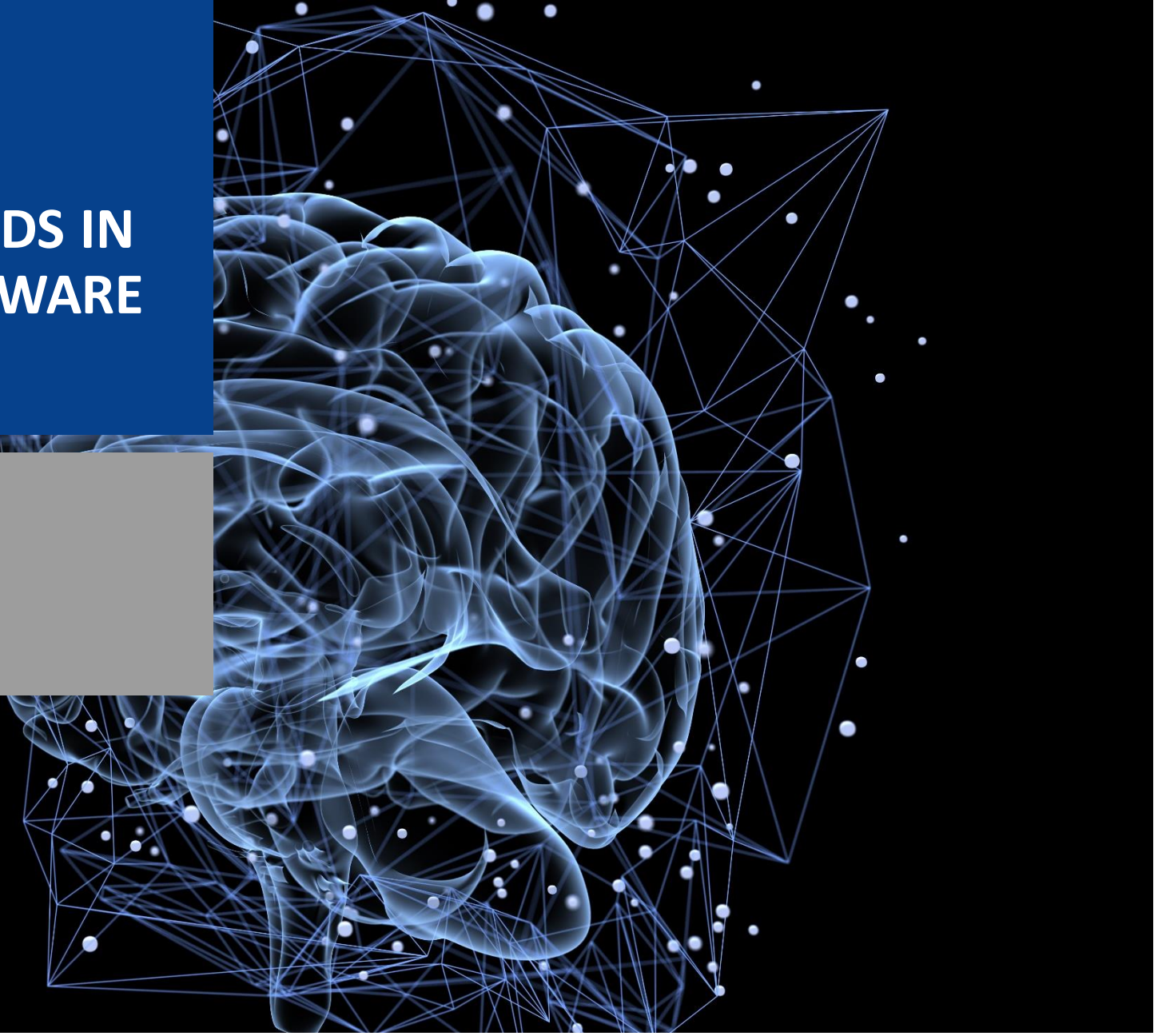
CHALLENGES AND TRENDS IN MODERN EDGE AI SOFTWARE DEVELOPMENT



MANDALI KHALESI
VICE PRESIDENT, AI PLATFORM & TECHNOLOGY
5 NOVEMBER 2025

RENESAS ELECTRONICS

WHAT IS EDGE AI?

"Edge artificial intelligence refers to the deployment of AI algorithms and AI models directly on local edge devices such as sensors or Internet of Things (IoT) devices, which enables real-time data processing and analysis without constant reliance on cloud infrastructure." - IBM

RAPID GROWTH OF EDGE AI SHIPMENTS

2.3B Edge AI devices shipped in 2024

Vs 1.2B Mobile devices shipped in 2024

=> 6B Edge AI devices to ship in 2030

- Research and markets, Statista

CHALLENGES IN EDGE AI SW: AI INFERENCE VS AI TRAINING

"In the world of artificial intelligence, much of the spotlight has been focused on the training of massive models... [that] require vast computational resources and months of training on specialized hardware. Yet, for all the attention paid to training, the most pressing challenge in AI today lies elsewhere: **inference**.

Solving these challenges demands a rethinking of how we design models, optimize hardware, and architect systems. The future of AI depends on our ability to master inference at the edge."

- HPC Wire (April 2025)

Why Edge AI is the next great computing challenge

CHALLENGES IN EDGE AI SW: AI ON MICROCONTROLLERS

"To run that tiny neural network, a microcontroller also needs a lean inference engine. A typical inference engine carries some dead weight — instructions for tasks it may rarely run. The extra code poses no problem for a laptop or smartphone, but it could easily overwhelm a microcontroller.

“It doesn’t have off-chip memory, and it doesn’t have a disk. Everything put together is just one megabyte of flash, so we have to really carefully manage such a small resource.”"

- MIT News (Nov 2020)

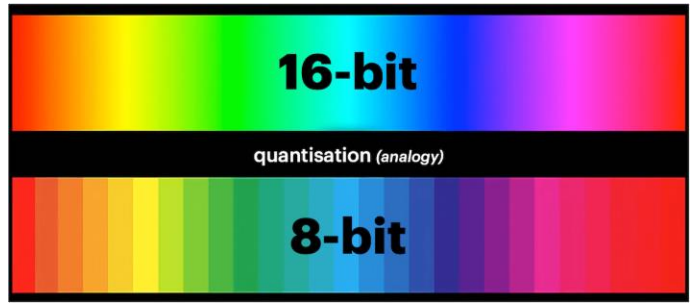
System brings deep learning to IoT devices

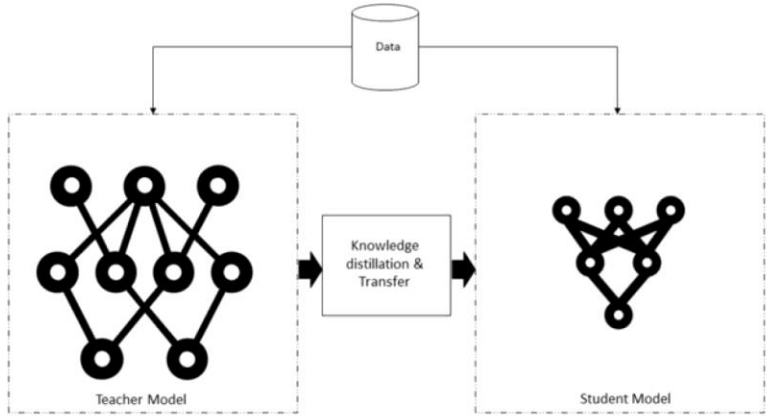
TECH TRENDS IN EDGE AI SW: QUANTIZATION & DISTILLATION

To reduce the size of a model, at last year's conference I shared novel techniques for **low-bit quantization**. Structured compression of the AI model from 32-bits to 8-bits or less while retaining its logical fidelity.

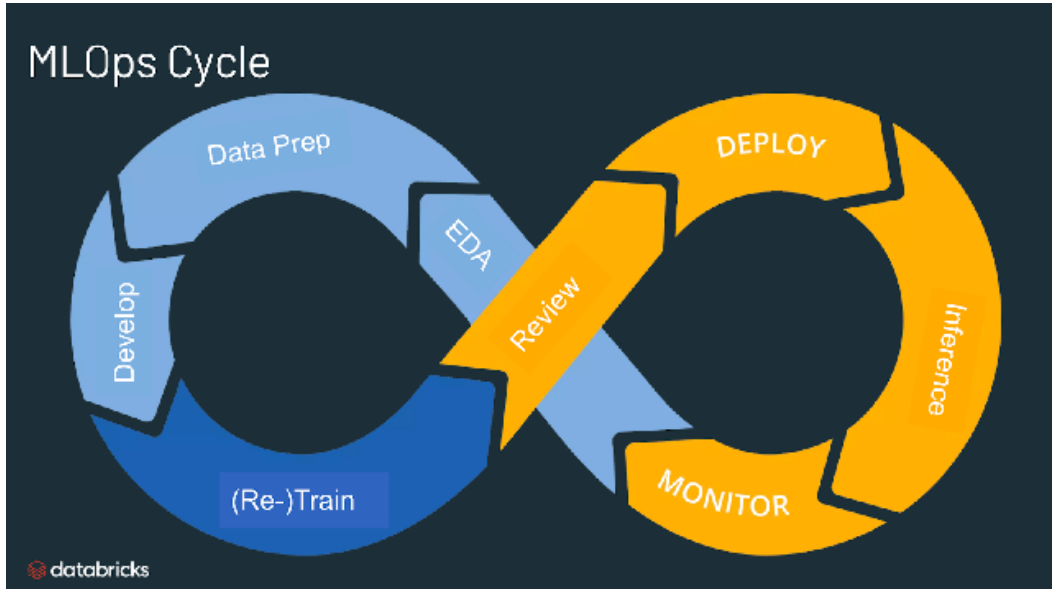
Model knowledge distillation is another technique which retrains the model into a smaller form factor by learning from a "teacher" model.

Both techniques are orthogonal to each other. Recent efforts by companies to bring quantization and distillation together with positive results: Quantization Aware Knowledge Distillation (2019), QUADS (2025)... 2026?





TECH TRENDS IN EDGE AI SW: ML OPS VS LLM OPS



deepseek

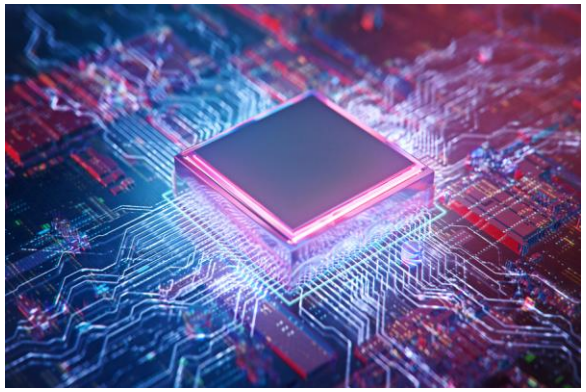
LLaMA

Gemma

Qwen

Combining the machine learning and large language models is necessary in most modern AI applications … but currently it's a challenge

TECH TRENDS IN EDGE AI: AI HARDWARE & AI SOFTWARE CO-DEV



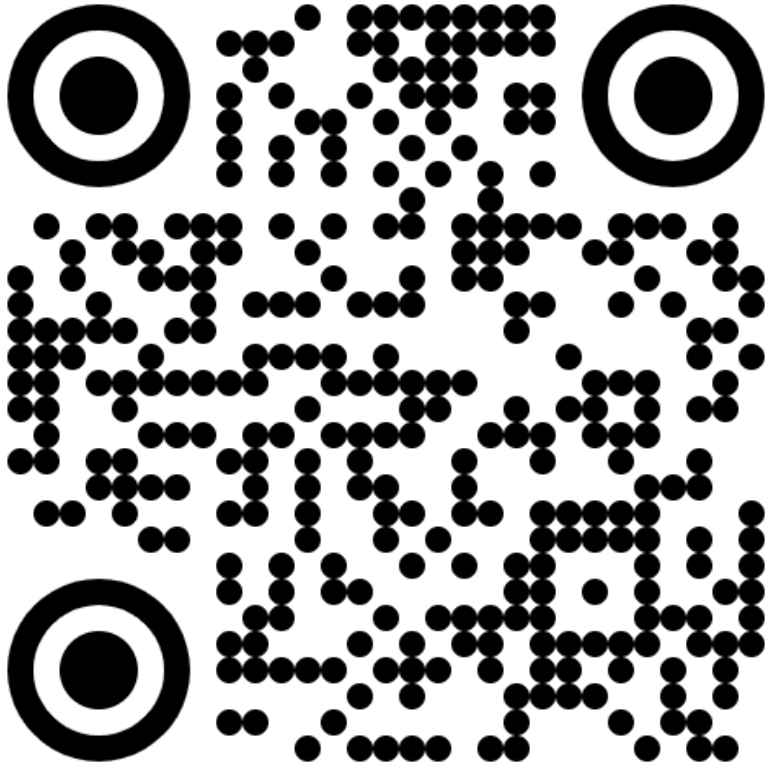
Lack of collaboration between AI hardware development and
AI software development to drive user benefits
… need to work together

ANNOUNCING: NEW GSA AI SOFTWARE INTEREST GROUP

- GSA is launching a new AI Software Interest Group, currently limited to APAC participants until end 2025

- Goal: share best practices in emerging areas of:
  - Optimization Techniques
  - ML Ops and LLM Ops
  - AI Agents
  - Collaborative Tooling, and more!

- Get in touch today via our AI SIG, or connect with me via LinkedIn, to join our community and help accelerate adoption across industries



(links to: https://www.gsaglobal.org/asig)