2.1

Cross-Sectional Technologies

EDGE COMPUTING AND EMBEDDED ARTIFICIAL INTELLIGENCE

2.1 Edge computing and embedded Artificial Intelligence

2.1.1 Summary

Edge computing and embedded AI are crucial for advancing digital technologies while addressing energy efficiency, system complexity, and sustainability. The integration of AI into edge devices offers significant benefits across various sectors, contributing to a more efficient and resilient digital infrastructure, keeping privacy by processing sensible data locally. Distributed computing forms a continuum from edge to cloud, with edge computing processing data close to its source to improve performance, reduce data transmission, latency and bandwidth, enhance safety, security and decrease global power consumption. This directly impacts the features of edge systems.

AI, especially embedded intelligence and Agentic AI, significantly influences various sectors such as productivity, environmental preservation, and transportation, enabling for example autonomous vehicles. The availability of new hardware technologies drives AI sustainability. Open-source initiatives are crucial for innovation, cost reduction, and security.

Embedded AI hardware was principally developed for perception tasks (vision, audio, signal processing) with high energy efficiency, but generative AI is emerging also at the edge, first fueled by smartphones and computers (Copilot+PC, Apple Intelligence) with the need to be able to process most of the data locally and adapting to the user's habits (fine tuning performed at the edge), and will extend to other edge applications (robotics, interfaces, high level perception of the environment). This drives new constraints not only for computing parts, but also to improve memory efficiency.

The Major Challenges 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.

2. System Complexity Management: Addressing the complexity of embedded systems through interoperability, modularity, and dynamic resource allocation in a safe and secure way. Web technologies cascade to edge (containerization, WASM, protocols, …) forming a continuum of computing resources. Using a federation of small models in a Mixture of Agents or Agentic AI instead of a very large model allows to better manage complexity and modularity while using less computing resources.

3. Lifespan of Devices: Enhancing hardware support for software upgradability, interoperability, and second-life applications. This will require hardware that can support future software updates, increasing memory capabilities, and communication stacks. Aggregation of various devices into a "virtual device" will allow older devices to be still useful in the pool.

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). Europe should master all steps for new AI technologies, especially the ones based on collaboration of AI agents.

2.1.2 Scope

This chapter focuses on computing components, and more specifically on embedded architectures, edge computing devices and systems using Artificial Intelligence (AI) at the edge. These elements rely on process technology and embedded software, and have constraints on quality, reliability, safety, and security. They also rely on system composition (systems of systems) and design and tools techniques to fulfill the requirements of the various application domains.

Furthermore, this chapter focuses on the trade-off between performances and power consumption reduction, and managing complexity (including security, safety, and privacy[1]) for embedded architectures to be used in different applications areas, which will spread edge computing and AI use and their contribution to European sustainability[2].

This chapter mainly covers the elements foreseen to be used to compose AI or edge systems:

- Processors (CPU) with high energy efficiency,
- Accelerators (for AI and for other tasks, such as security):

[1] Security, safety, and privacy will be covered in the Chapter about "Quality, reliability, safety and security"

[2] The scope of this Chapter is therefore to cover the hardware architectures and their realizations (Systems on Chip, Embedded architectures), mainly for edge and "near the user" devices such as IoT devices, cars, ICT for factories and local processing and servers (computing "on-premises"). Data centers and electronic components for data centers are not the main focus of the chapter, except when the components can be used in local processing units or local servers (local clouds, swarm, fog computing, etc.). We therefore also cover this "edge" side of the "continuum of computing" and the synergies with the cloud leading the following sections to discuss

- innovation which is needed to transfer HPC data center capabilities to edge and/or embedded environment, and
- innovation which directly addresses the embedded market (as opposed to technologies first developed for controlled, data center like environments, then later moved to embedded domain).

The technological aspects, at system level (PCB, assembly, system architecture, etc.), and embedded and application software are not part of this chapter as they are covered in other chapters. Software is important for these programmable or configurable embedded devices but will be handled in the "embedded software" chapter. However, we will still discuss architectures that efficiently execute software ("hardware friendly" software).

In the previous editions of the SRIA, this chapter dealt on Open Source and sustainability: these are transversal topics, to be addressed in the introductory chapter of the SRIA. In particular, some of the points currently covered by this chapter 2.1 regarding sustainability could be moved to other chapters:

- How to limit CO2 emissions incurred during ECS manufacturing is covered by chapters 1.1 and 1.2.
- Architectural approaches to limit energy consumption could be covered by chapter 2.3.

However, the RISC-V architecture and related IPs will still be covered in this chapter.

  - 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 for synaptic functions, and In/Near Memory Computing), etc.
- Power management.
- …[3]

In a nutshell, the main recommendation of this chapter is a paradigm shift towards distributed low power architectures/topologies, from cloud to edge (what is called the "*continuum of computing*") and using AI at the edge (but not necessarily only) leading to distributed intelligence.

2.1.3 Introduction

2.1.3.1 Positioning edge and cloud solutions

In the recent months, we saw more and more the emergence of what we can call the "continuum of computing". The "continuum of computing" is a paradigm shift that merges edge computing and cloud computing into a cohesive, synergistic system. Rather than opposing each other, these two approaches complement one another, working together to optimize computational efficiency and resource allocation. The main idea behind this continuum is to perform computation where it is most efficient. In its advanced evolution, the location of computing is not static, but dynamic allowing a smooth migration of tasks or services to where the best trade-off , according to criteria such as latency, bandwidth, cost, processing power and energy, are available. It also permits to create "virtual meta devices" by interconnecting various devices with various characteristics into a more global system where each individual device (or

[3] Of course, all the elements to build a SoC are also necessary, but not specifically in the scope of this chapter:

- Security infrastructure (e.g. Secure Enclave) with placeholder for customer-specific secure elements (PUF, cryptographic IPs…). Security requirements are dealt with details in the corresponding chapter. The appearance of LLMs / Generative AI calls for security measures, e.g. proof of origin/authenticity etc. will have an impact on the hardware. They should also run efficiently, in a protected environment without consuming too many resources.
- Field connectivity IPs (see connectivity Chapter, but the focus here is on field connectivity) (all kinds, wired, wireless, optical), ensuring interoperability.
- Integration using chiplet and interposer interfacing units are detailed in the chapters 1.1 and 1.2.
- And all other elements such as coherent cache infrastructure for many-cores, scratchpad memories, smart DMA, NoC with on-chip interfaces at router level to connect cores (coherent), memory (cache or not) and IOs (IO coherent or not), SerDes, high speed peripherals (PCIe controllers and switches, etc.), trace and debug hardware and low/medium speed peripherals (I2C, UART, SPI etc.).

part of a device) is executing a part of the global task. This has impact on various aspects of each device, for example increasing their lifetime by coupling them with others when they are not powerful enough to perform alone the requested task. Of course, this introduces complexity in the orchestration of the devices and the splitting of the global task into small sub-tasks, each device should be interconnected with secure protocol and authentication is key to ensure a trustable use of all devices into this federation.

An example for "continuum of computing" is Apple’s approach to AI, as illustrated by the announcement of Apple Intelligence. In this system, AI tasks are performed locally on the device when the local processing resources are sufficient. This ensures rapid responses and minimizes the need for data to be transmitted over networks, which can save energy and protect user privacy. For instance, simple AI tasks like voice recognition or routine summarizing, organizing text can be handled efficiently by the device’s onboard accelerators (NPU).

However, when a task exceeds the local device's processing capabilities, it is seamlessly offloaded to Apple's trusted cloud infrastructure. This hybrid approach allows for more complex computations to be performed without overwhelming the local device, ensuring a smooth user experience. Moreover, for tasks that require even more powerful AI, such as those involving large language models like Chat-GPT, the system can leverage the resources of cloud-based AI services. This layered strategy ensures that users benefit from the best possible performance and efficiency, irrespective of the complexity of the task.

The continuum of computing embodies a flexible, adaptive approach to resource utilization, ensuring that computational tasks are handled by the most appropriate platform[4]. This also has impact on the architecture of edge devices and processors, which need to be prepared to support the WEB and cloud protocols (and AI), with communication stacks, security and encryption, and containerization (executing “foreign” codes in secure sandboxes). We can observe that microcontrollers can now support the wired and wireless communication stacks (IP, Wifi, Bluetooth, Thread, …), have security IPs and encryption and can execute different OS on their different cores (e.g. a real-time OS and a Linux)[5].

[4] Some OS are designed for this kind of distribution, like HarmonyOS (https://developer.huawei.com/consumer/en/doc/harmonyos-guides-V3/harmonyos-overview-0000000000011903-V3) : « *The devices running HarmonyOS are aggregated at the system layer to form a super device, allowing flexible scaling of device hardware capabilities.*
With the support of HarmonyOS, users can integrate capabilities of their various smart devices, implementing ultra-fast connection, capability collaboration, and resource sharing among them. This way, services can be seamlessly transferred to the most suitable device, delivering smooth all-scenario experience. ». Some concepts are also developed in the open source Oniro project (https://oniroproject.org/)), also based on OpenHarmony (https://gitee.com/openharmony).

[5] The Chinese companies are pushing their low-cost MCU (often based on RISC-V architecture) with all these features, for example, the Sophgo SG2002 (https://en.sophgo.com/sophon-u/product/introduce/sg200x.html) or the Espressif ESP32-C6 (https://www.espressif.com/en/products/socs/esp32-c6).

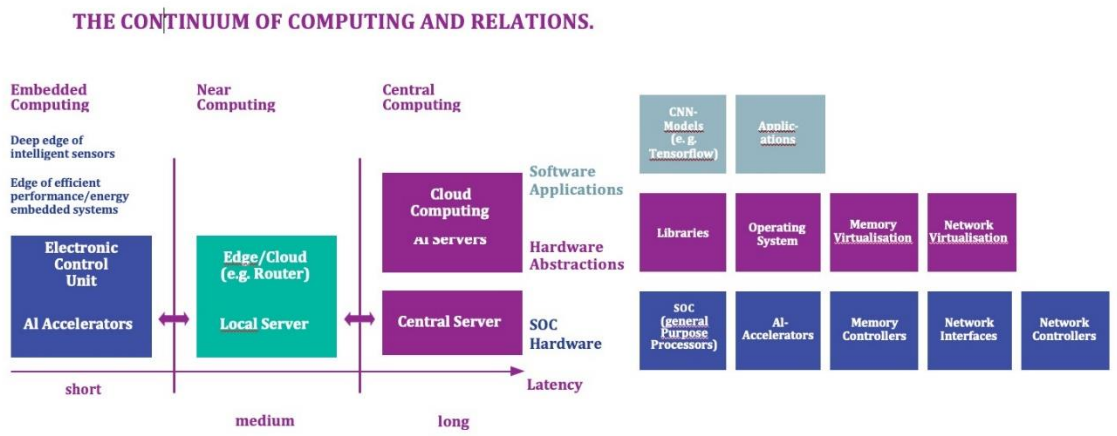


Figure 1: - The continuum of computing and relations between the elements constituting an embedded AI system (figure from Gerd Teepe)

Inside the "continuum of computing", edge computing involves processing data locally, close to where it is generated. This approach reduces latency, enhances real-time decision-making, and can improve data privacy and security by keeping sensitive information on local devices. Cloud computing, on the other hand, offers immense computational power and storage capacity, making it ideal for tasks that require extensive resources, such as large-scale data analysis and running complex AI models.

For intelligent embedded systems, the edge computing concept is reflected in the development of edge computing levels (micro, deep, meta) that covers the computing and intelligence continuum from the sensors/actuators, processing, units, controllers, gateways, on-premises servers to the interface with multi-access, fog, and cloud computing.

A description of the micro, deep and meta edge concepts is provided in the following paragraphs (as proposed by the AIoT community).

The **micro-edge** describes intelligent sensors, machine vision, and IIoT devices that generate insight data and are implemented using microcontrollers built around processors architectures such as ARM Cortex M4, or recently RISC-V, which are focused on minimizing costs and power consumption. The distance from the data source measured by the sensors is minimized. The compute resources process this raw data in line and produce insight data with minimal latency. The hardware devices of the micro-edge physical sensors/actuators generate from raw data insight data and/or actuate based on physical objects by integrating AI-based elements into these devices and running AI-based techniques for inference and self-training.

Intelligent micro-edge allows IoT real-time applications to become ubiquitous and merged into the environment where various IoT devices can sense their environments and react fast and intelligently with an excellent energy-efficient gain. Integrating AI capabilities into IoT devices significantly enhances their functionality, both by introducing entirely new capabilities, and, for example, by replacing accurate algorithmic implementations of complex tasks with AI-based approximations that are better embeddable. Overall, this can improve performance, reduce latency, and power consumption, and at the same time increase the devices usefulness, especially when the full power of these networked devices is harnessed – a trend called AI on edge.

The **deep-edge** comprises intelligent controllers PLCs, SCADA elements, connected machine vision embedded systems, networking equipment, gateways and computing units that aggregate data from the sensors/actuators of the IoT devices generating data. Deep edge processing resources are implemented with performant processors and microcontrollers such as Intel i-series, Atom, ARM M7+, etc., including CPUs, GPUs, TPUs, and ASICs. The system architecture, including the deep edge, depends on the envisioned functionality and deployment options considering that these devices' cores are controllers: PLCs, gateways with cognitive capabilities that can acquire, aggregate, understand, react to data, exchange, and distribute information.

The **meta-edge** integrates processing units, typically located on-premises, implemented with high-performance embedded computing units, edge machine vision systems, and edge servers (e.g. high-performance CPUs, GPUs, FPGAs, etc.) that are designed to handle compute-intensive tasks, such as processing, data analytics, AI-based functions, networking, and data storage.

This classification is closely related to the distance between the data source and the data processing, impacting overall latency. A high-level rough estimation of the communication latency and the distance from the data sources are as follows. With micro-edge the latency is below 1millisecond (ms), and the distances are from zero to max 15 meters (m). For deep-edge distances are under 1 km and latency below 2-5 ms, meta-edge shows latencies of under 10 ms and distances under 50 km, and up to 50 km (also) for fog computing. MEC concepts are combined with near-edge, with 10-20 ms latency and 100 km distance, while far-edge is 20-50ms and 200 km, and cloud and data centers are more than 50 ms and 1000 km.

	Latency	Distance
Micro-edge	Below 1ms	From 0 cm to 15 m
Deep-edge	Below 2-5 ms	Below 1km
Meta-edge	Below 10 ms	Below 50 km
Fog	10-20 ms	Up to 50 km
MEC[6] + near-edge	10-20 ms	100 km
Far-edge	20-50 ms	200 km
Cloud/data centres/HPC	More than 50 -100 ms	1000 km and beyond

[6] Multi-access edge computing (ETSI/ISG) : Multi-access edge computing (MEC) brings technology resources closer to the end user. Data is processed and stored at the network's edge, not at some distant data center, significantly reducing latency.

Deployments "at the edge" can contribute, thanks to its flexibility, to be adapted to the specific needs, to provide more energy-efficient processing solutions by integrating various types of computing architectures at the edge (e.g. neuromorphic, energy-efficient microcontrollers, AI processing units), reduce data traffic, data storage and the carbon footprint. One way to reduce the energy consumption is to know which data and why it is collected, which targets are achieved and to optimize all levels of processes, both at hardware and software levels, to achieve those targets, and finally to evaluate what is consumed to process the data.

In general, the edge (in the peripheral of a global network as the Internet) includes compute, storage, and networking resources, at different levels as described above, that may be shared by several users and applications using various forms of virtualization and abstraction of the resources, including standard APIs to support interoperability.

More specifically, an edge node covers the edge computing, communication, and data analytics capabilities that make it smart/intelligent. An edge node is built around the computing units (CPUs, GPUs/FPGAs, ASICs platforms, AI accelerators/processing), communication network, storage infrastructure and the applications (workloads) that run on it.

The edge can scale to several nodes, distributed in distinct locations and the location and the identity of the access links is essential. In edge computing, all nodes can be dynamic. They are physically separated and connected to each other by using wireless/wired connections in topologies such as mesh. The edge nodes can be functioning at remote locations and operate semi-autonomously using remote management administration tools.

The edge nodes are optimized based on the energy, connectivity, size, cost, and their computing resources are constrained by these parameters. In different application cases, it is required to provide isolation of edge computing from data centers in the cloud to limit the cloud domain interference and its impact on edge services.

Finally, the edge computing concept supports a dynamic pool of distributed nodes, using communication on partially unreliable network connections while distributing the computing tasks to resource-constrained nodes across the network.

2.1.3.2 Positioning Embedded Artificial Intelligence

Even if there will be a lot of applications that will not use AI, this field is currently on top of the hype curve and more and more embedded systems will be compatible with AI requirements. We can consider that the AI hardware landscape could be segmented into three categories, each defined by the processing power and application domains. These categories are Cloud AI, Embedded AI for high-performance needs, and Embedded AI for low-power applications. They also reflect the “continuum of computing” and computation should be done on all the 3 segments, where it is the most efficient according to a particular set of KPI (such a latency, energy cost, privacy preserving, communication bandwidth, global cost, …).

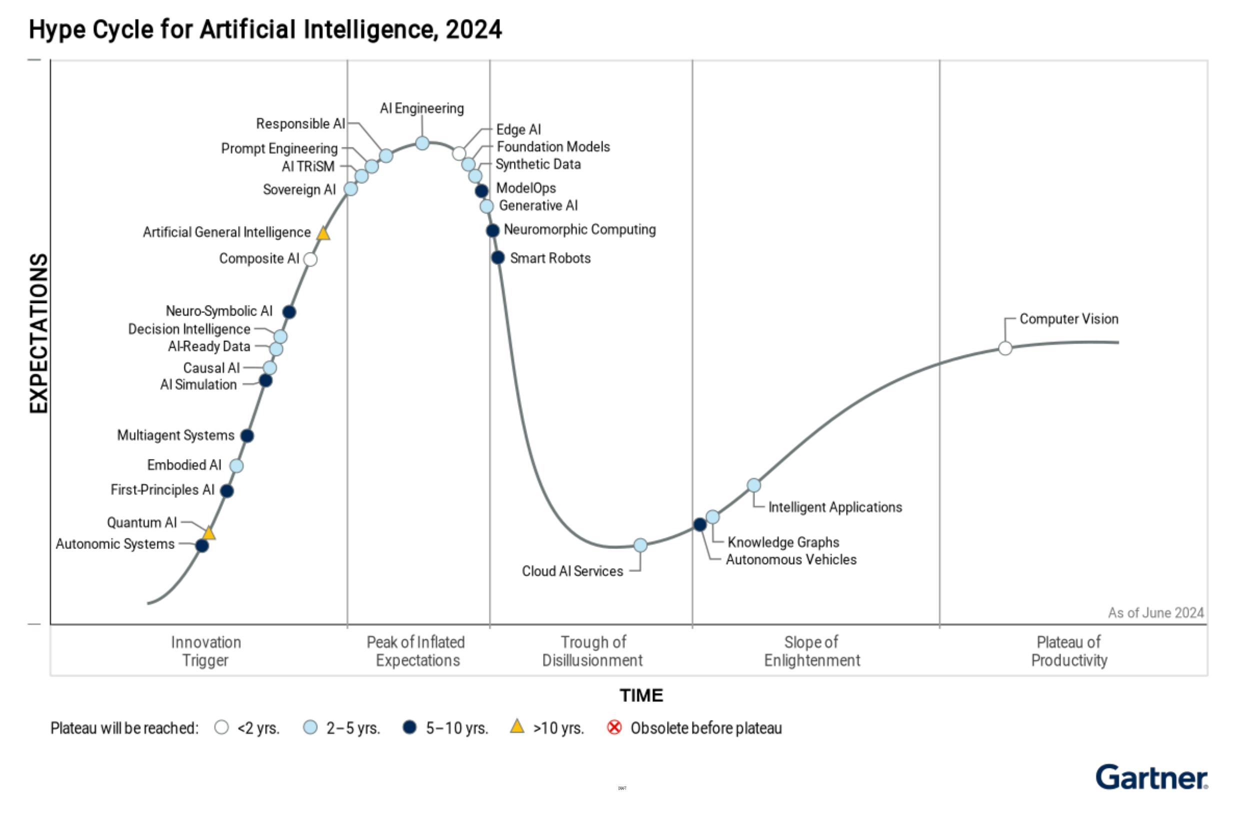


Figure 2: Hype Cycle for Emerging Technologies[7]

We can use the computing efficiency to differentiate the categories, but this is quite subjective and can change over time, but here is a classification into 3 clusters.

1. Cloud AI (~1 to 10 TOPS/W):

Cloud AI represents the most concentration of AI processing power, leveraging GPUs capable of exceeding 2000 TOPS (Tera Operations Per Second) and large language models (LLMs) with over 4 billion parameters. This category is primarily utilized in data centers and servers, where the focus is on both training and inferencing tasks. A system (for example in the Green500 system can reach efficiency up to more than 70 GFLOPS/W in FP32 and 4 PFLOPS for the FP8 on tensor cores, or roughly 4 TFLOPS/W in FP8).

2. Embedded AI (1 to ~50 TOPS/W):

Embedded AI in the range of 1 to 100 TOPS/W focuses on bringing powerful AI capabilities closer to the source of data generation. This category employs Neural Processing Units (NPUs) with performance ranging from 40 to 2000 TOPS with multiple GB of RRAM and smaller language models (SLMs) with 1

[7] *2024 from https://www.gartner.com/doc/reprints?id=1-2HV4OPON&ct=240618&st=sb*

million to 7 billion parameters. These systems are typically embedded in notebooks/PCs (for example, Copilot+PC from Microsoft[8]), smartphones (for example Apple Intelligence[9]) and will be more and more used in automotive, robotics, factory automation, etc. It is also where we found the majority of “perceptive AI”, used to directly analyse images, video, sound or signals, and typical AI processing methods are using Convolutional Neural Networks (CNNs) or related approaches.

For example, for automotive ADAS, Embedded AI hardware provides the necessary computational power to process sensor data in real-time, enabling advanced features like object detection, lane keeping, and adaptive cruise control. In consumer electronics such as notebooks and smartphones, Embedded AI enhances user experiences through features like voice recognition and generation, document analysis and synthesis, and intelligent photography. Additionally, in networking equipment like GPON and AP routers, Edge AI facilitates smarter data management and enhanced security measures.

3. Deep Embedded AI (>10 TOPS/W):

This category of Embedded AI is designed for applications requiring high efficiency and lower computational power. To reach high efficiency, the hardware is more specialized (ASIC), involving for example the use of Compute Near Memory (CNM) or Compute-In-Memory (CIM) architectures, and models like Convolutional Neural Networks (CNNs) or Spiking Neural Networks (SNNs) with far fewer than 4 million parameters. The processing techniques are simpler, not using Transformer based Neural Networks, but more CNNs, Bayesian approaches etc.

The scope of this chapter is on Embedded AI and Deep Embedded AI.

2.1.4 State of the Art

The key issues to the digital world are the availability of affordable computing resources and transfer of data to the computing node with an acceptable power budget. Computing systems are morphing from classical computers with a screen and a keyboard to smart phones and to deeply embedded systems in the fabric of things. This revolution on how we now interact with machines is mainly due to the advance in AI, more precisely of machine learning (ML) that allows machines to comprehend the world not only on the basis of various signal analysis but also on the level of cognitive sensing (vision and audio). Each computing device should be as efficient as possible and decrease the amount of energy used.

Low-power neural network accelerators will enable sensors to perform online, continuous learning and build complex information models of the world they perceive. Neuromorphic technologies such as spiking neural networks and compute-in-memory architectures are compelling choices to efficiently process and fuse streaming sensory data, especially when combined with event-based sensors. Event-based sensors, like the so-called retinomorphic cameras, are becoming extremely important especially in the case of edge computing where energy could be a very limited resource. Major issues for edge systems, and even more for AI-embedded systems, is energy efficiency and energy management. Implementation of intelligent

[8] https://blogs.microsoft.com/blog/2024/05/20/introducing-copilot-pcs/

[9] https://machinelearning.apple.com/research/introducing-apple-foundation-models

power/energy management policies are key for systems where AI techniques are part of processing sensor data and power management policies are needed to extend the battery life of the entire system.

As extracting useful information should happen on the (extreme) edge device, personal data protection must be achieved by design, and the amount of data traffic towards the cloud and the edge-cloud can be reduced to a minimum. Such intelligent sensors not only recognize low-level features but will be able to form higher level concepts as well as require only very little (or no) training. For example, whereas digital twins currently need to be hand-crafted and built bit-for-bit, so to speak, tomorrow's smart sensor systems will build digital twins autonomously by aggregating the sensory input that flows into them.

To achieve intelligent sensors with online learning capabilities, semiconductor technologies alone will not suffice. Neuroscience and information theory will continue to discover new ways[10] of transforming sensory data into knowledge. These theoretical frameworks help model the cortical code and will play an important role towards achieving real intelligence at the extreme edge.

Compared to the previous SRIA, as explained in the previous part, we can observe two new directions:

- The recent explosion of AI, and more precisely of Generative AI drove development of new solutions for embedded systems (for now mainly PCs and smartphones), which a potential very large market growth.
- RISC-V is coming to various range of products, and especially from ICs from China that support a lot a feature for connectivity (security), IOs and accelerators. Most of advanced MCU also have a small core, mostly always on, in charge to wake-up the rest of the system in case of event.

AI accelerator chips for embedded market have traditionally been designed to support convolutional neural networks (CNNs), which are particularly effective for image, audio, and signal analysis. CNNs excel at recognizing patterns and features within data, making them the backbone of applications such as facial recognition, speech recognition, and various forms of real-time data and signal processing[11]. These tasks require significant computational power and efficiency, which dedicated AI chips have been able to provide with new and specialized architectures (embedded NPU). **Most of current accelerators available today in edge computing systems are used and designed for perception applications and further developments are required to have further gains in efficiency.**

Generative AI at the edge (mainly for Embedded AI): the rise of (federation of) smaller models

New deep learning models are introduced at an increasing rate and one of the recent ones, with large applications potential, are transformers, which are the basis of LLMs. Based on the attention model[12], it is a "sequence-to-sequence architecture" that transforms a given sequence of elements into another sequence. Initially used for NLP (Natural Language Processing), where it can translate one sequence in a first language into another one, or complement the beginning of a text with potential follow-up, it is now extended to other domains such as video processing or elaborating a sequence of logical steps for robots. It is also a self-supervised approach: for learning it does not need labelled examples, but only part of the

[10] Even though our understanding of how the brain computes is still in its infancy, important breakthroughs in cortical (column) theory have been achieved in the last decade.

[11] For example, https://greenwaves-technologies.com/

[12] https://arxiv.org/abs/1706.03762

sequence, the remaining part being the "ground truth". The biggest models, such as GPT3, are based on this architecture. GPT3 was in the spotlights in May 2020 because of its potential use in many different applications (the context being given by the beginning sequence) such as generating new text, summarizing text, translating text, answering to questions and even generating code from specifications. This was even amplified by GPT4, and all those capabilities were made visible to the public in November 2022 with Chat-GPT, which triggered a maximum of hype and expectations. Even if today transformers are mainly used for cloud applications, this kind of architecture is rippling down in embedded system. Small to medium size (1, 3, 7 to 13 G parameters models) can be executed on single board computers such as Jetson Orin nano and even Raspberry PI. Quantization is a very important process to reduce the memory footprint of those models and 4-bit LLMs performs rather well. The new GPUs of NVIDIA support float8 in order to efficiently implement transformers. Supporting LLMs in a low-power and efficient way on edge devices is a new important challenge. However, most of these models are too big to run at the edge.

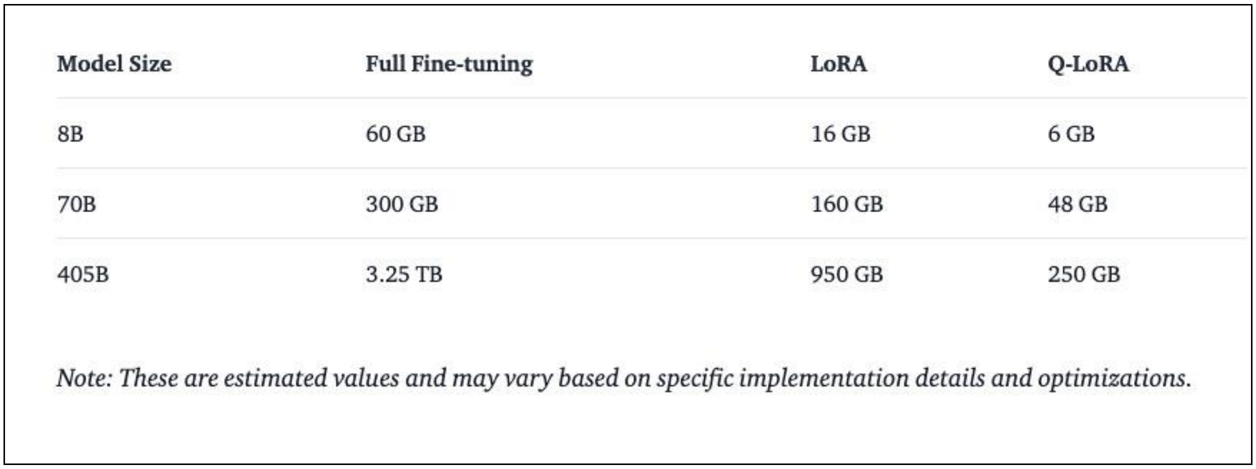
But we observe a trend in using several smaller specialized generative AI working together which could give comparable results than large monolithic LLMs. For example, the concept of Agentic AI, Mixture of Agents (MoA)[13], where a set of smaller Large Language Models (LLMs) works collaboratively. Smaller specialized LLMs can be trained and fine-tuned for specific tasks or domains. This specialization allows each model to become highly efficient and accurate in its particular area of expertise. By dividing the workload among these specialized agents[14], the system can leverage the strengths of each model, achieving a level of performance and accuracy that rivals or even surpasses that of a single, very large LLM. This targeted approach not only enhances the precision of the responses but also reduces the computational overhead associated with training and running a monolithic model. The MoA framework optimizes resource utilization by distributing tasks dynamically based on the specific strengths of each agent. This means that instead of overloading a single model with diverse and potentially conflicting requirements, the system can route tasks to the most suitable agent. Such an arrangement ensures that computational resources are used more efficiently, as each agent processes only the type of data it is best equipped to handle, and the other don't consume processing power, hence energy. Moreover, the MoA approach enhances scalability and maintainability and is more suited for edge devices, or a network of edge devices.

Training and updating a single enormous LLM is a complex and resource-intensive process, often requiring extensive computational power and time only available in the cloud or in large data centers. In contrast, updating smaller models can be more manageable and less resource-demanding. Additionally, smaller models can be incrementally improved or replaced without disrupting the entire system. This modularity allows for continuous enhancement and adaptation to new data and tasks, ensuring that the system remains current and effective over time. Furthermore, MoA systems inherently offer robustness and fault tolerance: with multiple smaller agents, the failure of one model does not cripple the entire system. This MoA (or similar) are very suited for Edge AI by distributing the specialized models in different (parts of) systems, and having them interconnected (hence the requirement for edge hardware to be have connectivity stack integrated, with all security requirements).

[13] https://www.together.ai/blog/together-moa

[14] For example, using appraoches like RouteLLM, see https://lmsys.org/blog/2024-07-01-routellm/

Concerning the training of large foundation models, it is clear that this is only possible now in cloud AI due to the very large training dataset and computing power required to train foundation models. However, there are at least two other ways to adapt the foundation model to a particular (local) context: fine tuning and large token context. Fine-tuning involves taking a pre-trained foundation model and further training it on a specific, smaller dataset that reflects the target application's domain or context. This process adjusts a very small proportion of the model's weights (using LoRA or Adapters[15]) based on the new data, allowing it to specialize and perform more accurately in the given local context. This can be done on the edge with local data, for example in PC or smartphone which have sufficiently memory for storing the training data, RAM for running the fine-tuning process and of course processing power. This can be done when the system is idle, e.g. during the night for PCs or smartphone.



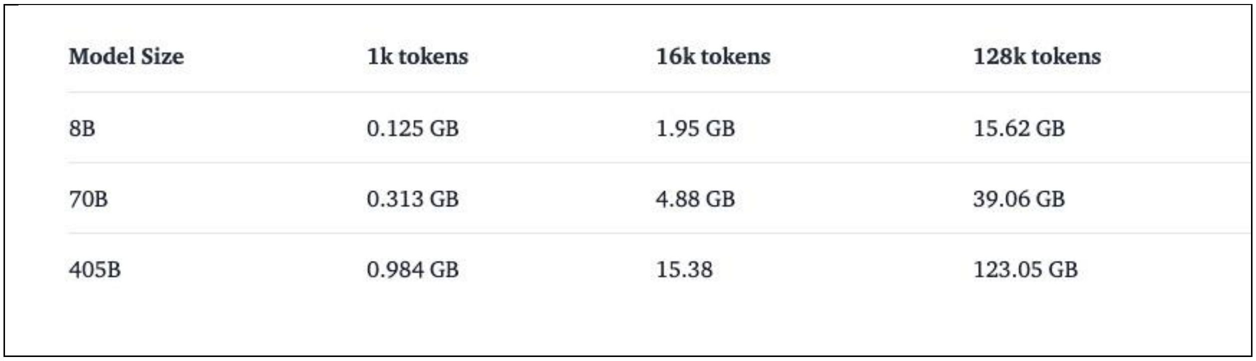
Model Size	Full Fine-tuning	LoRA	Q-LoRA
8B	60 GB	16 GB	6 GB
70B	300 GB	160 GB	48 GB
405B	3.25 TB	950 GB	250 GB

Note: These are estimated values and may vary based on specific implementation details and optimizations.

Figure 3: This table outlines the approximate memory requirements for training Llama 3.1 models using different techniques, from https://huggingface.co/blog/llama31

Leveraging a large token context refers to the model's ability to process and understand long sequences of text, providing it with a broader context window. This extended context allows the model to capture more nuanced dependencies and relationships within the data, improving its comprehension and relevance to specific local contexts. By accommodating a larger context, the model can maintain coherence and produce more accurate and contextually appropriate outputs over longer spans of text, making it particularly useful for applications that require in-depth understanding and continuity, such as detailed document summarization or extended conversational AI systems. Together, these methods enable foundation models to be effectively adapted and optimized for specialized applications, enhancing their utility and performance in specific scenarios and they can be usable (in a near future) at the Edge in Embedded AI systems. Enlarging the context mainly implies more memory, as shown in Figure 4.

[15] https://github.com/AGI-Edgerunners/LLM-Adapters



Model Size	1k tokens	16k tokens	128k tokens
8B	0.125 GB	1.95 GB	15.62 GB
70B	0.313 GB	4.88 GB	39.06 GB
405B	0.984 GB	15.38	123.05 GB

Figure 4: requirements for the KV cache memory of Llama 3.1 model s(FP16), from https://huggingface.co/blog/llama31

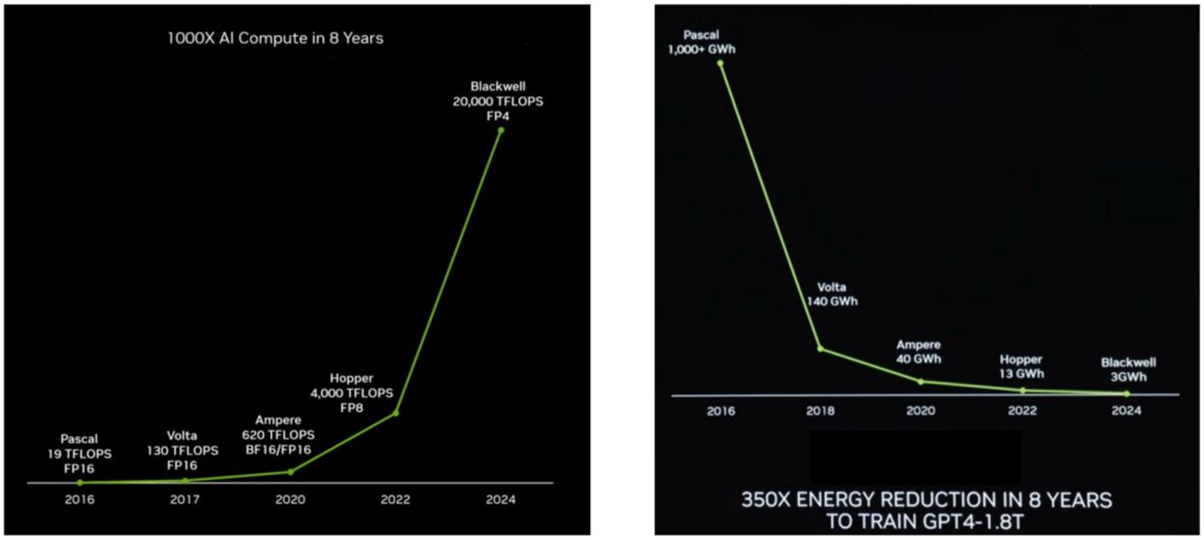
Energy efficiency of AI training:

Training AI models can be very energy demanding. As an example, according to a recent study, the model training process for natural-language processing (NLP, that is, the sub-field of AI focused on teaching machines to handle human language) could end emitting as much carbon as five cars in their lifetimes[16], [17]. However, if the inference of that trained model is executed billions of times (e.g. by billion users' smartphones), its carbon footprint could even offset the training one. Another analysis[18], published by the OpenAI association, unveils a dangerous trend: "since 2012, the amount of compute used in the largest AI training runs has been increasing exponentially with a 3.5 month-doubling time (by comparison, Moore's law had a 2-years doubling period)". These studies reveal that the need for computing power (and associated power consumption) for training AI models is dramatically widening. Consequently, the AI training processes need to turn greener and more energy efficient.

[16] https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/

[17] However, technology and algorithms improved a lot over time, and training a similar model (Bloom) 3 years later required 21 times less CO_2 emissions.

[18] https://openai.com/index/ai-and-compute/



From Nvidia, Computex 2024

From https://youtu.be/pKXDVsWZmUU?t=3981

Figure 5: Increase of performance for AI

In fact, we can say that it is the available performance in operation per Watt of AI accelerators (GPUs) that drive the evolution of AI based on neural networks, including generative AI: the Figure 5 shows that GPT-4, launched on March 14, 2023, was only possible because the cost of the energy to train it in 2022 was acceptable and it would not have -been possible to train it before. the Figure 5 also shows that GPU performances increased by 3 decades in 8 years, both thanks to new architecture, reduced data type (from FP16 to FP4) and smaller technology nodes. In this same period of time, the energy reduction was 350x.

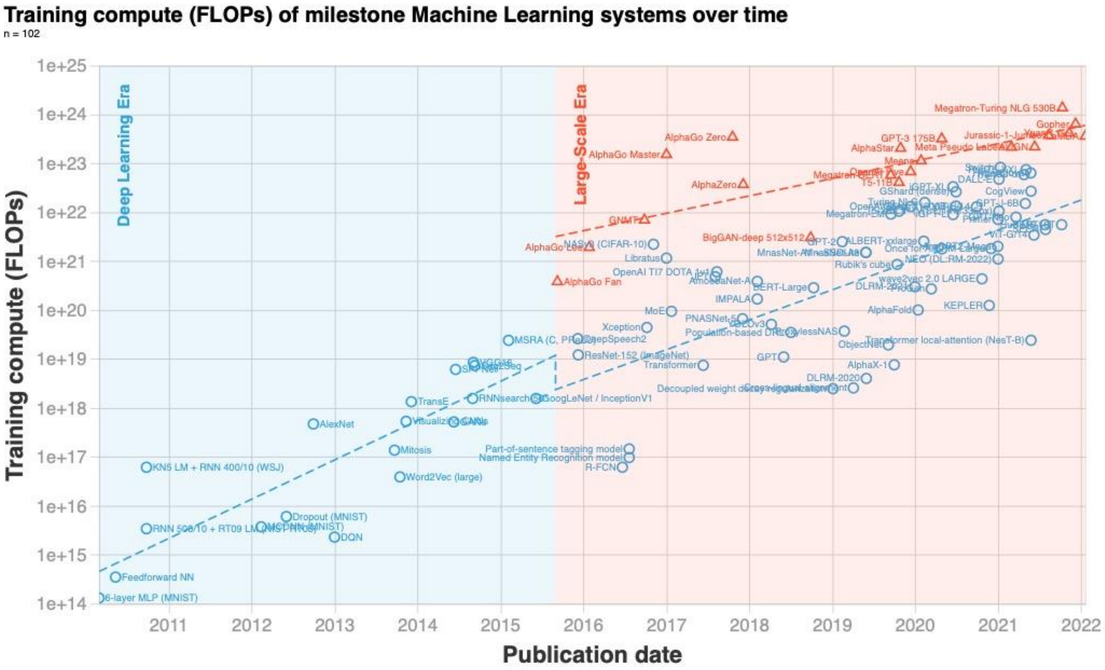


Figure 6:Evolution of the size of the most advanced deep learning networks (from https://arxiv.org/abs/2202.05924)

For a given use-case, the search for the optimal solution should meet multi-objective trade-offs among accuracy of the trained model, its latency, safety, security, and the overall energy cost of the associated solution. The latter means not only the energy consumed during the inference phase but also considering the frequency of use of the inference model and the energy needed to train it.

In addition, novel learning paradigms such as transfer learning, federated learning, self-supervised learning, online/continual/incremental learning, local and context adaptation, etc., should be preferred not only to increase the effectiveness of the inference models but also as an attempt to decrease the energy cost of the learning scheme. Indeed, these schemes avoid retraining models from scratch all the times or reduce the number and size of the model parameters to transmit back and forth during the distributed training phase.

It is also important to be able to support LLMs at the edge, in a low-cost and low-energy way, to benefit from their features (natural language processing, multimodality, few shot learning, etc..). Applications using transformers (such as LLMs) can run with 4 bit – or less - for storing each parameter, allowing to reduce the amount of memory required to use them in inference mode.

Although significant efforts have been focused in the past to enable ANN-based inference on less powerful computing integrated circuits with lower memory size, today, a considerable challenge to overcome is that non-trivial DL-based inference requires significantly more than the 0.5-1 MB of SRAM, that is the typical memory size integrated on top of microcontroller devices. Several approaches and methodologies to artificially reduce the size of a DL model exist, such as quantizing the neural weights and biases or pruning the network layers. These approaches are fundamental also to reduce the power consumption of the inference devices, but clearly, they cannot represent the definitive solution of the future.

We witness great development activity of computing systems explicitly supporting novel AI-oriented use cases, spanning different implementations, from chips to modules and systems. Moreover, as depicted in the following figure, it covers large ranges of performance and power, from high-end servers to ultra-low power IoT devices.

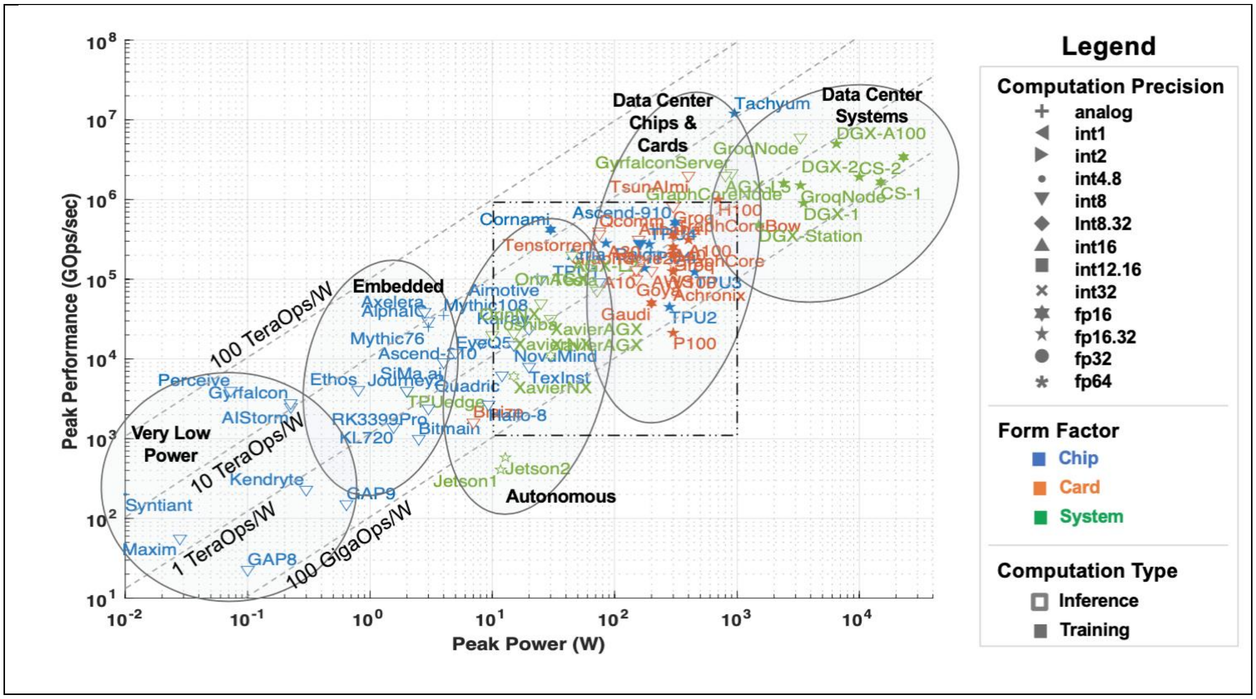


Figure 7: Landscape of AI chips according to their peak power consumption and peak performance[19]*.*

To efficiently support new AI-related applications, for both, the server and the client on the edge side, new accelerators need to be developed. For example, DL does not usually need a 32/64/128-bit floating point for its learning phase, but rather variable precision including dedicated formats such as bfloats. However, a close connection between the compute and storage parts are required (Neural Networks are an ideal "compute in memory" approach). Storage also needs to be adapted to support AI requirements (specific data accesses, co-location compute and storage), memory hierarchy, local vs. cloud storage. This is particularly important for LLMs which (still) have a large number of parameters (few billions) to be efficient. Quantization into 4 to 2 bits, new memories and clever architectures are required for their efficient execution at the edge.

Similarly, at the edge side, accelerators for AI applications will particularly require real-time inference, in view to reduce the power consumption. For DL applications, arithmetic operations are simple (mainly multiply-accumulate) but they are done on data sets with a very large set of data and the data access is therefore challenging. In addition, clever data processing schemes are required to reuse data in the case

[19] AI Accelerator Survey and Trends, Albert Reuther, Peter Michaleas, Michael Jones, Vijay Gadepally, Siddharth Samsi, Jeremy Kepner, October 2022 https://arxiv.org/abs/2210.04055

of convolutional neural networks or in systems with shared weights. Computing and storage are deeply intertwined. And of course, all the accelerators should fit efficiently with more conventional systems.

Reducing the size of the neural networks and the precision of computation is key to allow complex deep neural networks to run on embedded devices. This can be achieved either by pruning the topology of the networks, and/or by reducing the number of bits storing values of weight and neuron values. These processes can be done during the learning phase, or just after a full precision learning phase, or can be done (with less performance) independently of the learning phase (example: post-training quantization). The pruning principle is to eliminate nodes that have a low contribution to the final result. Quantization consists either in decreasing the precision of the representation (from float32 to float16 or even float8, as supported by the NVIDIA GPUs mainly for transformer networks), or to change the representation from float to integers. For the inference phase, current techniques allow to use 8-bit representations with a minimal loss of performances, and sometimes to reduce the number of bits further, with an acceptable reduction of performance or small increase of the size of the network (LLMs still seem to have a good performance with a 4-bit quantization). Most major developments environments (TensorFlow Lite[20], Aidge[21], etc.) support post-training quantization, and the Tiny ML community is actively using it. Supporting better tools and algorithms to reduce size and computational complexity of Deep Neural Networks is of paramount importance for allowing efficient AI applications to be executed at the edge.

Fixing and optimizing some parts of the processing (for example feature extraction for CNNs) leads to specialized architectures with very high-performance, as exemplified in the ANDANTE project.

[20] https://www.tensorflow.org/lite/performance/post_training_quantization

[21] https://projects.eclipse.org/projects/technology.aidge

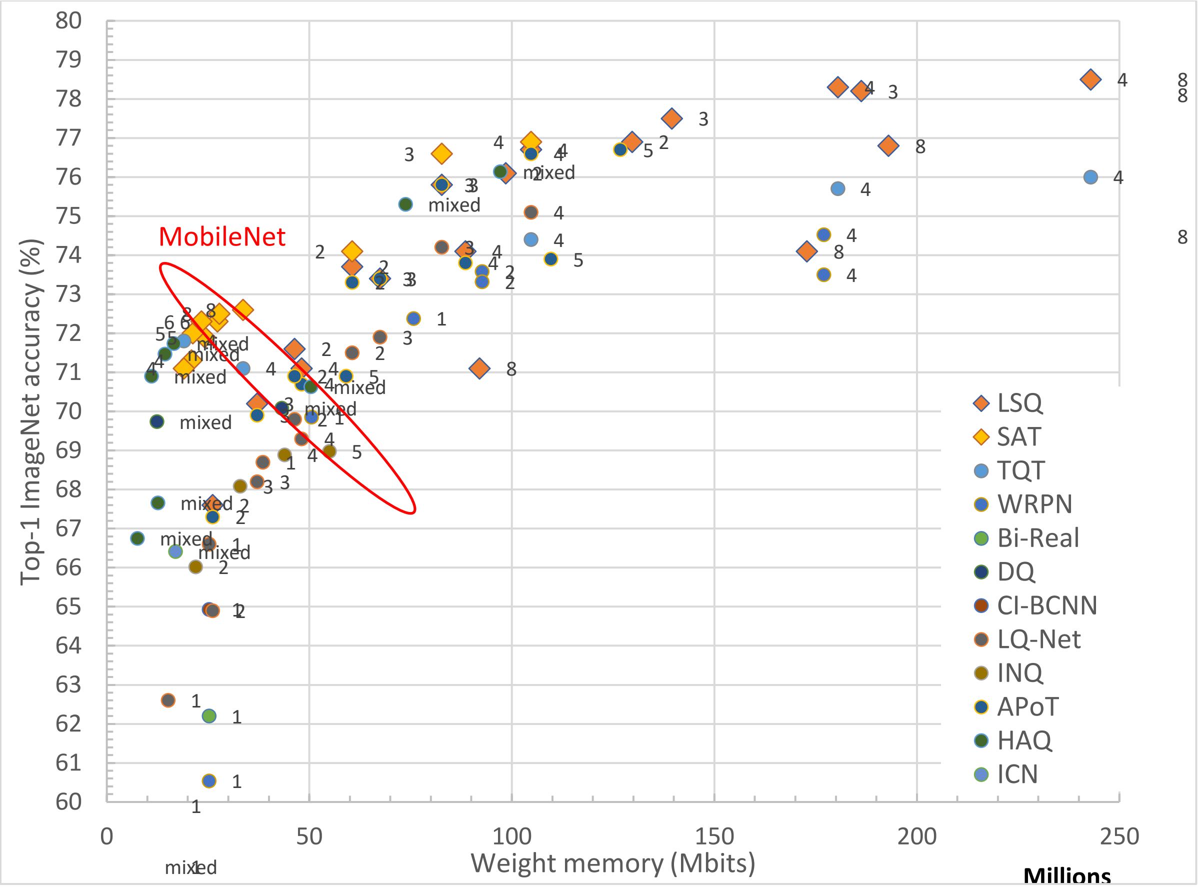


Figure 8: Results of various quantization methods versus Top-1 ImageNet accuracy

2.1.4.1 Impact of AI and embedded intelligence in sustainable development

Recently, the attention to the identification of sustainable computing solutions in modern digitalization processes has significantly increased. Climate changes and an initiative like the European Green Deal[22] are generating more sensitivity to sustainability topics, highlighting the need to always consider the technology impact on our planet, which has a delicate equilibrium with limited natural resources[23]. The computing approaches available today, as cloud computing, are in the list of the technologies that could potentially lead to unsustainable impacts. A study[24] has clearly confirmed the importance of edge computing for sustainability but, at the same time, highlighted the necessity of increasing the emphasis on sustainability, remarking that "research and development should include sustainability concerns in

[22] https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_en

[23] Nardi, B., Tomlinson, B., Patterson, D.J., Chen, J., Pargman, D., Raghavan, B., Penzenstadler, B.: Computing within limits. Commun. ACM. 61, 86–93 (2018)

[24] Hamm, Andrea & Willner, Alexander & Schieferdecker, Ina. (2020). Edge Computing: A Comprehensive Survey of Current Initiatives and a Roadmap for a Sustainable Edge Computing Development. 10.30844/wi_2020_g1-hamm.

their work routine" and that "sustainable developments generally receive too little attention within the framework of edge computing". The study identifies three sustainability dimensions (societal, ecological, and economical) and proposes a roadmap for sustainable edge computing development where the three dimensions are addressed in terms of security/privacy, real-time aspects, embedded intelligence and management capabilities.

AI and particularly embedded intelligence, with its ubiquity and its high integration level having the capability "to disappear" in the environment (ambient intelligence), is significantly influencing many aspects of our daily life, our society, the environment, the organizations in which we work, etc. AI is already impacting several heterogeneous and disparate sectors, such as companies' productivity[25], environmental areas like nature resources and biodiversity preservation[26], society in terms gender discrimination and inclusion[27] [28], smarter transportation systems[29], etc. just to mention a few examples. The adoption of AI in these sectors is expected to generate both positive and negative effects on the sustainability of AI itself, of the solutions based on AI and on their users[30] [31]. It is difficult to extensively assess these effects and there is not, to date, a comprehensive analysis of their impact on sustainability. A study[32] has tried to fill this gap, analyzing AI from the perspective of 17 Sustainable Development Goals (SDGs) and 169 targets internationally agreed in the 2030 Agenda for Sustainable Development[33]. From the study it emerges that AI can enable the accomplishment of 134 targets, but it may also inhibit 59 targets in the areas of society, education, health care, green energy production, sustainable cities, and communities.

From a technological perspective AI sustainability depends, at first instance, on the availability of new hardware and software technologies. From the application perspective, automotive, computing and healthcare are propelling the large demand of AI semiconductor components and, depending on the application domains, of components for embedded intelligence and edge AI. This is well illustrated by car factories being on hold because of the shortage of electronic components. Research and industry organizations are trying to provide new technologies that lead to sustainable solutions redefining

[25] Acemoglu, D. & Restrepo, P. Artificial Intelligence, Automation, and Work. NBER Working Paper No. 24196 (National Bureau of Economic Research, 2018). https://futurium.ec.europa.eu/en/connect-university/events/next-computing-paradigm-hipeac-2024

[26] Norouzzadeh, M. S. et al. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proc. Natl Acad. Sci. USA 115, E5716–E5725 (2018).

[27] Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V. & Kalai, A. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Adv. Neural Inf. Process. Syst. 29, 4349–4357 (2016).

[28] Tegmark, M. Life 3.0: Being Human in the Age of Artificial Intelligence (Random House Audio Publishing Group, 2017)

[29] Adeli, H. & Jiang, X. Intelligent Infrastructure: Neural Networks, Wavelets, and Chaos Theory for Intelligent Transportation Systems and Smart Structures (CRC Press, 2008).

[30] Jean, N. et al. Combining satellite imagery and machine learning to predict poverty. Science (80-.) 353, 790–794 (2016)

[31] Courtland, R. Bias detectives: the researchers striving to make algorithms fair. Nature 558, 357–360 (2018).

[32] Vinuesa, R., Azizpour, H., Leite, I. et al. The role of artificial intelligence in achieving the Sustainable Development Goals. Nat Commun 11, 233 (2020).

[33] UN General Assembly (UNGA). A/RES/70/1Transforming our world: the 2030 Agenda for Sustainable Development. Resolut 25, 1–35 (2015).

traditional processor architectures and memory structure. We already saw that computing near, or in-memory, can lead to parallel and high-efficient processing to ensure sustainability.

The second important component of AI that impacts sustainability concerns software and involves the engineering tools adopted to design and develop AI algorithms, frameworks, and applications. The majority of AI software and engineering tools adopt an open-source approach to ensure performance, lower development costs, time-to-market, more innovative solutions, higher design quality and software engineering sustainability. However, the entire European community should contribute and share the engineering efforts at reducing costs, improving the quality and variety of the results, increasing the security and robustness of the designs, supporting certification, etc.

The report on "Recommendations and roadmap for European sovereignty on open-source hardware, software and RISC-V Technologies"[34] discusses these aspects in more details.

Eventually, open-source initiatives (being so numerous, heterogeneous, and adopting different technologies) provide a rich set of potential solutions, allowing to select the most sustainable one depending on the vertical application. At the same time, open source is a strong attractor for applications developers as it gathers their efforts around the same kind of solutions for given use cases, democratizes those solutions and speeds up their development. However, some initiatives should be developed, at European level, to create a common framework to easily develop different types of AI architectures (CNN, ANN, SNN, LLM, etc.). This initiative should follow the examples of GAMAM (Google, Amazon, Meta, Apple, Microsoft). GAMAM have greatly understood its value and elaborated business models in line with open source, representing a sustainable development approach to support their frameworks[35]. It should be noted that open-source hardware should not only cover the processors and accelerators, but also all the required infrastructure IPs to create embedded architectures. It should be ensured that all IPs are interoperable and well documented, are delivered with a verification suite, and remain maintained constantly to keep up with errata from the field and to incorporate newer requirements. The availability of automated SoC composition solutions, allowing to build embedded architectures design from IP libraries in a turnkey fashion, is also a desired feature to quickly transform innovation into PoC (Proof of Concept) and to bring productivity gains and shorter time-to-market for industrial projects.

The extended GAMAM and the BATX[36] also have large in-house databases required for the training and the computing facilities. In addition, almost all of them are developing their chips for DL (e.g. Google with its line of TPUs) or made announcements that they will. The US and Chinese governments have also started initiatives in this field to ensure that they will remain prominent players in the field, and it is a domain of competition.

Sustainability through open technologies extends also to open data, rules engines[37] and libraries. The publication of open data and datasets is facilitating the work of researchers and developers for ML and DL, with the existence of numerous images, audio and text databases that are used to train the models

[34] https://digital-strategy.ec.europa.eu/en/library/recommendations-and-roadmap-european-sovereignty-open-source-hardware-software-and-risc-v

[35] DL networks with Tensorflow at Google, PyTorch / Caffe at Facebook, CNTK at Microsoft, Watson at IBM, DSSTNE at Amazon

[36] BATX is an acronym standing for Baidu, Alibaba, Tencent, and Xiaomi, the four biggest tech firms in China

[37] Clips, Drools distributed by red Hat, DTRules by Java, Gandalf on PH

and become benchmarks[38]. Reusable open-source libraries[39] allow to solve recurrent development problems, hiding the technical details and simplifying the access to AI technologies for developers and SMEs, maintaining high-quality results, reducing time to market and costs.

The Tiny ML community (https://www.tinyml.org/) is bringing Deep Learning to microcontrollers with limited resources and at ultra-low energy budget. The MLPerf allows to benchmark devices on similar applications (https://github.com/mlcommons/tiny), because it is nearly impossible to compare performances on figures given by chips providers. Other Open source initiatives are gearing towards helping the adaptation of AI algorithms to embedded systems, such as the Eclipse Aidge[40] , supported by the EU founded project Neurokit2E[41].

From the application perspective, impact of AI and embedded intelligence on sustainable development could be important. By integrating advanced technologies into systems and processes, AI and embedded intelligence enhance efficiency, reduce resource consumption, and promote sustainable practices, ultimately contributing to the achievement of the United Nations' Sustainable Development Goals (SDGs).

AI-powered systems could be used to optimizing energy consumption and improving the efficiency of renewable energy sources. For example, smart grids utilize AI to balance supply and demand dynamically, integrating renewable energy sources like solar and wind more effectively. Predictive analytics enable better forecasting of energy production and consumption patterns, reducing waste and enhancing the reliability of renewable energy. Additionally, AI-driven energy management systems in buildings and industrial facilities can significantly reduce energy usage by optimizing heating, cooling, and lighting based on real-time data.

Finally, embedded intelligence in consumer products promotes sustainability by enabling smarter usage and longer lifespans. For example, smart appliances can optimize their operations to reduce energy consumption, while predictive maintenance in industrial machinery can prevent breakdowns and extend equipment life, reducing the need for new resources and lowering overall environmental impact.

2.1.4.2 Market perspectives

Several market studies, although they don't give the same values, show the huge market perspectives for AI use in the next years.

According to ABI Research, it is expected that 1.2 billion devices capable of on-device AI inference have been shipped in 2023, with 70% of them coming from mobile devices and wearables. The market size for ASICs responsible for edge inference is expected to reach US$4.3 billion by 2024 including embedded architectures with integrated AI chipset, discrete ASICs, and hardware accelerators.

The market for semiconductors powering inference systems will likely remain fragmented because potential use cases (e.g. facial recognition, robotics, factory automation, autonomous driving, and

[38] A few examples are ImageNet (14 million images in open data), MNIST or WordNet (English linguistic basis)
[39] Nvidia Rapids, Amazon Comprehend, Google NLU Libraries
[40] https://projects.eclipse.org/projects/technology.aidge
[41] https://www.neurokit2e.eu/

surveillance) will require tailored solutions. In comparison, training systems will be primarily based on traditional CPUs, GPUs, FPGAs infrastructures and ASICs.

According to McKinsey, it is expected by 2025 that AI-related semiconductors could account for almost 20 percent of all demand, which would translate into about $65 billion in revenue with opportunities emerging at both data centers and the edge.

According to a recent study, the global AI chip market was estimated to USD 9.29 billion in 2019 and it is expected to grow to USD 253.30 billion by 2030, with a CAGR of 35.0% from 2020-2030.

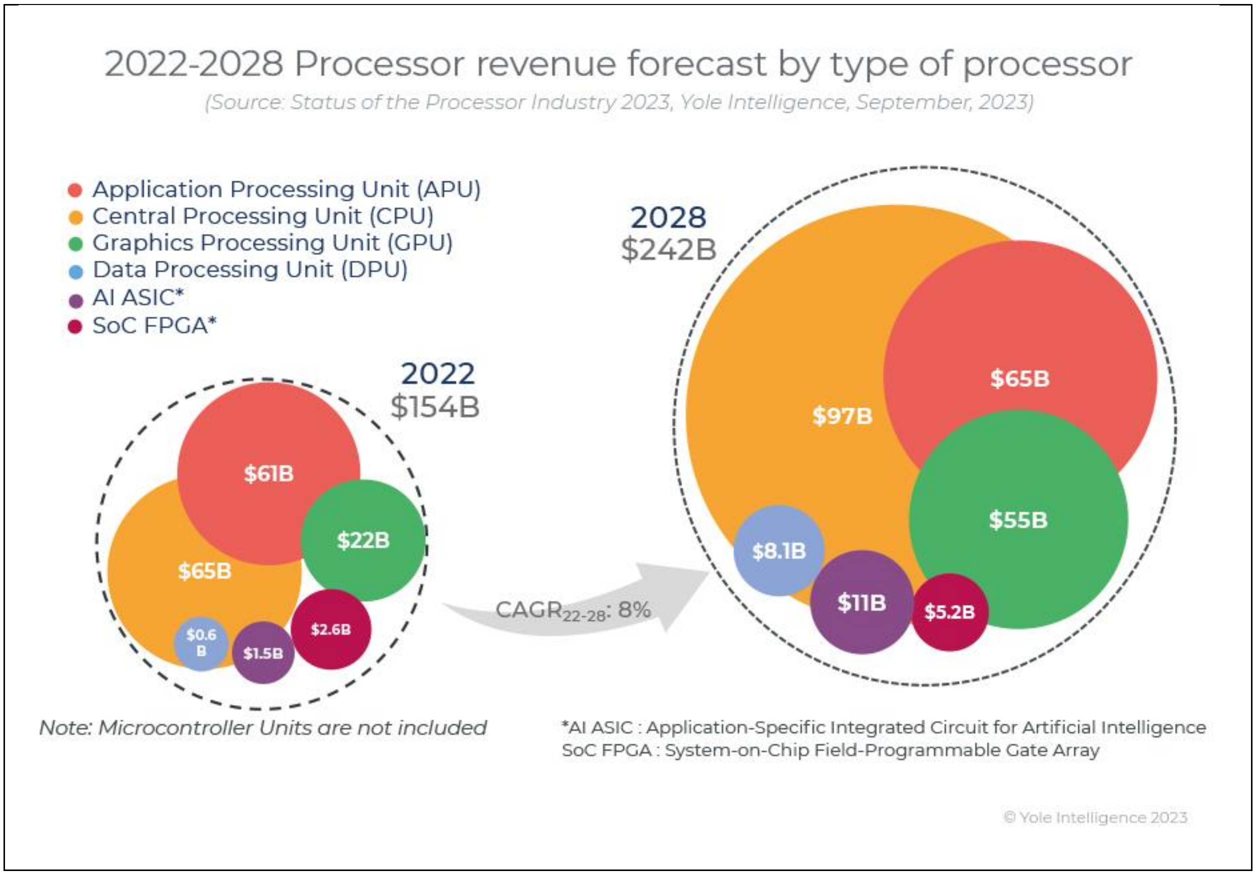


Figure 9: Market of high-end chips from https://www.yolegroup.com/product/report/status-of-the-processor-industry-2023/

There is a large increase in value of the market in the field of chip for IA, mainly for server/datacenters, but the market for Edge IA is also forecasted to have a large increase in the next years. The exact figures vary according to the market forecast company, but there is a common agreement for the large growth.

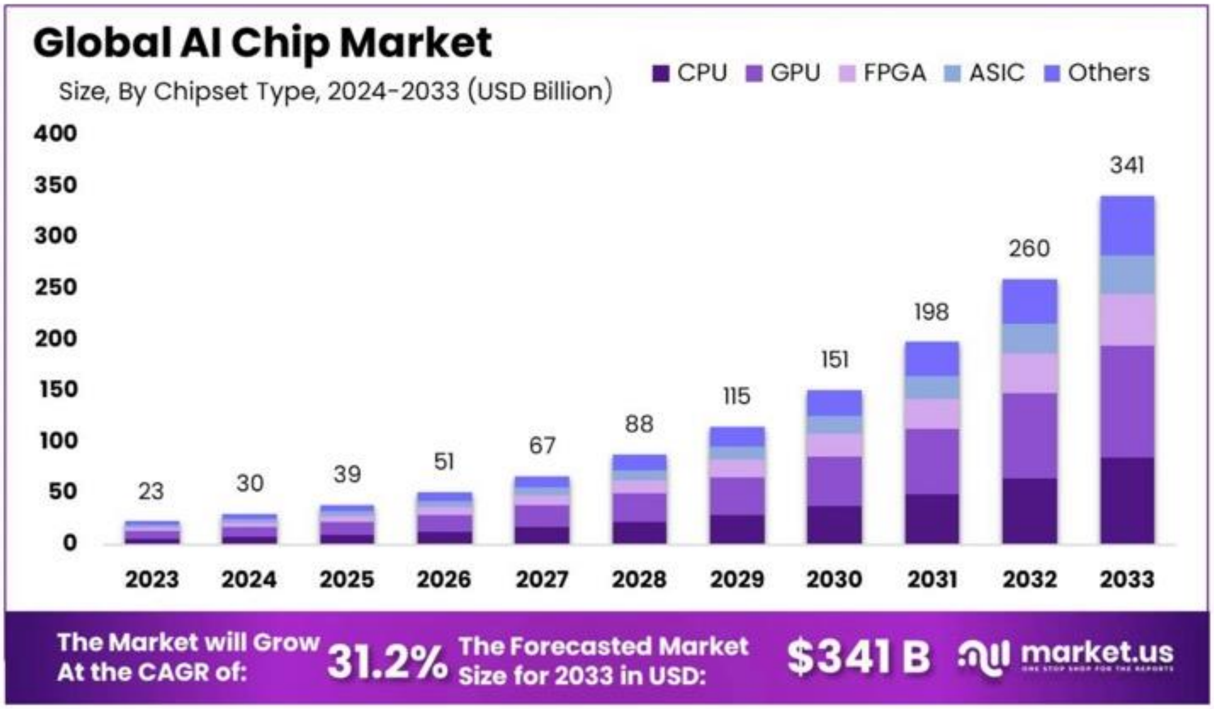


Figure 10: Market forecast of AI chips, from https://market.us/report/ai-chip-market/

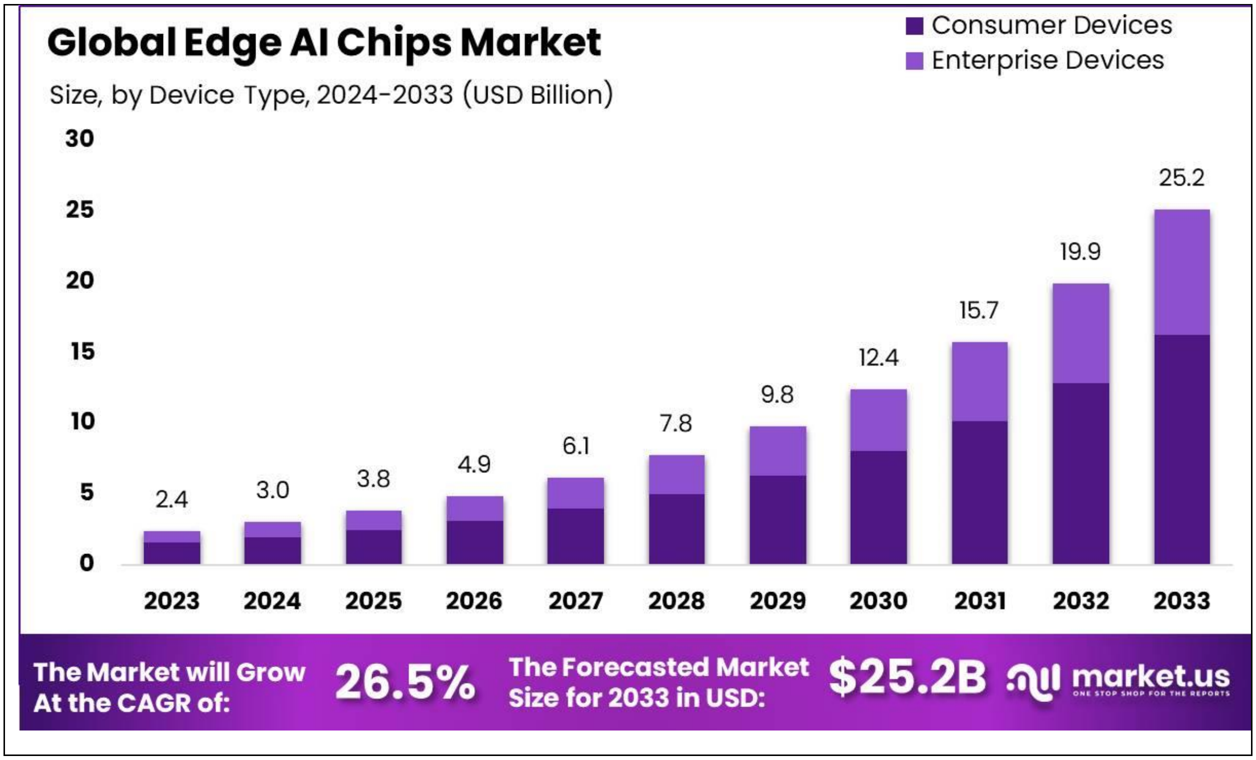


Figure 11: Edge AI market growth 2023-2033, from https://market.us/report/edge-ai-chips-market/

However, there is now an increasing trend towards AI chips that can execute generative AI models at the edge, and this trend is clearly increasing in 2024. These models can generate new content, predict complex sequences, and enhance decision-making processes in more sophisticated ways. Initially, the integration of these capabilities is seen in the PC and smartphone markets (Copilot+PC, with currently Snapdragon X series of chip, smartphones with Snapdragon 8 Gen 3 able to support generative AI models with up to 10 billion parameters[42], or the Mediatek Dimensity 9300[43]) , where there is a growing demand for advanced AI applications that require both high performance and low latency. For instance, new smartphones and PCs are being equipped with AI chips capable of running generative models locally, enabling features such as real-time language translation, advanced virtual assistants, and enhanced creative tools like automated photo and video or text editing[44].

European semiconductor companies are not in this market of chipset for smartphone or PCs, but **it is expected that, with the progress in the size reduction and specialization of large Foundation Models, generative AI at the edge will be used in the near (?) future in more and mode domains, including the ones primarily targeted by European chip manufacturers.** We also observe emerging tentative (not very successful yet) to develop new devices and use cases in the domain of edge devices (often connected to the cloud) such as Rabbit R1, Ai-pin, etc.

Another recent trend is that the **RISC-V instruction set based chips are also developing rapidly**, especially in China or Taiwan. For example, the Chinese company Sophgo delivers a cheap microcontroller with 2 RISC-V (one running at 1GHz, the other at 700 MHz) and a 1 TOPS (8bit) NPU, allowing to build systems running both Linux and a RTOS for less than 5$. On the cheap side, the CH32V003 is a 32-bit RISC-V microcontroller running at 48 MHz with 16kB of flash and 2 kB or SRAM which is sold at retail price (not for large quantities) below 10 cents. On the other side of the spectrum, various Chinese companies are announcing high performance RISC-V chips for servers, for example Alibaba with its C930[45]. US companies are also present in this market (SiFive, Esperanto, Meta – MTIA chip, etc). European companies are also involved in developing or using RISC-V cores (Codasip, Greenwaves, Bosch, Infineon, NXP, Nordic semiconductors, …). Research firm Semico projects a staggering 73.6 percent annual growth in the number of chips incorporating RISC-V technology, with a forecast of 25 billion AI chips by 2027 generating a revenue of US $291 billion.

2.1.4.3 AI components vendors

In the next few years, the hardware is serving as a differentiator in AI, and AI-related components will constitute a significant portion of future demand for different applications.

Qualcomm has launched a generation of Snapdragon processors, with NPU, allowing to run models un to 10 billion parameters in smartphones (such as Samsung Galaxy S24) or in PCs (Copilot+PCs), MediaTek

[42] https://www.qualcomm.com/products/mobile/snapdragon/smartphones/snapdragon-8-series-mobile-platforms/snapdragon-8-gen-3-mobile-platform#Overview
[43] https://www.mediatek.com/blog/whats-new-in-the-mediatek-dimensity-9300
[44] https://www.apple.com/apple-intelligence/
[45] https://www.tomshardware.com/pc-components/cpus/alibaba-claims-it-will-launch-a-server-grade-risc-v-processor-this-year

also has chips allowing to run generative AI on smartphones and Google, with its Tensor G4 (for the pixel 9) also introduces chips able to run small models in consumer grade (embedded) devices. Apple also introduces generative AI in their devices, starting from the iPhone 15 pro and device equipped with the M series of processors. Of course, the requirement is to have powerful NPUs (Copilot+PCs require NPU of 40 TOPs and 16 GB of RAM), but also a large amount of RAM able to host the model and the OS.

Huawei and MediaTek incorporate their embedded architectures into IoT gateways and home entertainment, and Xilinx finds its niche in machine vision through its Versal ACAP SoC. NVIDIA has advanced the developments based on the GPU architecture, NVIDIA Jetson AGX platform, a high performance SoC that features GPU, ARM-based CPU, DL accelerators and image signal processors. NXP and STMicroelectronics have begun adding AI HW accelerators and enablement SW to several of their microprocessors and microcontrollers.

ARM is developing core for machine learning applications and used in combination with the Ethos-U85[46] or Ethos-N78 AI accelerator. Both are designed for resource-constrained environments. The new ARM's cores are designed for customized extensions and for ultra-low power machine learning.

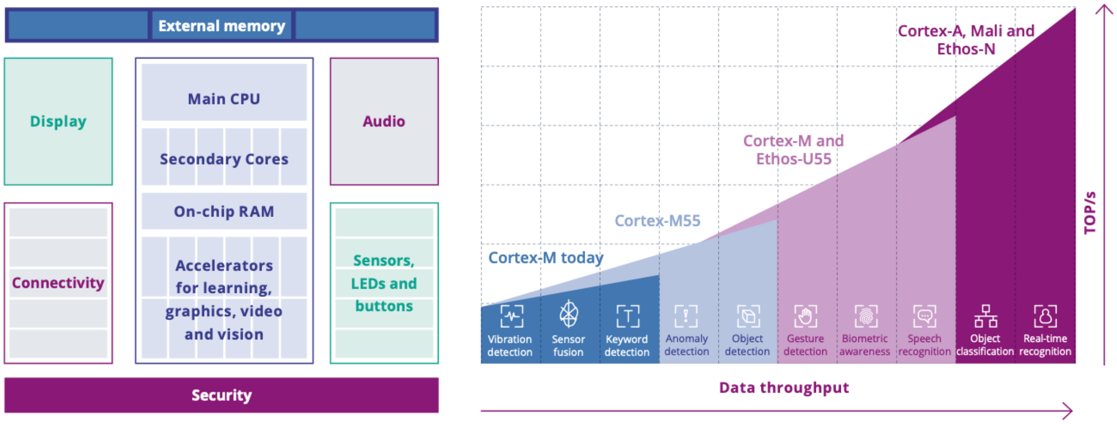


Figure 12: - Example of architecture of a modern SoC (from Paolo Azzoni, see also Chapter 1.3) / Arm's Cortex-M55 and Ethos-U55 Tandem. Provide processing power for gesture recognition, biometrics, and speech recognition applications (Source: Arm).

Companies like Google, Gyrfalcon, Mythic, NXP, STMicroelectronics and Syntiant are developing custom silicon for the edge. As an example, Google was releasing Edge TPU, a custom processor to run TensorFlow Lite models on edge devices. NVIDIA is releasing the Jetson Orin Nano range of products, allowing to perform up to 40 TOPS of sparce neural networks within a 15W power range[47].

[46] https://newsroom.arm.com/news/iot-reference-design-platform-2024

[47] https://developer.nvidia.com/blog/solving-entry-level-edge-ai-challenges-with-nvidia-jetson-orin-nano/

Open-source hardware, championed by RISC-V, will bring forth a new generation of open-source chipsets designed for specific ML and DL applications at the edge. French start-up GreenWaves is one of European companies using RISC-V cores to target the ultra-low power machine learning space. Its devices, GAP8 and GAP9, use 8- and 9-core compute clusters, the custom extensions give its cores a 3.6x improvement in energy consumption compared to unmodified RISC-V cores.

The development of the neuromorphic architectures is accelerated as the global neuromorphic AI semiconductor market size is expected to grow.

The major European semiconductor companies are already active and competitive in the domain of AI at the edge:

- Infineon is well positioned to fully realize AI's potential in different tech domains. By adding AI to its sensors, e.g. utilizing its PSOC microcontrollers and its Modus toolbox, Infineon opens the doors to a range of application fields in edge computing and IoT. First, Predictive Maintenance: Infineon's sensor-based condition monitoring makes IoT work. The solutions detect anomalies in heating, ventilation, and air conditioning (HVAC) equipment as well as motors, fans, drives, compressors, and refrigeration. They help to reduce breakdowns, maintenance costs and extend the lifetime of technical equipment. Second, Smart Homes and Buildings: Infineon's solutions make buildings smart on all levels with AI-enabled technologies, e.g. building's domains such as HVAC, lighting or access control become smarter with presence detection, air quality monitoring, default detection and many other use cases. Infineon's portfolio of sensors, microcontrollers, actuators, and connectivity solutions enables buildings to collect meaningful data, create insights and take better decisions to optimize its operations according to its occupants' needs. Third, Health and Wearables: the next generation health and wellness technology is enabled to utilize sophisticated AI at the edge and is empowered with sensor, compute, security, connectivity, and power management solutions, forming the basis for health-monitoring algorithms in lifestyle and medical wearable devices supplying highest precision sensing of altitude, location, vital signs, and sound while also enabling lowest power consumption. Fourth, Automotive: AI is enabled for innovative areas such as eMobility, automated driving and vehicle motion. The latest microcontroller generation AURIX™ TC4x with its Parallel Processing Unit (PPU) provides affordable embedded AI and safety for the future connected, eco-friendly vehicle.

- NXP, a semiconductor manufacturer with strong European roots, has begun adding Al HW accelerators and enablement SW to several of their microprocessors and microcontrollers targeting the automotive, consumer, health, and industrial market. For automotive applications, embedded AI systems process data coming from the onboard cameras and other sensors to detect and track traffic signs, road users and other important cues. In the consumer space the rising demand for voice interfaces led to ultra-efficient implementations of keyword spotters, whereas in the health sector AI is used to efficiently process data in hearing aids and smartwatches. The industrial market calls for efficient AI implementations for visual inspection of goods, early onset fault detection in moving machinery and a wide range of customer specific applications. These diverse requirements are met by pairing custom accelerators, multipurpose and efficient CPUs with a flexible SW tooling to support engineers implementing their system solution.

- STMicroelectronics integrated edge AI as one of the main pillars of its product strategy plan. By combining AI-ready features in its hardware products to a comprehensive ecosystem of software and tools, ST ambitions to overcome the uphill challenge of AI: opening technology access to all and for a broad range of applications. STMicroelectronics delivers a stack of Hardware and Software specifically designed to enable Neural Networks and Machine Learning inferences in an extremely low energy environment. Most recently, STMicroelectronics has started to bring to market MCUs and MPUs with Neural accelerators, able to handle workloads that were not conceivable in an edge device a few years ago. These devices rely on NPUs (Neural Processing Units) that execute Neural Network tasks up to 100 times faster than in traditional high end MCUs. The STM32MP2 MPU has been made available in the second quarter of 2024, it will be followed by a full family of STM32 MCUs leveraging STMicroelectronics home grown NPU technology, Neural Art. The first MCU leveraging this technology is already in the hands of tens of clients for workloads such as visual events recognition, people and objects recognition, all executed at the edge in the device. On the software side, to adapt algorithms to small CPU and memory footprints, STMicroelectronics has delivered the ST Edge AI Suite a toolset specifically designed to address all the needs of embedded developers from ideation with a rich Model zoo, to datalogging to optimization for the embedded world of pre-created Neural Networks to creation of predictive maintenance algorithms. Two tools are particularly interesting: 1) STM32Cube.ai, an optimizer tool which enables a drastic reduction of the power consumption while maintaining the accuracy of the prediction. 2) NanoEdge AI studio, an Auto-ML software for edge-AI, that automatically finds and configures the best AI library for STM32 microcontroller or smart MEMS that contain ST's embedded Intelligent Sensor Processing Unit (ISPU). NanoEdge AI algorithms are widely used in projects such arc-fault or technical equipment failure detection and extends the lifetime of industrial machines.

2.1.5 Applications breakthroughs

2.1.5.1 Example of AI application ranging from deep-edge to cloud

One example of an application that can leverage all three categories of AI hardware simultaneously in a true example of "continuum of computing" is an advanced autonomous driving system in smart cities.

In the cloud, massive amounts of data from numerous autonomous vehicles, traffic cameras, and other city infrastructure are aggregated and used to train large-scale AI models. Data centers equipped with GPUs handle the complex tasks of training models with large language models (LLMs). These models analyze patterns, improve decision-making algorithms, and enhance predictive capabilities. They also ensure that the autonomous driving algorithms continuously learn and adapt from real-world data. This centralized processing allows for the development of highly sophisticated models that can then be deployed to vehicles for local inferencing, ensuring they are always operating with the latest intelligence.

In the vehicles themselves, high-performance NPUs handle real-time inferencing tasks based on smaller models. These processors power the Advanced Driver Assistance Systems (ADAS) within each vehicle, enabling real-time perception, decision-making, and control. This includes tasks such as object detection, lane keeping, obstacle avoidance, and adaptive cruise control. This on-vehicle processing ensures that the car can react instantly to its environment without the latency that would be introduced by relying solely on cloud-based computations. They also ensure safety by working without a permanent connection to the cloud, which cannot be always guaranteed.

Within the vehicle, low-power AI hardware is used for continuous monitoring and processing of data from various sensors. These sensors include in-cabin cameras and audio systems, external cameras, LIDAR, radar, and ultrasonic sensors. The low-power AI systems process data for applications like driver behavior monitoring, passenger safety features, and environmental awareness (e.g., detecting nearby pedestrians or cyclists). By using low-power AI hardware, the vehicle can efficiently manage sensor data without draining the battery, which is crucial for electric vehicles. This also ensures continuous monitoring and safety features remain active without interruption.

The integration of these three AI hardware categories creates a robust autonomous driving ecosystem. Cloud AI ensures that the overall traffic management in the city/countryside is well managed, that the "intelligence" of the autonomous driving system is continuously improving through extensive data analysis and model training. High-Performance Embedded AI within the vehicle handles immediate, critical decision-making and control functions, allowing the car to operate safely and effectively in real time. Low-Power Embedded AI maintains constant environmental and internal monitoring, ensuring that all sensor data is processed efficiently, and that the vehicle can respond to changing conditions without excessive power consumption, and can prevent failures by continuous monitoring of the various functionalities of the vehicle.

It seems clear that advanced AI capabilities can enhance autonomous driving systems, in-car entertainment, and real-time vehicle diagnostics. Generative AI can be used to predict traffic patterns, generate detailed maps, and simulate various driving scenarios to improve safety and efficiency, and could improve the interface between the car and the user.

2.1.5.2 Other application domains

What we observe in automotive domain can also be applied to other domains of Edge computing and embedded Artificial Intelligence, such as systems for factories, for homes, etc. As we have seen, **there is more and more convergence between edge computing and embedded (generative) AI, but still a lot of edge applications will be without AI (for now?), so AI support/accelerators are required only in few systems.**

Another significant market is the healthcare sector. AI chips capable of executing advanced AI models can be used in medical devices for helping for diagnostics (with continuous monitoring, and forecasting potential problems), personalized treatment planning, and real-time monitoring of patient health. These applications require the ability to process and analyze data locally, providing immediate and accurate insights without relying on cloud-based solutions that may introduce latency or privacy concerns.

Important examples are its contribution in the recovery from the Covid-19 pandemic as well as its potential to ensure the required resilience in future crises[48].

Industrial automation and robotics also stand to benefit from this technology. Generative AI can enable more advanced predictive maintenance, optimize manufacturing processes, and allow robots to learn and adapt to new tasks more efficiently. By embedding AI chips with generative capabilities, industries can improve productivity, reduce downtime, and enhance operational flexibility. We see the first impressive results of using generative IA in robotics, not only for interfacing in natural ways with human, but also for scene analysis and action planning.

In Internet of Things (IoT), smart home devices, wearable technology, and even agriculture technology can leverage generative AI for a variety of innovative applications. These include creating more intelligent home automation systems, developing wearables that provide more personalized health and fitness insights, and implementing smart farming techniques that can predict crop yields and optimize resource use.

While AI accelerated chips have traditionally focused on supporting CNNs for image, audio, and signal analysis, the rise of generative AI capabilities is driving a new wave of AI chips designed for edge computing. This evolution is starting in the PC and smartphone markets but will spread across a range of embedded markets, including automotive, healthcare, industrial automation, robotics, and IoT, significantly broadening the scope and impact of edge AI technologies. Therefore, **it is important that Europe continue to be recognized as a player in the market of embedded MCU and MPU by being prepared to introduce generative AI accelerators in their products, even cheap ones.**

2.1.6 Major challenges

The convergence between edge computing and embedded generative AI is becoming increasingly evident, but many edge applications currently operate without AI integration. Consequently, AI support and accelerators are only required in a limited number of systems at present. Most of the accelerators available today in edge computing systems are designed for perception applications, such as image and audio processing, which require significant advancements to achieve further efficiency gains. As technology progresses, it is expected that the size reduction and specialization of large foundation models will make generative AI at the edge feasible in a growing number of domains. This trend includes areas primarily targeted by European chip manufacturers, signaling a shift towards broader AI integration in various applications.

Europe must maintain its status as a key player in the embedded MCU (Microcontroller Unit) and MPU (Microprocessor Unit) markets by preparing to introduce generative AI accelerators in their products, even in more affordable devices. These systems should also be ready for software upgradability, and connectivity allowing modularity and collaborative performances. This readiness will ensure that European manufacturers can meet the future demand for edge AI capabilities. Additionally, the rapid development of RISC-V instruction set-based chips presents a significant opportunity for innovation and

[48] https://www.eenewsembedded.com/news/nxp-developing-neural-networks-identify-covid-19

competitiveness in the global semiconductor market. Embracing these advancements will be crucial for Europe to stay at the forefront of the evolving edge computing landscape, where the integration of generative AI and specialized accelerators will become increasingly commonplace.

Europe should also be ready for the smooth integration of the devices into the computing continuum. This will not only require extra feature in the embedded devices (such as connectivity, covered in the connectivity chapter, security, covered in the security chapter) but also a global software stack certainly based on As a Service approach, the development of smart orchestrators and a global architecture view at system level. As such, the evolution towards the computing continuum is not a major challenge of this chapter, but is more a global challenge. We will only list here some aspects that need to be further developed:

Four Major Challenges have been identified for Europe to be an important player in the further development of computing systems, especially in the field of embedded AI architectures and edge computing:

1. **Increasing Energy Efficiency**
2. **Managing the Increasing Complexity of Systems**
3. **Supporting the Increasing Lifespan of Devices and Systems**
4. **Ensuring European Sustainability**

2.1.6.1 Major Challenge 1: Increasing the energy efficiency of computing systems and embedded intelligence

2.1.6.1.1 State of the art

Increasing energy efficiency is the results of progress at multiple levels:

Technology level:

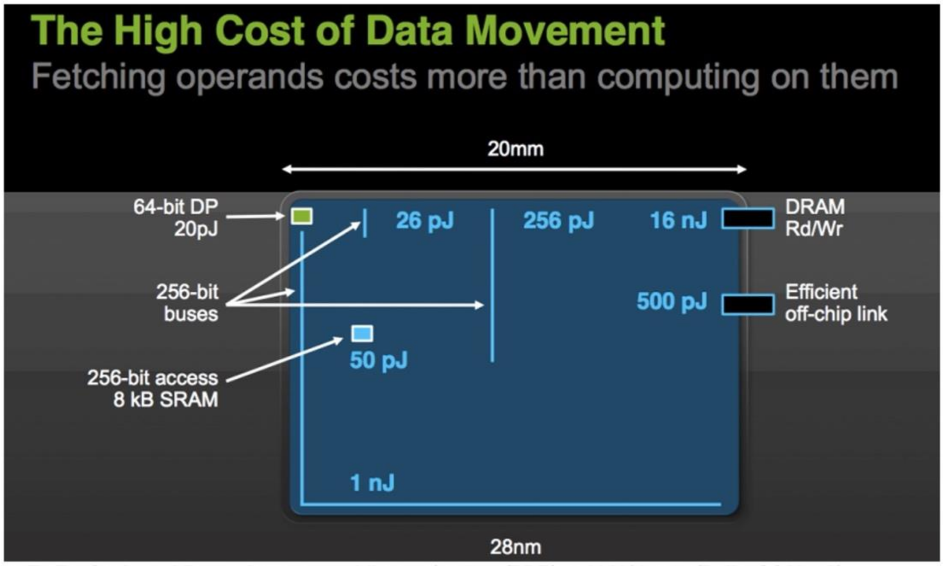
At technology level (FinFet, FDSOI, silicon nanowires or nanosheets), technologies are pushing the limits to be ultra-low power. Technologies related to advanced integration and packaging have also recently emerged (2.5D, chiplets, active interposers, etc.) that open innovative design possibilities, particularly for what concerns tighter sensor-compute and memory-compute integration.

Device level:

At device level, several type of architectures are currently developed worldwide. The list is moving from the well-known CPU to some more and more dedicated accelerators integrated in Embedded architectures (GPU, NPU, DPU, TPU, XPU, etc.) providing accelerated data processing and management capabilities, which are implemented going from fully digital to mixed or full analog solutions:

- Fully digital solutions have addressed the needs of emerging application loads such as AI/DL workloads using a combination of parallel computing (e.g. SMP[49] and GPU) and accelerated hardware primitives (such as systolic arrays), often combined in heterogeneous Embedded architectures. Low-bit-precision (8-bit integer, 8 or 4 bits floating representations, …) computation as well as sparsity-aware acceleration have been shown as effective strategies to minimize the energy consumption per each elementary operation in regular AI/DL inference workloads; on the other hand, many challenges remain in terms of hardware capable of opportunistically exploiting the characteristics of more irregular mixed-precision networks. Some applications also require further development due to their need for more flexibility and precision in numerical representation (32- or 16-bit floating point), which puts a limit to the amount of hardware efficiency that can be achieved on the compute side.
- **Avoiding moving data:** this is crucial because the access energy of any off-chip memory is currently 10-100x more expensive than access to on-chip memory. Emerging non-volatile memory technologies such as MRAM, with asymmetric read/write energy cost, could provide a potential solution to relieve this issue, by means of their greater density at the same technology node. Near-Memory Computing (NMC) and In-Memory Computing (IMC) techniques move part of the computation near or inside memory, respectively, further offsetting this problem. While IMC in particular is extremely promising, careful optimization at the system level is required to really take advantage of the theoretical peak efficiency potential. Figure 13 shows that moving data requires order of magnitude more energy than processing them (an operation on 64 bit data requires about 20pJ in 28nm technology, while getting the same 64 bit data from external DRAM takes 16 nJ – without the energy of the external DRAM).

[49] Symmetric multiprocessing or shared-memory multiprocessing (SMP) involves a multiprocessor computer hardware and software architecture where two or more identical processors are connected to a single, shared main memory, from https://en.wikipedia.org/wiki/Symmetric_multiprocessing .



The High Cost of Data Movement

Fetching operands costs more than computing on them

Source: Bill Dally, « To ExaScale and Beyond » www.nvidia.com/content/PDF/sc_2010/theater/Dally_SC10.pdf

Figure 13:- Energy for compute and data movement. This explains the order of magnitude of the problem of data movement, and this problem is still relevant in all technology nodes.

This figure shows that in a modern system, large parts of the energy are dissipated in moving data from one place to another. For this reason, new architectures are required, such as computing in or near memory, neuromorphic architectures (also where the physics of the NVM - PCM, CBRAM, MRAM, OXRAM, ReRAM, FeFET, etc. - technology can be used also to compute) and lower bit-count processing are of primary importance. Not only the memories itself, e.g. bitcells, are needed but the complementing libraries and IP for IMC or NMC as well.

		Flash reference	**MRAM type**	**PCM type**	**ReRAM type**	**FeFET**	**Hf based FeRAM 1T1C**
Performance	**Programming power**	<200pJ/bit - # 100pJ/bit (eSTM)	~20pJ/bit	~90pJ/bit	~100pJ/bit	<~20pJ/bit	<pJ/bit
	Reading access time	HV devices ~15ns	Core oxide device ~1ns	No HV devices ~5ns	No HV devices ~5ns	No HV devices ~5ns	Write after read (Destr. Read)
	Erasing granularity	FN mechanis	bit-2-bit erasing	bit-2-bit erasing	bit-2-bit erasing	bit-2-bit? Depend	bit-2-bit

		m Full page erasing	Fine granularity	Fine granularity	Fine granularity	on archi	erasing
Reliability	**Endurance**	Mature technology	High capability 10^15 ?	500Kcy	10^5 trade Off with BER	10^5 Gate stress sensibility	10^11 #10^6 with write after read
	Retention	Mature technology	Main weakness Trade-off with Taa	150°C auto compliant	Demonstrated Trade Off with power	To be proven	To be proven
	Soldering reflow	Mature technology	High risk To be proven	pass	possible	To be proven	To be proven
Cost	**Extra masks**	Very high (>10)	Limited (3-5)	Limited (3-5)	Limited (3-5)	Low (1-3)	Low (1-3)
	Process flow	Complex	Complex	Simple	Simple	Simple	Simple
	New assets vs CMOS	Shared	New _ manufacturable	New _manufacturable	BE High-k material	FE High k material	BE High-k material

Figure 14: eNVM technologies, strengths and challenges (from Andante: CPS & IoT summer school, Budva, Montenegro, June 6th-10th, 2023

System level:

- micro-edge computing near sensors (i.e. integrating processing inside or very close to the sensors or into local control) will allow embedded architectures to operate in the range of 10 mW (milliwatt) to 100 mW with an estimated energy efficiency in the order of 100s of GOPs/Watt up to a few TOPs/Watt in the next 5 years. This could be negligible compared to the energy consumption of the sensor (for example, a MEMS microphone can consume a few mW). On top, the device itself can go in standby or in sleep mode when not used, and the connectivity must not be permanent. Devices currently deployed on the edge rarely process data 24/7 like data centers: to minimize global energy, a key requirement for future edge Embedded architectures is to combine high performance "nominal" operating modes with lower-voltage high compute efficiency modes and, most importantly, with ultra-low-power sleep states, consuming well below

1 mW in fully state-retentive sleep, and less than 1-10 µW in deep sleep. The possibility to leave embedded architectures in an ultra-low power state for most of the time has a significant impact on the global energy consumed. The possibility to orchestrate and manage edge devices becomes fundamental from this perspective and should be supported by design. On the contrary, data servers are currently always on even if they are loaded only at 60% of their computing capability.

- **Unified (or shared) memory**, which allows CPUs and GPUs (or NPUs) to share a common memory pool, is gaining traction in first for high end systems such a Mx series of systems from Apple. This approach not only streamlines memory management but also provides flexibility by enabling the allocation of this unified memory pool to Neural Processing Units (NPUs) as needed. This is particularly advantageous for handling the diverse and dynamic workloads typical of AI applications and reduce the amount of energy required to move data between different parts of the system.
- **Maximum utilization and energy proportionality:** Energy efficiency in cloud servers rely on the principle that silicon is utilized at its maximum efficiency, ideally operating at 100% load. This maximizes the computational throughput per watt of power consumed, optimizing the energy efficiency of the server infrastructure. In cloud environments, high utilization rates are easier to achieve because workloads from multiple users can be aggregated and balanced across a vast number of servers. This dynamic allocation of resources ensures that the hardware is consistently operating near its capacity, minimizing idle periods. Considering processing at the edge, achieving a clear net benefit in terms of energy efficiency presents additional challenges. Unlike centralized cloud servers, edge devices often face fluctuating and unpredictable workloads. These devices must be capable of maintaining high efficiency across a wide range of utilization levels, not just at peak performance. To accomplish this, edge hardware needs to be designed with adaptive power management features that allow it to scale power usage dynamically in response to varying workloads. This means the **hardware should be as efficient at low workloads as it is at 100% utilization**, ensuring that energy is not wasted during idle or low-demand periods. Energy proportionality aims to scale computing power according to the computational demand at any given moment. This concept is crucial for modern computing environments, which must handle a diverse range of workloads, from the intensive processing needs of self-driving vehicles to the minimal computational requirements of IoT devices. The ability to dynamically adapt power usage ensures efficiency and effectiveness across this spectrum of applications. The **scalability of processing power is fundamental to achieving energy proportionality**. In high-demand scenarios, such as those faced by self-driving cars, the computational power must scale up to handle complex algorithms for navigation, sensor fusion, and real-time decision-making. Conversely, IoT devices often require minimal computational power, as they are designed to perform specific, limited functions with a few lines of code. This wide range of computational needs presents a significant challenge because the requirements and hardware for each application are vastly different.

Data level:

At data level, **memory hierarchies** will have to be designed considering the data reuse characteristics and access patterns of algorithms, which strongly impact load and store access rate and hence, the energy necessary to access each memory in the hierarchy and avoiding to duplicate data. For example (but not only), weights and activations in a Deep Neural Network have very different access patterns and can be deployed to entirely separate hierarchies exploiting different combinations of external Flash, DRAM, non-volatile on-chip memory (MRAM, FRAM, etc.) and SRAM.

The **right data type** should be also supported to perform correct (or as correct as acceptable) computation with the minimum number of bits and if possible, in integer (the floating-point representation as specified in IEEE 754 is very costly to implement, especially with the particular cases). AI based on transformers and generative AI could use a far simpler representation for the inference phase, such a bfloat16 (which could also be used for training) or even float 8 or 4 and even in some cases, using less than 4 bits integer representation. It is obvious that storing and computing a weight coded with INT4 is about an order of magnitude more efficient than using a Float32 representation.

Tools level:

At tools level, **HW/SW co-design of system** and their associated algorithms are mandatory to minimize the data moves and optimally exploit hardware resources, particularly if accelerators are available, and thus optimize the power consumption. New AI-based HW/SW platforms with increased dependability, optimized for increased energy efficiency, low cost, compactness and providing balanced mechanisms between performance and interoperability should be further developed to support the integration into various applications across the industrial sectors of AI and other accelerators[50].

2.1.6.1.2 Key focus areas

Therefore, there are several key axes to further develop in order to increase energy efficiency of computing and AI systems:

Processing Data Locally and reducing data movements:

Increasing the energy efficiency of computing systems in general is a multifaceted challenge that requires innovations across various domains, including hardware, software, and data management. Indeed, power consumption should not only be seen at the level of the device, but at the levels composing the full technology stack. One effective strategy is **processing data locally where it is created**, which reduces the energy consumption associated with data transmission to centralized locations, and also reducing latency required in real-time applications, such as self-driving vehicle or other industrial commands. This localized data processing not only enhances energy efficiency but also helps in securing the data by keeping it within the local environment, thus ensuring privacy. Additionally, **the co-design of architectures, hardware, software, and topologies can significantly optimize system performance and energy usage**. This is closely linked to the concept of "local-first" software, as articulated by M. Kleppmann, A. Wiggins, P. Van Hardenberg, and M. McGranaghan in their 2019 paper, which presents a paradigm shift that emphasizes user autonomy and data ownership, even in the era of pervasive cloud computing. Local-first software is designed to ensure that users maintain control over their data by storing it primarily on their local devices, thus enabling full functionality even without an active internet connection. In the local-first paradigm, the software architecture is designed to prioritize local operations, using Conflict-free Replicated Data Types (CRDTs) and other synchronization mechanisms to manage concurrent data updates and ensure consistency across distributed systems. This approach aligns with the capabilities of modern edge hardware, which can handle complex data structures and algorithms locally, reducing the reliance on

[50] An example is the platform developed in the project Neurokit2E, see https://www.neurokit2e.eu/

centralized cloud infrastructures. As shown in the beginning of this chapter, this concept is clearly involved in Apple Intelligence and in HarmonyOS[51].

Co-Design of Algorithms, Hardware, Software:

The co-design of algorithms, hardware, software represents a holistic approach to optimizing system performance and energy usage. This integrated design process can help to identify and eliminate bottlenecks, streamline operations, and reduce overall energy consumption. It is also crucial to have a co-optimization of the software and hardware to explore more advanced trade-offs. At tools level, HW/SW co-design of system and their associated algorithms are mandatory to minimize the data moves and optimally exploit hardware resources, particularly if accelerators are available, and thus optimize the power consumption.

Efficient Management of storage resources:

In the landscape of **edge AI, the efficient management of memory resources is becoming increasingly critical to guarantee both a sustainable energy consumption profile and the required performances.** Local execution of smaller language models (SLMs) often encounters bottlenecks primarily at the memory level. This is particularly evident with the need for cheap and efficient memory solutions able not only to store but allowing to be used directly for computation, avoiding the transfer of the weight of the model from flash to standard DRAM – as done today – which is expansive. Research is only starting in this field[52]. Memory technology itself is advancing swiftly, with smartwatches now boasting up to 8GB of memory, showing the trend towards more substantial memory capacities in compact devices.

Moreover, the cost of memory remains a crucial factor for deploying generative AI at the edge. Efficient memory usage is imperative to keep costs manageable, especially as generative models require substantial memory for optimal performance**. Innovations in compressing weight** (quantization from FP16 to FP4 or even INT4 or lower), **pruning the network or advances in memory technology are therefore essential** to minimize RAM wastage and enhance the feasibility of deploying advanced AI capabilities on edge devices. In most NPUs, the weights have to be transferred from non-volatile memory to faster RAM, leading to energy waste and duplication of storage resources. This involves strategic allocation (between Flash, DRAM or other kind of memories) and utilization of memory to ensure that AI tasks are handled efficiently without unnecessary data copy (which consumes energy and requires more resources).

Concept of Unified Memory:

The concept of unified memory, where a common memory pool is shared between the CPU and GPU/NPU, offers significant advantages for energy efficiency. This approach allows for more flexible and dynamic allocation of memory resources, ensuring that the system can adapt to varying workloads. Unified memory can also help to reduce data duplication and transfer overheads, leading to more efficient processing.

[51] Some concepts are also developed in the open source Oniro project (https://oniroproject.org/), also based on OpenHarmony (https://gitee.com/openharmony).

[52] See https://arxiv.org/abs/2312.11514 for example

Innovations in Memory Technology:

Innovations in compressing weights, pruning neural networks, and advances in memory technology are essential for increasing energy efficiency. Techniques such as weight compression and network pruning reduce the amount of data that needs to be processed and stored, thereby lowering energy consumption. Advances in memory technology, such as the development of more efficient and high-capacity non-volatile memory solutions, that can be used directly for computing, further contribute to this goal.

Energy Proportionality:

Energy proportionality is a critical principle for hardware design, ensuring that devices are as efficient at low workloads as they are at 100% utilization. Addressing the scalability challenge involves **designing systems that can efficiently adjust their power usage without compromising performance**. This means developing hardware that is capable of scaling its power consumption dynamically and software that can efficiently distribute computational loads. The goal is to achieve a balance where each device, whether a powerful autonomous vehicle or a simple IoT sensor, operates at optimal energy efficiency.

Ultra-Low Standby Current:

Maintaining an ultra-low standby current is vital for devices that spend significant amounts of time in idle or low-power states, waiting to perform tasks triggered by specific events or intervals. This is particularly important for battery-operated and portable devices, which may include sensors, mobile devices, and IoT appliances, where energy efficiency is a key concern for these devices that spend significant amounts of time in standby mode. Achieving ultra-low standby current requires special power management techniques, such as power gating and dynamic voltage scaling, which can shut down non-essential components and reduce power supply to idle parts of the circuit without compromising the device's readiness to wake up and perform tasks promptly.

Leveraging Physical Phenomena for Computation:

One innovative approach to reducing energy consumption in computing involves **leveraging physical phenomena to perform computations**, effectively using the inherent properties of physical systems to solve complex problems. This method relies on the principle that certain physical processes can naturally execute the same types of equations and operations that we aim to solve computationally. This is the case for neuromorphic computing. This term covers various kind or realization, using sparce information coding method of "spikes", using RRAMs (resistive RAMs) to make essentially the weighted sum which is the basis of artificial neural network-based AI, including LLMs. Ohm's law can be used to make the (analog) product, and either Kirchhoff's law or adding charges in a capacitor used for the summation. The value of RRAM can be also locally modified, allowing to implement some training algorithm very efficiently. However, despite their promising potential, neuromorphic computing technologies are still largely in the research phase. Scalability to large networks, dispersion of properties, interfacing with other systems, etc. still need to improve. **Extensive efforts are needed to refine these systems and validate their effectiveness** in real-world, product-ready solutions.

2.1.6.2 Major Challenge 2: Managing the increasing complexity of systems and embedded intelligence

2.1.6.2.1 State of the art

Managing the increasing complexity of systems is another crucial aspect. It is driven by the diversity and dependencies of its components as well as by the requested functions and performance, affecting concepts, architectures and operation at all levels of the system hierarchy. Therefore, the reference architectures for future AI-based systems need to provide modular and scalable solutions that support interoperability and interfaces among platforms that can collaborate, exchange information and share computing resources to allow the functional evolution of the silicon-born embedded systems, ready for the computing continuum evolution.

Still, it is observed that the strategical backbone technologies to realize such new architectures are not available. These strategical backbone technologies include smart and scalable electronic components and systems, the AI accelerator and control hardware and software, the security engines, and the connectivity technologies.

To achieve this, balanced mechanisms must be developed to ensure performance while maintaining interoperability between diverse systems. AI techniques play a vital role in complexity management by enabling self-optimizing, self-reconfiguring, and self-managing systems. These self-X capabilities allow systems to adapt dynamically to changing conditions, enhancing their reliability and resilience but they are not available to the required extent.

The current SoCs are reaching a very high complexity by integrating a lot of functions and different accelerators on the same die. A decomposition of the SoCs into chiplets for more flexibility and scalability (e.g., mix and match) may help dealing with the complexity but for that an ecosystem which facilitates the design and also provides plug and play capable, interoperable chiplets is requried.

A holistic end-to-end approach is necessary to manage the increasing complexity of systems, to remain competitive and to continuously innovate the European electronic components and systems ecosystem. This approach involves technologies described in other chapters such as AI for system conception described in methods and tools chapter. Besides theses, here is a short list of themes are need to be promoted in order to reduce global complexity of systems in general, and systems supporting AI techniques in particular:

- **Tools and techniques leveraging AI to help in the management of complexity**, e.g. tools to adapt LLMs to edge / embedded targets are required. For example, the development of AI based HW/SW for multi-tasking and providing techniques to adapt the trained model to produce close or expected outputs when provided with a different but related set of data. The new solutions must provide dynamic transfer learning, by assuring the transfer of training instance, feature representation, parameters, and relational knowledge from the existing trained AI model to a new one that addresses the new target task. They should also ensure hyperparameter tuning for automated machine learning scenarios, etc. Furthermore, environments for **modeling how the various parts interact together should be considered since topic is still a challenge,** since having accurate digital twins of complex systems with the right level of abstraction for the various uses (conception, runtime optimization, etc.) is still an unsolved challenge.
- New design and architecture concepts including HW-SW codesign with corresponding HW/SW platforms for AI born embedded systems should be promoted to facilitate trust by providing the

dependable design techniques, that enable the end-to-end AI systems to be scalable, make correct decisions in a repetitive manner, provide mechanisms to be transparent, explainable, interpretable, and able to achieve interpretable results and embed features for AI model's and interfaces' interpretability. Linked to the previous point, infrastructure for the secure and safe execution of AI should be created.

- Enabling factors to **manage the complexity of the continuum of computing** are **the use and development of standardized APIs for hardware** and software tool chains, methods and interoperability across different system layers like sensors, gateways, on-premise servers and edge processing units. Additionally, interoperability concepts for AI edge-based platforms for data tagging, training, deployment, and analysis should be supported, allowing the development of distributed edge computing architecture with AI models running on distributed devices, servers, or gateways away from data centers or cloud servers. This implies scalable and hardware agnostic AI models capable of delivering comparable performance on different computing platforms, (e.g. Intel, AMD or ARM, Risc-V architectures) and seamless and secure integration at HW/SW embedded systems with the AI models integrated in the SW/HW and APIs to support configurable data integrated with enterprise authentication technologies through standards-based methods.
- **Deterministic behaviors, low latency and reliable communications** are also important for other vertical applications, such as connected cars, where edge computing and AI represent "the" enabling technology, independently from the sustainability aspects. The evolution of 5G is strongly dependent on edge computing and multi-access edge computing (MEC) developments.

In this chapter, five Key Focus Areas have been identified to tackle this complexity challenge:

- 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

2.1.6.2.2 Key focus areas

Complexity management utilizing AI

In the domain of edge AI, the concept of a **Mixture of Agents** (MoA) or Agentic AI, where a set of smaller Large Language Models (LLMs) called agents works collaboratively, presents a promising approach to managing complexity more effectively than relying on very large "universal" foundation models. This strategy involves deploying multiple specialized models, each trained to handle specific tasks or domains, and orchestrating their cooperation to achieve comprehensive AI functionalities. By distributing the workload across several specialized agents, the system can maintain high levels of efficiency and accuracy, tailored to the particularities of different tasks. The MoA framework enhances scalability and flexibility, enabling more precise and context-aware responses by leveraging the strengths of each specialized model. This approach also allows for tailored optimization, enhancing the performance of AI applications across diverse domains. Additionally, it reduces the computational and memory overhead typically associated with very large models (only the relevant agents are active, not all of them), making it more feasible to implement advanced AI capabilities on edge devices with limited resources. Consequently, the MoA paradigm not only simplifies the management of AI complexity but also optimizes performance and resource utilization in edge computing environments, paving the way for more responsive and efficient AI applications.

Decomposition of Complex SoCs into Chiplets and Interposers

The decomposition of complex SoCs into chiplets and interposers is an emerging strategy to manage system complexity. This approach involves breaking down a monolithic SoC into smaller, modular components (chiplets) that are interconnected using interposers. This architecture allows for greater flexibility in design and manufacturing, enabling more efficient scaling and customization of systems, e.g. adaptation to domains/applications. Additionally, it facilitates better management of thermal and power characteristics, improving overall system performance and it may help to optimize cost. By adopting this decomposition approach featuring an ecosystem which includes interoperable chiplets, we can more effectively address the challenges posed by the increasing complexity of systems. However, for chiplets and interposer ecosystems to really emerge, an agreed standard for interconnect (such as UCIe) will be required, together with physical specifications allowing interoperability between providers.

Ensuring Programmability and Interoperability

To handle the increasing complexity of systems, it is essential to ensure programmability and interoperability for example using techniques and adherence to standards that originate from the web and cloud computing (but not necessarily involving the cloud). Technologies such as containers, orchestration, and WebAssembly (WASM) maintain high flexibility and adaptability and enable seamless integration and operation across different platforms managing the diversity and dynamic range of modern computing environments. These tools allow developers to manage and update systems more efficiently, ensuring that various components can communicate and function together effectively while developer's productivity can be helped by generative AI (chapter "Embedded Software" of this SRIA), e.g., for the exact syntax of the API for driving a particular function, or even generating templates to use a part of the hardware.

Combining Processing Devices to Work Together

The combination of various processing devices working together is another key focus area for managing system complexity. By combining CPUs, GPUs, NPUs, and other specialized processors systems can distribute tasks according to the strengths of each device. The combination may be either locally (integration at chip level, at the package level using for example chiplets or across distributed chips) or as separate components of a system, like in a car, or distributed on physically separated systems, each part having a special task to do, thus separating the concerns and helping the management of the complexity by splitting it explicitly. This collaborative approach ensures optimal performance and energy efficiency, as each processor type can handle specific aspects of the workload more effectively. Coordinating these devices requires sophisticated management and scheduling techniques, but it leads to more robust and scalable systems. Together with interoperability, the ability of devices to work together is the foundation of the computing continuum approach.

Modeling Interactions Among System Components

Modeling the interactions among various parts of a system remains a significant challenge in managing complexity. Understanding how different components interact, communicate, and influence each other is critical for optimizing system performance and reliability. Advanced modeling techniques and simulation tools are necessary to accurately represent these interactions and predict system behavior under different conditions. By improving our ability to model these interactions, we can better design and manage complex systems, ensuring they operate smoothly and efficiently.

2.1.6.3 Major Challenge 3: Supporting the increasing lifespan of devices and systems and embedded intelligence

2.1.6.3.1 State of the art

With the power of embedded AI, we can now unleash the self-X concept which didn’t took off in the past. We can now design HW/SW techniques and architectures that can dynamically monitor their behavior, predict default (predictive maintenance), allowing devices to self-optimize, reconfigure and self-manage the resource demands (e.g. memory management, power consumption, AI model selection, etc…). This will have a direct effect in increasing the lifetime of devices and increasing their up-time.

Supporting the increasing lifespan of devices and systems can mainly be based on three main principles:

- Upgradability
- Modularity
- Reuse

By extension, this means that devices and systems have to increase their scalability and interoperability.

Upgradability

It requires hardware that can accommodate for example software upgrades, ensuring that **devices remain functional and up to date over extended periods**. This is the drive for approaches such as Software Defined Vehicle.

The first level of lifetime extension is clearly the upgrade to avoid replacing an object but instead improving its features and performance through either hardware or software update. This concept is not new as it is already applied in several industrial domains for dozens of years.

Hardware requirements for edge processing devices that support software upgrades must address several critical factors to ensure that these devices remain functional and up to date over extended periods. The hardware must have sufficient computational power and memory to handle the evolving complexity of software applications. Storage capacity must be flexible and expandable, allowing for the installation of future updates and new applications without compromising performance. The devices must also incorporate robust connectivity options, such as advanced wireless protocols and multiple I/O ports, to ensure seamless integration with new peripherals and communication technologies that may emerge over time. This has an initial cost because the hardware is perhaps over dimensioned for the exact initial requirements but is compensated for the user by its potential increased lifetime.

Secondly, hardware for edge processing devices must be compatible with various operating systems, RTOS and software ecosystems, providing the flexibility needed to adapt to different applications and use cases over time. This compatibility ensures that the devices can support a wide range of software upgrades and maintain interoperability with other systems. On the smartphone level, some companies (Google, Samsung, are guaranteeing that their device will be able to support software upgrade for 7 years). This is an important point that should also happen in the other domains of edge computing. But new disruptive innovations (like generative IA) can invalidate this long lifetime because of new drastic hardware requirements (mainly memory and requirement of a NPU for embedded generative IA).

Modularity

Enhancing interoperability and modularity allows different generations of devices to work together, extending their usability and reducing electronic waste. It would lead to systems which are more eco-conceived and with an augmented sustainability. The hardware should be **designed with modularity in mind**, enabling easy replacement or upgrading of individual components. This implies a long-lasting chip to chip communication protocol between units (like I_2C, SPI, PCI-e, USB, Ethernet, MIPI, AXI, etc.). **Modularity at the level of chiplets is not yet here** (adding/removing/upgrading chiplets). This modular approach not only extends the device's operational life by allowing for incremental hardware improvements but also reduces electronic waste and the overall cost of ownership.

Interoperability, modularity, scalability, virtualization, upgradability are well-known in embedded systems and are already widely applied. But they are brand new in AI and nearly non-existent in edge AI. On top, self-x (learning/training, configuration or reconfiguration, adaptation, etc.) are very promising but still under research or low level of development. Federative learning and prediction on the fly will certainly take a large place in future edge AI systems where many similar equipment collect data (Smartphone, electrical vehicles, etc.) and could be improved and refreshed continuously.

Reuse

The last aspect of increasing lifetime is to reuse a system in an application framework less demanding in terms of performance, power consumption, safety, etc. Developing the concept of a second life for components further contributes to sustainability by repurposing older hardware for new applications. Having hardware compatible with standard protocols (such as the one used for the web or cloud) could improve to find a new purpose for the hardware.

2.1.6.3.2 Key focus areas

Ensuring Long-Term Functionality and Up-to-Date Operation

A fundamental aspect of supporting the increasing lifespan of devices is ensuring that they remain functional and up to date over extended periods. This involves not only maintaining their operational capabilities but also providing ongoing support for software updates and new features. By keeping devices updated with the latest advancements and security patches, manufacturers can extend their usability and relevance, thereby maximizing their value to users and reducing electronic waste. This may incur over-specification to be able to be future proof.

- Dynamic reconfiguration and Self-X techs

  Intelligent reconfigurable concepts are an essential key technology for increasing the re-use and service life of hardware and software components. Such modular solutions on system level require the consideration of different quality or development stages of sensors, software, or AI solutions. If the resulting uncertainties (measurements, predictions, estimates by virtual sensors, etc.) are considered in networked control concepts, the interoperability of agents/objects of different generations can be designed in an optimal way.

- Dynamic reconfiguration: a critical feature of the AI circuits is to dynamically change their functions in real-time to match the computing needs of the software, AI algorithms and the data available, and create software-defined AI circuits and virtualise AI functions on different computing platforms. The use of reconfigurable computing technology for IoT devices with AI capabilities allows hardware architecture and functions to change with software providing scalability, flexibility, high performance, and low power consumption for the hardware. The reconfigurable computing architectures, integrated into AI-based circuits can support several AI algorithms (e.g. convolutional neural network (CNN), fully connected neural network, recursive neural network (RNN), etc.) and increase the accuracy, performance and energy efficiency of the algorithms that are integrated as part of software-defined functions.
- Realizing self-X (adaptation, reconfiguration, etc.): for embedded systems self-adaptation, self-reconfiguration has an enormous potential in many applications. Usually in self-reorganizing systems the major issue is how to self-reorganise while preserving the key parameters of a system (performance, power consumption, real time constraints, etc.). For any system, there is an operating area which is defined in the multi-dimensional operating parameter space and coherent with the requirements. Of course, very often the real operating conditions are not always covering the whole operating domain for which the system was initially designed. Thanks to AI, when some malfunctioning parts are identified it could be possible to decide, relying on AI and the data accumulated during system operation, if it affects the behaviors of the system regarding its real operating conditions. If this is not the case, it could be considered that the system can continue to work, with maybe some limitations, but which are not vital regarding normal operation. It would then extend its lifetime "in place". The second case is to better understand the degraded part of a system and then its new operating space. This can be used to decide how it could be integrated in another application making sure that the new operating space of the new part is compatible with the operating requirements of the new hosting system.
- Self-learning techniques are promising. Prediction on Natural Language Understanding (NLU) on the fly or keyboard typing, predictive maintenance on mechanical systems (e.g. motors) are more and more studied. Many domains can benefit of the AI in mobility, smart building, and communication infrastructure. LLMs shows interesting properties in this field, such as few-shot learning.

- Monitoring the devices and systems health

Dynamic reconfiguration and Self-X Techs also need data to be as efficient as possible. It means that monitoring of devices and systems health are necessary to acquire those data. Several techniques exist.

- Distributed monitoring: continuous monitoring and diagnosis also play a crucial role for the optimization of product lifetime. Where a large amount of data is collected during daily life operation (e.g. usage, environment, sensor data), big data analysis techniques can be used to predictively manipulate the operational strategy, e.g. to extend service life. Similarly, an increase in power efficiency can be achieved by adjusting the calibration in individual agents. For example, consider a fuel cell electric vehicle where the operation strategy decisively determines durability and service life. Distributed monitoring collects data from various interconnected agents in real-time (e.g. a truck platoon, an aircraft swarm, a smart electricity distribution network, a fleet of electric vehicles) and uses these data to draw conclusions about the state of the overall system (e.g. the state of health or state of function). This allows to detect shifting behavior or faulty conditions in the systems and to even isolate them by attributing causes to changes in individual agents in the network or even ageing of individual objects and components. Such detection should be accomplished by analyzing the continuous data stream that is available in the network of agents. A statistical or model-based comparison of the individual objects with each other provides additional insights. Thus, for example, early failures of individual systems could be predicted in advance. This monitoring should also cover the performance of the semiconductor devices themselves, especially to characterize and adjust to ageing and environmental effects and adjust operations accordingly.

- Distributed predictive optimization is possible, whenever information about future events in a complex system is available. Examples are load predictions in networked traffic control or demand forecasts in smart energy supply networks. In automation, a concept dual to control is monitoring and state observation, leading to safety-aware and reconfigurable automation systems. Naturally, all these concepts, as they concern complex distributed systems, must rely on the availability of vast data, which is commonly associated with the term big data. Note that in distributed systems the information content of big data is mostly processed, condensed, and evaluated locally thus relieving both communication and computational infrastructure.

- The other area of lifetime extension is how AI could identify very low signal in a noisy data environment. In the case of predictive maintenance for instance it is difficult for complex machinery to identify a potential failing part early in advance. The more complex the machinery, less possible is to have a complete analytic view of the system which would allow for simulation and thus identify potential problems in advance. Thanks to AI and collecting large datasets it is possible to extract some very complex patterns which could allow very early identification of parts with a potential problem. AI could not only identify these parts but also give some advice regarding when an exchange of a part is needed before failure, and then help in maintenance task planning.

- Artificial Intelligence challenges

Secure software upgradability is necessary in nearly all systems now and hardware should be able to support future updates. AI introduces additional constraints compared to previous systems. Multiplicity of AI approaches (Machine learning, DL, semantic, symbolic, etc.), multiplicity of neural network architectures based on a huge diversity of neuron types (CNN, RNN, LLMs, etc.), potential complete reconfiguration of neural networks for a same system (linked to a same use case) with a retraining phase based on an adapted set of data make upgradability much more complex. This this why HW/SW, related stacks, tools, data sets compatible with the edge AI system must be developed in synergy. HW/SW plasticity is necessary whatever the AI background principle of each system is, to make them as much as possible upgradable and interoperable and to extend the system lifetime. HW virtualization will help to achieve this, as well as standardization. The key point is that lifespan extensions, like power management, are requirements which must be considered from day one of the design of the system. It is impossible to introduce them near the end without a strong rework.

One challenge for the AI edge model is upgradability of the firmware and new learning/training algorithms for edge devices. This includes the updates over-the-air and the device management of the updating of AI/ML algorithms based on the training and retraining of the networks (e.g. neural networks, etc.) that for IoT devices at the edge is very much distributed and is adapted to the various devices. The challenge of the AI edge inference model is to gather data for training to refine the inference model as there is no continuous feedback loop for providing this data. The related security questions regarding model confidentiality, data privacy etc. need to be addressed specifically for such fleets of devices.

The novelty with AI systems is to upgrade while preserving and guaranteeing the same level of safety and performance. For previous systems based on conventional algorithmic approaches, the behavior of the system could be evaluated offline in validating the upgrade with a predefined data set representative enough for the operating conditions, knowing that, more than the data themselves, the way they are processed is important. In the case of AI, things are completely different, as the way data are processed is not typically immediately understandable, but what is key are the data set itself and the results it produces. In these conditions it is important to have frameworks where people could reasonably validate their modification, whether it is hardware or software, in order to guarantee the adequate level of performance and safety, especially for systems which are human-life-critical. Another upgrade-related challenge is that of designing systems with a sufficient degree of architectural heterogeneity to cope with the performance demands of AI and machine learning algorithms, but at the same time to be flexible enough to adapt to the fast-moving constraints of AI algorithms. Whereas the design of a new Embedded architecture or electronic device, even of moderate complexity, takes typically 1-3 years, AI models such as Deep Neural Networks are outdated in just months by new networks. Often, new AI models employ different algorithmic strategies from older ones, outdating fixed-function hardware accelerators and necessitating the design of hardware whose functionality can be updated.

Designing with Modularity and Extensibility in Mind

To achieve long-term functionality, devices need to be designed with modularity and extensibility as core principles. Modularity allows to extend the functionality of a chip for example by externally adding another support chip. For this to be practically usable, that means that a specific chip to chip interface should be added and internally connected to ensure this modularity at board/system level. This design philosophy supports easier maintenance and enables incremental improvements, which can significantly extend the device's lifespan. Extensibility ensures that the device can adapt to new technologies and requirements over time at system level, adding extra features with new supporting chips. This seems at the cost of integration and performance, but new techniques could be developed that overcome most of the disadvantages. For example, Compression Attached Memory Module (CAMM)[53] still allows to replace the memories in a system like a PC, while the recent trend is to integrate memory directly on a (passive) interposer, thus preventing any memory upgrade.

- Advancing Modularity at the Chiplet Level

  While the concept of modularity is well-established at board and system level, achieving it at the level of chiplets is an emerging challenge that has not yet been fully realized. Chiplet-based architectures involve breaking down a monolithic System on Chip (SoC) into smaller, interconnected modules (chiplets) that can be individually upgraded or replaced. This approach promises greater flexibility, improved scalability, and easier integration of new technologies. However, the industry is still in the early stages of developing standardized, interoperable chiplet designs. Advancing this modularity at the chiplet level is crucial for enabling devices to support prolonged lifespans.

- Virtual Modularity

  Virtual Modularity refers to the dynamic integration and seamless interaction between multiple devices within an interconnected ecosystem, enabling them to function as a unified system. This concept allows different hardware components and devices to pool their resources and capabilities, creating a flexible, scalable environment where services and tasks can be effortlessly transferred or shared across the network. As a result, users experience a cohesive, uninterrupted operation, regardless of which device they are using, as the underlying system intelligently manages and allocates resources to optimize performance and deliver a consistent user experience across all connected devices[54].

  Virtual Modularity can increase the lifespan of devices by distributing workloads and optimizing resource usage across a network of connected devices, rather than relying on a single device to handle all tasks. Since different devices in the ecosystem can share and collaborate on tasks, the wear and tear on individual devices is reduced. For example, a smartphone may offload demanding processing tasks to a more powerful device, such as a tablet or computer, thereby reducing its own strain and energy consumption. This collaborative approach helps prevent any one device from being overworked, which can extend its operational life. Additionally, as new devices are added to the system, older devices can continue to contribute and function effectively within the network, further extending their useful life by leveraging the collective power and capabilities of the entire ecosystem. Older devices can have their lifetime increased thanks to the extra computing, storage resources delivered by other devices in the pool.

Reuse of components or systems in a downgraded or requalified use case

Re-use: One concept called the "2nd life" is actually the re-use of parts of systems.   Such re-use could be adapted to edge AI as far as some basic rules are followed. First, it is possible to extract the edge AI HW/SW module which is performing a set of functions. For example, this module performs classification for images, movement detection, sound recognition, etc. Second, the edge AI module can be requalified and

[53] https://en.wikipedia.org/wiki/CAMM_(memory_module)

[54] This concept is developed for example in the Computing Continuum approach, and examples are the Continuity feature on Apple devices or the "super device" in HarmonyOS.

recertified downgrading its quality level. A module implemented in aeronautic systems could be reused in automotive or industrial applications. A module used in industrial could be reused in consumer applications. Third, an AI system may be re-trained[55] to fit the “2nd life” similar use case, going for example from smart manufacturing to smart home. Last, the business model will be affordable only if such “2nd life” use is on a significant volume scale. A specific edge AI embedded module integrated in tens of thousands of cars could be removed and transferred in a new consumer product being sold on the market.

2.1.6.4 Major Challenge 4: Ensuring European sustainability of embedded intelligence

2.1.6.4.1 State of the art

Edge computing and Artificial Intelligence are quickly becoming more and more tied together, converging towards the unified concept of Embedded Artificial Intelligence. To ensure the growth of this emerging technologies it is crucial that the European ecosystem of companies and academia can strategically cover all the steps of the associated value chain. This means covering the entire stack starting from the Embedded Artificial Intelligence hardware, including all the software layers required to exploit the hardware functionalities and to develop the final applications, the engineering the tools for AI development and the data sets, with a trustable and certifiable environment.

Economic and environmental sustainability represents to key factors when defining actions and investments promoting the coverage of the value chain and of the technology stack. Technology is strongly affected by sustainability that, very often, tips the scale between the ones that are promising, but not practically usable, and the ones making the difference. E.g. cloud computing, based on data centers, plays a fundamental role for the digitalization process. However, data centers consume a lot of resources (energy[56], water, etc.), they are responsible for significant carbon emissions during their entire lifecycle and generate a lot of electronic and chemical waste.

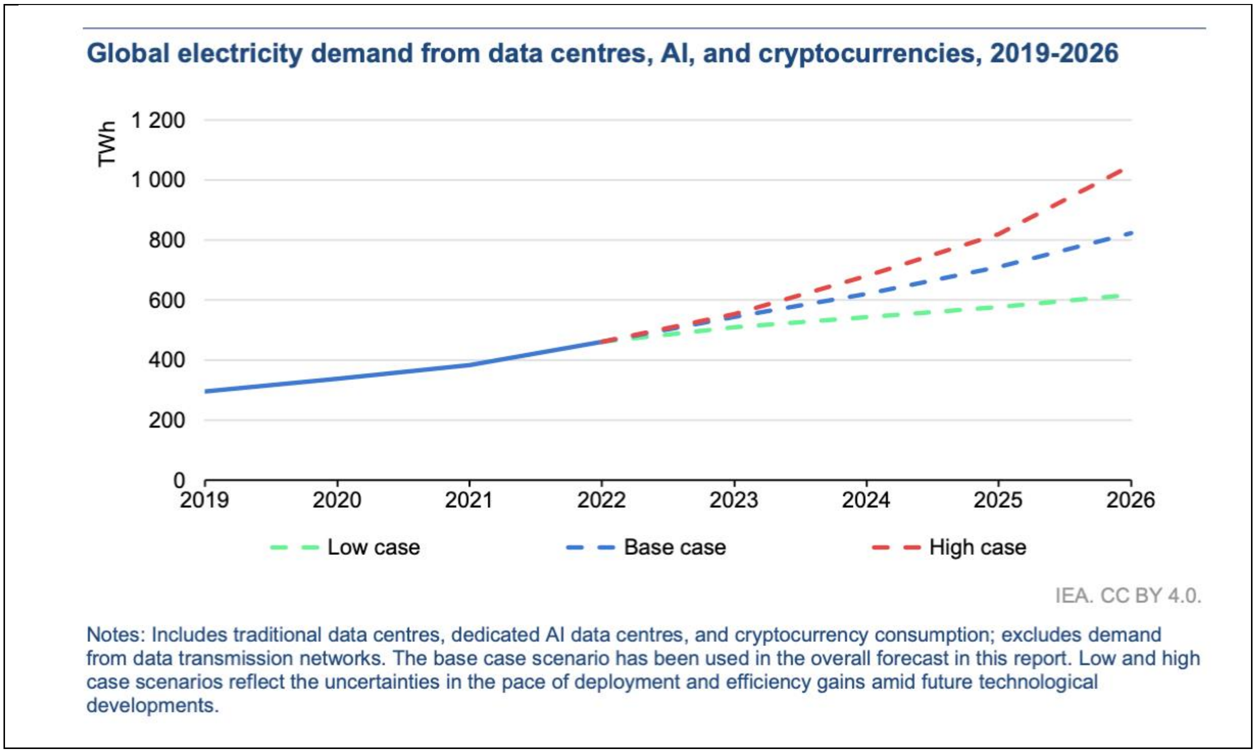
Today, the percentage of worldwide electricity consumed by datacenters and data transmission networks is 2 to 3%. “*Rapid improvements in energy efficiency have, however, helped moderate growth in energy demand. Data centres and data transmission networks are responsible for 1% of energy-related GHG emissions*” [57]. A report from January 2024 of IEA[58] indicates that “Global electricity demand from data centers could double towards 2026” (see Figure 15). Interestingly, the electricity demand to process crypto-currencies seems still to surpass the consumption related to AI in 2026 (see Figure 16).

[55] In the domain of AI using LLMs, the extension of lifetime of foundational models could be performed by several approaches, e.g by evolution, i.e. fine-tuning of the model. It might also be possible to “retune” the use of the LLMs in the latent space, i.e. without changing the parameters of the networks, but only by changing the context of the “prompts”. This needs further analysis for being effectively applicable for practical reuse.

[56] Andrae, Anders. (2017). Total Consumer Power Consumption Forecast

[57] https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks#tracking, captured in August 2024

[58] https://iea.blob.core.windows.net/assets/6b2fd954-2017-408e-bf08-952fdd62118a/Electricity2024-Analysisandforecastto2026.pdf

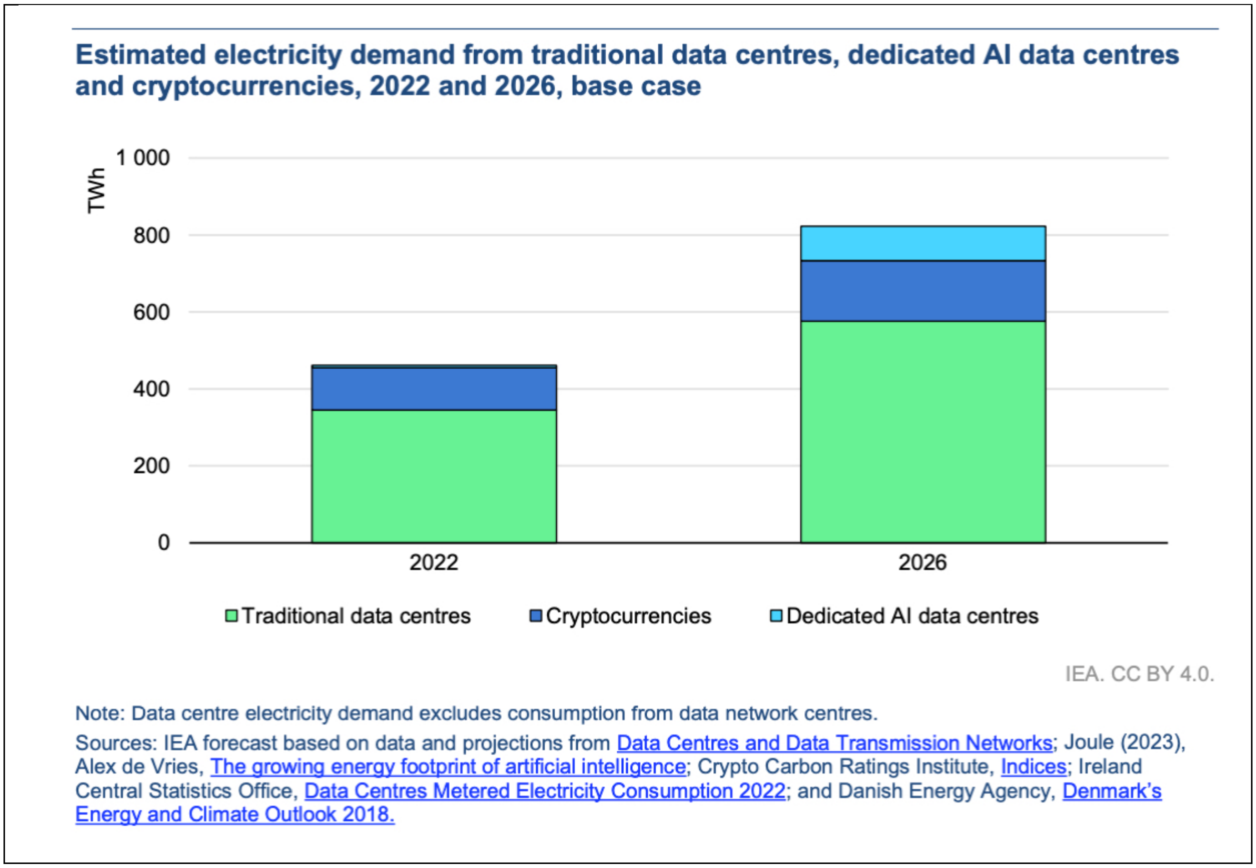


Global electricity demand from data centres, AI, and cryptocurrencies, 2019-2026

IEA. CC BY 4.0.

Notes: Includes traditional data centres, dedicated AI data centres, and cryptocurrency consumption; excludes demand from data transmission networks. The base case scenario has been used in the overall forecast in this report. Low and high case scenarios reflect the uncertainties in the pace of deployment and efficiency gains amid future technological developments.

Figure 15: Forecast of growth of global electricity demand from data centers, from [72]



Estimated electricity demand from traditional data centres, dedicated AI data centres and cryptocurrencies, 2022 and 2026, base case

IEA. CC BY 4.0.

Note: Data centre electricity demand excludes consumption from data network centres.

Sources: IEA forecast based on data and projections from Data Centres and Data Transmission Networks; Joule (2023), Alex de Vries, The growing energy footprint of artificial intelligence; Crypto Carbon Ratings Institute, Indices; Ireland Central Statistics Office, Data Centres Metered Electricity Consumption 2022; and Danish Energy Agency, Denmark's Energy and Climate Outlook 2018.

Figure 16: growth of cryptocurrency and AI centric data centers between 2022 and 2026, from [72]

Data centers are progressively becoming more efficient, but shifting the computing to the edge, for example, allows to temporally reduce data traffic, data centers storage and processing. However, only a new computing paradigm could significantly reduce their environmental footprint and ensure sustainability. Edge Computing could contribute to reach this goal by the introduction of ultra-low power and efficient computing solutions, and Embedded Intelligence (and collaborative AI agents approach) can significantly contribute to optimize their operations and, directly and indirectly, their overall energy consumption profile.

Shifting to green energy is certainly a complementary approach to ensure sustainability, but the conjunction of AI and edge computing, the Embedded AI, has the potential to provide intrinsically sustainable solutions with a wider and more consolidated impact. Indeed, a more effective and longer-term approach to sustainable digitalization implies reconsidering the current models adopted for data storage, filtering, analysis, processing (including AI training itself), and communication. By embracing Edge Computing, for example, it is possible to significantly reduce the amount of useless and wasteful data flowing to and from the cloud and data centers, because it means adopting architectural and structural solutions that are more efficient and that permanently reduce the overall power consumption. These solutions bring also other important benefits such as real-time data analysis, reduced memory and storage capacity and better data protection. The Edge Computing paradigm also makes AI more sustainable: it is evident that cloud-based machine learning inference is characterized by a huge network load, with a serious impact on power consumption and huge costs for organizations. Transferring machine learning inference and data pruning to the edge, for example, could exponentially decrease the digitization costs and enable sustainable businesses. To avoid this type of drawbacks, new AI algorithm specifically tailored for embedded systems, in conjunction with new HW components (e.g. based on neuromorphic architectures), should be design and developed and, considering the application areas, in some cases, this could lead to more specialized and extremely efficient solutions.

Sustainability of Edge Computing and AI is affected by many technological factors, which require investments to consolidate and strengthen EU global positioning, generating a potential positive fallback on the sustainability of future digitalization solutions and related applications. GAMAM already master these technologies and are already controlling the complete value chain and technology stack associated with them: to achieve strategic autonomy in this field, Europe must cover the entire value chain, consolidating the existing segments, identifying/filling the technology gaps, and leveraging on a strong cooperation between the European stakeholders, with a particular attention to SMEs (which generate a large part of European GDP). From this perspective, European coordination to develop AI, edge computing and Embedded AI technologies is fundamental to create a sustainable value chain, based on solid alliances, and capable to support the European key vertical applications.

Ensuring sustainability for Embedded AI in Europe involves several key initiatives: open-source hardware can drive innovation and reduce costs, while energy efficiency improvements are critical for sustainable development. Engineering support represents the third key factor and is necessary to advance sustainable Embedded AI from several perspectives: optimize the energy profile of Embedded AI solutions (technical and environmental levels), tackle the lack of expertise and of human resources (operational/engineering level), keep the costs of these solutions sustainable (economic). For example, by mastering the adoption of new Embedded AI techniques, such as transformers and large language models (LLMs) with a focus on federation of smaller specialized models developed in Europe[59], and integrating them with efficient accelerators and algorithms, Europe can maintain a competitive edge. This is crucial for many vertical domains that are key for Europe: for example, in the domain of autonomous systems, Europe has the potential to lead in the development of autonomous vehicles and robots, leveraging advanced AI and embedded intelligence. The necessity to ensure an efficient engineering process, with consequent short time to market, is motivated by the speed of global market evolution: for example, the emergence of cheap Chinese RISC-V microcontrollers emphasizes the necessity for fast development, robust, cost-effective computing solutions made in Europe. Another example from a vertical application perspective is the partitioning of complex systems on a chip (SoC) into chiplets and interposers, which requires the development of an ecosystem of interoperable chiplets, comprehensive architecture and engineering solution to support this approach: the automotive market can significantly benefit from this, maintaining Europe at the global market forefront.

Sustainability can be measured also in term of democratizing technology use, to make it more accessible to a broader community of users and facilitate market entry to startups and SMEs: natural language interfaces for example could contribute to facilitate the access to digital technologies. This requires that the associated AI models are trained with European data covering European languages, respecting European ethics. Additionally, democratizing technology means also supporting the diversity of European computing systems/chips, simplifying their programmability and consolidating this diversity through interoperability. This is also crucial to support the research into new computing paradigms, like neuromorphic and analog computing and to ensure they can reach high TRL levels (i.e. 7, 8 and 9).

One of the major challenges that need to be accounted for in the next few years is related to the design of progressively more complex electronic systems to support advanced functionalities such as AI and cognitive functionality, especially on the edge. This is particularly challenging in the European landscape, which is dominated by small and medium enterprises (SMEs) with only some large actors that can invest and support larger-scale projects. To consolidate and strengthen European competitiveness and ensure sustainability in advanced Embedded AI solutions it is therefore crucial to create an ecosystem covering the entire value chain, in which SMEs can cooperate and increase their level of innovation and productivity. The definition of open industrial standards and a market of Intellectual Properties (IPs) is crucial to accelerate the design and development of these solution, guaranteeing EU competitiveness and increasing the market dimensions. Open-source software, hardware and engineering tools can play an extremely important role in this regard, because they allow to significantly reduce engineering costs for licensing and verification, lowering the entry barrier to design innovative products.

2.1.6.4.2 Key focus areas

Leveraging Open-source Hardware and Software for Innovation and Cost Reduction

Open-source Hardware and Software are drivers of innovation and cost reduction in the fields of Edge Computing and Embedded AI, allowing democratization of design and creating a collaborative environment where innovation can emerge and making possible the development of cutting-edge

[59] For example, see https://mistral.ai/

solutions that are both affordable and highly effective. Open-source also facilitates transparency and security, which are essential factors for building trust and compliance with European standards.

Developing and Federating Smaller Specialized AI Models

The federation of smaller specialized AI models developed in Europe represents a strategic approach to harnessing localized expertise and enhancing the efficiency of AI applications at the edge. These models, when federated (in "*Agentic AI*"), can work together to tackle complex tasks as, or even more effectively than monolithic models. Promoting the development of these models in Europe contributes to the alignment of technology with European technologies and applications necessities, leveraging regional strengths in AI research and development. Specialized agents can be derived from few foundation models (made by few European companies or from Open Source) by fine tuning that can be achieved by SMEs or start-ups, as this specialization requires less computing power and data than the creation of a foundation model.

Training Models with European Data and Ethical Compliance

For AI models to be truly sustainable and beneficial to Europe, they must be trained on data that covers European languages and reflects the diverse cultural and social contexts of the continent. This ensures that AI applications are more accurate, inclusive, and useful across different European regions. Additionally, adherence to European ethical standards in AI development is crucial. This can be achieved:

- Finding solutions which ensure privacy, fairness, and transparency in AI systems, thereby aligning technological advancements with the core values and regulations of Europe.
- Allowing those AI to work at the edge, hence promoting the development of specialized AI chips for edge AI, this will facilitate access to digital technologies to a broad audience.
- Supporting natural language interfaces which enable users to interact with technology in intuitive ways, breaking down barriers to adoption and use, in a continent characterized by cultural, societal and language diversity.
- Increasing the re-use and sharing of knowledge and models generated by embedded intelligence, improving the energy- and cost-efficiency of AI training, and adopting new benchmarking AI approach oriented to sustainability.

Deploying Efficient and Sustainable Embedded AI-oriented ECS

Efficient ECS like new memories, accelerators and algorithms specifically conceived for Embedded AI are fundamental to the performance and sustainability of domain. Developing and deploying these technologies within Europe ensures that AI applications can operate with high efficiency, reducing energy consumption and enhancing overall system performance. Most current designs involve perception-based AI accelerators, but Europe should be at the forefront to use new disruptive technologies such as generative AI and ensuring its sovereignty by developing the HW and SW that will enable it at the edge. This will require for example:

- New materials and substrates oriented to low power consumption.
- New generations of embedded ultra-low power electronic components.
- 3D-based device scaling for low power consumption and high level of integration.
- Chiplet-based solutions promoting modularity and reuse (sustainability by design).
- Strategies for self-powering nodes/systems on the edge and efficient cooling solutions.
- Policies and operational algorithms for power consumption at edge computing level.

- Efficient and secure code mobility.
- Advanced Neuromorphic components.
- Generative AI for the edge.
- Inclusion of existing embedded systems on the edge (huge market opportunity).

Accelerating Development of Robust, Cost-Effective Solutions

There is a pressing need for the rapid development of robust, cost-effective computing solutions made in Europe. These solutions must be designed to meet the specific demands of European markets and regulatory environments. By focusing on speed of realization, reliability, and cost-effectiveness, European manufacturers can remain competitive and responsive to global and local market needs. This involves developing the full spectrum of technologies, from hardware innovations to sophisticated software solutions, to new tools allowing to increase productivity ensuring that European products stand out in the global marketplace. For example, this must include:

- Engineering process automation for full lifecycle support of Embedded AI solutions.
- Edge Computing and Embedded AI security and sustainability by design.
- Engineering support for Embedded AI verification and certification, addressing end-to-end edge solutions.
- Support for education and professional training to enable, facilitate, improve and speed-up the design, development and deploy of Embedded AI.
- Engineering process automation based on generative AI.

2.1.7 Timeline

As the field concerning Artificial Intelligence is evolving very rapidly, all predictions are subjective, especially for 2035 and beyond (the foundation paper about Transformers, which lead to the current LLMs, was only published in 2017[60], so in 2015 nobody could have predicted what we are seeing today in this field). It is why the long-term column (2035 and beyond, is rather empty).

Legend:

- (EC): edge computing
- (eAI:) Embedded Artificial Intelligence

[60] https://arxiv.org/abs/1706.03762

MAJOR CHALLENGE	TOPIC	SHORT TERM	MEDIUM TERM	LONG TERM
		2025-2029	2030-2034	2035 AND BEYOND
Major Challenge 1: increasing the energy efficiency of computing systems	**Processing data locally and reducing data movements: towards the computing continuum**	Move towards the continuum of computing: orchestrators can select computing where it is the most efficient (statically). Using the same software development infrastructure from deep edge to edge and possibly HPC applications. Use of similar building blocks from deep edge to edge devices. Developing open architectures (for fast development) with maximum reuse of tools and frameworks Development of edge (ex: fog) type of computing (peer to peer), development of distributed systems e.g. for federated learning or fine tuning of LLMs reducing need to exchange data. Unified memory: avoid copying data from the CPU memory to the accelerator(s) memory(ies).	Advance storage management: towards distributed storage supporting the continuum of computing approach also for storage. Move towards the continuum of computing: orchestrators can select or move computing where it is the most efficient (dynamically) avoiding data movements.	
	Co-Design of algorithms, hardware, software	AI can be used to help this codesign. This is mainly a topic for the "methodology and tools" chapter. Tools allowing fast realization of hardware accelerators when a new paradigm emerges (e.g. Transformers/generative AI). Tools allowing semi-automatic design exploration of the space of configurations, including variants of algorithms, computing paradigms, hardware performances, etc. Complete 2.5D (interposers and chiplets) ecosystem, with tools increasing productivity and reuse of chiplets in different designs Automatic adaptation of complex neural networks or emerging AI algorithms to embedded systems with a	Systems can generate hardware (orchestrators, accelerators, storage and communication) and the associated low-level software from high level specifications. Auto-configuration of a distributed set of resources to satisfy the application requirements (functional and non-functional) Supporting tools integrating multiple computing paradigms in the same package.	

		minimum loss of performances		
	Efficient management of storage resources	Compressing weights (from FP32 to INT4 or less), pruning neural networks. Federations of multiples devices into “meta-devices” that share storage resources. Dynamic use of adapters/LoRA to switch the specialization of a foundation model (changing a small proportion of its weights[61]). Easily upgradable LLM accelerators, fine tuning “on premises”. Innovation in memory technology: Using directly the parameters of neural networks from non-volatile memory without transferring them to RAM, computing near/in memory Create gateways between various solutions, beyond ONNX (for eAI)	New memory technology with last cost, fast access and certain level of non-volatility allowing the redefine completely the memory hierarchy.	
	Energy proportionality	Scalable accelerators where parts can be switched off. Refining dynamic power management techniques to ensure that energy consumption closely matches the workload, minimizing power wastage during low activity periods. Advancements in adaptive voltage scaling, power gating, and more efficient power distribution networks.	Use AI to dynamically manage the repartition of active parts in a chip.	
	Ultra-low standby current	Improving energy efficiency at low energy level, reducing leakage currents in devices, optimizing power management	Pushes further towards near-zero standby power. Overcoming the limitations of current materials.	

[61] It seems that Apple is doing something similar in its « Apple Intelligence »

		at the software level. Develop always-on units that can react to information in order to wake-up the rest of the system.	Developing new semiconductor technologies that can operate at ultra-low power levels. Integrating advanced energy harvesting techniques to sustain device functionality even in standby mode.	
	Development of innovative hardware architectures	Development of computing paradigms (e.g. using physics to perform computing, e.g. neuromorphic). Use of other technologies than silicon (e.g. photonics) Use of 2.5D, interposers and chiplets, with efficient interconnection network, e.g. using photonics) Creating an ecosystem around interposers and chiplets, with interoperability standards New In-memory computing accelerators Supporting software for all these hardware innovations	Dynamic instantiation of multi-paradigm computing resources according to the specifications of the task to be performed. New (de facto) standard allowing automatic interfacing, discovery, and configuration of resources (computing, storage) Global reconfiguration of the resources to satisfy the functional and non-functional requirements (latency, energy, etc.) Development of hybrid architectures, with smooth integration of various processing paradigms (classical, neuromorphic, deep learning), including new OSs supporting distributed computing of multiple computing paradigms Advanced In-memory computing accelerators	Integration in the same package of multiple computing paradigms (classical, Deep Learning, neuromorphic, photonic, quantum, etc.) Exploring potential use of quantum computing in Artificial Intelligence?

Major Challenge 2: managing the increasing complexity of systems	**Balanced mechanisms between performance and interoperability**	Exposing the non-functional characteristic of devices/blocks and off-line optimization when combining the devices/blocks Explore AI techniques for Self-x	On-line (dynamic) reconfiguration of the system to fulfil the requirements that can dynamically change (Self-x) Use of AI techniques for Self-x.	Drive partitioning through standards
	Developing distributed edge computing systems	Introduce standard APIs comprising Web based technologies; support of a layered AI architecture See items above in *Increasing the energy efficiency of computing systems*	Development by AI, e.g. of communication and functional collaboration, orchestration. Agentic (AI) approach See items above in *Increasing the energy efficiency of computing systems*	See items above in *Increasing the energy efficiency of computing systems*
	Scalable and Modular AI	Decomposition of SoCs, e.g., into chiplets See items above in *Increasing the energy efficiency of computing systems*	Ecosystem of chiplets for embedded control and embedded AI established Data and learning driven circuits design See also items above in *Increasing the energy efficiency of computing systems*	See items above in *Increasing the energy efficiency of computing systems*
	Easy adaptation of models	Development of efficient and automated transfer learning: only partial relearning required to adapt to a new application (Ex: Federative learning) Create a European training reference database for same class of applications/use	Optimization of the Neural Network topology from a generically learned networks to an application specific one.	Generic model based digital AI development system

		cases network learning		
	Easy adaptation of modules	Easy migration of application on different computing platforms (different CPU – x86, ARM, RISC-V, different accelerators)	Use of HW virtualization Automatic transcoding of application for a particular hardware instance (à la Rosetta 2)	Generic model based digital development system
	Realizing self-X **Self-optimize, reconfiguration and self-management**	Add self-assessment features to edge devices Explore what AI techniques (such as LLMs) can do	Automatic reconfiguration of operational resources following the self-assessment to fulfil the goal in the most efficient way Deploy AI based approaches for self-optimization	Modelling simulation tools for scalable digital twins
	Using AI techniques to help in complexity management	Using AI techniques for the assessment of solutions and decrease the design space exploration	Automatic generation of architecture according to a certain set of requirements (in a specific domain)	Modelling simulation tools for scalable digital twins
	Using AI techniques modeling of interactions among system components	Explore usage of AI-based methods, tools and environments for the modeling and simulating of the system designs and their components	Leveraging networks of agents (LLMs and other AI-based programs) for automatized modeling and simulations	

Major Challenge 3: supporting the increasing lifespan of devices and systems	**HW supporting software upgradability** **(eAI)**	Create European training reference databases for same class of applications/use cases network learning Develop European training benchmarks (Methods and methodologies) Build framework tools for HW/SW for fast validation and qualification Establish interfaces standards compatible with most of AI approaches	HW virtualization based on AI algorithms Generic AI functions virtualization European training standards (Compliance/Certification) Certifiable AI (and paths towards explainability and interpretability) Interoperability and extensibility within computing continuum	Explainable AI Universal control
	Realizing self-X Also partially in ***Managing the increasing complexity of systems*** ***(eAI)***	Unsupervised learning technics Development of efficient and automated transfer learning: only partial relearning required to adapt to a new application (Ex: Federative learning)	HW virtualization based on AI algorithms Generic AI functions virtualization Certifiable AI (and paths towards explainability and interpretability) Use of Mixed of AI Agents and/or Transformers for generative AI on self-X Techs	Explainable AI Universal control
	Improving interoperability (with the same class of application) and between classes, modularity, and complementarity between generations of	Developing open architectures (to quickly develop) with maximum reuse of tools and frameworks Interfaces standards (more than solutions) (could help explainability move from	Generic functions modules by class of applications/use cases + virtualization Use of Mixed of AI Agents and/or Transformers for generative AI Chiplets	

	devices. **(EC)** Also, partially in ***Increasing the energy efficiency of computing systems***	black to grey boxes)	Virtual Modularity	
	Improving interoperability of AI functions (with the same class of application) and between classes, modularity, and complementarity between generations of devices. **(eAI)** Also, partially in ***Increasing the energy efficiency of computing systems***	Developing open AI architectures (to fast develop) with maximum reuse of tools and frameworks Interfaces standards (more than solutions) (could help explainability of AI with a move from black to grey boxes) Clarified requirements for embedded AI in industry Applications and Apps working simultaneously through different equipments	Generic AI functions modules by class of applications/use cases + virtualization	Universal control
	Developing the concept of 2nd life for components **(EC)** (Link with sustainability)	Inclusion of existing embedded systems on the edge (huge market opportunity)	Generic set of functions for multi-applications/use cases Library of generic set of functions (Standardization) Basic data collection for predictive maintenance Global data collections for predictive maintenance by applications/use cases	Standardize flow for HW/SW qualification of generic set of functions (including re-training) which are used in a downgraded application/use case Full chain of reuse / Ecodesign conception

Major Challenge 4: ensuring European sustainability of embedded intelligence	**Leveraging Open-source Hardware and Software for Innovation and Cost Reduction**	Open-source eAI software Open-source training datasets for eAI Open-source foundation models for eAI	Open-source eAI hardware	
	Developing and Federating Smaller Specialized AI Models	AI specialized models tailored for eAI eAI models for specific applications eAI models modularity and reuse Partial eAI models federation	Interdisciplinary eAI models eAI models covering cross-domain applications Extensive eAI models federation	
	Training Models with European Data and Ethical Compliance	European-native training data sets Methods and HW/SW solutions respecting privacy, fairness, and transparency Natural language support in embedded systems Energy- and cost-efficiency of eAI training Specialized eAI chip to facilitate market entry	European-native training data sets Methods and HW/SW solutions respecting privacy, fairness, and transparency Natural language support in embedded systems Energy- and cost-efficiency of eAI training	
	Deploying Efficient and Sustainable Embedded AI-oriented ECS	Materials and electronic components oriented to low and ultralow power solutions Strategies for self-powering nodes/systems on the edge Efficient cooling solutions Chiplet-based solutions for modularity and reuse (sustainability by design) Inclusion of legacy	3D-based device scaling for low energy consumption Efficient and secure code mobility. Advanced Neuromorphic components.	

	embedded systems Neuromorphic components Integrated power consumption management		
Accelerating Development of Robust, Cost-Effective Solutions	Engineering process automation Continuous engineering across the product life cycle eAI security and sustainability by design eAI HW and SW modularity Adoption of generative AI in the eAI engineering process	Holistic development environment for eAI Engineering support for verification and certification eAI security and sustainability by design Reuse of knowledge and models generated by embedded intelligence	