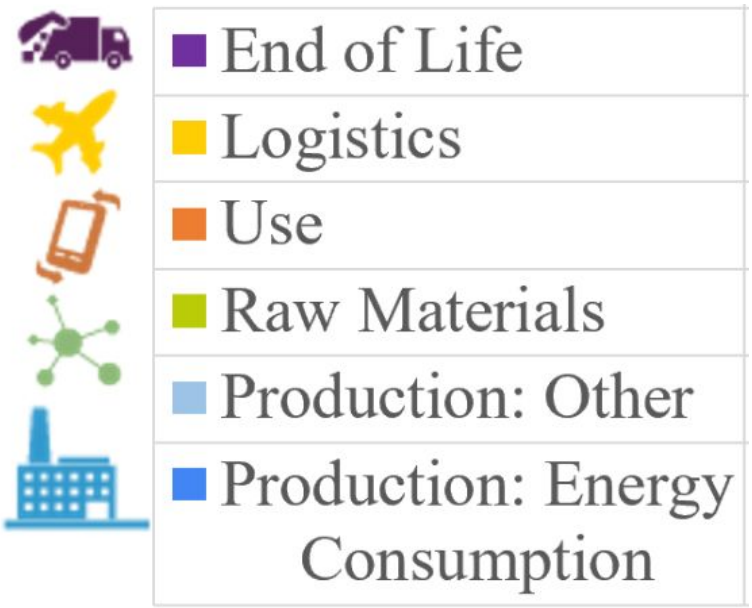

THE FUTURE OF MACHINE LEARNING
IS TINY AND BRIGHT
CS249R

Udit Gupta

Prof. Udit Gupta is from the Department of Electrical and Computer Engineering at Cornell Tech and the Jacob's Technion-Cornell Institute. His research lies at the intersection of computer architecture, systems for machine learning, and sustainable computing. His research focuses on discovering and demonstrating new ways to design systems and hardware to improve the performance, efficiency, and environmental sustainability of next-generation computing platforms and emerging applications. During his PhD in computer science at Harvard University he was also a Visiting Research Scientist at Meta AI. His work has been evaluated at-scale in industry use cases, open-sourced, featured in news articles from outlets like Bloomberg and CNBC, and received the IEEE MICRO Top Picks Award (2022) and Honorable Mention (2021).

Environmental Impact of an Individual MCU

How might you be able to quantify the environmental impact of an MCU?

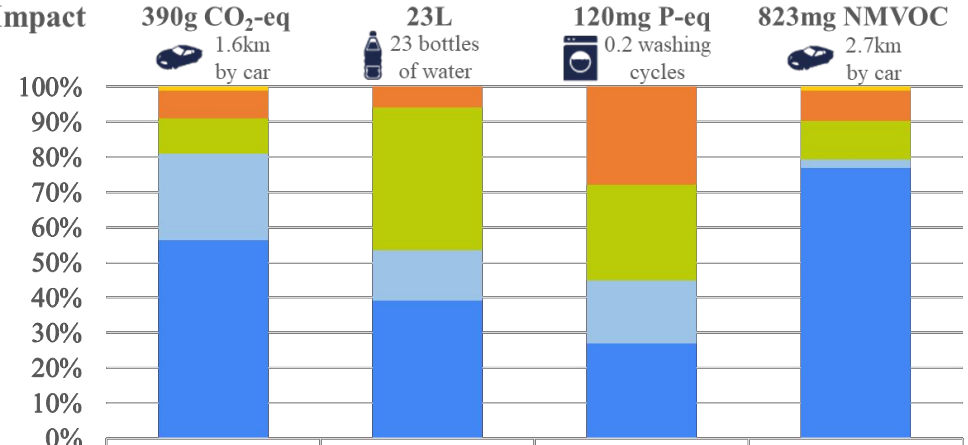


life.augmented

https://www.st.com/content/st_com/en/about/sustainability/sustainable-technology.html

Energy Consumption During Production Dominates the Small Footprint

Reuse of old MCUs would greatly improve their sustainability!



	Climate Change	Water Demand	Freshwater Eutrophication	Protochemical Oxidant Formation
End of Life	<1%	<1%	<1%	<1%
Logistics	1%	<1%	<1%	1%
Use	8%	6%	28%	8%
Raw Materials	10%	41%	27%	10%
Production: Other	25%	15%	18%	2%
Production: Energy Consumption	56%	39%	27%	71%

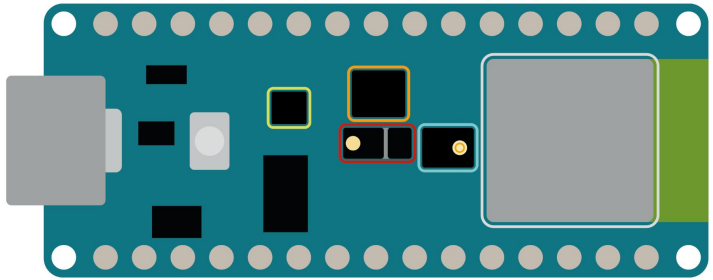
Environmental Footprint of TinyML Systems

Real TinyML Systems are more than just an MCU!

What else is in a TinyML System?

Sensors, Casing, Power Supply, and more!

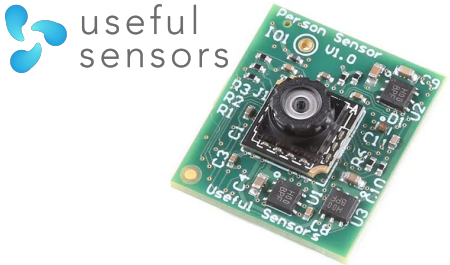
Real TinyML Systems are more than just an MCU!



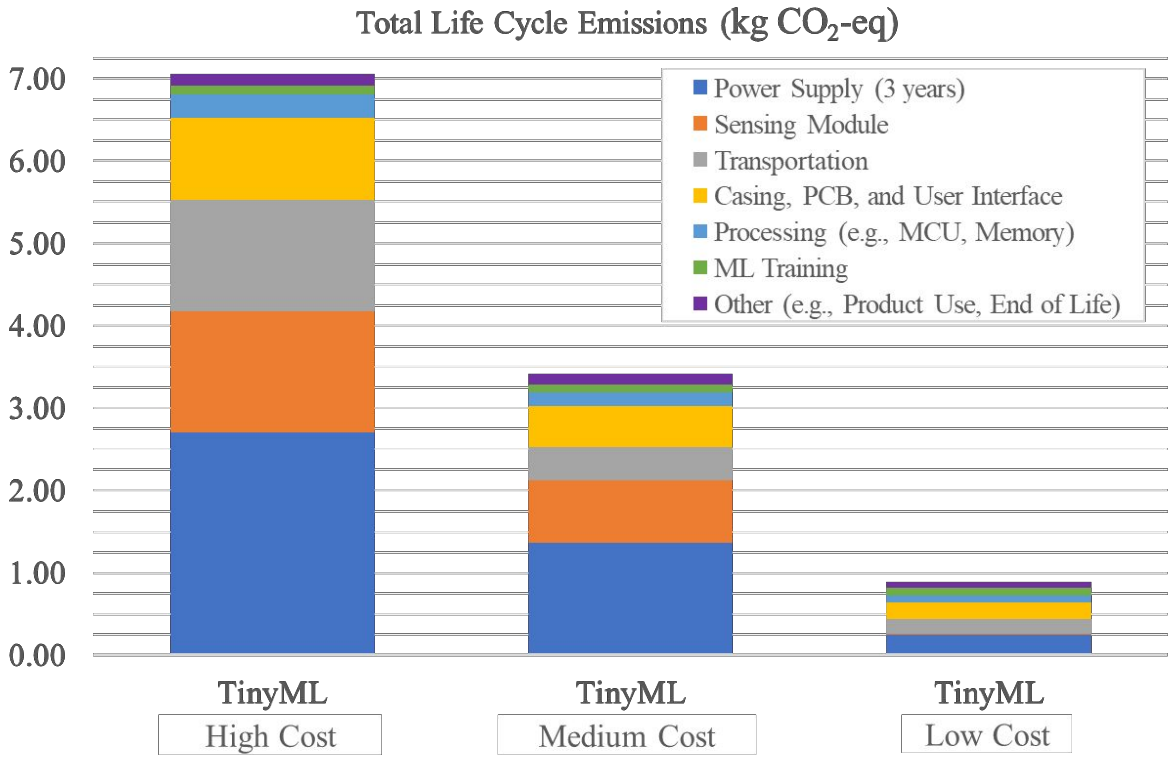
- Color, brightness, proximity and gesture sensor
- Digital microphone
- Motion, vibration and orientation sensor
- Temperature, humidity and pressure sensor
- Arm Cortex-M4 microcontroller and BLE module

Building Representative Systems

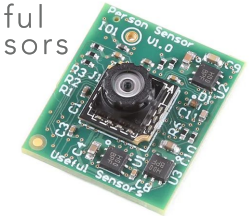
Cost Level	High Cost	Medium Cost	Low Cost
Application	Image Classification		Keyword Spotting
Size	Large	Compact	Compact



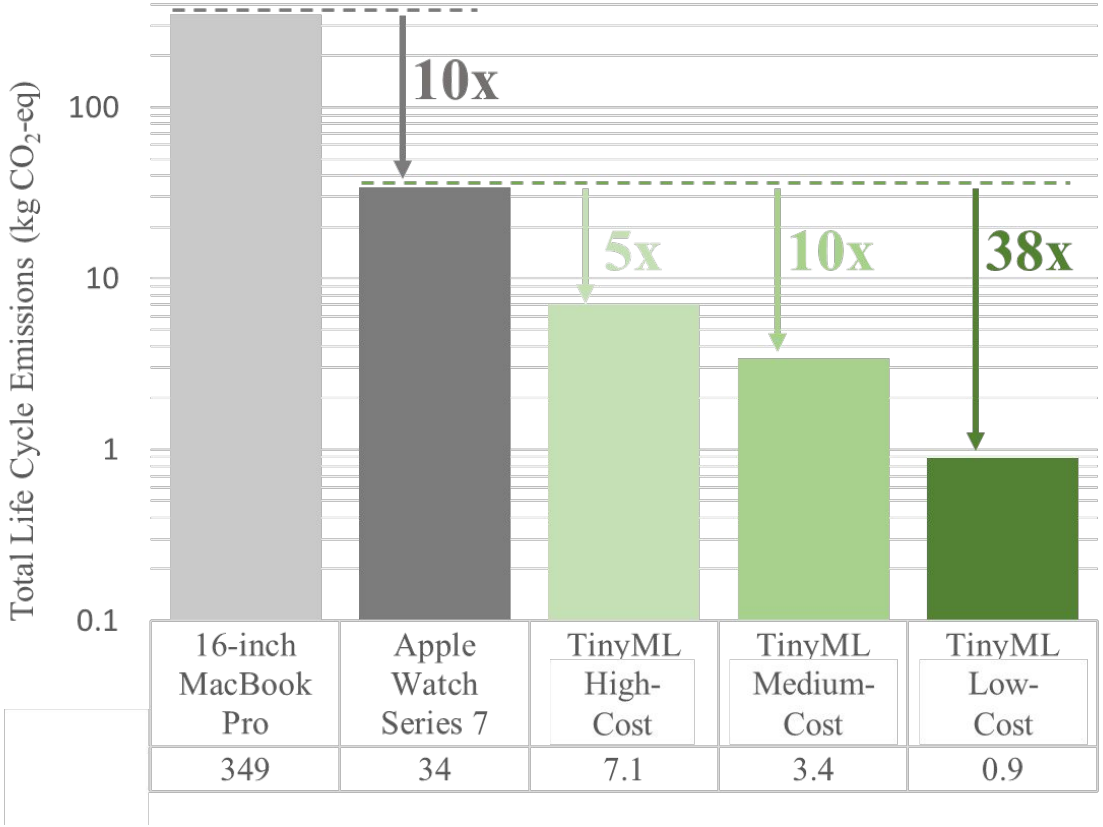
Building Representative Systems



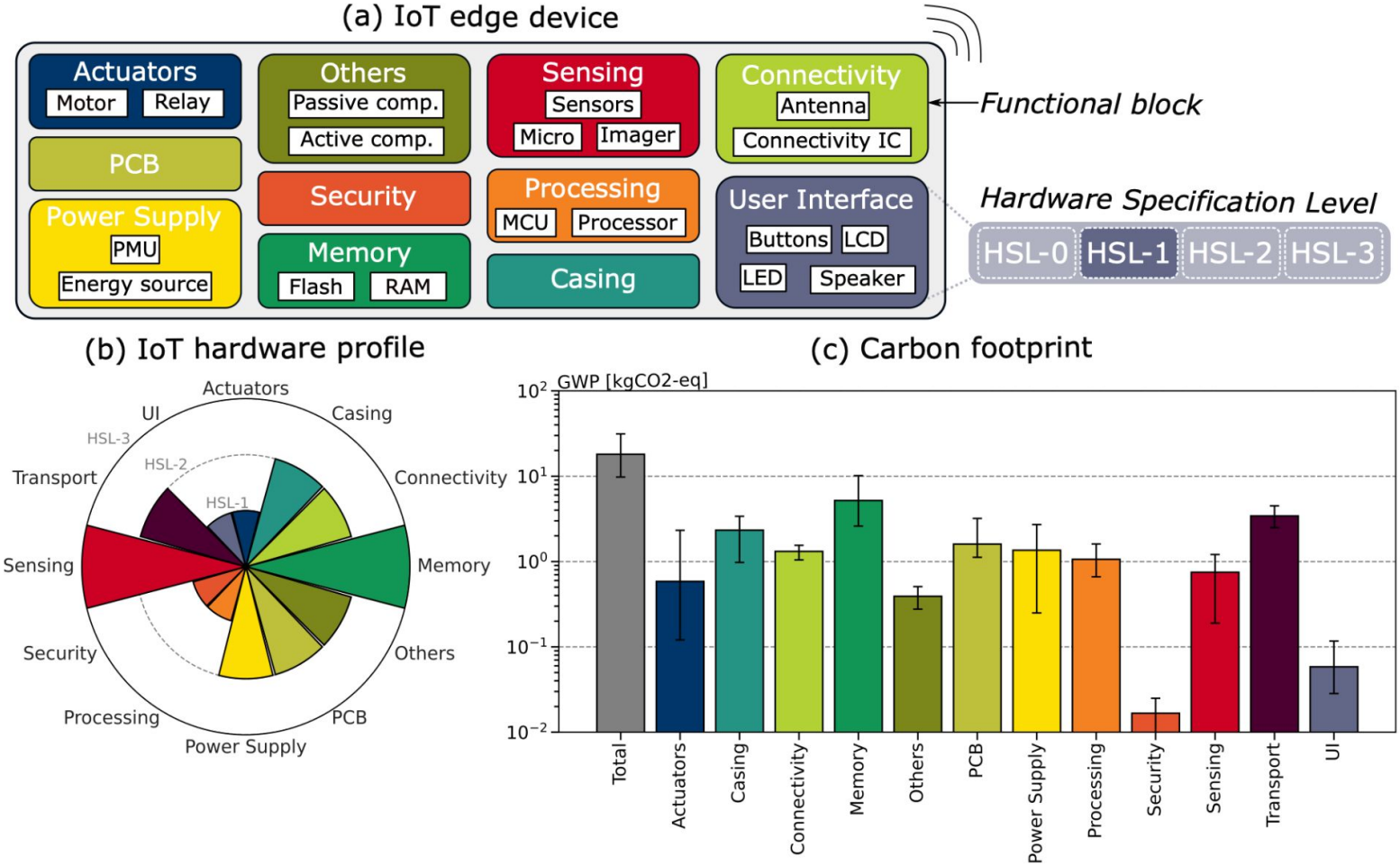
useful sensors



TinyML Systems in Context



5x to 38x Savings over a 3-year lifespan!


(a) IoT edge device
Actuators
Motor
Relay
PCB
Power Supply
PMU
Energy source
Others
Passive comp.
Active comp.
Security
Memory
Flash
RAM
Sensing
Sensors
Micro
Imager
Processing
MCU
Processor
Casing
Connectivity
Antenna
Connectivity IC
User Interface
Buttons
LCD
LED
Speaker
Functional block
Hardware Specification Level
HSL-0
HSL-1
HSL-2
HSL-3
(b) IoT hardware profile
Actuators
Casing
Connectivity
Memory
Others
PCB
Power Supply
Processing
Security
Sensing
Transport
UI
HSL-3
HSL-2
HSL-1
(c) Carbon footprint
GWP [kgCO2-eq]
10^2
10^1
10^0
10^{-1}
10^{-2}
Total
Actuators
Casing
Connectivity
Memory
Others
PCB
Power Supply
Processing
Security
Sensing
Transport
UI

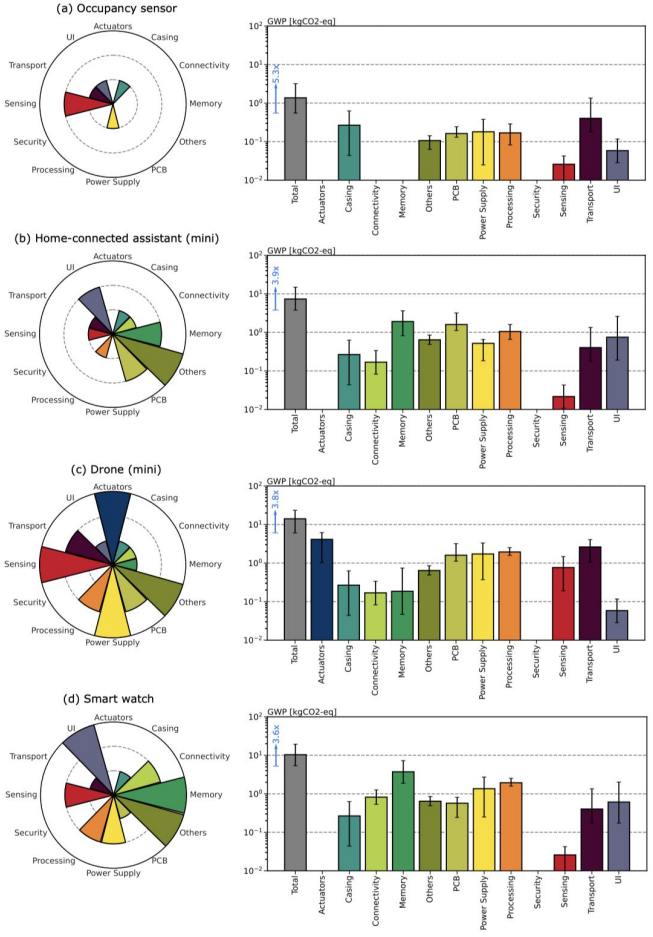


Figure 3: Carbon footprint results based on IoT hardware profiles for the use cases considered in this study: (a) a Philips HUE occupancy sensor, (b) a Google Home mini home connected assistant, (c) a DJI MAVIC mini light-weight drone and (d) a smart watch from Apple and Garmin.

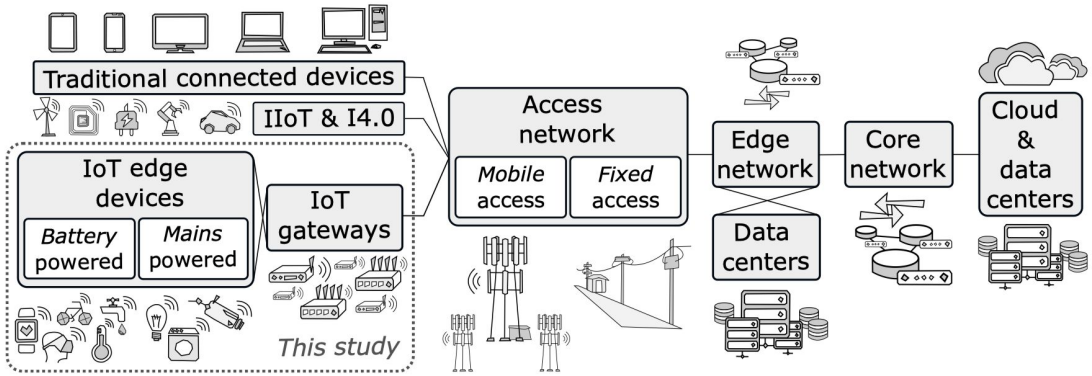
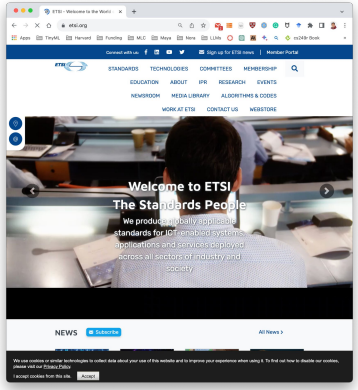


Figure 2: Schematic IoT network representation, adapted from [IEA, 2016, Gray, 2018, Samie et al., 2016]. This study focuses on the IoT edge devices and gateways.

Table 1: Detailed life-cycle inventory (LCI) for IoT hardware profiles. For each *Hardware Specification Level*, lower/typical/upper parameters considered are given.

Functional Block	Hardware Specification Level (HSL) HSL-0	HSL-1	HSL-2	HSL-3
Actuators	No actuator	Vibration motor (1g) 1/2/4	Relay (SSR) 1/2/4	DC motor (50g) 1/4/6 Motor driver ‡ 1/2/5 mm^2 Motor driver transistor 1/4/6
Casing	No casing	ABS plastic granulate 10/50/100 g Aluminium 1/10/30 g Steel 1/10/30 g	ABS plastic granulate 200/400/500 g Aluminium 20/80/150 g Steel 20/80/150 g	ABS plastic granulate 700/800/900 g Aluminium 70/160/300 g Steel 70/160/300 g
Connectivity	Embedded in *Processing* (share of the die area) Printed antenna (embedded in *PCB*)	Connectivity IC * 5/10/20 mm^2 Printed antenna (embedded in *PCB*)	Connectivity IC ▲ 20/30/45 mm^2 External whip-like antenna 10/15/30 g	Connectivity IC ▲ 45/50/60 mm^2 External whip-like antenna 10/15/30 g
Memory	Embedded in *Processing*, Flash + RAM ($\simeq$ kB)	DRAM ◇ (32/128/512 MB) 2/7.9/31.5 mm^2 Flash † (32/128/512 MB) 0.2/0.8/3.2 mm^2	DRAM ◇ (0.5/1/2 GB) 31.5/61.5/123.1 mm^2 Flash † (1/4/8 GB) 6.3/25/50 mm^2	DRAM ◇ (0.5/1/2 GB) 31.5/61.5/123.1 mm^2 Flash † (8/16/32 GB) 50/100/200 mm^2
Others	Capacitors and resistors 5/10/15 Diodes 2/2/2, transistors 1/2/3 Tantalum capacitors 0/0/2, crystals 0/1/1	Capacitors and resistors 15/20/25 Diodes 2/4/6, transistors 2/4/6 Tantalum capacitors 0/0/3, crystals 1/1/2 Steel metal shield 0.5/1/2 g, cables 1/2/5 cm	Capacitors and resistors 40/50/60 Diodes 2/4/6, transistors 4/7/9 Tantalum capacitors 0/0/4, crystals 1/2/4 Steel metal shield 0.5/1/2 g , cables 1/2/5 cm	Capacitors and resistors 75/85/100 Diodes 2/6/10, transistors 7/10/15 Tantalum capacitors 0/2/4, crystals 1/2/4 Steel metal shield 0.5/1/2 g, cables 1/2/5 cm
PCB	FR4 (4 layers) 8/10/15 cm^2 Solder Paste (SAC305) 4/8/13 mg	FR4 (4 layers) 15/35/50 cm^2 Solder Paste (SAC305) 28/53/98 mg	FR4 (8 layers) 35/50/100 cm^2 Solder Paste (SAC305) 99/155/249 mg	FR4 (8 layers) 80/120/150 cm^2 Solder Paste (SAC305) 178/265/454 mg
Power Supply	Mains powered Power transistor 2/3/4 Diodes power 0/1/2, radial capacitor 2/3/4 Miniature coil 2/3/4, ring core coil 0/1/1 Power cable 0.5/1/1.5 m CEE 7/4 Schuko plug 0/1/1	1 Coin cell Li-Po/2 AAA alkaline/2 AA alkaline	Li-ion battery 10/50/100 g Power transistor 0/1/2 Diodes power 0/1/2, radial capacitor 0/1/2 Miniature coil 0/1/2	Li-ion battery 10/50/100 g Power transistor 0/1/2 Diodes power 0/1/2, radial capacitor 0/1/2 Miniature coil 0/1/2 External IC ‡ 5/15/25 mm^2
Processing	MCU * 5/10/17 mm^2	Application processor △ 20/30/45 mm^2 Auxiliary MCU * 5/10/17 mm^2	Application processor △ 50/60/75 mm^2 Auxiliary MCU * 5/10/17 mm^2	Application processor △ 75/100/125 mm^2 Auxiliary MCU * 5/10/17 mm^2
Security	Embedded in *Processing* or non-existent	External IC ‡ 1/2/3 mm^2	N/A	N/A
Sensing	No sensor	Electret microphone 0.05/0.1/0.2 g	Single-multiple sensors ◇ 0/3/5 mm^2	Single-multiple sensors ◇ 0/3/5 mm^2 Single CMOS imager ‡ (1/4" to 2/3") 8/30/58 mm^2
Transport	No transport	Transport from China to Europe Truck distance : 100/300/600 km Plane distance : 6100/6775/7400 km Total weight = 50/100/300 g	Transport from China to Europe Truck distance : 100/300/600 km Plane distance : 6100/6775/7400 km Total weight = 300/650/900 g	Transport from China to Europe Truck distance : 600/900/1200 km Plane distance : 6100/6775/7400 km Total weight = 900/1500/2000 g
User Interface	No user interface	Switch-button 0/1/2 LED 1/2/4	Switch-button 0/2/3 LED 2/4/6, LED driver ‡ 0/1/2 mm^2 Speaker 2/10/40 g, audio driver ‡ 1/2/5 mm^2	Switch-button 2/3/4 LED 3/5/8, LED driver ‡ 0/1/2 mm^2 1 LCD screen 5/25/100 cm^2, driver ‡ 0/1/2 mm^2.

CMOS 0.25μm ‡ CMOS 0.13μm * CMOS 90nm ▲ CMOS 22nm △ CMOS 14 nm † Flash 45nm ◇ DRAM 57nm



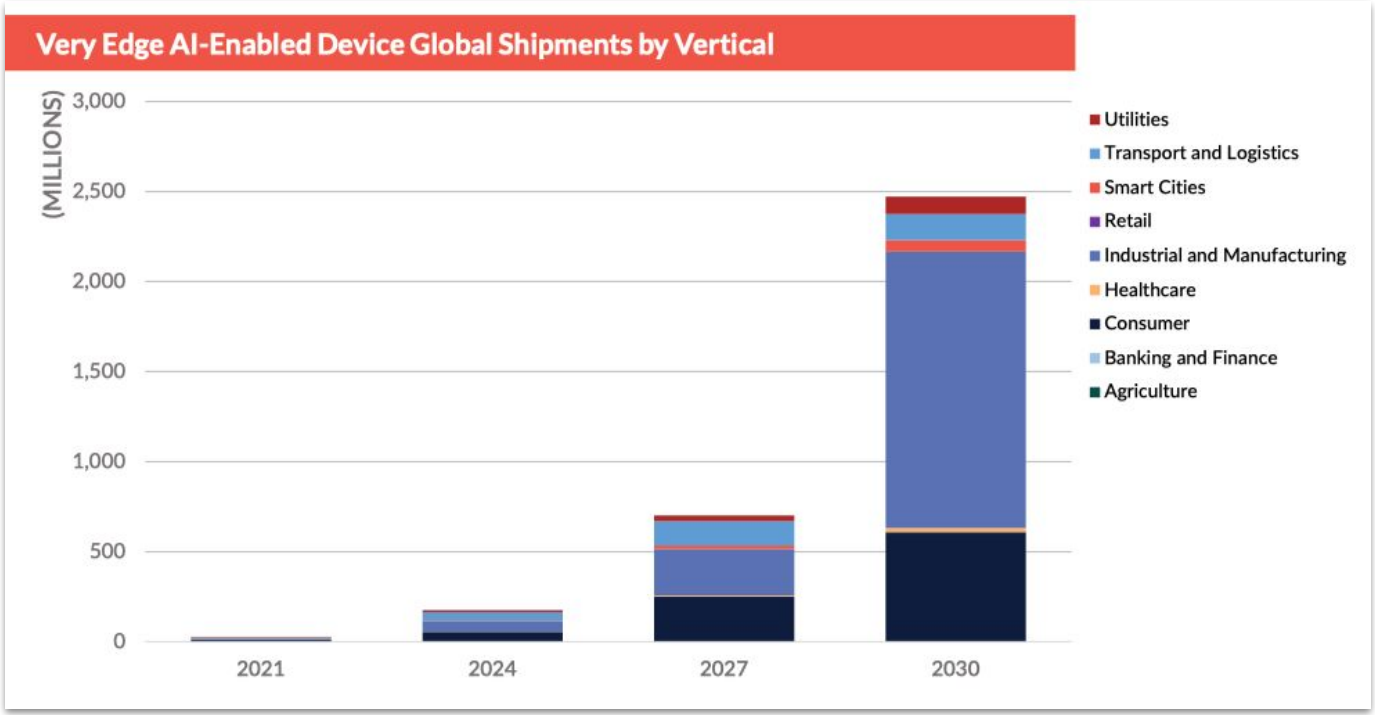
(Pirson and Bol, 2021)

(Forte et al.)

Block / Function	Hardware Specification Level (HSL)											
	HSL-0			HSL-1			HSL-2			HSL-3		
	low	typical	up	low	typical	up	low	typical	up	low	typical	up
Actuators	0.00	0.00	0.00	0.03	0.06	0.12	0.17	0.33	0.66	1.03	4.12	6.19
Casing	0.00	0.00	0.00	0.04	0.27	0.63	0.88	2.13	3.14	3.06	4.26	5.93
Connectivity	0.00	0.00	0.00	0.08	0.17	0.34	0.53	0.82	1.26	1.21	1.36	1.63
Memory	0.00	0.00	0.00	0.05	0.18	0.74	0.82	1.91	3.63	1.88	3.73	7.27
Others	0.06	0.11	0.14	0.14	0.22	0.31	0.28	0.41	0.54	0.49	0.64	0.85
PCB	0.13	0.16	0.24	0.24	0.57	0.81	1.12	1.60	3.20	2.56	3.84	4.80
Power Supply	0.18	0.52	0.66	0.02	0.18	0.38	0.25	1.36	2.71	0.37	1.72	3.32
Processing	0.08	0.17	0.29	0.66	1.05	1.60	1.58	1.94	2.52	2.31	3.13	3.98
Security	0.00	0.00	0.00	0.01	0.02	0.03	N/A	N/A	N/A	N/A	N/A	N/A
Sensing	0.00	0.00	0.00	0.01	0.02	0.04	0.00	0.03	0.04	0.19	0.77	1.47
Transport	0.00	0.00	0.00	0.18	0.40	1.35	1.07	2.62	4.05	3.34	6.30	9.34
UI	0.00	0.00	0.00	0.03	0.06	0.12	0.19	0.75	2.60	0.17	0.61	2.01
Total	0.46	0.96	1.33	1.49	3.20	6.47	6.89	13.88	24.36	16.62	30.47	46.80

(Pirson and Bol, 2021)

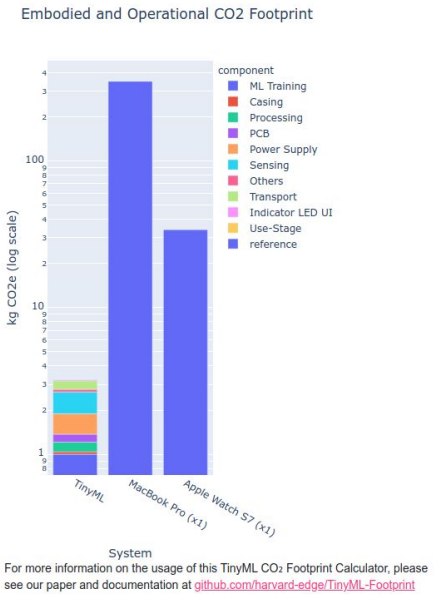
TinyML Market Forecast

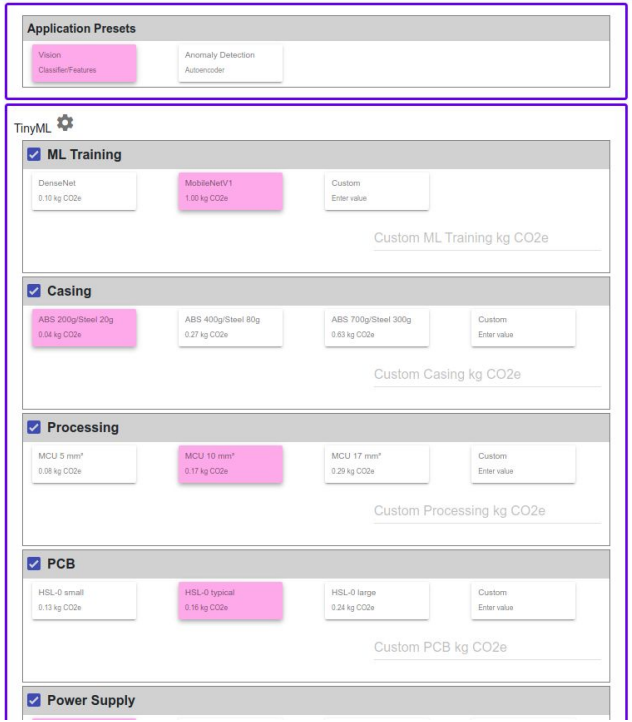


Source: ABI Research: TinyML

harvard-edge.github.io/TinyML-Footprint/

TinyML CO_2 Footprint Calculator





Questions for Students

- Which component contributes the most to environmental footprint?
- Is the camera the only way to go about calculating occupancy? How else might we achieve this?
- How much impact does overprovisioning our system have on the environmental impact?
- What are the pros and cons of using an ML-based approach vs. a non-ML approach for this application?

Limitations and Areas for Future Study

What about the net impact of factors **beyond carbon**?

What about **Jevons' Paradox**?

What about the **human costs**?

How can **emerging technologies** help?