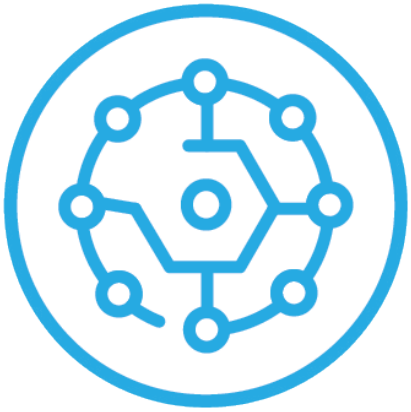
EPoSS.
European Association on
Smart Systems Integration



Joint workshop by EPoSS and
INSIDE Industry Associations

The Future of Innovation in Edge AI

Online, April 04, 2025

Notes from the Workshop

Part II: Introduction: The Landscape of Edge AI

Presented by Paolo Azzoni, INSIDE

Three Layers of Limitation

Deploying AI at the edge introduces a unique set of challenges that fall into three broad categories:

- **Device-level limitations**
- **AI model and software constraints**
- **Environmental and financial pressures**

Each of these introduces its own set of barriers that must be addressed for successful implementation.

1. Device-Level Constraints

Edge devices often operate with very limited memory, storage, and processing power. Unlike centralized cloud infrastructure, these systems cannot depend on scalable resources.

The limitations are compounded by energy restrictions – many edge devices are battery-powered and must be highly energy-efficient. Furthermore, these devices rarely serve a single purpose. They typically support sensors, actuators, communication modules, and general computing tasks, meaning that AI workloads must compete for already-limited resources.

2. Software and AI Model Challenges

Even the most optimized hardware cannot overcome the demands of poorly adapted AI models. Edge-based AI systems face several software-related challenges:

- **Precision requirements:** High-precision models consume more resources.
- **Model optimization:** Pruning and quantization techniques are often necessary to make models fit the hardware, sometimes at the expense of performance.
- **Preprocessing overhead:** Real-world data needs cleaning and formatting, which adds processing time and power use.
- **Latency vs. accuracy:** A highly accurate model is ineffective if its inference time is too slow for real-time use.

These issues mean developers must constantly balance performance, resource consumption, and responsiveness when designing edge AI systems.

3. Environmental and Financial Barriers

The edge often means remote, rugged, and resource-constrained environments. Many devices operate in locations that are physically difficult to access. This limits maintenance options, upgrades, and even physical replacements.

Moreover, physical requirements such as ruggedization or waterproofing limit the design flexibility of edge devices. The cost factor is equally important – incorporating AI increases both the complexity and the overall expense of devices, creating additional barriers to widespread adoption.

Reaching the Limits of Hardware

As technology advances, the industry is beginning to reach physical and thermal constraints. The historical momentum of Moore's Law is slowing. In response, new technologies are emerging – such as tensor processing units (TPUs), 3D chip integration, integrated photonics, and memory-based architectures. While these technologies hold promise, they also increase system complexity, cost, and face the challenge of lacking standardized frameworks.

The Memory Wall and Data Bottlenecks

One of the most persistent technical challenges is the so-called **"memory wall"** – the bottleneck created by moving data between processors and memory. A substantial portion of the processing time in AI execution is spent simply transferring data, not analyzing it.

Solutions like stacked memory, faster buses, and new memory hierarchies have been proposed, but these are not yet widely adopted. Until they are, the memory wall remains a key inefficiency in edge AI systems.

The Energy Dilemma

Energy efficiency continues to be a major hurdle. Even as processors become more powerful, energy consumption does not always scale accordingly. On the edge, where energy sources are limited, even fast processors are constrained by the available power budget. This remains one of the most significant limitations to building more capable and responsive edge systems.

Toward Vertical Integration and System-Level Thinking

Meeting these challenges requires more than technical fixes – it calls for a strategic shift in how systems are designed. Leading tech companies like Nvidia and major cloud providers have adopted **vertical integration**: designing their own chips, building their own systems, and optimizing every layer from hardware to software.

This approach allows for seamless performance, better scalability, and full-stack optimization. For edge AI, a similar strategy is necessary. Holistic design – from the chip level to the application layer – is essential to create systems that are not only powerful but also efficient and adaptable to real-world conditions.

Conclusion: Engineering the Future of Edge AI

Edge AI stands at the intersection of immense potential and significant constraint. From limited device resources and strict energy requirements to software trade-offs and cost barriers, the road ahead is complex. However, with a shift toward co-design, system-level engineering, and vertical integration, these challenges can be transformed into opportunities.

Future success in this domain will belong to those who approach the technology stack as a whole – optimizing from the ground up to deliver AI that truly works at the edge.