TinyML Workshop

Building Efficient AI for the Edge.

Bostan Khan | Obed Mogaka

HERO Lab, Malardalens University

November 24, 2025

Workshop Presenters

Obed Mogaka
PhD student, IDT
Research Focus: Accelerating Deep Learning Inference on FPGAs.
obed.mogaka@mdu.se

Bostan Khan
PhD student, IDT
Research Focus: Federated Learning and Neural Architecture Search.
bostan.khan@mdu.se

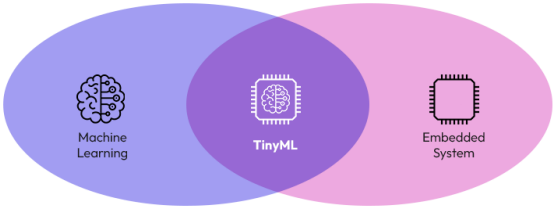
Workshop Agenda

Time	Session	Description	Presenter	Duration (Min)
9:00	Introduction To Seminar	Welcome & Seminar Overview	Bostan	15
9:15	Quantization	Core techniques for reducing model size & latency.	Obed	60
10:15	Coffee Break	Networking & Refreshments	-	15
10:30	Pruning + Hands-on Demo	Model pruning theory & practical compression lab.	Obed	60
12:00	Lunch Break	-	-	60
13:00	Knowledge Distillation	Using “teacher” models to create small “student” models.	Bostan	60
14:00	Neural Architecture Search	Automatically designing efficient model architectures.	Bostan	60
15:00	Coffee Break	Networking & Refreshments	-	15
15:15	Hardware Deployment Demo	Deploying a final model on a real edge device.	Obed	60
16:30	End of Seminar	Closing Remarks & Q&A	-	-

Edge AI

What is Edge-AI, TinyML?

- The deployment of AI algorithms and AI models directly on local edge devices such as sensors or Internet of Things (IoT) devices.
- The combination of edge computing and artificial intelligence to perform machine learning tasks directly on interconnected edge devices.[1]



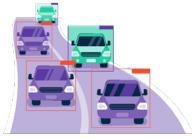
[1] https://www.ibm.com/think/topics/edge-ai

AI is everywhere!

On-device or Edge-AI has enabled transformative applications on several domains.

Medical

Wearables



Computer Vision

Robotics

IoT and Smart Cities

The Power of Edge AI: Key Advantages

Benefit 1: Diminished Latency

- **Instantaneous Response:** On-device processing eliminates the round-trip delay of sending data to a distant server.
- **Real-Time Experience:** Users benefit from rapid, seamless interactions without network lag.
- **Critical for Speed:** Essential for applications where immediate action is required (e.g., autonomous vehicles, industrial robotics).

Benefit 2: Decreased Bandwidth Usage

- **Local Data Processing:** Minimizes the amount of data transmitted over the internet.
- **Network Efficiency:** Preserves internet bandwidth, allowing the network to handle more devices and traffic simultaneously.

The Power of Edge AI: Key Advantages

Benefit 3: Real-Time Analytics & Operation

- **Immediate Insights:** Perform data processing and analysis directly on the device, without needing to connect to a central system.
- **Offline Capability:** Enables continuous operation even with intermittent or no internet connectivity.
- **Consideration:** For massive-scale AI, a hybrid approach with the cloud may be needed to leverage its superior resources.

Benefit 4: Enhanced Data Privacy & Security

- **Data Stays Local:** Sensitive information is processed and stored on the device, not transferred over vulnerable networks.
- **Reduced Risk:** Minimizes exposure to cyberattacks and potential data mishandling during transmission.
- **Compliance Ready:** Helps meet data sovereignty regulations (like GDPR) by keeping data within designated jurisdictions.

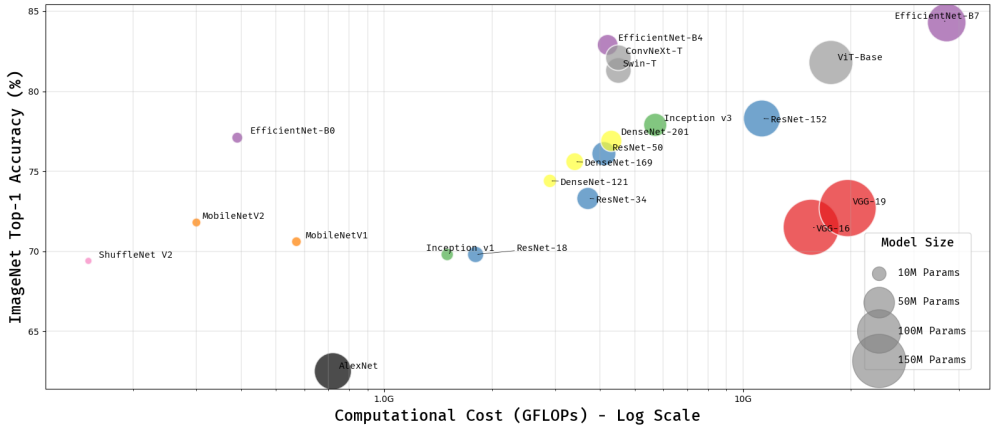
Benefit 5: Improved Scalability & Reliability

- **Increased Resilience:** Local networks can maintain functionality even if the central cloud or other nodes go down.
- **Decentralized Strength:** Reduces dependency on a single point of failure, creating a more robust system.

The Edge has got the edge!

- **FASTER:** Delivers real-time responses by eliminating network latency.
- **MORE EFFICIENT:** Saves bandwidth and lowers cloud computing costs.
- **MORE SECURE:** Enhances data privacy by keeping sensitive information local.
- **MORE RELIABLE:** Ensures continuous operation, even when offline, and scales with ease.

Challenge 1 & 2: Computation Cost and Model size



- AI models often comes at the cost of high computation and memory requirements.

Challenge 3: Energy Efficiency

- AI is responsible for 100 terawatt-hours (TWh) of electricity globally in 2025, with worst-case forecasts predicting a surge to 1,370 TWh by 2035. [a]
- Today, AI systems account for 0.1 percent of global greenhouse gas emissions, equivalent to the annual emissions of Sweden. [b]

[a] https://huggingface.github.io/AIEnergyScore

[b] https://www.ri.se/en/how-energy-efficient-ai-can-reduce-climate-impact

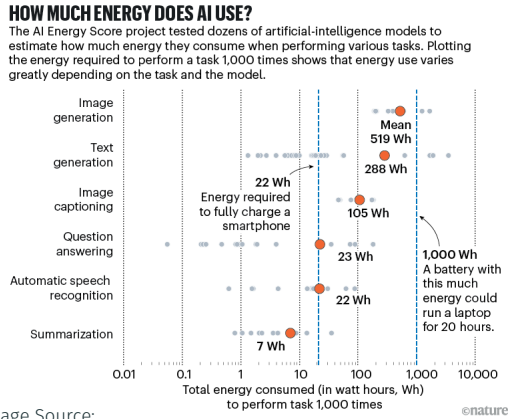
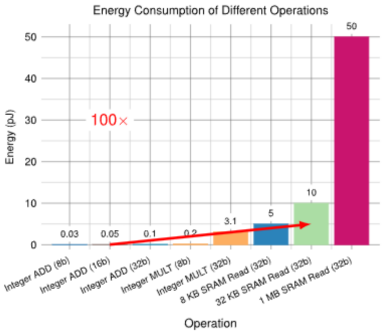


Image Source:
https://www.nature.com/articles/d41586-025-00616-z

Where is the energy consumed?

- Different **numerical precisions** are associated with energy costs.
- Off-chip DRAM, is far more energy-intensive than performing arithmetic operations.

	Operation	Energy_pJ
1	Integer ADD (8b)	0.03
2	Integer ADD (16b)	0.05
3	Integer ADD (32b)	0.10
4	Integer MULT (8b)	0.20
5	Integer MULT (32b)	3.10
6	8 KB SRAM Read (32b)	5.00
7	32 KB SRAM Read (32b)	10.00
8	1 MB SRAM Read (32b)	50.00

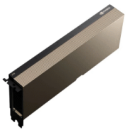


Cloud AI Hardware

NVIDIA

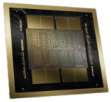
P100 (2016)

V100 (2017)

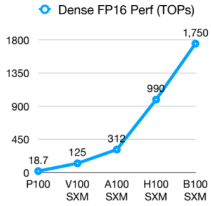


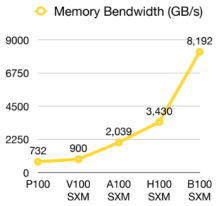
A100 (2020)

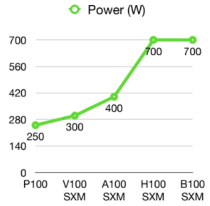
H100 (2022)

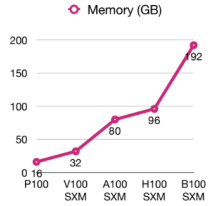


B100 (2024)





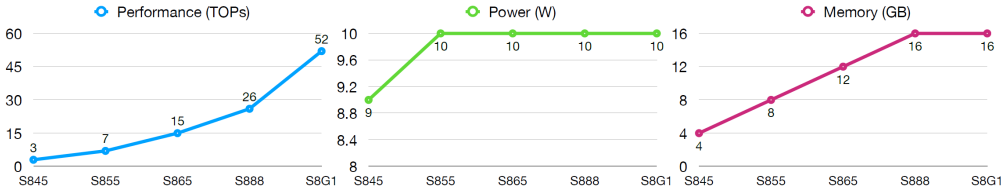




Edge AI Hardware

Qualcomm Hexagon DSP

Qualcomm Hexagon is a family of digital signal processor (DSP) products by Qualcomm. It is designed to deliver performance with low power over a variety of applications.



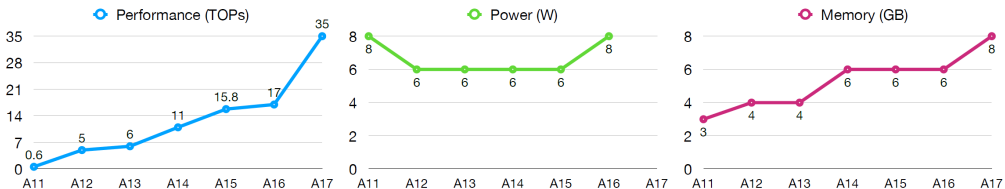
Wikipedia [2]

[2] https://en.wikipedia.org/wiki/List_of_Qualcomm_Snapdragon_processors

Edge AI Hardware

Apple Neural Engine

The Apple Neural Engine (ANE) is an energy-efficient and high-throughput engine for ML inference on Apple silicon.



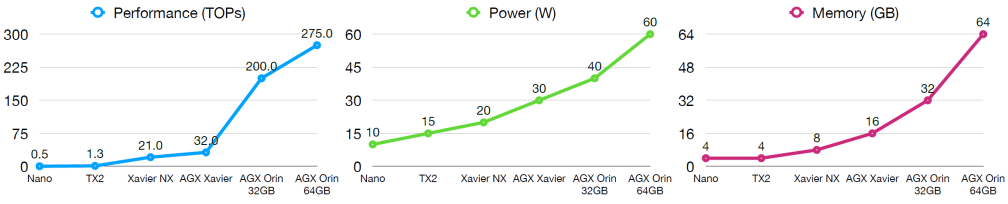
NanoReview [3]

[3] https://nanoreview.net/en

Edge AI Hardware

NVIDIA Jetson

NVIDIA Jetson is a complete System on Module (SOM) that includes a GPU, CPU, memory, power management, high-speed interfaces, and more.

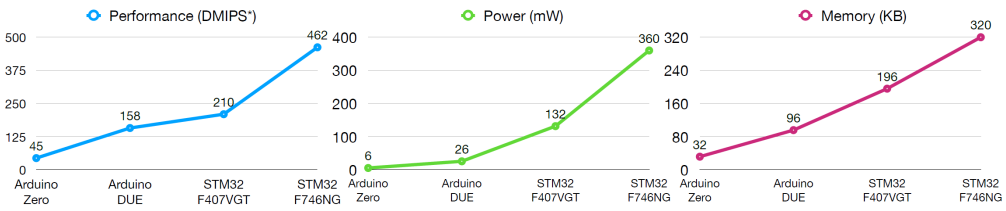


NanoReview [4]

[4] https://connecttech.com/jetson/jetson-module-comparison/

Microcontrollers (MCU)

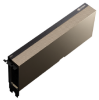
A microcontroller is a compact integrated circuit designed for embedded systems. A typical microcontroller includes a processor, memory and input/output (I/O) peripherals on a single chip.



** Dhrystone Million Instructions Per Second (DMIPs) is an index for integer computation.*

Edge AI Hardware

Edge AI devices have a huge gap to cloud processors.

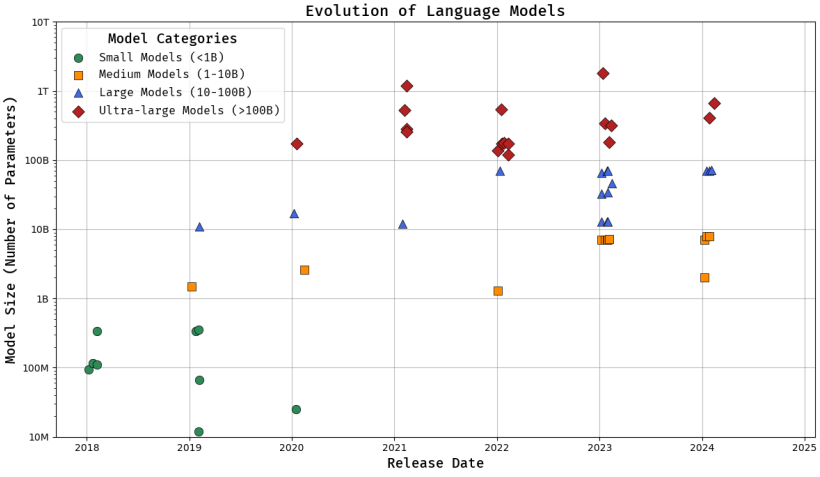


	Cloud AI	Mobile AI	Tiny AI
Memory (Activation)	80GB	4GB	320kB
Storage (Weights)	~TB/PB	256GB	1MB

Edge devices are resource constrained:

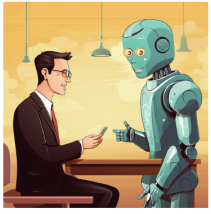
- LIMITED battery life
- LIMITED memory capacity
- LIMITED computing power

Model size of language models is growing fast.



Running LLMs on the edge is also very important.






- Deploying LLM on the edge is useful: running copilot services (code completion, office, game chat) locally on laptops, cars, robots, and more.
- These devices are resource-constrained, low-power and sometimes do not have access to the Internet.
- Data privacy is important. Users do not want to upload personal data to the cloud.

Deployment Readiness Matrix for Edge Generative AI

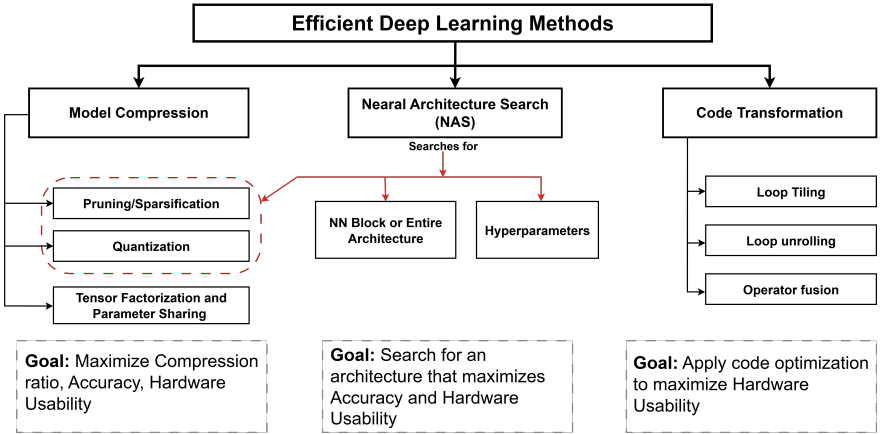
	IOT/ MICROCONTROLLER	EDGE DEVICE (PHONE/WEARABLE)	FOG/ENTERPRISE (EDGE SERVER)	CLOUD (DATA CENTER)
SMALL (<1B PARAMETERS) ~100MB-2GB MODEL SIZE	CHALLENGING Heavily quantized. Specialized arch. Limited capabilities.	IDEAL 8-bit quantization. 100-500ms latency. Modern NPU required.	IDEAL Can serve multiple users/devices. Efficient & responsive.	IDEAL Massive batching. Low-latency API. Energy efficient.
MEDIUM (1-10B PARAMETERS) ~2-20GB MODEL SIZE	INFEASIBLE Exceeds memory constraints.	CHALLENGING High-end phones only. 4-bit quantization. Thermal limitations.	IDEAL GPU/TPU accelerated. Good inference speed. Multi-user capable.	IDEAL High throughput. Cost effective. Scalable.
LARGE (10-100B PARAMETERS) ~20-200GB MODEL SIZE	INFEASIBLE Exceeds memory constraints.	INFEASIBLE Exceeds memory & compute constraints.	CHALLENGING Multiple GPUs. Higher latency. Power constraints.	IDEAL Standard deployment. GPU/TPU clusters. Elastic scaling.
ULTRA-LARGE (100B+ PARAMETERS) ~200GB+ MODEL SIZE	INFEASIBLE Exceeds memory constraints.	INFEASIBLE Exceeds memory & compute constraints.	INFEASIBLE Exceeds memory & compute constraints.	IDEAL Distributed inference. High throughput. Dynamic workload balancing.

Source:

- Color coding indicates deployment feasibility: green (ideal), yellow (challenging but possible), and red (infeasible).
- Small models (<1B parameters) are viable across most environments, while ultra-large models (100B+ parameters) are restricted primarily to cloud deployments.

Workshop Proper

Efficient Deep Learning Methods



Workshop Agenda

Time	Session	Description	Presenter	Duration (Min)
9:00	Introduction To Seminar	Welcome & Seminar Overview	Bostan	15
9:15	Quantization	Core techniques for reducing model size & latency.	Obed	60
10:15	Coffee Break	Networking & Refreshments	-	15
10:30	Pruning + Hands-on Demo	Model pruning theory & practical compression lab.	Obed	60
12:00	Lunch Break	-	-	60
13:00	Knowledge Distillation	Using “teacher” models to create small “student” models.	Bostan	60
14:00	Neural Architecture Search	Automatically designing efficient model architectures.	Bostan	60
15:00	Coffee Break	Networking & Refreshments	-	15
15:15	Hardware Deployment Demo	Deploying a final model on a real edge device.	Obed	60
16:30	End of Seminar	Closing Remarks & Q&A	-	-