# Chapter 2.1
# EDGE COMPUTING AND EMBEDDED ARTIFICIAL INTELLIGENCE

Marc Duranton, CEA

E C S

*Strategic Research and Innovation Agenda 2025*

Aeneas

EPoSS
European Association on Smart Systems Integration

Inside
Industry Association

# Scope

This chapter focuses on **computing components**, and more specifically on **embedded architectures, edge computing devices and systems using artificial intelligence at the edge**, including, but not exclusively:

- Processors (CPU, MPU) with high energy efficiency,
- Accelerators (for AI and for other tasks, such as security):
- GPU (and their generic usage),
- NPU (Neural processing unit)
- DPU (Data processing Unit, e.g. logging and collecting information for automotive and other systems) and processing data early (decreasing the load on processors/accelerators),
- Other accelerators xPU (FPU, IPU, TPU, XPU, …)
- **Memories** and associated controllers, specialized for low power and/or for processing data locally (e.g. using non-volatile memories such as PCRAM, CBRAM, MRAM, and ***In/Near Memory Computing***), etc.
- **Power management systems** and techniques.

# Key trends

New

More and more convergence between edge computing and embedded (generative) AI, but ***still a lot of edge will be without AI.***

Emergence of ***Gen-AI at the edge*** (Copilot+PC, Apple Intelligence), for automatizing tasks locally, processing local data, for natural user interface, but also for ***image interpretation and robotics***.

<u>New Recommendations:</u>

- Managing diversity and the ***dynamic range of computing*** -> A system becomes an ***orchestration of federated services, distributed or centralized*** (Software Defined X).

Impact on architecture and of the programmability and interoperability (cloud techniques used: containers (silo), orchestration, WASM, interoperability, “Web standards”) therefore their hardware support, including for networking.

- Disaggregation of complex SoC into chiplet + interposers, for example potentially for the automotive market, but still no ecosystem of interoperable chiplets and overall architecture.

# Key trends

New

- Generative AI is not only for the cloud, Intelligent Agents (federations of Small Action Models) or ***Agentic AI*** will emerge at the edge, and not only in smartphones.

- ***Fine tuning*** of models should be possible at the edge only with local data

- ***Memory cost is crucial for generative AI at the edge***. New innovations required to avoid to waste RAM

- Emergence of (very) cheap Chinese Risc-V microcontrollers

- Further ***reducing standby power*** and fast on operation (stop and go for chips?) and ***proportionality of power/load*** for reducing overall power consumption

- Still research required for ***new computing paradigms*** (neuromorphic, ***using physics to make computation*** – analog computing -, etc) and their ***validation*** in product ready solutions.

# Major challenges

The Major Challenges (still) are :

- 1. ***Energy Efficiency***: Developing innovative hardware architectures and minimizing data movement are critical for energy-efficient computing systems. **Memory is becoming an important challenge** as we are moving from a computing centric paradigm to a data centric (driven by AI). **Zero standby energy and energy proportionality** to load is essential for edge devices.

New

- 2. ***System Complexity Management***: Addressing the complexity of embedded systems through interoperability, modularity, and ***dynamic resource allocation*** in a (***distributed***,) safe and secure way. ***Web technologies cascade to edge*** (containerization, WASM, protocols, …) forming a ***continuum of computing resources***

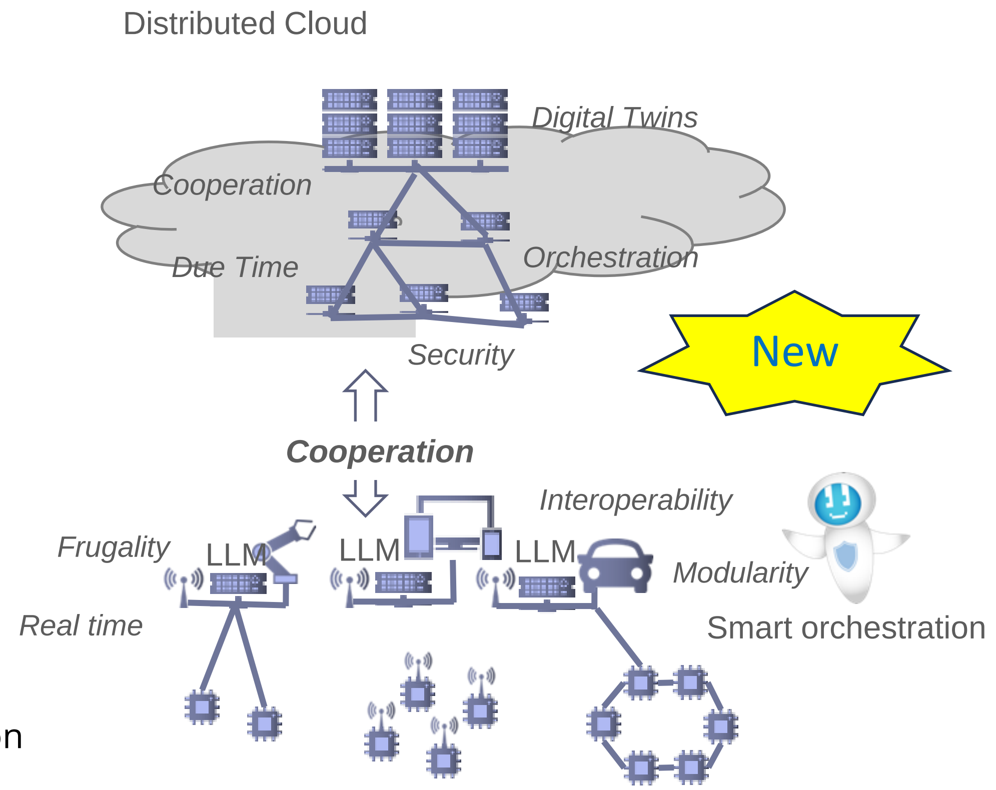
New

- 3. ***Lifespan of Devices***: Enhancing hardware support for software upgradability, interoperability, and second-life applications.

- 4. ***Sustainability***: Ensuring European sustainability by developing solutions aligned with ethical principles (for embedded AI) and transforming innovations into commercial successes (for example, based on open standards, such as Risc-V, and for innovative solutions such as neuromorphic computing)

# R&I focus areas

- Processing data locally and reducing data movements (towards the computing continuum)
- Co-design of algorithms, hardware and software
- Efficient management of storage resources
    - Unified memory
    - Innovations in memory technology
- Energy proportionality
- Ultra-low standby current
- Leveraging physical phenomena for computation
- Complexity management utilizing AI
- Decomposition of complex SoCs into chiplets and interposers
- Ensuring programmability and interoperability
- Combining processing devices to work together
- Modeling interactions among system components
- Ensuring long-term functionality and up to date operation
- Designing with modularity and extensibility in mind
- Reuse of components or systems in a downgraded or requalified use case
- Leveraging open-source hardware and software for innovation and cost reduction
- Developing and federating smaller specialized AI models
- Training models with European data and ethical compliance
- Deploying efficient and sustainable embedded AI-oriented ECS
- Accelerating development of robust, cost-effective solutions




LLM: Large Language Model