INSIDE
Industry Association

EPoSS.
European Association on
Smart Systems Integration

ARTIFICIAL INTELLIGENCE AT THE EDGE

A JOINT EUROPEAN ROADMAP FOR EDGE AI

October 2025

IMPRINT

EPoSS e. V. European Association on Smart Systems Integration
Steinplatz 1, 10623 Berlin, Germany

www.smart-systems-integration.org

Layout
Juliane Lenz – Berlin, Germany

Contact
Inessa Seifert – VDI/VDE Innovation + Technik GmbH | EPoSS
Inessa.Seifert@vdivde-it.de

Content

1 Introduction

In recent years, digitalisation, the availability of data and the possibilities for applying Artificial Intelligence (AI) have become important business drivers for Europe's key industrial sectors. In our understanding, AI is a technical system that has the ability to mimic human intelligence, which is characterised by behaviours such as sensing, learning, understanding, decision-making and acting. Due to the availability of powerful computing hardware (graphics processing units (GPUs) and specialised architectures) and large amounts of data, AI solutions – in particular Machine Learning (ML), and more specifically Deep Learning (DL) – have found numerous and widespread applications over the last two decades (including image recognition, fault detection and automated driving functions).

Low latency, privacy, connectivity limits and distributed applications have driven research in Edge AI, which enables processing and decision-making near data sources – across cloud, edge, and Internet of Things (IoT) devices. It involves training AI models in the cloud and deploying them on edge devices.

In 2021, the EPoSS Edge AI Working Group published a white paper called "AI at the Edge" [1], which provided a broad overview of AI methods and techniques, together with technological milestones to guide the research and innovation over the next few years.

Following the publication of this white paper, two industry associations – EPoSS and INSIDE – joined forces. The joint Edge AI Working Group is a community of hardware and software experts from industry and academia who drive research and innovation for both national and EU-funded projects, and contribute their insights and views concerning the future of Edge AI.

Recent breakthroughs, and in particular in the domain of Generative AI (GenAI), have driven a clear need to revise our roadmap, including the technology milestones, to better understand and exploit the potential of GenAI in the computing continuum, including at the edge. ***Figure 1.1*** shows how to read our refined and updated Vision.

1 *EPoSS Whitepaper, 2021, "AI at the Edge" (available at https://www.smart-systems-integration.org/publication-eposs-ai-white-paper)*

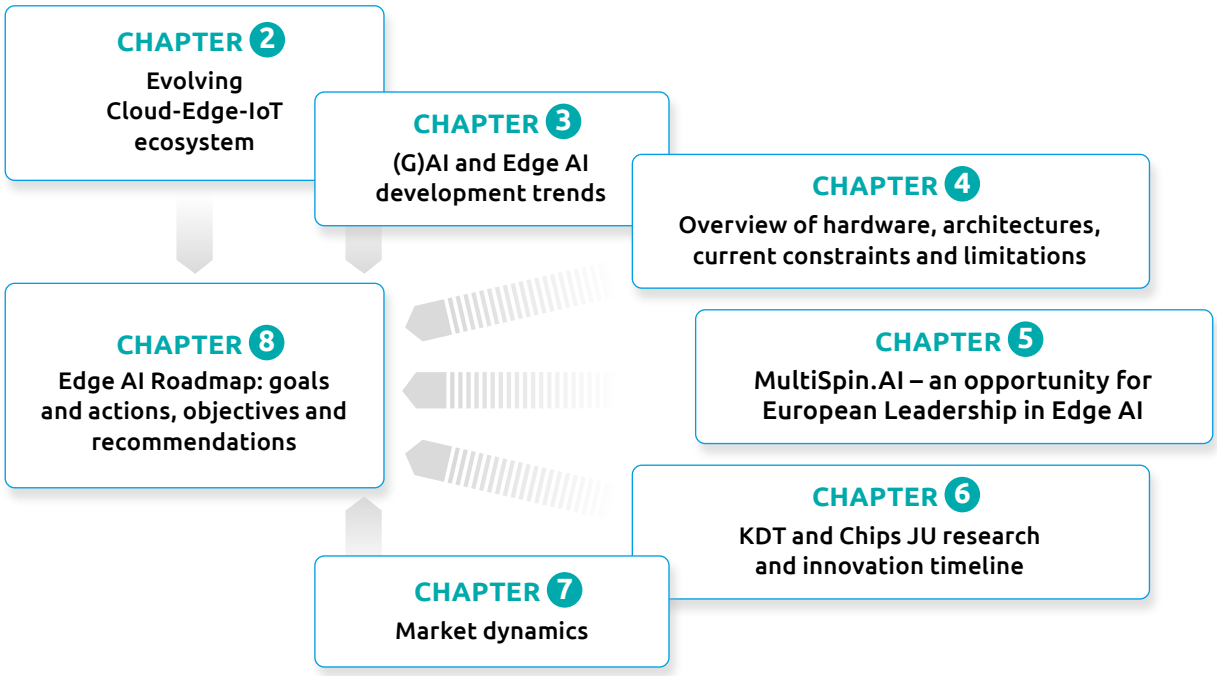


Figure 1.1: *How to read this document*

This white paper begins with an overview of the evolving cloud-edge-IoT ecosystem, highlighting the critical role of intelligent, resource-constrained devices that interact with both humans and machines. ***Chapter 3*** then explores the current AI trajectory, including the five levels of Artificial General Intelligence (AGI) coined by OpenAI CEO Sam Altman[2]. ***Chapter 4*** dives into cutting-edge hardware architectures, while ***Chapter 5*** examines the many challenges, constraints and limitations around innovation in hardware for Edge AI development. ***Chapter 6*** introduces a novel spintronics-based solution that addresses the memory-wall issue with impressive energy efficiency and performance. ***Chapter 7*** outlines the timeline and expected outcomes of KDT and Chips Joint Undertaking (Chips JU) projects in the coming years. ***Chapter 8*** analyses global market trends, spotlighting Europe's Edge AI landscape and NVIDIA's growing dominance in the ecosystem. The final chapter outlines important goals, objectives and recommendations for action that will boost the competitiveness of European companies, building on the insights from earlier chapters.

2 https://www.forbes.com/sites/jodiecook/2024/07/16/openais-5-levels-of-super-ai-agi-to-outperform-human-capability

2 Evolving Cloud-Edge-IoT Infrastructures and Data-driven Value Chains

The distributed and resource-constrained nature of edge computing presents challenges that are different from those of centralised computing. Deploying AI on edge devices presents significant technical challenges, largely due to heterogeneity: the variety of hardware platforms, real-time operating systems, sensor types, and AI workloads. While classic AI is now effectively deployed at the edge, GenAI has introduced new complexities. From around 2014 with the rise of Generative Adversarial Networks (GANs) and popularised by breakthroughs such as Transformers (2017), GenAI workloads have become increasingly hyperparameterised and resource-intensive.

The findings collected in the study "Transitioning from TinyML to Edge GenAI: A Review"[3] underscore the growing interest in deploying Edge GenAI models specifically on smartphones. For instance, imagine a hypothetical service designed specifically for Gen Z smartphone users. It comes with a USD15 monthly subscription and sets a strict performance expectation: latency must not exceed five seconds. Meeting these demands at scale presents significant challenges, raising questions about the readiness of the current infrastructure for widespread deployment.

A case study with Qwen2-VL-7B-Instruct[4], a cutting-edge multimodal GenAI model, highlights some of the key scalability challenges. With modest usage assumptions (60 tokens per user per query, and a five-second latency limit), serving all 5.16 billion smartphone Gen Z users would demand:

- over 40,000 AI superclusters (each on the scale of NVIDIA's Cortex AI cluster[5]);
- power infrastructure of up to 130 MW per cluster; and
- unfeasible levels of acceleration and cost.

In short, large-scale GenAI deployment via the cloud is neither economically nor environmentally sustainable. However, for training GenAI models, cloud computing remains essential; to preserve data privacy and sovereignty, on premises AI training is also a promising direction to attain some relief from cloud dependency.

3 https://www.mdpi.com/2504-2289/9/3/61

4 https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct

5 https://technologymagazine.com/ai-and-machine-learning/a-first-look-at-elon-musks-new-cortex-ai-supercluster

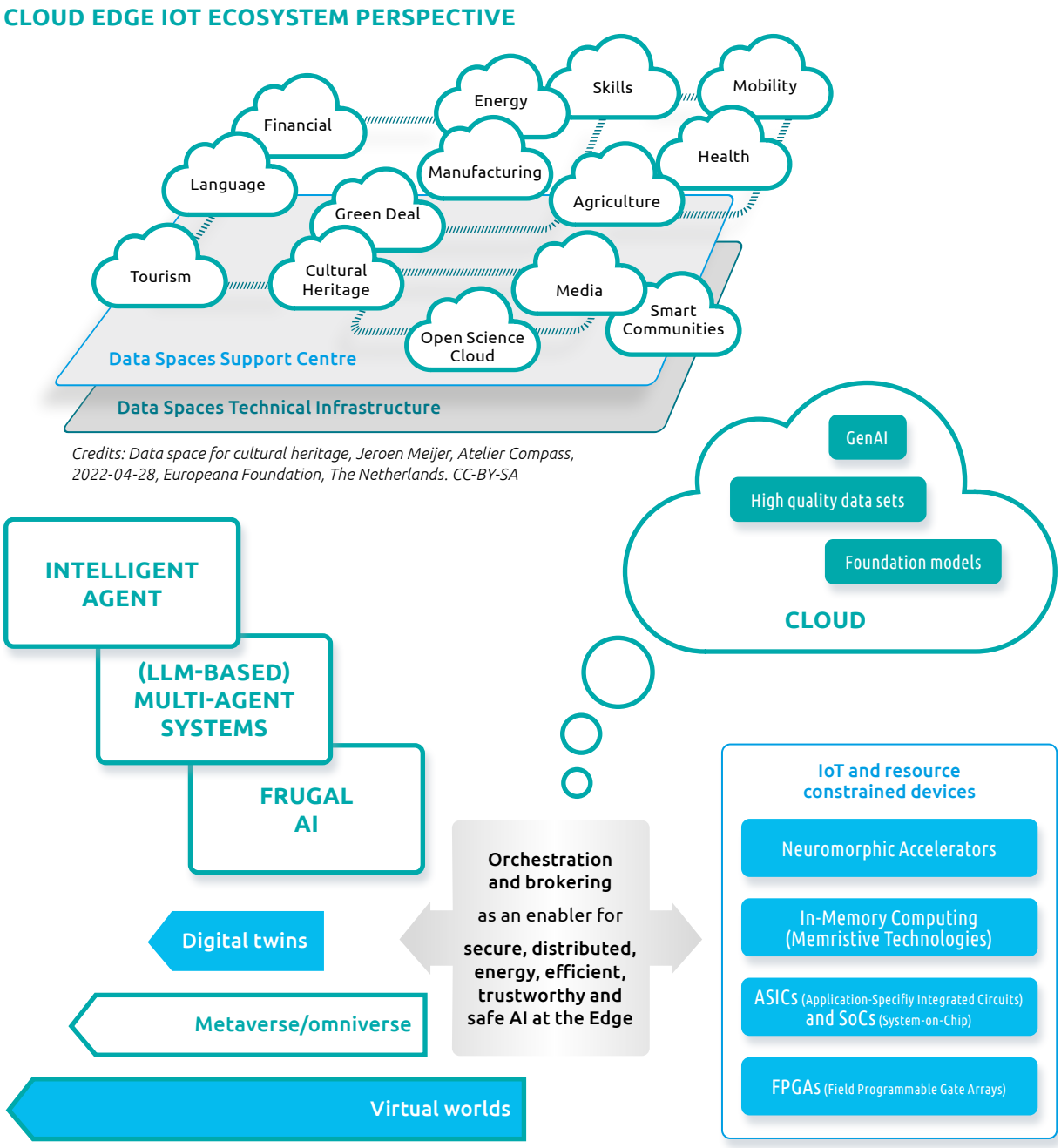


Figure 2.1: *Cloud-Edge-IoT ecosystem view*

In Edge AI systems, data is both collected and processed locally at or near the edge of the network, leveraging IoT devices and resource-constrained hardware. Cloud-edge-IoT infrastructures must be highly adaptable to accommodate varying data volumes, velocities, and privacy and security requirements. The data journey begins with collection at tiny sensors, data generators, and micro-devices. Based on the application's needs and privacy considerations (*see* ***Table 7.1***), the data is either processed locally or transferred to cloud or high-performance computing infrastructures for advanced optimisation and decision-making tasks.

The tech stack for data-driven Edge AI consists of several interconnected layers that enable the collection, processing and application of data. The key building blocks include the following.

- **GenAI, foundation models, high-quality datasets and data spaces:** Robust AI solutions at the edge rely on foundation models and high-quality datasets. Common European Data Spaces offer the infrastructure for federated, distributed sharing of these datasets.
- **Multi-agent systems**: Powered by specialised Large Language Models (LLMs) and foundation models, these agents deliver high performance while being optimised for resource-constrained devices such as smartphones. They enable advanced AI functionalities directly at the edge.
- **Digital twins, metaverse/omniverse, and virtual worlds:** Virtual models of physical objects use real-time sensor data to simulate behaviour, monitor operations, and optimise performance throughout their lifecycle.
- **Neural architecture search:** To automatically devise AI models to solve edge problems by leveraging on-premises AI energy-efficient computing and data availability.
- **Orchestration and brokering:** Automating the configuration, management and coordination of systems, applications, services and devices for streamlined operations.
- **Trust and security:** Incorporating software and hardware components to ensure system reliability, privacy, robustness, dependability, safety and performance, all critical for secure deployments.

Each of these building blocks represents an innovation area together with market opportunities, with emerging or established players driving innovation to accelerate Edge AI adoption across the computing continuum.

The broader view aims to illustrate key interactions within the ecosystem, revealing the complexity of dependencies together with associated challenges and potential risks. In this context, ***Chapter 4***, "Overview of New Hardware Architectures", focuses on the specific challenges of running Edge AI on resource-constrained devices. This ecosystem perspective offers a strategic lens to understand the research and innovation activities of KDT and Chips JU projects described in ***Chapter 7***, as well as market structure with the positioning of dominant players explored in ***Chapter 8***. The next chapter, "AI and Edge AI Development Trends: Setting the Scene", explores the evolution of AI, highlighting key trends that are shaping the future of innovation in Edge AI.

3 AI and Edge AI Development Trends: Setting the Scene

3.1 Most discussed Edge AI topics

AI is the most rapidly developing technologies that is affecting and challenging the current technological landscape. According to Gartner's Hype Cycle[6], Edge AI has surpassed its peak and is expected to reach a "plateau of productivity" within two years. This signifies the technology's transition through its initial phases of hype, disillusionment and experimentation, ultimately becoming a standard and reliable tool for various use cases. Furthermore, according to the Bank of America, the Taiwan Semiconductor Manufacturing Co (TSMC) will enable USD1 trillion in manufacturing digital chips by 2030[7] for its driving customers through AI computer servers, including on-premises AI, Edge AI, tiny, and in particular agentic, including humanoid robots[8].

GenAI introduces new challenges, particularly in the context of distributed computing environments. The training of generative AI models, especially LLMs, requires a huge amount of computing power and energy, usually provided by cloud computing infrastructures and efficient data centres. According to Yann LeCun, modern LLMs are trained with 20 trillion tokens, with each token comprising three bytes – so that's 10^{14} tokens! In the first four years of life, the brain receives 16,000 hours of visual information at 2 MB/s. This is the equivalent amount of information needed to train an LLM. Therefore, for the foreseeable future, we will be very far (perhaps light years) from achieving superhuman intelligence. It remains to be seen what computing and energy resources would be required to power such a computer should humanity ever reach that point.

High-quality datasets are fundamental to the training of LLMs as they ensure the development of accurate, unbiased and comprehensive representations of language. These datasets minimise the propagation of errors and biases, thereby enhancing the model's generalisation capabilities and reliability. High-quality data collected from IoT devices and sensor networks reduces noise during training, enabling the model to focus on meaningful patterns and relationships for more efficient learning. This ensures that LLMs achieve higher performance, particularly in real-world applications and complex tasks requiring contextual understanding and domain-specific expertise. Consequently, the quality of training data directly influences the trustworthiness, applicability and ethical deployment of LLMs across diverse fields. The objective of Common European Data Spaces[9] is to establish uniform data infrastructures and governance frameworks that enable data pooling, access and sharing. This allows them to provide high-quality resources for data-driven AI-based applications.

The recent breakthroughs in AI technologies have had a significant impact on the technology landscape. The most intensively discussed areas in the Edge AI community are currently the following.

- **LLMs** enable machines to understand, reason and generate human-like language, revolutionising natural language processing (NLP) tasks.
- **GenAI** enables the creation of novel content such as images, music and text using advanced transformer and other architectures of generative models.

6 https://www.gartner.com/en/articles/hype-cycle-for-artificial-intelligence

7 https://www.investing.com/news/stock-market-news/how-tsmc-is-enabling-1-trillion-semiconductor-era-4010839

8 https://www.forbes.com/sites/johnkoetsier/2025/04/30/humanoid-robot-mass-adoption-will-start-in-2028-says-bank-of-america

9 https://digital-strategy.ec.europa.eu/en/policies/data-spaces

- **Responsible AI** focuses on building trustworthy AI systems that prioritise ethical decision-making, fairness and societal well-being. It also promotes transparency and accountability in AI processes. However, it requires the creation of governance frameworks and regulatory policies to align AI development with the principles of responsibility, sustainability and social impact.
- **Multi-agent AI systems (MAS)** are composed of multiple intelligent agents that can sense, search information, learn and act autonomously to achieve individual and collective goals. Powered by artificial reasoning intelligence, these systems demonstrate building sequences of thought capabilities by being flexible, scalable and robust to enable broader real-world impact across industries. MAS involve multiple interacting agents – software or hardware entities – that work together to solve complex problems beyond their individual capabilities.
- **Embodied (physical) AI** refers to the use of AI techniques to solve problems that involve direct interaction with the physical world – for example, by observing the world through sensors or modifying the world through actuators. It integrates AI into physical systems, and is increasingly combined with digital twins and simulations to improve performance and decision-making in various industries.
- **AI and quantum computing** is still an emerging technology, promising breakthroughs in optimisation, cryptography and drug discovery through quantum speed-ups. It has accelerated the need for hybrid AI-quantum algorithms, and novel computational and open programming frameworks.

One of the most debated emerging topics in AI is **Artificial General Intelligence (AGI)**[10]. According to Gartner, AGI refers to AI that can understand, learn and apply knowledge across a wide range of tasks and domains. Unlike narrow AI, which is designed for specific applications, AGI possesses cognitive flexibility, adaptability and general problem-solving skills.

AGI is defined as AI capable of surpassing human performance in most tasks. Sam Altman, CEO of Open AI, a leading force in GenAI and the creator of ChatGPT, uses a **five-tier scale** to measure progress toward this goal[11]:

1. Conversational AI (current stage): At this level, AI interacts with users in natural language. Think of customer service chatbots, AI writing assistants such as ChatGPT, or AI coaches. Most businesses today leverage AI at this stage.

2. Reasoning AI (near future): This stage introduces "reasoners" – that is, AI capable of sequences of thought to achieve problem-solving at a level comparable to a PhD graduate, but without external tools.

3. Autonomous AI: Here, AI "agents" can operate independently for days, managing tasks without human intervention. Unlike today's automations, which require monitoring, future AI at this level will be self-correcting, ensuring reliability with minimal oversight. This may include autonomous learning, in addition to inference.

4. Innovating AI: Known as "innovators", these systems go beyond executing tasks – they improve them. Instead of just following rules, they critically analyse processes to enhance efficiency and effectiveness.

5. Organisational AI (super AI): At the final stage, AI functions as an entire organisation, managing all roles, optimising processes and collaborating autonomously – without human involvement.

He predicts we could reach level five within **10 years** (see *Figure 3.1*), while others estimate it may take up to **50 years**. The exact timeline remains uncertain, but the rapid pace of AI advancement is undeniable.

10 https://www.gartner.com/en/information-technology/glossary/artificial-general-intelligence-agi

11 https://www.forbes.com/sites/jodiecook/2024/07/16/openais-5-levels-of-super-ai-agi-to-outperform-human-capability

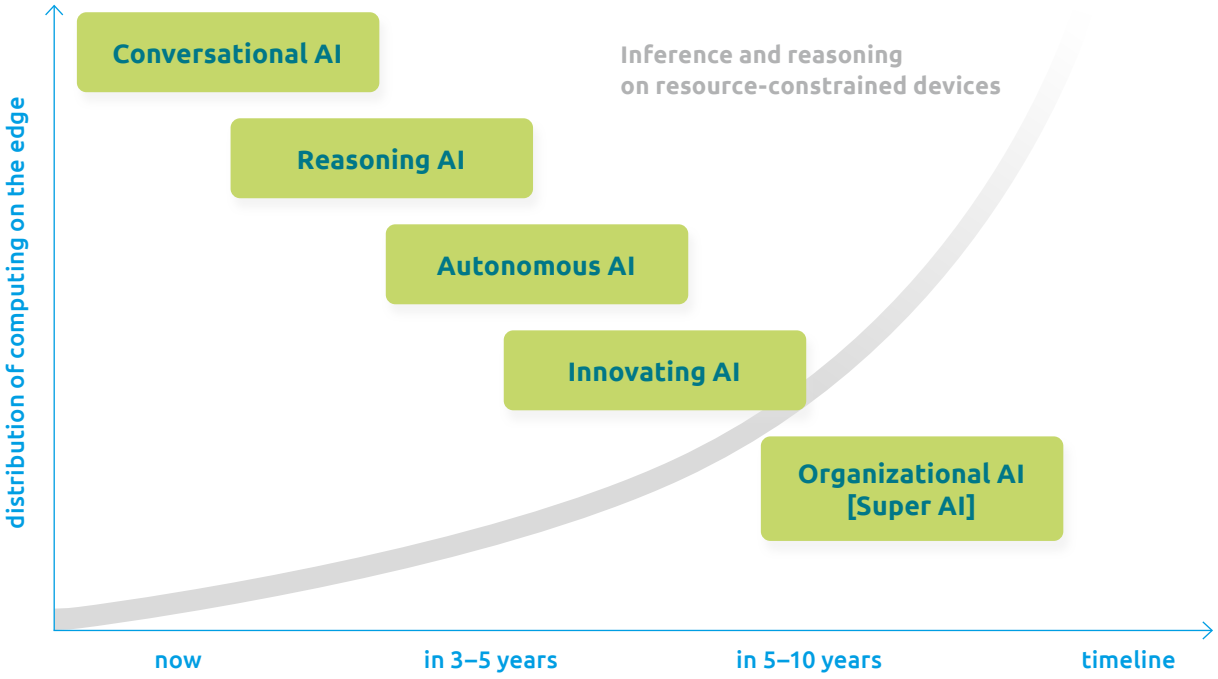


Figure 3.1: *Timeline with evolving AI trends with implications on Edge AI*

GenAI will inevitably have a significant impact on Edge AI that will bring real-time decision-making capabilities to resource-constrained devices such as IoT, sensors and smartphones. It will push advances in hardware optimisation and lightweight AI models to reshape edge computing paradigms. The rapid progress of GenAI presents both challenges and opportunities for the semiconductor research and innovation community, requiring a strategic reassessment of its R&I trajectory.

4 Overview of New Hardware Architectures

Deep Neural Network (DNN) algorithms achieve high-performance results for various applications – autonomous driving, smart health, smart home, smart agriculture, etc. However, these algorithms require high computational power for both training and inference. The field of high-performance DNN accelerators has been largely dominated by cloud platforms using NVIDIA GPUs and Google tensor processing units (TPUs), and the general trend has been to provide flexibility and performance to serve a wide range of DNN applications – without much concern for power consumption.

In contrast to monolithic accelerators such as the Google TPU, GPUs are modular by design and hence can scale from high-performance computing systems to edge devices. For example, NVIDIA's Ampere microarchitecture powers big A100 cores in data centres but also the Jetson Orin chips. A similar approach is taken by AMD, whose AI Engine Architecture is a scalable array of vector processors that accelerates AI inference workloads in laptop chips, 5G/6G communication infrastructure, as well as automotive edge devices. The advantage of edge and smart sensor AI solutions is the use of inference accelerators for tiny neural network models that offer low power, high throughput and low latency, opening up the possibility of moving processing closer to the sensor and sensor nodes.

4.1 SNN-based accelerators

Spiking Neural Networks (SNNs) represent an evolution in artificial neural networks (ANNs) incorporating principles inspired by the workings of biological brains. Unlike ANNs, which process data continuously, SNNs utilise discrete spikes as communication signals, introducing a time dimension to neuron activity. This makes SNNs uniquely capable of modelling the temporal dynamics of biological neurons, such as the timing of spikes and inter-neuronal dependencies. By leveraging event-driven computation, SNNs achieve remarkable energy efficiency, particularly when implemented on specialised neuromorphic hardware like Intel's Loihi or IBM's TrueNorth.

In neuromorphic hardware, their efficient computation paradigms make them ideal for low power environments such as edge devices. In robotics and sensory processing, their capacity for real-time, temporal pattern recognition allows for advanced control systems and adaptive behaviours. There are also applications in fields such as autonomous systems, speech recognition and time-series analysis, where SNNs can naturally encode and process sequential data. Despite their flexibility, SNNs adoption presents challenges such as the complexity of training methods, the need for specialised hardware, and difficulties in analysing their temporal activity patterns. Training SNNs is currently a complex task, often relying on approximations or hybrid approaches involving traditional neural networks.

SNNs require specialised hardware to fully realise their potential, as general-purpose GPUs or CPUs struggle with the sparse and temporal nature of spiking activity. SNN accelerators are designed to efficiently handle SNN highly parallel event-driven operations and temporal characteristics, with the advantage of energy efficiency and low-latency computation. Chips like Intel's Loihi and IBM's TrueNorth have set benchmarks in this field by integrating programmable synaptic plasticity, on-chip learning, and support for large-scale spiking networks. Intel's Loihi, for example, has pioneered the inclusion of biologically inspired learning rules such as Spike-Timing-Dependent Plasticity (STDP), enabling real-time adaptability. Similarly, IBM's TrueNorth chip offers ultra-low power operation with its million-neuron architecture, demonstrating the scalability of neuromorphic systems.

Recent advances in neuromorphic hardware have focused on enhancing scalability, enabling chips to support larger and more complex networks – for example, with the adoption of improved memory architectures and

3D-stacked designs to overcome data bandwidth limitations[12] [13]. This also reduces latency and enables real-time processing of high-dimensional data[14] [15].

Energy efficiency remains a primary objective as temporal sparsity and event-driven computation minimises unnecessary activity, significantly reducing power consumption. For Edge AI devices and advanced memory technologies such as memristors and resistive RAM (ReRAM), this represents a promising evolution. Moreover, the combination of SNN accelerators with energy-harvesting technologies could contribute to the diffusion of energy-autonomous systems, enabling devices to operate indefinitely in remote or resource-constrained environments; in IoT applications, this could represent a game changer.

Another trend is the integration of SNNs with traditional deep learning frameworks, creating hybrid architectures that combine the strengths of both paradigms: these systems can switch between continuous and event-driven computation, optimising workloads dynamically for a wide range of applications.

The next generation of SNN accelerators will require novel materials, enhancing computational density, such as phase-change memory and memristors, to replicate synaptic functions with greater efficiency. They will allow the simulation of more biologically accurate neural dynamics, while a futuristic evolution could involve the fusion of quantum computing with neuromorphic principles. Such quantum systems, with their inherent parallelism and superposition capabilities, offer a new dimension for processing spike-based computations. Hybrid quantum-SNN architectures could also accelerate learning and inference processes, tackling optimisation problems that are currently infeasible with classical systems[16].

From an architectural perspective, SNN accelerators will include cognitive-level processing, enabling chips to perform higher-order tasks such as reasoning, abstraction and multi-task learning. By incorporating hierarchical and modular architectures, these systems will approximate the layered complexity of biological brains, making them suitable for applications in AGI.

4.2 RISC-V based accelerators

RISC-V is very frequently adopted to develop Edge AI accelerators due to their flexibility and modularity, which enables the customisation of processors tailored to specific workloads and applications. Current RISC-V based accelerators are characterised by their ability to balance performance and power efficiency, crucial for Edge AI systems operating in resource-constrained environments – such as IoT devices, autonomous sensors and robotics. For example, the integration of domain-specific extensions within RISC-V cores, enabling accelerators to handle specialised tasks such as matrix multiplications, Convolutional Neural Network (CNN) inference, and vectorised computations, have been adopted to develop lightweight accelerators with a reduced energy consumption profile while maintaining high throughput in machine-learning tasks. A practical implementation is the Parallel Ultra-Low Power (PULP) platform, which builds on RISC-V cores to deliver ultra-low power AI solutions. The PULP project emphasises fine-grained parallelism and energy-efficient computation, leveraging custom extensions for machine-learning inference, to enable efficient data movement and computation, key factors for Edge AI tasks.

12 *Indiveri, G., & Liu, S. C. (2015). "Memory and information processing in neuromorphic systems." Proceedings of the IEEE, 103(8), 1379–1397.*

13 *Prezioso, M. et al. (2015). "Training and operation of an integrated neuromorphic network based on metal-oxide memristors." Nature, 521(7550), 61–64.*

14 *BrainChip. (2022). "Akida: Neuromorphic Processing at the Edge." [white paper].*

15 *Zidan, M. A. et al. (2018). "The future of electronics based on memristive systems." Nature Electronics, 1(1), 22–29.*

16 *Marković, D. et al. (2020). "Physics for neuromorphic computing." Nature Reviews Physics, 2(9), 499–510.*

A different approach adopts vectorised processing units in RISC-V accelerators to process multiple data elements simultaneously, significantly improving the performance of neural network operations. For example, the RISC-V Vector Extension (RVV) standard enables scalable vector processing, making it particularly effective for handling the parallel nature of deep learning algorithms.

Emerging trends emphasise the use of heterogeneous architectures, where RISC-V cores work synergistically with specialised AI processing units. This approach leverages the programmability of RISC-V for control tasks while delegating computation-heavy operations to AI-specific accelerators. Such architectures enable a more efficient division of computing load, reducing power consumption and latency in real-time applications.

The integration of approximate computing is another frontier for these accelerators, paving the way for accelerators that strike a balance between accuracy and efficiency. By exploiting the inherent tolerance of AI algorithms to computational noise, approximate computing techniques reduce precision levels in arithmetic operations, thereby enhancing energy efficiency.

Moreover, the combination of RISC-V with emerging memory technologies like ReRAM and 3D-stacked memory is anticipated to address the memory bottleneck in AI workloads. These technologies enable faster and more energy-efficient data access, which is critical for large-scale AI models at the edge. Future accelerators may integrate these memory systems with RISC-V cores to enhance the processing of data-intensive tasks.

Another promising direction involves the use of RISC-V in neuromorphic computing, where accelerators are designed to emulate biological neural networks. By leveraging RISC-V's modularity, developers can implement spiking neural network accelerators that combine biological plausibility with energy efficiency.

4.3 Photonic/optical-based accelerators

Photonics and optical technologies offer an alternative for high-speed and efficient AI tasks. These technologies exploit the unique properties of light, such as high bandwidth, low latency and minimal energy dissipation, to perform computations that would be prohibitively slow or power-intensive on conventional electronic hardware. As Edge AI applications demand compact, energy-efficient systems capable of processing massive data streams in real time, photonics-based accelerators are emerging as a promising solution.

At the forefront of this field are photonic neural networks, which leverage optical components such as waveguides, modulators and resonators to execute AI workloads, drastically reducing latency and power consumption. These solutions use optical interference to compute in parallel and efficiently perform matrix multiplications[17].

Silicon photonics, a mature and scalable technology, has enabled the integration of photonic accelerators into edge devices, combining the precision of photonics with the practicality of CMOS-compatible manufacturing, paving the way for cost-effective deployment. In this context, the use of optical memory, such as phase-change memory, allows the storage of data in light-sensitive materials, enabling ultra-fast read/write cycles. Similarly, optical interconnects eliminate bottlenecks associated with electronic data transfer, allowing accelerators to handle high-throughput tasks with minimal latency. These innovations are particularly beneficial for edge scenarios involving real-time data analytics and autonomous decision-making[18].

Photonics-based AI accelerators present several challenges, specifically in the integration of optical and electronic components, as hybrid systems often encounter inefficiencies at the interface. Additionally, scaling photonic architectures for more complex neural networks requires innovations in device miniaturisation and photonic circuit design. Co-packaged photonic processors, where optical and electronic components share a common substrate,

17 *Shen, Y., et al. (2017). "Deep learning with coherent nanophotonic circuits." Nature Photonics, 11(7), 441–446.*

18 *Feldmann, J., et al. (2021). "Parallel convolutional processing using an integrated photonic tensor core." Nature, 589(7840), 52–58.*

will eliminate inefficiencies at the interface, enabling seamless communication between light and electrons. This will be crucial for scaling photonic accelerators to support large, complex neural networks in edge devices.

Beyond these advances, the future of photonic accelerators could lie in the use of novel materials such as two-dimensional semiconductors and meta-surfaces to enhance the efficiency and scalability of photonic devices. These materials allow for more compact, denser (nanoscale precision) and energy-efficient photonic circuits, making them suitable for deployment in constrained edge environments.

Quantum photonics is another transformative direction, as it offers the potential to harness quantum phenomena such as superposition and entanglement for AI computations. Hybrid quantum-photonic systems could drastically enhance the parallelism and speed of AI accelerators, particularly for tasks such as optimisation and pattern recognition.

Merging neuromorphic and photonics could also be an alternative, with photonic implementations in SNNs, which could enhance tasks requiring temporal data processing, such as speech recognition and autonomous navigation.

4.4 Biological processors

Biological processors and organoids represent an emerging frontier in AI hardware, where biological systems are employed to perform computation. This paradigm diverges significantly from traditional silicon-based processors, leveraging the unique properties of biological materials, such as adaptability, energy efficiency and self-organisation. As Edge AI demands compact and efficient systems capable of real-time processing, biological processors and organoids present promising solutions by mimicking the unparalleled computational capabilities of biological brains.

Biological processors, particularly those based on synthetic biology and engineered genetic circuits, use living cells or biomolecules to process inputs and generate outputs. For example, bacterial cells can be programmed to function as logic gates, responding to chemical signals with specific outputs. These systems demonstrate the potential for massive parallelism, as billions of cells can work simultaneously to process complex datasets. Recent advances[19] highlight the development of molecular logic circuits capable of performing computations similar to traditional electronics, but with far lower energy requirements.

Organoids, three-dimensional cellular structures that mimic the architecture and functionality of the brain, represents another alternative for neuromorphic computation. Brain organoids, in particular, are cultivated from stem cells to replicate certain aspects of neural processing. Recent research has demonstrated the ability of brain organoids to exhibit spontaneous electrical activity, resembling primitive forms of neural computation. Organoids hold potential for Edge AI, as they can perform real-time processing in a biologically realistic manner, with minimal energy consumption.

While these technologies are still in their infancy, their unique features make them well-suited for Edge AI applications, especially as biological processors excel in energy efficiency and adaptability, qualities critical for remote or autonomous systems. Organoids, on the other hand, offer unparalleled parallelism and plasticity, enabling them to learn and adapt to new data, much like biological brains.

Despite these advantages, challenges remain. Biological systems are inherently less predictable than electronic circuits, and their integration with existing AI infrastructures poses significant hurdles. Additionally, scaling these technologies for practical applications requires breakthroughs in bioengineering and computational frameworks.

One promising direction is the development of hybrid bioelectronic systems, where biological components interface seamlessly with traditional electronics. Advances in bioelectronic interfaces are enabling real-time

19 *Qian, L., et al. (2011). "Neural network computation with DNA strand displacement cascades." Nature, 475(7356), 368–372.*

communication between living cells and silicon-based processors. This hybrid approach combines the adaptability of biological systems with the precision and scalability of electronics, creating versatile platforms for Edge AI.

Organoids are also being adopted in neuromorphic computing, as by cultivating larger and more complex brain organoids researchers aim to replicate higher-order cognitive functions such as decision-making and pattern recognition. Recently, organoids have been trained to control robotic systems, suggesting their potential for real-time autonomous operations at the edge.

Furthermore, synthetic biology is driving innovations in the programmability of biological processors. Techniques such as CRISPR-Cas9 gene editing are enabling the design of genetic circuits with greater complexity and specificity. With this technology, engineered bacterial systems have been able to process spatial and temporal data, opening new possibilities for applications in environmental monitoring and healthcare.

4.5 Chiplets

Chiplets are small integrated circuit (IC) die that are designed to work together within a single package to form a complete system. Instead of having one large, monolithic die, a system is split into multiple smaller die, or chiplets, each performing specific functions. These chiplets are interconnected using advanced packaging technologies to create a cohesive system-on-a-chip (SoC).

This technology promises enhanced performance, flexibility, scalability and power efficiency, as well as improved yield and cost-reduction due to the modularity it enables for SoCs. This modularity enables the reuse of chiplets and their optimisation for specific tasks. All these advantages make chiplets an interesting approach for many markets such as IoT devices or automotive applications.

Naturally, they are also applicable to Edge AI aspects of these areas. However, before chiplets can find widespread adoption, challenges such as standardisation, power distribution management and the linking of different chiplets need to be resolved. To tackle these issues, groups such the ASRA group in Japan and the IMEC automotive chiplet program in Europe were formed.

4.6 In-memory computing (memristive technologies)

In-memory computing integrates computation and data storage within the same physical components, significantly reducing the need to transfer data between separate processors and memory units. Memristive technologies – including spin-orbit torque MRAM (SOT-MRAM), phase-change RAM (PCRAM) and oxide-based resistive RAM (OxRAM) – enable memory cells to perform logic or analogue computations directly. By substantially reducing data movement, in-memory computing greatly enhances the speed and energy efficiency of AI inference. Traditional deep learning hardware often spends more time and energy moving data (weights and activations) between off-chip dynamic random-access memory (DRAM), on-chip static random-access memory (SRAM), and computational units than executing arithmetic operations[20].

Emerging technologies such as SOT-MRAM, PCRAM and OxRAM integrate memory and processing functions, significantly reducing data transfer latency. By minimising bottlenecks between the CPU and memory, these architectures boost inference speed – an essential advantage for real-time AI applications. Their low-latency performance makes them especially well-suited for Edge AI, where fast on-device processing is critical.

20 https://semiengineering.com/increasing-ai-energy-efficiency-with-compute-in-memory

4.7 ASICs, SoCs and microcontrollers

Application-specific integrated circuits (ASICs) and AI-centric SoCs are custom-engineered to deliver highly efficient deep learning inference. Unlike general-purpose CPUs or GPUs, which are designed for a wide range of tasks, these chips incorporate specialised circuits such as tensor engines and neural processing units, components that are finely tuned for the types of matrix operations and neural network computations that underpin modern AI models.

The result of this specialisation is a significant boost in both performance and energy efficiency. ASICs and SoCs can achieve extremely high throughput – often measured in trillions of operations per second (TOPS) – while maintaining a low power footprint. For instance, a neural processing unit embedded in a smartphone can perform several TOPS of inference while consuming only a few hundred milliwatts of power, a level of efficiency that conventional CPUs or GPUs cannot sustain. However, this high level of optimisation comes with a trade-off: these chips are typically limited in flexibility, and are best suited for specific tasks rather than general-purpose computing.

In many AI systems, particularly those operating at the edge, microcontrollers (MCUs) are integrated alongside ASICs or within SoCs to handle tasks that require low power and real-time responsiveness. While MCUs lack the processing muscle for intensive inference, they are essential for coordinating sensor input, triggering inference operations, and managing communication between different components of the system. In certain ultra-low power scenarios, such as TinyML applications, even simple neural networks can be deployed directly on microcontrollers, enabling basic AI functionality directly on the device without relying on cloud resources.

As AI continues to expand into embedded and autonomous systems, ASICs and SoCs are becoming increasingly vital. Their ability to deliver high-performance, low-latency inference makes them well-suited for demanding applications such as voice recognition, computer vision, autonomous vehicles and industrial automation.

4.8 FPGAs

Field programmable gate arrays (FPGAs) provide a unique and powerful platform for accelerating AI models by offering reconfigurable hardware fabrics that enable massive parallelism. At their core, FPGAs consist of an array of configurable logic blocks interconnected in a way that allows designers to create custom datapaths and computational units. This flexibility is particularly valuable for AI workloads, where operations such as multiply-accumulates, adders and control logic can be spatially mapped and optimised to match the structure of a given neural network.

Unlike ASICs, which are fixed-function chips tailored for specific tasks, FPGAs can be reprogrammed to support new or evolving model architectures. This reconfigurability makes them ideal for AI applications that require frequent updates or experimentation, such as in Edge AI deployments or during the prototyping phase of development. Engineers can fine-tune hardware characteristics – including dataflows, memory hierarchies and bit-widths – to match the demands of each model, thereby enhancing both performance and efficiency.

One of the key strengths of FPGAs lies in their ability to adapt to a wide range of AI models while maintaining moderate power consumption. Their architecture supports extremely low-precision computing, with some designs utilising quantisation down to just one or two bits. This not only accelerates computation but also drastically reduces power usage – an essential advantage for power-sensitive environments.

As the landscape of AI continues to evolve rapidly, FPGAs offer the agility and customisation required to stay aligned with the latest advances, making them a compelling choice for developers building cutting-edge adaptive AI solutions.

4.9 ECHO gateway for AI processing

A standardised, automated interface framework enabling seamless chip-to-cloud (such as machine-to-machine) communication is essential. Automated access from Edge CHip to clOud (ECHO) should enable fast AI processing on the cloud without any access to external world to offer trustworthiness, and ensure privacy and secured AI processing.

Bridging the gap between edge devices and cloud infrastructure at the hardware level minimises the fragmentation of operating systems and communication protocols, as highlighted by a CEUR-WS paper[21]. To maintain a secure data flow from edge to cloud, direct hardware-level access within cloud platforms – such as AWS, IONOS and Azure – must be enabled via secure application programming interfaces (APIs), independent of application-specific knowledge. For futuristic multi-core edge processors, message queuing telemetry transport (MQTT) and constrained application protocol (CoAP) are not efficient when there is a hardware-based API communication as then channels are scalable, and it enables a priority-based channel for uplink and downlink (also easy to port on 5/6G).

To mitigate potential security threats, the system must implement end-to-end encryption, strong authentication mechanisms, zero trust, a time stamp, and enforce consistent security policies across the entire datapath – from edge devices to cloud infrastructure. This architecture ensures that no intermediate software layer can access or tamper with the data during transfer, enabling secure AI training and inference in the cloud.

To address the different application needs, the cloud can offer improved scalability and hardware-level flexibility to accommodate a wide range of application requirements, facilitating seamless ECHO integration[22].

A key benefit of this hardware-centric gateway approach is reduced latency and faster AI model training. It also supports in-memory computing and facilitates the integration of deep neural networks directly within data pipelines, enabling AI processing closer to the source without overloading higher-tier AI accelerators[23].

Ultimately, this architecture reduces dependency on software configuration, minimises manual handovers, and simplifies secure cloud access – paving the way for highly efficient and secure AI-driven systems.

21 *Stanko, A. et al. (2024). "Artificial intelligence of things (AIoT): Integration challenges, and security issues" (https://ceur-ws.org/Vol-3842/paper6.pdf).*

22 *See PwC, (2024) "2024 cloud and AI business survey." (https://www.pwc.com/us/en/tech-effect/cloud/cloud-ai-business-survey.html).*

23 *Jhang et al. (2021) "Challenges and trends of SRAM-based computing-in-memory for AI edge devices." IEEE Transactions on Circuits and Systems. 68(5). 1773–1786 (https://ieeexplore.ieee.org/document/9382915).*

4.10 Conclusion

There is much evidence of a shifting paradigm toward Edge AI. Traditional DNNs dominate high-performance cloud-based applications but face scalability issues at the edge due to high power and computing demands. There is a growing need for energy-efficient, real-time AI solutions closer to data sources, which is fuelling innovation in edge-focused hardware.

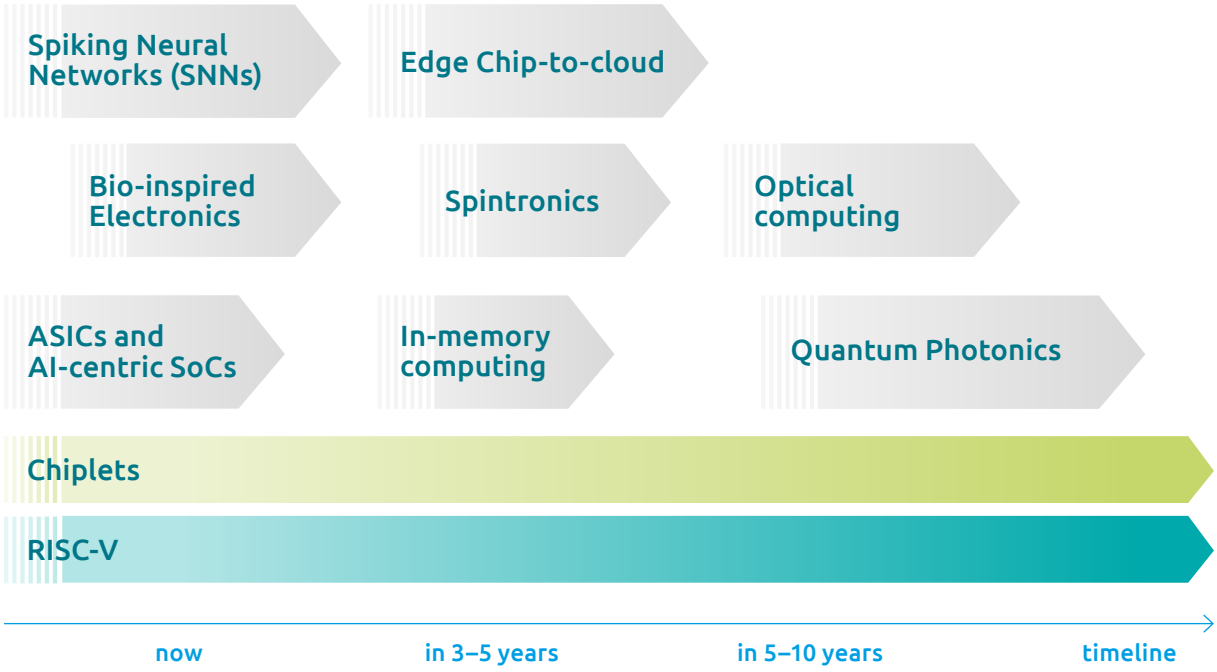


Figure 4.1: *Timeline for the emerging hardware architectures*

SNNs, inspired by biological neurons, offer ultra-low-power and real-time processing, particularly suitable for robotics, time-series data and sensory applications. Neuromorphic chips such as Intel's Loihi and IBM's TrueNorth showcase the potential of on-chip learning and energy efficiency. However, challenges remain in training complexity and hardware specialisation.

RISC-V's modularity makes it ideal for customising AI accelerators for edge devices. Platforms like PULP and vector extensions (RVV) enable efficient processing of ML workloads. The integration of heterogeneous computing and approximate computing further enhances power efficiency and performance in constrained environments.

Optical computing offers significant advantages in speed, parallelism and energy efficiency. Photonic neural networks and silicon photonics reduce latency and power usage, making them well-suited for high-throughput edge applications. Future advances will rely on hybrid photonic–electronic systems, new materials, and potentially quantum photonics for extreme acceleration.

Chiplets enable flexible, scalable, and cost-effective AI hardware by modularising specific functions within a chip package. Their reuse and task-specific optimisation make them ideal for Edge AI in domains including IoT and automotive. Widespread adoption depends on overcoming standardisation and integration challenges.

Although they are still experimental, biological computing systems (eg, brain organoids and synthetic bio-processors) show promise for ultra-energy-efficient, adaptive AI at the edge. Hybrid bioelectronic interfaces are also emerging, with the aim of combining biological adaptability with electronic control for the next-generation of intelligent systems[24].

In-memory computing technologies (eg, SOT-MRAM, PCRAM, OxRAM) drastically reduce data movement, enhancing speed and power efficiency. This is particularly important for edge devices that require fast, local AI inference. These architectures address memory bottlenecks and support real-time AI processing.

ASICs and AI-centric SoCs are highly specialised for deep learning inference, offering maximum performance and energy efficiency for specific tasks. In contrast, FPGAs provide a reconfigurable platform that trades some efficiency for flexibility, making them ideal for evolving or frequently changing AI models. The choice between them reflects a trade-off between performance optimisation and hardware adaptability.

The ECHO architecture provides a highly efficient and secure foundation for next-generation AI systems. It simplifies cloud access, minimises manual configuration, and delivers the flexibility and scalability needed to accommodate diverse application requirements – ultimately setting a new standard for secure, hardware-level chip-to-cloud integration.

24 *Boufidis, D. et al. (2025) "Bio-inspired electronics: Soft, biohybrid, and 'living' neural interfaces." Nature Communications. 16 (https://www.nature.com/articles/s41467-025-57016-0).*

5 Challenges, Constraints and Limitations Drive Innovation in Hardware Solutions for Edge AI

As Edge AI continues to evolve, it brings with it a unique set of challenges, constraints and limitations that demand a fresh wave of innovation in hardware design. This chapter explores the pressing technical, environmental, computational and specific AI model-related hurdles that require innovation in Edge AI hardware solutions.

5.1 Edge device constraints

Deploying AI algorithms on edge devices presents several constraints that must be carefully managed to ensure optimal performance.

- **Processing power and speed:** AI algorithms require substantial computational resources to execute within acceptable timeframes. Edge devices often have limited processing capabilities, making it challenging to run complex models efficiently. Specialised hardware accelerators, such as neural processing units (NPUs), can enhance performance by offloading AI-specific tasks from general-purpose CPUs.
- **Available memory:** Sufficient onboard memory is essential for temporarily storing and retrieving data during AI model execution. The size and speed of this memory directly impacts processing speed, energy consumption and overall efficiency. Techniques such as model quantisation and pruning can reduce memory requirements, enabling the deployment of AI models on devices with constrained resources. AI models must be stored on the device, and storage limitations can restrict the complexity and size of deployable models. Efficient model compression methods are crucial to fit models within the storage constraints of edge devices without significantly compromising performance.
- **Energy consumption:** Processing and data movement in AI tasks consume power, and larger models typically lead to higher energy consumption, reducing device autonomy. Energy-efficient model architectures and hardware accelerators can mitigate this issue by optimising power usage during inference.
- **Processing support**: Traditional processors (CPUs or microcontrollers) often complement AI accelerators in edge devices, handling tasks that are not well-suited for specialised hardware. However, this collaboration can further reduce device autonomy due to increased energy consumption. Balancing the workload between general-purpose and specialised processors is essential to maintain efficiency.
- **Connectivity:** Edge nodes are typically connected to external resources, typically to send sensory data or receive commands, and to interact with cloud resources. However, they suffer from unreliable connectivity, and can also be unable to deliver the data rates and latency required by the application. Introducing connectivity management and local AI capabilities (in particular with distributed or split AI approaches) significantly increases the robustness and performance of the deployed application.
- **Hardware deterioration**: Edge devices are exposed to a much wider range of sources of hardware deterioration (including different kinds of weather) than processing hardware in cloud servers. The deterioration of the underlying hardware leads to a reduction in the performance of AI models deployed on edge devices. Hence, it is essential that Edge AI models are robust and flexible, and that edge application systems include mechanisms for performance monitoring and updates to deal with the deterioration, which will increase the lifetime and sustainability of AI-based edge products.
- **Security and safety:** Edge devices are often much easier to access than a cloud server. This makes them vulnerable to wider range of attacks, especially physical ones. Hence, AI models that are used for safety-critical processes need to be deployed on certified edge hardware with security and safety components and mechanisms.

- **Device resource sharing:** The adoption of multiple AI models of the same device generally involves the concurrent use of its limited recourses, reducing their availability and negatively impacting on performances.

Addressing these constraints requires a holistic approach, combining hardware advancements with software optimisation techniques to enable effective AI deployment on edge devices.

5.2 Edge model and application constraints

Software–hardware co-design is essential for Edge AI, tightly integrating hardware capabilities with software demands to optimise efficiency, performance and power usage – critical aspects for edge applications. Edge devices typically face stringent power constraints; co-design ensures software algorithms leverage hardware strengths to significantly reduce energy consumption. By tailoring hardware acceleration specifically to AI models, co-design enables faster, responsive and real-time processing.

Aligning software requirements with hardware execution minimises data movement and latency, which is crucial for real-time performance. Additionally, this approach supports adaptable and future-proof hardware architectures that can evolve alongside emerging software techniques and increasing AI model complexity. Ultimately, software–hardware co-design effectively bridges algorithm innovation and hardware functionality, creating efficient, powerful and responsive Edge AI solutions.

Optimising AI models and applications for edge devices involves addressing several key constraints.

- **Model size:** Large models demand more computational power and memory, which can lead to slower operations on resource-limited edge devices. Techniques such as model pruning and quantisation can reduce model size, enhancing performance without significantly compromising accuracy.
- **Model accuracy and precision:** The level of precision used in data representation affects hardware resource requirements, and consequently the performance and accuracy of AI models. Balancing precision and resource utilisation is crucial for efficient edge deployment.
- **Model architecture:** The design and parameter interconnections within a neural network influence computational efficiency, memory usage and processing speed. Selecting architectures optimised for edge environments is essential for effective deployment.
- **Model training and inference:** In the context of Edge AI, it is important to differentiate between training and inference (deployment). Typically, AI models undergo resource-intensive training processes in cloud environments, where substantial computational resources are available. Once trained, these optimised models are deployed to edge devices, where inference occurs. This separation ensures computationally demanding training tasks do not burden resource-constrained edge hardware, while still enabling efficient, real-time, on-device AI. Training models directly on low-power devices is still a cutting-edge area, one that comes with a host of challenges – both technical and practical; however, with breakthroughs in software and data-centric strategies, federated learning and hardware, it is becoming more feasible.
- **Application speed requirements:** Edge devices may struggle to meet the speed demands of applications due to resource constraints, affecting their ability to ingest data and perform inference in a timely manner. Optimising both hardware and software is necessary to achieve the required performance levels.

- **Data volume versus resource availability:** Handling large datasets or high-resolution inputs can quickly exceed the available resources of edge devices, hindering application performance. Implementing data compression and efficient data-handling strategies can mitigate this issue. However, edge devices may have limited or intermittent access to labelled data (essential for supervised training), and this has led to new strategies such as federated learning, self-supervised learning, and active learning techniques.
- **Raw data preprocessing:** Preprocessing raw data before feeding it into AI models often requires substantial computing and memory resources. Efficient preprocessing pipelines are necessary to manage resource consumption effectively.
- **Robustness:** Unforeseen events and hardware deterioration can arise in all application contexts. At the same time, retraining and updates are more difficult at the edge due to the limited resources. Hence, Edge AI models need to be made robust to deal with these issues to a degree.

Addressing these constraints requires a comprehensive approach, one that combines model optimisation techniques with efficient data handling and hardware considerations to ensure effective AI deployment on edge devices.

5.3 Environmental, operating and financial constraints

Deploying edge devices involves navigating a range of environmental, operational and financial constraints.

- **Device form factor:** Edge devices must adhere to specific size and weight limitations, which can be challenging due to the need for components such as cooling systems, interfaces and batteries. Balancing these requirements is essential to meet form factor constraints.
- **Environmental considerations:** Edge devices often operate in harsh conditions, such as extreme temperatures, humidity, dust, or radiation. Ensuring high reliability in these environments may necessitate specialised hardware, which can be less performant and more costly.
- **Safety and security:** In safety-critical applications, hardware redundancy is typically necessary to enhance reliability, although it can increase costs and introduce additional design constraints. Additionally, securing data communication is essential when deploying Edge AI applications in public or remote environments to protect against potential vulnerabilities and ensure privacy.
- **Accessibility:** Accessing edge devices can be difficult, especially in remote or hard-to-reach locations, making maintenance and updates challenging and expensive.
- **Deployment and commissioning:** The process of deploying and commissioning edge devices is often complex and costly, particularly when dealing with large-scale or geographically dispersed installations.
- **Maintenance and evolution:** The ongoing operation, management, updates, maintenance, replacement and eventual decommissioning of edge devices represents a significant cost over the device lifecycle. Ensuring that AI model updates have been correctly implemented and are working as intended is critical. Techniques such as runtime behaviour analysis and provenance tracking can be used to verify model integrity.
- **Standards for protocols and interfaces:** Due to the diverse nature of edge devices – ranging from small IoT sensors to complex autonomous systems – establishing standards, protocols and interfaces becomes crucial. Standards and protocols ensure interoperability between various hardware and software components, facilitating seamless integration, scalability and communication across different platforms. Well-defined interfaces enable efficient data exchange, software reuse and simplified development, ultimately reducing complexity and cost.

Addressing these constraints requires careful planning and consideration of trade-offs to ensure that edge deployments are both effective and sustainable.

5.4 Safety, security and privacy technologies

Edge AI refers to deploying AI algorithms directly at the point of data processing and decision-making, such as an IoT device or an integrated module in a modern car (eg, a pedestrian detector for collision warnings). While research has focused on making AI training more robust, reliable and secure by reducing reliance on third-party cloud services, Edge AI introduces unique challenges.

For instance, in the case of a connected car fleet, model retraining may be necessary to enhance performance. Since on-device training is typically impractical, collected data must be transferred to a powerful server. Once retrained, updated models must then be deployed back to edge devices. This shift from traditional AI pipelines raises key safety and security concerns, including the following.

- **Functional safety:** How can we ensure that IoT devices operate correctly, addressing hardware issues (eg, bit flips, loose cables) and maintaining software integrity?
- **Physical intrusion:** How can we prevent tampering that could compromise device stability or expose it to external threats?
- **Security:** How can we protect on-device data – whether gathered, processed or stored – from unauthorised access?
- **Transmission integrity:** How can we guarantee the security and integrity of training data sent to servers and new models deployed back to devices?

Addressing these concerns is crucial for building secure, reliable, and efficient Edge AI systems that can operate independently while ensuring data privacy and system stability.

5.5 Technology challenges for computation

Advances in computing performance have historically relied on transistor miniaturisation and architectural improvements. However, as we approach the physical limits of transistor scaling, alternative strategies are essential to overcome emerging challenges such as the memory wall and energy inefficiency.

The continuous shrinking of transistors faces significant obstacles.

- **Thermodynamic constraints:** As transistors approach atomic scales, quantum effects like electron tunnelling become prominent, hindering further miniaturisation.
- **Manufacturing challenges:** Photolithography faces challenges at nanometre scales, making advanced chip production more complex. Another key issue is identifying the optimal combination of technologies for the diverse functions in an Edge AI component. In this context, chiplets offer a promising solution.

To address these limitations, several approaches are under exploration.

- **3D integration and heterogeneous architectures:** Stacking chips vertically and integrating diverse components can enhance performance and mitigate space constraints.
- **Specialised hardware:** Developing ASICs tailored for particular tasks can offer efficiency gains over general-purpose processors.
- **Alternative technologies:** Exploring new materials and devices, such as memristors and integrated photonics, holds promise for surpassing current transistor limitations.

5.6 Memory wall challenge

A significant portion of processing time is consumed by data transfer between memory and processors, leading to inefficiencies.

- **Data transfer bottlenecks:** In large-scale AI models, substantial time is spent moving data, which doesn't scale efficiently with increased processing power. Ensuring that AI models run efficiently across diverse hardware environment – from IoT devices to smartphones – adds complexity. Variations in hardware capabilities necessitate tailored optimisation strategies to maintain performance[25].

To overcome the memory wall, strategies such as implementing memory hierarchies are key. For this, the following approaches may be useful.

- **Compute-in-memory (CIM) architectures:** Integrating processing capabilities within memory units reduces data movement, enhancing speed and energy efficiency.
- **3D memory technologies:** Expanding memory bandwidth through vertical stacking can alleviate data transfer limitations.

5.7 Energy efficiency

Energy efficiency has become a critical concern in the computing industry due to the significant environmental and economic challenges posed by the escalating power consumption of data centres and high-performance computing systems. The growing energy demands of advanced computing systems pose sustainability challenges.

- **High power consumption:** Traditional architectures consume substantial energy, leading to increased operational costs and environmental impact.
- **Specialised low-power hardware:** Designing chips optimised for specific tasks can significantly reduce energy consumption.
- **Algorithmic optimisation:** Developing more efficient algorithms can decrease computational load and associated energy use.

25 https://www.wevolver.com/article/challenges-and-opportunities-in-edge-based-generative-ai

5.8 Modularity and interoperability of the technology stack

In the rapidly evolving computing landscape, hyperscalers – large-scale cloud service providers – recognise that mere hardware advancements are insufficient to meet escalating application demands. Their distinctive advantage lies in a holistic approach known as "verticalisation", emphasising comprehensive control over the entire technology stack. This strategy integrates hardware design, alternative materials and optimised algorithms to sustain progress in computing performance. By managing both hardware and software components, hyperscalers can tailor solutions that enhance efficiency, scalability and innovation, setting them apart in the competitive cloud services market[26].

This strategy is rooted in "system thinking", and involves the following.

- **Iterative co-design and co-optimisation:** By continuously refining and aligning system requirements down to the hardware level, and spanning all layers of the technology stack, hyperscalers ensure that each component is optimised in harmony with the others. This process, often referred to as system technology co-optimisation (STCO), enables architectural and technology trade-offs early in the system design process to achieve high-performance, cost-effective solutions in a reduced timeframe.
- **Multidisciplinary collaboration:** Leveraging expertise across diverse fields allows for innovative solutions that address complex challenges, ensuring that the final product meets client needs effectively. This holistic co-design approach tends to break the barrier across the vertical layers (devices, circuits, architecture and systems, algorithms, and applications), and therefore achieve global optimisation.

By embracing this vertically integrated methodology, hyperscalers can deliver cloud solutions that not only meet but often exceed client expectations, minimising the effort required to build on hardware and ensuring seamless, efficient performance.

5.9 Software and data challenges in on-device training

Training machine-learning models directly on edge devices introduces a range of complex challenges that go far beyond hardware limitations. From a software and data perspective, the core difficulties stem from adapting conventional training paradigms – originally designed for data centre-scale environments – to extremely resource-constrained, heterogeneous, and often dynamic, edge environments.

One of the most fundamental training paradigms is backpropagation, which requires the storage of intermediate activations across all layers of a network. On standard servers or GPUs, this is not a problem – but on edge devices it is a major constraint. Efficient gradient computation thus becomes a bottleneck. Developers must rely on strategies such as reduced precision gradients to squeeze training processes into these limited environments; however, these workarounds introduce trade-offs in terms of convergence speed and numerical stability.

Another critical factor is the batch size. Modern training workflows depend on mini-batch gradient descent to stabilise updates and efficiently utilise vectorised operations. On the edge, the available memory usually allows for processing only one or a few samples at a time. This severely increases the noise in gradient estimates, slows convergence, and makes it harder for the model to generalise. As a result, optimisers that adapt quickly to sparse or noisy gradients are more suitable, although they bring their own overhead that must be managed carefully on-device.

26 https://www.nextplatform.com/2020/02/03/vertical-integration-is-eating-the-datacenter-part-two

Another challenge is often a lack of labelled data. Edge devices typically collect vast streams of raw data – sensor readings, images, audio snippets – but without associated ground-truth labels. This makes traditional supervised learning infeasible in most real-world edge scenarios. Developers must lean on self-supervised or semi-supervised learning techniques, such as contrastive learning or pseudo-labelling, methods that reduce dependence on annotated data but require careful calibration to avoid reinforcing model bias or overfitting to incorrect signals.

Moreover, training on edge devices is almost always continual in nature. Rather than training once on a fixed dataset, the model is exposed to a constantly evolving data stream. This leads to the well-known problem of catastrophic forgetting, where learning new data causes the model to lose previously acquired knowledge. Resolving this requires implementing continual learning techniques, memory replay buffers, or regularisation-based strategies – all of which need to be implemented in lightweight and memory-efficient ways that are compatible with the device's constraints.

The challenge is compounded by data drift. The input distribution seen by an edge device often changes over time – i.e., users behave differently and/or hardware may degrade. Unlike in the cloud, there is no centralised retraining pipeline or data validation loop. Models must be able to adapt locally, ideally using online learning or meta-learning techniques that support fast adaptation. Nevertheless, without access to large-scale metrics or test sets, it is difficult to even know whether the model is still performing well.

Finally, there is the issue of infrastructure. The ML software stack at the edge is fragmented and immature when it comes to training. Most available tools are designed strictly for inference, not training. Often, teams must write their own training loops from scratch, manually handling forward and backward passes, memory allocation, and serialisation.

Altogether, these challenges make on-device training a highly specialised area of research and development. While inference on the edge has become increasingly practical, training still requires a nuanced blend of algorithmic adaptation, software engineering and clever approximation techniques. However, as interest in Edge AI grows the need to solve these training bottlenecks becomes more urgent (and more rewarding).

5.10 Engineering tools for designing Edge AI-driven products

When developing AI-driven products, it is important to consider the entire technology stack to ensure seamless integration, optimal performance and adaptability. This comprehensive approach encompasses several layers, from data ingestion and processing to model training, deployment and user interfaces. By addressing each component, engineers can harmonise the interactions between hardware and software, resulting in efficient resource utilisation and improved system performance. In addition, a holistic perspective enables the implementation of robust security measures at every level, protecting against vulnerabilities and ensuring data integrity. This strategy not only streamlines the development process, but also facilitates the creation of AI-driven products that are robust, efficient, secure, and able to meet the complex demands of today's applications.

- **Integrating AI into smart system products:** Developing AI-driven smart systems is an interdisciplinary challenge, requiring seamless collaboration between data scientists, system architects, verification engineers, and specialists in mechanics, electronics, semiconductors and software. Implementation decisions are shaped by key product requirements such as power consumption, size, thermal dissipation, and real-time performance, as well as economic factors like production cost and time-to-market.

5.10.1 CHALLENGES IN AI-DRIVEN SMART PRODUCT DEVELOPMENT

AI-based products offer a wide range of implementation technologies, making architecture decisions critical. Poor analysis can lead to excessive costs, power consumption, or hardware resource constraints. Traditional **domain-specific design methodologies** struggle to handle this multi-dimensional design space, often leading to miscommunication between teams using different terminologies, delays, or even product failure.

A **holistic, scalable methodology and tooling** is needed to manage development – from simple IoT devices to complex system-of-subsystems (eg, vehicles). Key here is **hierarchical design phases and tooling.** For this, AI-driven smart product development follows **five interconnected design phases:**

- requirements capture and management;
- AI algorithm development and training;
- architecture exploration;
- implementation architecture validation; and
- domain-specific implementation paths.

Each phase **propagates requirements and feedback** to ensure continuous refinement. We will now examine each of these in turn.

1. Requirements capture and management

This phase involves well-established requirements management tools that integrate with subsequent design workflows.

2. AI algorithm development and training

Neural network development relies on tools such as **TensorFlow, PyTorch, Keras, and Apache MXNet**, mostly open-source and Python-based. The tooling must support importing models from multiple AI frameworks.

3. Architecture exploration

At this stage, potential **implementation technologies** are evaluated. AI models are mapped onto processing elements and accelerators in **abstract performance simulations** to analyse key metrics:

- processing time (latency);
- interconnect utilisation;
- storage usage; and
- power consumption.

The goal is to **narrow down viable architectures** for detailed analysis. To accommodate diverse hardware platforms, architecture exploration must support **hierarchical virtual modelling**, targeting:

- off-the-shelf electronic control units (ECUs);
- custom ECUs with standard processors/SoCs;
- pre-built SoCs with internal accelerators;
- custom SoCs or 3D ICs; and
- hybrid solutions combining off-the-shelf and configurable components.

A **parametric simulation model** enables rapid architecture adjustments and design sweeps. If the analysis shows feasibility constraints, either the **algorithms or requirements** must be adjusted.

4. Validation of implementation architecture

With the solution space reduced, the next step is **functional and performance validation** using **virtual platform technology – a bit-accurate, timing-approximate simulation** that runs real software on modelled processors. This offers:

- more precise timing, power and interconnect/memory utilisation analysis than prior simulation models;
- confidence that the architecture meets constraints; and
- integration with a full digital twin for real-world validation.

5. Domain-specific implementation paths

Once the architecture is finalised, it is handed off to domain-specific development teams using specialised design tools:

- electronic design automation (EDA) tools for printed circuit board (PCB), IC, and three-dimensional IC (3D IC) design;
- vendor-specific tools for FPGA, NPU and custom SoC implementation; and
- conventional software development tools for firmware and application software.

6. Access to tools

To support small and mid-sized companies, development tools must be:

- affordable with low entry barriers;
- easily accessible, such as cloud-based solutions with pre-installed toolchains and secure remote access; and
- supported professionally, as open-source tools require expertise to handle the complexity of AI system design

To summarise, AI-driven smart product development demands an integrated, multi-phase approach with scalable methodologies and toolchains. By addressing implementation challenges early, companies can accelerate time-to-market, optimise performance, and control costs.

5.11 Conclusion: Challenges driving innovation in Edge AI hardware

Edge AI faces significant constraints in processing power, memory, energy and connectivity, demanding specialised, efficient hardware and optimised AI models. Software–hardware co-design is essential to align performance, power and latency requirements. Harsh operating environments, limited access, and the need for robust, secure systems further complicate deployment.

Energy efficiency is a critical driver, pushing innovation in low-power architectures, in-memory computing, and neuromorphic hardware. As traditional transistor scaling nears its limits, new solutions such as chiplets, 3D integration, and emerging technologies (eg, photonics, memristors, biological processors) are gaining traction. Standardisation, modularity, and advanced design tools are crucial to manage complexity, ensure interoperability and accelerate development. Finally, lifecycle sustainability – through efficient updates, monitoring and maintenance – is key to enabling scalable, long-term Edge AI deployment.

6 MultiSpin.AI: An Opportunity for Europe to Lead the Field of Edge AI Computation Hardware

The increasing demand for real-time, energy-efficient AI processing has driven the development of dedicated hardware architectures for Edge AI applications. Traditional digital computing hardware based on von Neumann architectures cannot keep up with these AI requirements, leading to the development of novel computing schemes. In this chapter, we explore the evolution of Edge AI hardware, focusing on spintronic-based analogue AI platforms such as MultiSpin.AI[27], which could play an important role in the development of novel Edge AI hardware in Europe.

6.1 Requirements on Edge AI hardware driving innovation in spintronics

The transition from cloud-based AI processing to Edge AI has been significantly accelerated by emerging industries such as autonomous vehicles, which necessitate real-time processing, minimal latency, and reduced energy consumption. This industry-specific shift has led to increased demand for specialised AI hardware solutions characterised by low power usage and high computational efficiency.

6.2 Spintronic AI platforms

Spintronic technologies exploit the intrinsic quantum mechanical property of electron spin for information storage and computation. These technologies have emerged as promising foundations for both general-purpose neuromorphic systems and specialised analogue in-memory coprocessors. Spintronic-based systems offer significant advantages, including ultra-low power consumption, improved scalability, and resilience to miniaturisation effects, making them ideally suited for compact, energy-sensitive Edge AI applications.

Collectively, these technological innovations represent a transformative step in AI hardware, providing solutions specifically tailored for edge computing environments where energy efficiency, speed and real-time responsiveness are paramount.

27 https://multispinai.eu

6.3 Comparison of Edge AI hardware platforms

HARDWARE TYPE	POWER EFFICIENCY (TOPS/W)	PROCESSING DENSITY (TOPS/MM²)	SUITABILITY FOR AI INFERENCE
GPGPUs (eg, NVIDIA Jetson)	10–20	0.2–0.5	High
Neuromorphic chips (Intel's Loihi)[28]	50–100	0.3–0.6	Moderate, application specific
In-memory coprocessor (PCM, ReRAM)[29]	75–150	0.6–1.2	High
Spintronics-based in-memory coprocessor (MultiSpin.AI)[30]	1,000+	2.0+	Very high

Table 6.1: *Comparing hardware types*

6.3.1 THE ROLE OF SPINTRONICS IN AI HARDWARE EVOLUTION

Spintronics significantly enhances traditional charge-based electronics by exploiting the spin property of electrons in addition to charge transport. While conventional electronics rely primarily on charge to generate voltages, currents and define resistance, spintronics leverages electron spin – a quantum mechanical property representing intrinsic angular momentum – to achieve more sophisticated functionalities.

This dual utilisation of electron charge and spin opens pathways to advanced technologies and new paradigms in computing and data storage. The additional functionalities offered by spin-based effects include the following.

- **Non-volatility:** Spin-based devices, such as MRAM, retain stored information even in the absence of power, eliminating the need for continuous energy supply. This inherent memory retention capability facilitates durable and persistent data storage, crucial for reducing boot-up time and enhancing reliability in electronic devices.
- **Energy efficiency:** Spintronic devices drastically reduce power consumption compared to traditional electronics. This efficiency arises from the minimal energy required to manipulate electron spin states compared to moving charges through resistive channels. Spintronics thus significantly lowers energy dissipation, potentially reducing power usage by orders of magnitude, contributing to longer battery life and more sustainable electronic systems.

28 *The Loihi 2 chip by Intel consists of six embedded microprocessor cores (Lakemont x86) and 128 fully asynchronous neuron cores connected by a network-on-chip (see https://open-neuromorphic.org/neuromorphic-computing/hardware/loihi-2-intel/). Intel claims Loihi is about 1,000 times more energy efficient than general-purpose computing systems used to train neural networks (see https://en.wikipedia.org/wiki/Cognitive_computer).*

29 *Information on specific performance metrics for in-memory coprocessors using phase-change memory (PCM) or ReRAM varies based on implementation. mPower consumption details for these technologies are implementation-specific (see https://www.spintronics-info.com/new-eu-funded-project-applies-spintronics-field-artificial-intelligence). The flexibility of PCM and ReRAM-based in-memory coprocessors depends on their design and application.*

30 *The MultiSpin.AI project aims to develop an AI coprocessor based on a crossbar of multi-level magnetic tunnel junctions (M2TJ) cells, enabling n-ary state cells (see https://researchportal.vub.be/en/projects/multispinai-n-ary-spintronics-based-edge-computing-co-processor-f). The MultiSpin.AI project is designed to enhance neuromorphic computing by integrating spintronic hardware and AI, aiming for significant advancements in AI development.*

- **Scalability**: The intrinsic nature of electron spin allows spintronic technologies to be integrated at very high densities, facilitating scalability to smaller dimensions without compromising device performance. This feature is critical for developing the ultra-high-density storage solutions and compact computing architectures necessary for next-generation electronics, including quantum computing and advanced ICs.
- **Low bit-to-bit variability:** Spin-based technologies exhibit inherently low variability between individual bits, ensuring consistently high-accuracy performance, especially crucial for AI workloads. Reduced variability enhances computational precision, reliability and reproducibility in critical applications, such as neural network inference, machine-learning accelerators, and precise computational tasks requiring stable and repeatable results.

In summary, spintronics not only complements but significantly advances traditional electronic approaches by enabling more efficient, robust, scalable and reliable computing systems, and is therefore poised to address future technological challenges.

The key spintronic technologies used in MultiSpin.AI are as follows.

- **SOT devices:** Spin-orbit torque devices utilise spin-orbit coupling to rapidly switch magnetic states. This enables high-speed, energy-efficient computation, ideal for advanced computing and AI applications, significantly reducing power consumption and improving device reliability.
- **Multi-level magnetic tunnel junctions (M^2TJ):** M^2TJ support multiple magnetic states per cell, enabling n-ary logic operations. This enhances computational efficiency, reduces energy usage, and increases accuracy in AI workloads, providing reliable and efficient processing capabilities.

6.4 MultiSpin.AI: A paradigm shift in Edge AI processing

MultiSpin.AI has advantages over conventional AI hardware. For instance, it introduces an ***n*-ary spintronic AI coprocessor**, overcoming the limitations of existing AI accelerators. Key benefits include:

- bypassing the von Neumann bottleneck with memory-integrated AI processing;
- reducing energy consumption by over 1,000 times compared to conventional digital AI chips; and
- enabling high-density, real-time AI inference for edge applications.

Performance comparison

TECHNOLOGY	ENERGY EFFICIENCY (TOPS/W)	LATENCY (NS)	SCALABILITY
CPUs[31]	1–5	1,000+	Low
GPGPUs[32]	10–20	500–1000	Moderate
Neuromorphic chips[33]	50–100	100-500	High
MultiSpin.AI coprocessor[34]	1,000+	<10	Very high

6.4.1 THE STRATEGIC IMPORTANCE OF MULTISPIN.AI FOR EUROPEAN AI HARDWARE

So why should MultiSpin.AI be monitored and developed in Europe?

- **European innovation leadership:** Europe currently lacks a major player in AI semiconductor technologies. Investing in MultiSpin.AI aligns with the European Chips JU initiative, boosting Europe's sovereignty in chip development and enhancing competitiveness in next-generation AI computing.
- **Alignment with sustainability goals:** AI workloads consume increasing amounts of energy. MultiSpin.AI's ultra-low power spintronic technology directly supports the European Green Deal, significantly cutting energy use and aiding sustainable digital transformation.
- **Strategic Edge AI applications:** MultiSpin.AI's technology benefits critical European sectors such as automotive, healthcare and industrial automation, enabling efficient, real-time and low-energy AI processing. This drives innovation, sustainability and competitiveness in key industries.

31 Traditional CPUs typically exhibit energy efficiencies ranging from one to five TOPS/W, depending on the specific architecture and workload. CPUs generally have latencies exceeding 1,000 nanoseconds, influenced by factors such as instruction processing and memory access times. CPUs face scalability challenges due to limitations in parallel processing capabilities and increasing power consumption with added cores.

32 GPGPUs offer energy efficiencies between 10 and 20 TOPS/W, leveraging parallel architectures for enhanced performance. GPGPUs typically exhibit latencies ranging from 500 to 1,000 nanoseconds, depending on the specific architecture and workload. GPGPUs provide moderate scalability, effectively handling parallel tasks, but encounter challenges with memory bandwidth and power consumption as the number of cores increases.

33 Neuromorphic chips, such as IBM's NorthPole, have achieved significant energy-efficiency improvements, outperforming traditional GPUs in certain tasks (see https://research.ibm.com/blog/northpole-llm-inference-results). Neuromorphic systems are engineered to process information in a highly parallel and energy-efficient manner, making them ideally suited for applications requiring low latency (see https://aditya-sunjava.medium.com/innovative-alternatives-to-gpu-computing-for-parallel-processing-8340f91e1a79). Neuromorphic architectures are designed for high scalability, enabling efficient parallel processing and adaptability to complex computational tasks.

34 The MultiSpin.AI project aims to develop a spintronics-based edge computing coprocessor, targeting up to 1,000 times higher energy efficiency compared to traditional architectures (see https://multispinai.eu/the-project/). This technology is designed for ultra-low latency responses, making it ideal for applications requiring instant processing in energy-constrained environments. The spintronics-based design of the MultiSpin.AI coprocessor offers very high scalability, facilitating efficient parallel processing and integration into various computing environments.

6.5 Sustaining the future of spintronic AI hardware

The sustainable future of spintronic AI hardware includes the following pillars.

- **Funding and policy support:** Securing dedicated funding and policy support is essential to position Europe as a leader in spintronic AI. Integrating MultiSpin.AI into Horizon Europe will provide necessary resources and enable strategic planning and implementation.
- **Industry collaborations:** Collaborations with semiconductor companies such as STMicroelectronics, Infineon Technologies, NXP Semiconductors, and research institutions like imec are key to commercialising spintronic technologies. Their design and fabrication expertise can expedite product development and market entry.
- **Academic research:** Expanding academic research in spintronic neuromorphic computing will enhance Europe's position in next-generation AI hardware. Supporting research on spintronic materials, devices and algorithms is crucial for innovation and intellectual property creation.
- **Addressing growing AI demand:** Rising AI demand necessitates energy-efficient, high-performance accelerators. MultiSpin.AI provides an opportunity to advance sustainable AI hardware innovation. Collaboration among policymakers, academia and industry is vital to develop and commercialise this technology effectively.

6.6 Conclusion

The growing demand for real-time, energy-efficient AI processing is outpacing the capabilities of traditional digital hardware, prompting a shift toward novel architectures tailored for Edge AI. Spintronic technologies – exemplified by platforms such as MultiSpin.AI – offer a promising alternative by enabling ultra-low power, scalable and high-performance AI inference. With advantages such as in-memory processing, non-volatility and quantum-level efficiency, spintronic systems address key Edge AI challenges, including latency, energy consumption and device miniaturisation.

MultiSpin.AI, in particular, represents a paradigm shift, delivering over 1,000x energy efficiency improvements compared to conventional processors. It also aligns with Europe's strategic goals in sustainability, digital sovereignty and industrial competitiveness. To fully realise this potential, continued investment, cross-sector collaboration and targeted research are essential. Spintronic AI hardware not only meets the technical demands of edge computing, but also positions Europe as a leader in next-generation AI innovation.

7 KDT and Chips JU Research and Innovation Timeline

Chips JU (formerly Key Digital Technologies) is an industry-led initiative aimed at boosting Europe's semiconductor ecosystem by tackling critical technological and strategic challenges. Its core focus, "Advanced Chip Design", targets next-generation architectures for AI, IoT and edge computing. For this chapter, we will intentionally narrow our scope to highlight some prominent trajectories in the current scientific Edge AI landscape, distilling key insights and outcomes. While other programmes, such as Horizon Europe and national initiatives, also drive progress in AI and Edge AI, they fall outside our data collection scope.

7.1 Data collection

Projects were sourced from the CORDIS database (European Commission) [35] and the Chips JU website[36]. After reviewing their goals and objectives, we categorised them into two groups:

- projects focused on innovative Edge AI hardware and use cases; and
- projects centred on ecosystem development, tools, and engineering platforms.

For each project, we recorded key dimensions, including name, objectives, use cases, and information sources, including official websites. We used ChatGPT-4o (premium version with web search) to extract project goals and example use cases, verifying all results for accuracy.

Due to the confidential nature of many deliverables, and their strategic value to industrial partners, this analysis relies on publicly available data. For the first category, we examined example use cases featured on public project pages. For the second, we gathered insights on hardware strategies and platforms. This allowed us to identify the expected outcomes and contributions of each project based on accessible information.

The KDT JU launched AI4DI (Artificial Intelligence for Digitizing Industry) in Europe in 2019, followed by ANDANTE (AI for New Devices and Technologies at the Edge) in 2020. Both projects have now concluded and delivered tangible results. Meanwhile, later projects are still ongoing, and their full impact will become evident in the coming years.

35 https://cordis.europa.eu

36 https://www.chips-ju.europa.eu

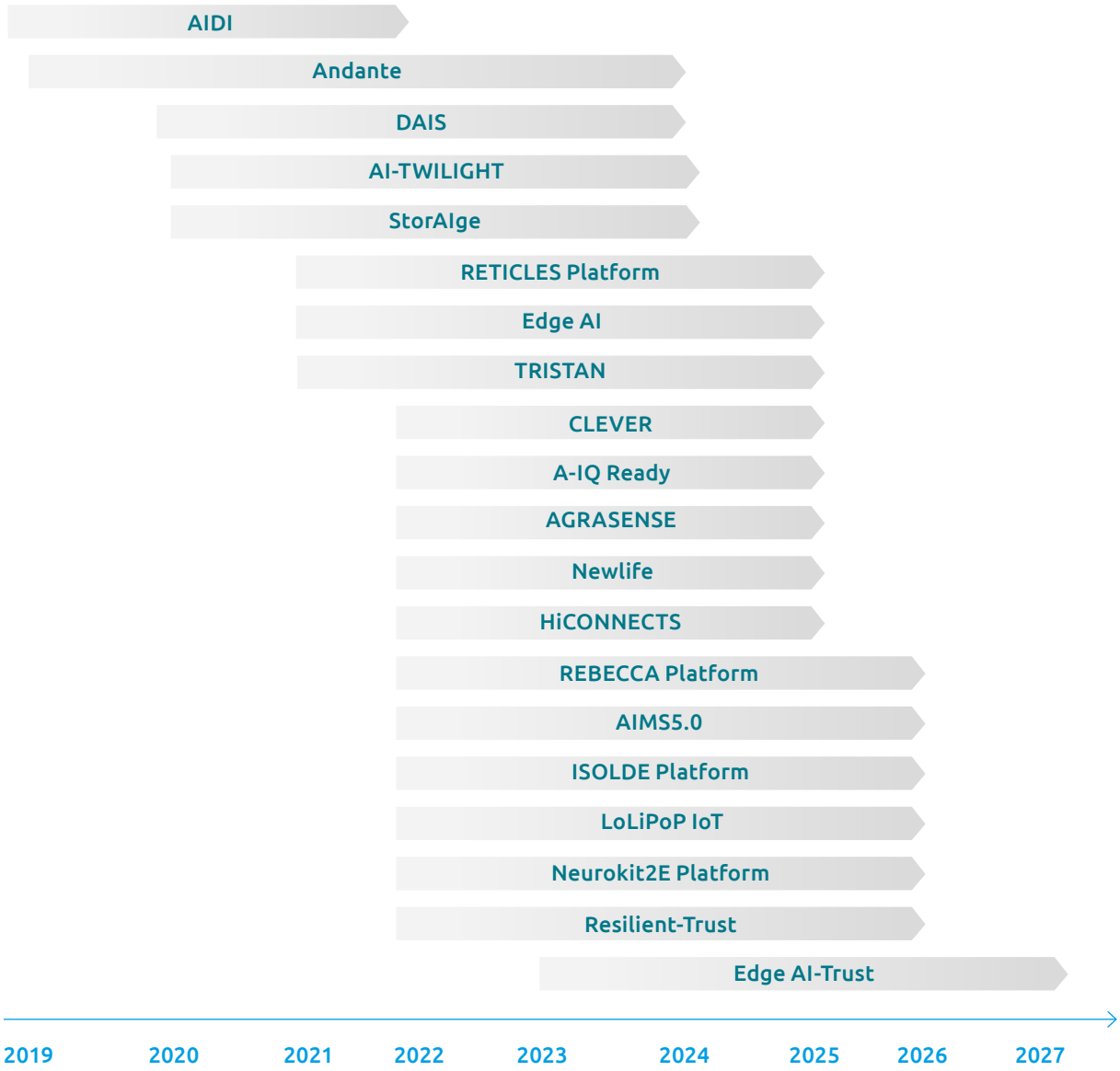


Figure 7.1: *Timeline of the KDT and Chips JU projects*

Table 7.1 provides an overview of the project goals, objectives and example use cases.

Table 7.1: Project goals, objectives and example use cases

PROJECT	GOALS AND OBJECTIVES	EXAMPLE USE CASES	WEBSITE
AI4DI	▪ Advance Moore's Law by developing innovative edge processing technologies. ▪ Bridge AI from centralised cloud solutions to distributed edge solutions to increase efficiency.	**Automotive:** AI-based logistics solutions for optimising assembly processes, development of autonomous reconfigurable battery systems, virtual AI platforms for training, deployment of autonomous mobile robotic agents for operational efficiency, and predictive maintenance using digital twins. **Semiconductor**: Enhancing wafer inspection using AI-based vision systems, automating semiconductor process inspection, and improving MEMS sensor predictions through neural networks.	https://ai4di.eu
ANDANTE	▪ Develop innovative hardware/software platforms leveraging neuromorphic and SNN architectures for IoT and edge devices.	**Digital industry:** Indoor positioning systems for real-time monitoring, quality control using AI-based edge computing. **Digital farming:** AI-driven systems for pest and disease prediction in crops, autonomous weeding robots for sustainable farming. **Transport and smart mobility:** Autonomous drone systems, acoustic signal classification for underwater applications, robust autonomous vehicle landing, and multi-modal path planning. Healthcare: AI-driven medical imaging analysis and glucose monitoring systems.	https://www.andante-ai.eu
DAIS	▪ Create distributed AI systems that provide faster, secure and energy-efficient data processing. ▪ Ensure connectivity and interoperability in distributed Edge AI systems.	**Digital industry:** Deployment of distributed AI to enable automation and efficiency in manufacturing. **Digital life:** Integration of Edge AI in smart home environments for enhanced user experience and real-time responsiveness. **Transport and smart mobility:** Leveraging Edge AI for improving autonomous vehicle perception and decision-making, ensuring secure and reliable communication between edge devices.	https://dais-project.eu
StorAIge	▪ Develop advanced embedded phase change memory (ePCM) and FDSOI 28nm technologies for high-performance edge applications.	**Automotive:** Enhanced automotive systems leveraging next-gen semiconductor memory for faster data processing. **Industrial applications:** High-reliability edge computing for industrial machinery. **Secure data processing:** Edge technologies designed to improve security and reduce latency in data transmission and storage.	https://storaige.eu
EdgeAI	▪ Build secure end-to-end hardware/software solutions for AI-driven edge platforms. ▪ Advance hybrid architecture designs for scalable and efficient AI systems.	**Digital industry:** Integration of advanced sensing, automated defect classification, and AI-enabled decision-making in production environments. **Energy sector:** Distributed AI for optimising energy usage in smart grids and industrial operations. **Agriculture and food:** Use of AI for predictive analytics, quality control, and precision farming. **Mobility:** Enhancing autonomous vehicle technologies with Edge AI. **Digital society:** AI-driven systems for activity and intention detection in real-world environments.	https://edge-ai-tech.eu

PROJECT	GOALS AND OBJECTIVES	EXAMPLE USE CASES	WEBSITE
CLEVER	▪ Develop an edge-cloud continuum for embedded AI solutions targeting futuristic industries and urban transformations.	**Fashion:** Deployment of virtual fitting rooms to improve online shopping experiences. **Smart environments:** Architectures to detect and adapt to concept drift in dynamic environments. **Smart cities:** Application of data-driven transformations for urban development, infrastructure optimisation, and citizen services enhancement.	https://www.cleverproject.eu
A-IQ Ready	▪ Innovate IoT systems by integrating quantum sensors and neuromorphic computing. ▪ Build edge-to-cloud solutions supporting the digital backbone for Society 5.0.	**Quantum technologies:** Integration of multi-physics (quantum) sensors to improve accuracy in complex environmental sensing applications. **IoT systems**: Development of edge-enabled, AI-integrated devices for a wide range of applications, including smart home systems, healthcare monitoring and industrial process automation.	https://www.aiqready.eu
AGRAR-SENSE	▪ Develop innovative microelectronics, photonics and packaging solutions tailored for agricultural and forestry applications. ▪ Advance ICT and data management systems to enable large-scale field demonstrations that address real-world industrial needs. ▪ Improve global food security and sustainability by deploying cutting-edge tools that increase agricultural efficiency and productivity.	▪ Implementation of automated tools for precision agriculture, such as robotic systems for planting and harvesting. ▪ Deployment of advanced sensor networks to monitor crop health, soil moisture, and environmental conditions, enabling data-driven decision-making for farmers.	https://www.agrarsense.eu
Newlife	▪ Design and develop comprehensive health-monitoring solutions that cover the entire pregnancy and neonatal period, ensuring the health and well-being of mothers and their babies. ▪ Employ non-invasive and early-detection methods to identify potential health risks, such as gestational diabetes or pre-eclampsia, before they become severe. ▪ Lower the incidence of pre-term births and related complications, leading to reduced healthcare costs and improved quality of life for families.	▪ Continuous monitoring of maternal vital signs, such as blood pressure and oxygen levels, using wearable devices and smart sensors. ▪ Development of non-invasive imaging and diagnostic tools to monitor foetal development and detect anomalies, ensuring timely medical interventions when necessary.	https://www.newlife-kdt.eu
AIMS5.0	▪ Strengthen Europe's technological and digital sovereignty by integrating advanced AI into sustainable production processes. ▪ Facilitate the transition from Industry 4.0, which focuses on automation and data exchange, to Industry 5.0, which emphasises human-centric, environmentally friendly and sustainable workplaces.	▪ Integration of AI algorithms to optimise energy consumption and material use in factories, reducing costs and environmental impact. ▪ Development of smart AI systems that assist workers with repetitive tasks, enhancing safety and ergonomics while maintaining high productivity levels.	https://aims50.eu

PROJECT	GOALS AND OBJECTIVES	EXAMPLE USE CASES	WEBSITE
	▪ Enhance the eco-efficiency of manufacturing by optimising resource usage and minimising waste through AI-driven tools.		
EdgeAI-Trust	▪ Create a secure and trustworthy ecosystem for Edge AI, focusing on the design of architectures, components and development tools that support edge devices. ▪ Enable real-time collaboration among heterogeneous edge devices, ensuring they operate securely and sustainably in decentralised networks. ▪ Advance AI applications for safety-critical systems, such as healthcare, autonomous transportation and cybersecurity, prioritising reliability and resilience.	▪ Deployment of federated learning models across distributed edge devices in healthcare, ensuring data privacy and compliance with regulations like GDPR. ▪ Implementation of real-time decision-making capabilities in autonomous vehicles using Edge AI to enhance safety, responsiveness and operational reliability.	https://www.edgeai-trust.eu
Resilient Trust	▪ Secure IoT 5.0 for small and medium-sized enterprises (SMEs): Develop an end-to-end security framework tailored for SMEs, addressing vulnerabilities due to lack of specialised security resources. ▪ Trust and resilience via hardware: Create specialised hardware components (IPs) that build system-level trust and protect against quantum-resistant and AI-based attacks. ▪ Threat modelling and architecture: Perform threat analysis, identify assets and risks, and define security requirements to shape the secure system design. ▪ Sustainable development and digital sovereignty: Strengthen Europe's independence in chip and IoT security, fostering societal and economic value.	**Multi-standard IoT communication:** Integrate a flexible transceiver into STM32 platforms supporting WLAN, UWB, DECT NR+, BLE, ZigBee – all in one chip. **Drone detection and jamming:** Use intelligent systems to detect and selectively jam drone signals for security-sensitive environments. **Secure supply chain (implied):** Ensure traceability of chip lifecycle via blockchain and physically unclonable functions (PUFs) to prevent IP theft and counterfeiting. **Ambient intelligence in offices (implied):** Trustworthy IoT integration for smart office environments, enhancing data privacy and system resilience.	https://tima.univ-grenoble-alpes.fr/research/amfors/research-projects/resilient-trust
AI Twilight	▪ **Develop trustworthy, low-power AI solutions:** AI Twilight focuses on designing and integrating AI at the edge in a manner that balances performance with energy efficiency. ▪ **Foster secure data handling:** Ensures end-to-end data privacy and integrity while enabling real-time analytics. ▪ **Strengthen European digital sovereignty:** Contributes to Europe's competitiveness by creating AI ecosystems that reduce dependency on external technologies.	**Industrial quality control:** AI-based inspection systems at the edge that rapidly detect defects on production lines without large-scale cloud dependencies. **Resource-efficient smart sensors:** Low-power sensors for monitoring critical infrastructure (eg, water treatment, public utilities) with on-device intelligence. **Smart healthcare devices:** AI-enabled patient-monitoring solutions that process vital signals locally, improving responsiveness and data security.	https://ai-twilight.eu

PROJECT	GOALS AND OBJECTIVES	EXAMPLE USE CASES	WEBSITE
hiCONNECTS	▪ Advance high-speed connectivity: Focuses on designing the next generation of secure, high-throughput and low-latency interconnect technologies. ▪ Optimise edge-to-cloud architectures: Bridges edge devices and data centres to enable seamless, scalable data processing. ▪ Promote interoperability and standards: Facilitates cooperation across diverse platforms to ensure broad adoption of high-performance connectivity solutions in Europe.	Automotive data networks: High-bandwidth interconnects for vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication, supporting autonomous driving and advanced driver-assistance systems. Smart city infrastructure: Robust data transmission between sensors, traffic lights and municipal services, improving urban mobility and resource management. Industry 4.0 connectivity: Reliable, real-time communications for factory automation and robotics, reducing latency and boosting productivity in manufacturing environments.	https://www.hiconnects.org

The projects surveyed collectively demonstrate a diverse range of Edge AI solutions that have the potential to transform industries such as automotive, manufacturing, healthcare and agriculture. Innovations in hardware (eg, embedded memory, neuromorphic chips), software (eg, federated learning, real-time analytics), and architectural design (eg, edge–cloud continuum) are unlocking new levels of performance, security and sustainability. Furthermore, these advancements are fostering cross-sector collaborations, enabling technology transfer and shared value creation. As Edge AI matures, it holds the promise of more human-centric, resilient and eco-friendly applications, setting the stage for widespread digital transformation that spans the entire economic and social landscape.

7.2 Design hardware platforms, engineering tools and ecosystems

KDT and Chips JU fund a series of projects focused on tools and platforms for hardware design, integration and engineering. *Table 7.2* introduces the innovative hardware approaches pursued in these projects.

Table 7.2: Hardware design, integration and engineering

PROJECT	GOALS AND OBJECTIVES	EXAMPLE APPROACHES FOR HARDWARE ENGINEERING TOOLS AND PLATFORMS	WEBSITE
RETICLES	▪ Develop specialised, high-performance reconfigurable hardware platforms and design flows. ▪ Provide tools that simplify system partitioning and integration for edge and cloud applications. ▪ Foster European sovereignty in next-generation computing and AI by promoting open standards and collaboration.	▪ Creation of reconfigurable accelerator IP blocks adaptable to multiple domains (eg, AI inference, security). ▪ Development of toolchains that automate partitioning across FPGA, ASIC and processor architectures. ▪ Utilisation of open-source hardware frameworks for faster prototyping and validation.	https://reticles.eu
Rebecca-Chip	▪ Drive innovations in chip architecture for next-gen machine learning and data analytics. ▪ Ensure energy-efficient design to meet edge constraints without sacrificing performance. ▪ Strengthen the European semiconductor ecosystem through joint research and pilot deployments.	▪ Exploration of heterogeneous SoC designs, combining CPU, GPU and specialised accelerators on a single chip. ▪ Use of advanced packaging and interconnect solutions to optimise bandwidth and reduce power consumption. ▪ Development of EDA tool suites that integrate AI-specific design libraries.	https://www.rebecca-chip.eu
TRISTAN	▪ Provide a trusted hardware platform for AI and high-performance computing (HPC) workloads in critical domains. ▪ Enhance security-by-design methodologies, including hardware-level encryption and attestation. ▪ Accelerate industrial uptake of secure and performance-optimised chipsets for high-assurance applications.	▪ Implementation of secure enclaves and cryptographic modules embedded at the silicon level. ▪ Development of verification workflows integrating formal methods to validate hardware security properties. ▪ Integration with hardware-based root of trust for mission-critical systems (eg, aerospace).	https://tristan-project.eu
ISOLDE	▪ Innovate in the design and verification of complex SoCs for AI, with a focus on low power consumption. ▪ Offer modular frameworks that shorten time-to-market for embedded and edge computing solutions. ▪ Promote standardisation and interoperability of EDA tools across industry partners.	▪ Development of multi-level simulation and debugging environments tailored to AI/ML hardware. ▪ Provision of IP blocks optimised for battery-powered devices, reducing leakage and dynamic power. ▪ Creation of cross-tool integration plugins for streamlined chip design and verification processes.	https://www.isolde-project.eu
LOLIPOP	▪ Advance low power IoT platforms through hardware-software co-design. ▪ Facilitate edge intelligence by incorporating lightweight AI accelerators on sensor devices. ▪ Strengthen the IoT ecosystem in Europe, targeting ultra-low power, long-lifetime embedded solutions.	▪ Use of custom ASIC accelerators for microcontrollers handling local AI tasks (eg, anomaly detection). ▪ Energy harvesting techniques combined with ultra-low-power silicon design for IoT nodes. ▪ Hardware toolkits enabling quick prototyping of smart sensor solutions.	https://www.lolipop-iot.eu
NeuroKit2e	▪ Research and implement neuromorphic hardware paradigms for high-efficiency computation. ▪ Enable event-driven processing and spike-based neural networks in real-world edge scenarios. ▪ Pioneer hardware-software toolchains that leverage bio-inspired architectures for AI at the sensor level.	▪ Integration of SNN cores with analogue/digital hybrid designs for real-time, low-power AI. ▪ Development of simulation and compiler frameworks to map conventional ML models onto neuromorphic hardware. ▪ Exploration of CMOS and emerging device approaches for spike-based computation.	https://www.neuro-kit2e.eu

These initiatives emphasise hardware-centric innovation, with each project aiming to push boundaries in semiconductor design, low power architectures, and security-by-design. They share a drive to refine or develop new toolchains and platforms that streamline the creation of advanced hardware solutions – whether for FPGAs, ASICs, neuromorphic chips, or secure SoCs.

Across all projects, energy efficiency and AI acceleration in resource-constrained environments remain central goals, often achieved by integrating trust anchors, encryption, and neuromorphic or event-driven paradigms at the silicon level. Lastly, the collective focus on collaborative development and European sovereignty highlights a broader ambition to bolster the continent's standing in semiconductor technology and AI innovation.[37]

37 *Concrete benchmarks cannot be disclosed due to the sensitivity and confidentiality of the project deliverables, in order to maintain the participating companies' competitive advantage.*

8 Market Dynamics

The market dynamics discussed in this chapter can be observed from the key players driving innovation in Edge AI and creating new applications in recently growing markets. Market boundaries are becoming blurred as most European semiconductor companies are now global. Infineon, for example, has the largest number of its employees in China/Asia.

Strategically important markets are shifting to emerging industrial countries such as India and Mexico, which are members of the BRICS[38] alliance. Europe also has a growing need to secure supply chains and gain technological autonomy in the face of current geopolitical tensions.

Global IT and AI players such as Google, AWS and Tesla have long recognised this global trend, and are building flexible cross-domain architectures that allow assets to be moved flexibly across domains and countries.

Table 8.1 presents an overview of the leading semiconductor companies, ranked by market capitalisation, as of February 14, 2025. Of course, market capitalisations are subject to change due to market fluctuations, and this data is based on the latest available information as of the specified date.

RANK	COMPANY	MARKET CAPITALISATION (USD)
1	**NVIDIA**	3.313 trillion
2	**Broadcom**	1.105 trillion
3	**TSMC**	1.046 trillion
4	**ASML**	305.51 billion
5	**Qualcomm**	190.39 billion
6	**Advanced Micro Devices (AMD)**	181.18 billion
7	**Texas Instruments**	164.92 billion
8	**Applied Materials**	149.75 billion
9	**Intel**	104.48 billion
10	**Lam Research**	106.92 billion
11	**Micron Technology**	106.58 billion
12	**KLA Corporation**	101.56 billion
13	**Marvell Technology**	89.55 billion
14	**Tokyo Electron**	73.93 billion
15	**NXP Semiconductors**	55.81 billion
16	**Infineon Technologies**	51.10 billion
17	**Analog Devices**	103.86 billion
18	**SK Hynix**	99.96 billion
19	**STMicroelectronics**	21.41 billion
20	**ON Semiconductor**	21.45 billion

***Table 8.1:** The leading semiconductor companies, ranked by market capitalisation*
(Source: https://disfold.com/industry/semiconductors/companies/#google_vignette)

38 *BRICS stands for Brazil, Russia, India, China and South Africa. Egypt, Ethiopia, Iran and the United Arab Emirates have also recently joined the alliance (see https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2024)760368).*

The biggest European chip vendor companies, such as NXP, Infineon and STMicroelectronics, are in the top 20; NVIDIA, by a large margin, is ahead of the pack. According to the recent report of market.us[39], the global Edge AI chips market is projected to grow from USD2.4 billion in 2023 to USD25.2 billion by 2033, reflecting a compound annual growth rate (CAGR) of 26.5% during the forecast period.

The growth in the Edge AI chips market is driven by several factors.

- **Reduced latency:** Processing data on-device minimises the delay associated with transmitting data to centralised cloud servers, leading to faster decision-making.
- **Enhanced privacy:** On-device processing ensures that sensitive data remains local, reducing the risk of data breaches and enhancing user privacy.
- **Improved efficiency**: By handling AI tasks locally, devices can operate more efficiently, conserving bandwidth and reducing reliance on constant internet connectivity.

These advantages are contributing to the rapid adoption of Edge AI solutions across various industries, including **consumer electronics, automotive, healthcare** and **manufacturing**. It is important to note that market projections can vary based on different research methodologies and data sources. For instance, some reports suggest that the global AI chips market, which includes both edge and cloud AI chips, could reach up to USD520.91 billion by 2033, growing at a CAGR of 37.77%[40].

In summary, the Edge AI chips market is poised for substantial growth, driven by the increasing demand for real-time processing, enhanced privacy and improved efficiency in AI applications across various sectors. As leading market player, NVIDIA offers a comprehensive suite of platforms and solutions tailored for Edge AI applications across various industries.

39 https://market.us/report/edge-ai-ics-market

40 https://www.cervicornconsulting.com/artificial-intelligence-chips-market

Table 8.2: NVIDIA Edge AI technologies (detailed)

TECHNOLOGY	DESCRIPTION	KEY FEATURES	WEBSITE
Jetson AGX Orin	High-performance AI module for advanced robotics, autonomous machines, and edge computing.	▪ 12-core Arm Cortex-A78AE CPU. ▪ 2048-core Ampere GPU with 64 Tensor Cores. ▪ Up to **275 TOPS** AI performance. ▪ Dual NVDLA deep learning accelerators. ▪ Supports **multi-camera** vision AI. ▪ Configurable power: **15 W – 60 W.**	https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-orin/
Jetson Orin NX	Mid-tier Edge AI module for robotics, smart cameras, and embedded vision.	▪ 8-core Arm Cortex-A78AE CPU. ▪ 1024-core Ampere GPU with 32 Tensor Cores. ▪ Up to **160 TOPS** AI performance. ▪ Single NVDLA accelerator. ▪ **10 W – 40 W** configurable power.	https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-orin/
Jetson Orin Nano	Entry-level AI module for small devices such as drones, IoT sensors and edge analytics.	▪ 6-core Arm Cortex-A78AE CPU. ▪ 512-core Ampere GPU with 16 Tensor Cores. ▪ Up to **67 TOP**S AI performance. ▪ Power-efficient: **7 W – 25 W.**	https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-orin/
NVIDIA A2 Tensor Core GPU	Low-profile Edge AI GPU for inference acceleration in small edge servers and network appliances.	▪ 1280 CUDA Cores, 40 Tensor Cores. ▪ 16 GB GDDR6 memory. ▪ **36 INT8 TOPS**, optimised for AI inference. ▪ PCIe Gen4, half-length, single-slot. ▪ **40 W – 60 W** power range.	https://www.nvidia.com/en-gb/data-center/products/a2/
NVIDIA L4 Tensor Core GPU	High-efficiency AI and video processing GPU for edge data centres and AI workloads.	▪ 7424 CUDA Cores, 24 GB GDDR6 memory. ▪ Up to **485 TOPS INT**8 inferencing. ▪ Dedicated AV1 hardware encoding/decoding. ▪ **72 W** power consumption.	https://www.nvidia.com/en-us/data-center/l4/
Jetson AGX Orin Developer Kit	Official development board for Jetson AGX Orin, designed for AI and robotics prototyping.	▪ Integrated Jetson AGX Orin module. ▪ Multiple I/O: PCIe, Ethernet, USB 3.2, MIPI CSI for cameras. ▪ Preloaded with **JetPack SDK** and TensorRT.	https://developer.nvidia.com/embedded/learn/get-started-jetson-agx-orin-devkit
Jetson Orin NX Developer Kit	Development board for Jetson Orin NX, enabling real-world AI testing.	▪ Compact design with full I/O support. ▪ AI-ready with DeepStream and TensorRT. ▪ Power-efficient form factor.	https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-orin/
NVIDIA JetPack SDK	Core software stack for Jetson platforms, including AI inference and vision-processing tools.	▪ Includes **CUDA, cuDNN, TensorRT.** ▪ Supports **Ubuntu-based Jetson Linux.** ▪ Cloud-native AI deployment (Docker, Kubernetes). ▪ Pre-optimised libraries for AI and vision.	https://developer.nvidia.com/embedded/jetpack
NVIDIA TensorRT	AI inference engine that optimises and accelerates deep learning models for real-time edge deployment.	▪ **4×–6×** faster inference versus unoptimised models. ▪ Optimised for Jetson and NVIDIA GPUs. ▪ Supports INT8, FP16 quantisation for efficiency. ▪ Works with PyTorch, TensorFlow, and ONNX.	https://developer.nvidia.com/tensorrt

TECHNOLOGY	DESCRIPTION	KEY FEATURES	WEBSITE
NVIDIA DeepStream SDK	AI-powered intelligent video analytics framework for edge applications.	▪ Accelerates vision AI applications. ▪ Supports multiple video streams for real-time processing. ▪ Integrates TensorRT for optimised inference. ▪ Used in **smart cities, security, and retail analytics.**	https://developer.nvidia.com/deepstream-sdk
NVIDIA TAO Toolkit	Low-code AI model training and optimisation tool for edge deployment.	▪ Fine-tunes pre-trained models with **transfer learning.** ▪ Requires minimal training data. ▪ Optimises models for **Jetson and TensorRT.** ▪ Supports vision AI (detection, segmentation, pose estimation).	https://developer.nvidia.com/tao-toolkit
NVIDIA Triton Inference Server	Open-source inference server for deploying AI models at the edge.	▪ Supports PyTorch, TensorFlow, ONNX, and TensorRT. ▪ Efficient model scheduling and batching. ▪ Runs on Jetson, edge GPUs, and data centres. ▪ Enables multi-tenant AI inference workloads.	https://developer.nvidia.com/dynamo

Key offerings include those detailed in *Table 8.2*, which provides a structured summary of **NVIDIA's latest Edge AI hardware and tools**, covering **Jetson modules, discrete GPUs, development kits and AI software**. These platforms are designed to bring AI capabilities directly to edge devices, enabling real-time processing, enhanced privacy, and improved efficiency across various applications. In addition, NVIDIA have designed two new platforms, DIGITS and Cosmos.

- **DIGITS**: Introduced by NVIDIA at CES 2025, DIGITS is a personal AI supercomputer designed to provide high-performance AI computing to individual developers, researchers and students. This compact system is powered by the new NVIDIA GB10 Grace Blackwell Superchip, delivering up to one petaflop of AI performance. This enables users to efficiently prototype, fine-tune and run large AI models directly on their desktops. Starting at USD3,000, Project DIGITS makes high-performance AI computing more accessible, reducing reliance on cloud services and associated costs. Its compact design allows it to operate using a standard electrical outlet, making it suitable for various work environments.
- **Cosmos:** NVIDIA's Cosmos is a platform designed to accelerate the development of physical AI systems, such as autonomous vehicles and robots. It offers generative world foundation models trained on extensive video data, enabling the generation of physics-aware simulations from various inputs. Cosmos includes advanced tokenisers for efficient data processing and guardrails to ensure safety and ethical standards. By providing these tools, Cosmos aims to make physical AI development more accessible and efficient for developers.

8.1.1 CHANCES AND OPPORTUNITIES FOR THE EUROPEAN VENDORS

European AI chip vendors distinguish themselves from NVIDIA through several unique selling points:

- **Energy efficiency:** Companies such as Arm Holdings have developed chip architectures renowned for their energy efficiency. Arm's designs are widely used in mobile devices and are increasingly adopted in data centres to reduce power consumption, offering a more sustainable alternative to NVIDIA's GPUs.
- **Trust, security and safety:** Operating within the European Union's regulatory framework, European vendors may benefit from policies aimed at promoting high safety and security standards in the tech industry. For instance, the Artificial Intelligence Act, introduced in 2024, contributes to the trustworthiness of European solutions. European companies' expertise in power management ICs and embedded security solutions also provides them with a competitive advantage in Edge AI. Security and safety are critical for deploying AI in regulated applications like automotive systems, areas often overlooked by others. Processing data locally at the edge enhances security and data protection, which is essential for applications such as autonomous vehicles. In terms of products, tools and platforms, the market for classical, functionally fixed Edge AI is quite mature. Tools for deploying lightweight, domain-specific GenAI models (NXP's eIQ GenAI Flow) also open the way to deploying GenAI at the edge.

Leading European chip and microelectronics companies, along with prominent research organisations, have formed the Edge AI Working Group. Together, they have outlined objectives to create a roadmap to guide the future of Edge AI development. This roadmap aims to sustain Europe's leadership in the field and to keep pace with rapid innovations.

8.1.2 STMICROELECTRONICS

STMicroelectronics (ST) offers a comprehensive suite of Edge AI technologies, combining advanced hardware and software solutions to enable efficient on-device AI across various applications. Tabkle 83 provides a summary of their key offerings.

***Table 8.3:** ST key offerings (Source: https://www.st.com/content/st_com/en/st-edge-ai-suite/tools.html)*

TOOL/SERVICE	DESCRIPTION	WEBSITE
ST AIoT Craft	An online tool that accelerates the development of sensor-to-cloud solutions using ST components with in-sensor AI capabilities. It enables users to create AI-enabled IoT nodes, program the machine learning core within MEMS sensors, and explore end-to-end project examples.	https://www.st.com/content/st_com/en/st-edge-ai-suite/tools.html
NanoEdge AI Studio	A free AutoML (Automatic Machine Learning) software that guides users step-by-step to integrate Edge AI into embedded projects. It supports over 1,000 Arm® Cortex®-M microcontrollers, offering an automatic machine-learning model generator and a user-friendly interface for end-to-end deployment.	https://www.st.com/en/development-tools/nanoedgeaistudio.html
ST Edge AI Developer Cloud	A free online platform that allows users to optimise and benchmark Edge AI models across various ST devices. Leveraging the ST Edge AI Core, it provides services such as online AI benchmarking, model optimisation and profiling, enabling users to run their AI models on ST's board farm.	https://www.st.com/content/st_com/en/st-edge-ai-suite/tools.html

TOOL/SERVICE	DESCRIPTION	WEBSITE
MEMS Studio	A comprehensive desktop software solution designed to enable Edge AI features on MEMS sensors. It facilitates the collection, labelling and analysis of sensor data, profiling and optimisation of neural network and machine-learning models for the intelligent sensor processing unit (ISPU), and configuration of the MEMS machine learning core.	https://www.st.com/en/development-tools/mems-studio.html
ST Edge AI Core	A command-line interface (CLI) tool that allows users to import AI models from popular machine-learning frameworks, perform detailed analyses, and optimise models for deployment on various ST devices, including sensors, microcontrollers and microprocessors.	https://www.st.com/en/development-tools/st-edgeai-core.html
ST Edge AI Model Zoo	A collection of reference Edge AI models optimised for execution on ST devices. Users can select from a variety of AI models, retrain them using provided datasets and scripts, and deploy them in their applications.	https://stm32ai.st.com/model-zoo/
STM32Cube.AI	A free STM32Cube expansion package (X-CUBE-AI) that enables developers to optimise, profile and evaluate neural network and machine-learning models specifically for STM32 platforms.	https://stm32ai.st.com/stm32-cube-ai/
High Speed Datalog	A tool designed to manage the acquisition and labelling of sensor data. It allows users to capture and monitor high-rate data, manage data using a Python SDK, and port projects across multiple MCU series.	https://www.st.com/en/embedded-software/fp-sns-datalog2.html
StellarStudioAI	An AI plugin for Stellar electrification (E) microcontrollers, it facilitates the conversion of AI models, creation and review of neural network performance reports, and automatic conversion of pretrained neural networks.	https://www.st.com/en/development-tools/stellar-studioai.html
AI for OpenSTLinux	The X-LINUX-AI is an STM32 MPU OpenSTLinux expansion package that supports various AI applications, including pose estimation (Yolov8n), semantic segmentation (DeepLabv3), and image classification (MobileNetv2).	https://stm32ai.st.com/ai-for-linux/
Hand Posture ToF AI	A hand posture recognition solution that detects a set of hand postures based on ST's multizone Time-of-Flight sensors, eliminating the need for a camera. It recognises seven predefined hand postures using data from an 8x8 ranging distance and signal rate matrix.	https://www.st.com/content/st_com/en/campaigns/st-gesture-and-hand-posture-recognition-imag-mc-ghpr.html#:~:text=Train%20your%20AI%20to%20create%20unlimited%20hand%20postures&text=Our%20Hand%20Posture%20ToF%20AI,Explore%20new%20possibilities%20today!

These tools collectively provide a robust ecosystem for developers aiming to implement Edge AI solutions across a wide range of applications, leveraging STMicroelectronics' hardware platforms.

8.1.3 NXP

NXP Semiconductors offers a broad **Edge AI portfolio** spanning high-performance application processors, efficient crossover processors, and even microcontrollers, all supported by a unified machine learning software environment. These solutions are designed to enable AI at the edge with low latency, privacy and energy efficiency[41].

Table 8.4 details NXP's newest processors, development boards, AI accelerators, and software tools for Edge AI – highlighting key specs, target applications, and innovations, along with how they support model deployment, optimisation and real-time inferencing at the edge.

***Table 8.4:** NXP Edge AI technologies*

TECHNOLOGY	DESCRIPTION	KEY FEATURES	WEBSITE
i.MX 95 Applications Processor	High-performance Edge AI processor with integrated NPU, GPU, and safety features.	▪ 6x Cortex-A55 (2 GHz) for application processing. ▪ Arm Mali GPU for advanced 2D/3D graphics. ▪ 2 TOPS eIQ Neutron NPU for AI inferencing. ▪ Dual ISP supporting up to 12MP sensors. ▪ ASIL-B/SIL-2 Safety Certification. ▪ EdgeLock Secure Enclave for hardware-based security. ▪ Dual GbE TSN for industrial and automotive networking. ▪ PCIe Gen3, USB 3.0.	https://www.nxp.com/docs/en/fact-sheet/IMX95FS.pdf
i.MX 93 Applications Processor	Efficient Edge AI processor with Arm Ethos-U65 microNPU for low-power AI applications.	▪ Dual Cortex-A55 (1.7 GHz) for Linux-based AI applications. ▪ Cortex-M33 for real-time tasks. ▪ 0.5 TOPS Ethos-U65 microNPU for AI inferencing. ▪ Energy Flex architecture for dynamic power control. ▪ EdgeLock Secure Enclave for encrypted data storage and authentication. ▪ Dual CAN-FD, GbE TSN. ▪ Low power operation for battery-powered Edge AI.	https://www.nxp.com/products/processors-and-microcontrollers/arm-processors/i-mx-applications-processors/i-mx-9-processors/i-mx-93-applications-processor-family-arm-cortex-a55-ml-acceleration-power-efficient-mpu:i.MX93
i.MX 8M Plus Applications Processor	AI-focused SoC with integrated NPU and dual ISP for vision and multimedia applications.	▪ Quad Cortex-A53 (1.8 GHz), Cortex-M7 for real-time control. ▪ 2.3 TOPS NPU for AI workloads. ▪ Dual ISP supporting 1080p60 video input. ▪ Hardware video encoding (H.265/H.264). ▪ DSP for audio processing and voice recognition. ▪ LPDDR4 RAM support. ▪ Industrial-grade temperature range (–40°C to 105°C).	https://www.nxp.com/products/
MCX N Series Microcontrollers	First NXP-designed MCU with integrated NPU for TinyML applications.	▪ Dual Cortex-M33 at 150 MHz. ▪ eIQ Neutron NPU (30× AI acceleration vs CPU-only). ▪ Integrated DSP for audio/signal processing. ▪ EdgeLock Secure Enclave. ▪ Ultra-low power consumption (<45 µA/MHz). Extensive analogue and digital peripherals for IoT.	https://www.nxp.com/products/processors-and-microcontrollers/arm-microcontrollers/general-purpose-mcus/mcx-arm-cortex-m/mcx-n-series-microcontrollers:MCX-N-SERIES

41 *https://www.nxp.com/company/about-nxp/newsroom/NB-NXP-EXPANDS-EDGE-AI-CAPABILITIES-EIQ#:~:text=Deploying%20AI%20at%20the%20edge,wider%20range%20of%20edge%20processors*

TECHNOLOGY	DESCRIPTION	KEY FEATURES	WEBSITE
i.MX 95 Evaluation Board	Early-access hardware platform for testing i.MX 95 features and AI acceleration.	▪ SoM with i.MX 95. ▪ Dual camera input. ▪ Multiple display interfaces. ▪ PCIe Gen3, USB 3.0. ▪ Ethernet, Audio, GPIO expansion. ▪ AI and vision demos included.	https://www.toradex.com/computer-on-modules/verdin-arm-family/nxp-imx95-evaluation-kit?utm_term=&utm_campaign=PMax-:+Toradex_EU_Smart_Shopping_Ads_20240205(UK)&utm_source=adwords&utm_medium=ppc&hsa_acc=5623819148&hsa_cam=20985550698&hsa_grp=&hsa_ad=&hsa_src=x&hsa_tgt=&hsa_kw=&hsa_mt=&hsa_net=adwords&hsa_ver=3&gad_source=1&gad_campaignid=21184289667&gbraid=0AAAAAD_Ks1XrlAkB_1VmI49AueKd78HYw&gclid=EAIaIQobChMIuPLE-IS6jgMVDJJQBh1DODeCEAAYASAAEgItO_D_BwE
i.MX 93 Evaluation Kit	Compact three-board setup to develop AI-powered applications with i.MX 93.	▪ Compute module with i.MX 93 SoC. ▪ Expansion boards for vision/audio interfaces. ▪ Supports AI inferencing on Ethos-U65 NPU. ▪ Pre-loaded machine-learning demos. ▪ Low power AI development-ready.	https://www.nxp.com/design/design-center/development-boards-and-designs/i.MX93EVK
MCX N9xx-EVK	Evaluation board for MCX N series MCUs with built-in TinyML support.	▪ Onboard sensors (accelerometer, microphone). ▪ AI-optimised power management. ▪ Pre-configured with eIQ ML demos. ▪ Secure boot and encryption support.	https://www.nxp.com/design/design-center/development-boards-and-designs/MCX-N9XX-EVK
eIQ Machine Learning Toolkit	Comprehensive software suite for AI model optimisation and deployment on NXP hardware.	▪ Supports TensorFlow Lite, Arm NN, Glow Compiler, DeepViewRT. ▪ Model Zoo with pre-trained models. ▪ Optimisation tools for NXP NPUs and MCUs. ▪ Secure AI model execution with EdgeLock.	https://www.nxp.com/design/design-center/software/eiq-ai-development-environment/eiq-toolkit-for-end-to-end-model-development-and-deployment:EIQ-TOOLKIT
eIQ Time Series Studio	Automated ML workflow tool for time-series sensor data, targeting MCU-class devices.	▪ No-code AI model training for industrial sensors and predictive maintenance. ▪ AutoML tools for anomaly detection. ▪ Low power AI model optimisation.	https://www.nxp.com/company/about-nxp/smarter-world-blog/BL-INTRODUCING-THE-EIQ-TIME-SERIES-STUDIO
eIQ GenAI Flow	Development tool for deploying small generative AI models on NXP edge processors.	▪ Supports domain-specific LLMs. ▪ Local natural language processing. ▪ Retrieval-Augmented Training (RAG) for edge inference. ▪ Optimised for i.MX 8/9 processors.	https://www.nxp.com/design/design-center/software/eiq-ai-development-environment:EIQ
EdgeReady Solutions	Turnkey AI hardware and software for facial recognition and voice control at the edge.	▪ i.MX RT106F MCU for AI facial recognition with liveness detection. ▪ i.MX RT106V for offline voice command processing. ▪ Low-latency, privacy-focused AI inferencing.	https://www.nxp.com/applications/technologies/edge-computing/edgeready:EDGEREADY
Kinara Acquisition by NXP	Kinara is a leading edge AI accelerator specialising in accelerating LLMs and multimodal AI applications.	▪ Up to 40 TOPS performance on the ARA-2 accelerator. ▪ Support for transformers with up to 30B parameters in INT4 precision. ▪ Design and customer wins in growing Edge AI markets like retail and AI PCs.	https://www.eetimes.com/nxp-acquires-ai-chip-startup-kinara/

Through these innovations, NXP empowers developers to create intelligent, safe, secure, certified and efficient edge applications.

8.1.4 INFINEON TECHNOLOGIES AG

Infineon is a leading global semiconductor manufacturer specialising in power systems and IoT solutions. Table 8.5 provides a summary of Infineon's key Edge AI technologies.

***Table 8.5:** Infineon's key Edge AI technologies*

TECHNOLOGY	DESCRIPTION	KEY FEATURES	WEBSITE
DEEPCRAFT™ Edge AI Solutions	A comprehensive software platform enabling rapid implementation of AI and ML functionalities in IoT edge devices.	▪ DEEPCRAFT™ Studio: Development environment for creating or optimising AI models. ▪ DEEPCRAFT™ Ready Models: Pre-trained, production-ready AI models optimised for Infineon's sensors and microcontrollers.	https://www.infineon.com/design-resources/embedded-software/deepcraft-edge-ai-solutions
PSoC™ Edge Microcontroller Family	A new generation of microcontrollers optimised for machine learning-based applications, offering scalable performance, features, and memory options.	▪ The PSOC™ Edge Family of Arm® Cortex®-M microcontrollers feature high-performance, low power, secured MCUs with hardware-assisted ML acceleration for next generation applications. ▪ They support an extensive set of peripheral sets, on-chip memories, timers, robust hardware security features and comprehensive connectivity options, built for a variety of consumer and industrial applications where device-based intelligent intuitive interaction is rapidly evolving. This includes appliances, speakers, wearables, robotics, and other smart home devices, some of which are also connected IoT products.	https://www.infineon.com/promo/next-generation-mcu?redirId=269245#family-overview
ModusToolbox™ Software	ModusToolbox™ software is a modern, extensible development environment supporting a wide range of Infineon microcontroller devices.	▪ ModusToolbox™ provides a flexible set of tools and a diverse, high-quality collection of application-focused software. These include configuration tools, low-level drivers, libraries, AI development tools and operating system support, most of which are compatible with Linux, macOS, and Windows-hosted environments.	https://www.infineon.com/design-resources/development-tools/sdk/modustoolbox-software
AURIX™ TC4x Family	Infineon's AURIX™ TC4x family of microcontrollers focuses on real-time safe and secure processing for edge applications.	▪ They are designed for next-generation eMobility, ADAS, automotive E/E architectures and affordable AI applications. ▪ AURIX™ Accelerator Suite: ▪ Parallel Processing Unit (PPU) enabling AI up to ASIL-D. ▪ Data Routing Engine (DRE), for efficient communication and data handling. ▪ cDSP: Programmable digital signal processing for the ADC signals. ▪ Signal Processing Unit (SPU): radar accelerator. ▪ Security Accelerators (CSRM/CSS): Hardware Crypto Acceleration.	https://www.infineon.com/products/microcontroller/32-bit-tricore/aurix-tc4x

These technologies empower developers to create efficient, intelligent edge devices tailored to a wide range of applications.

9 Goals, Objectives and Recommendations for Action

While the early focus in GenAI was on extremely large language models (100s to 1,000 billion parameters) that demanded substantial computing power, high-speed connectivity, wide bandwidth and huge training datasets, the field is now evolving toward more efficient and accessible approaches that can deliver strong performance on narrower use cases and with fewer resources. Today, major tech companies offer lightweight AI models designed to operate efficiently at the edge, even on low power devices with limited resources, unreliable connectivity, and stringent real-time or safety-critical requirements (such as for automotive applications). However, addressing device constraints and limitations demands innovative approaches and a paradigm shift in both hardware and software development for Edge AI.

The Edge AI Working Group has formulated the following actions.

- Achieving **strategic autonomy** for European business and manufacturing industry involves reducing dependency on external entities by fostering self-reliance in critical sectors such as technology, defence and energy. This approach enhances the EU's capacity to act independently, uphold democratic values, and strengthen its position as a global actor.
- Communicating a **clear vision** to the European Commission and relevant stakeholders is essential for aligning efforts toward common goals. A well-defined European strategy facilitates open collaboration, ensures policy coherence and mobilises resources effectively, thereby advancing initiatives that promote innovation, competitiveness and sustainability within the EU.
- Identifying **use cases from industry, especially SMEs**, is crucial for tailoring technological solutions to real-world challenges. By understanding the specific needs of SMEs, policies can be designed to support their integration of AI and other advanced technologies, fostering growth, competitiveness and democratisation of AI.
- Identifying **key enabling technologies and building blocks over a five-to-10 year period** with a reasonable market size involves **forecasting technological trends and market demands.** This foresight enables the EU to invest strategically in areas such as Edge AI, ensuring that emerging technologies align with European values and have the potential for significant economic impact.
- Identifying **dependencies and risks** is vital for ensuring technological autonomy. By assessing reliance on non-EU technologies and resources, the EU can develop strategies to mitigate risks, diversify supply chains and strengthen internal capabilities, thereby enhancing resilience against external shocks.
- Identifying **opportunities for collaboration between industry and research** fosters innovation and accelerates technological development. Partnerships between businesses and research institutions facilitate knowledge transfer, support the commercialisation of research outcomes, and enhance the EU's competitive edge in global markets.
- Identifying **cross-domain synergies, technology transformations, ecosystem and tool design** involves recognising overlaps between different sectors and technologies. Leveraging these synergies can lead to more efficient development processes, cost reductions, and the creation of versatile tools that serve multiple applications, thereby maximising the impact of technological advancements.
- Helping companies make **decisions about investments in technology development and strategic collaborations** requires providing them with insights into market trends, technological advancements and potential partnerships. This guidance enables businesses to allocate resources effectively, innovate, and remain competitive in a rapidly evolving technological landscape.

- Exploiting **enabler technologies** such as RISC-V and Green ICT involves adopting open-source hardware architectures and sustainable information and communication technologies. These technologies promote innovation, reduce costs and align with environmental goals, contributing to the EU's digital sovereignty and sustainability objectives.
- Making **Edge AI a strategic asset** for the Chips JU entails integrating AI capabilities directly into hardware components. This approach enhances processing efficiency, reduces latency, and supports the development of advanced applications and programming frameworks, thereby strengthening the EU's position in the semiconductor industry.
- Building European on-premises and at the edge AI computational capacities, including being able to run **AI learning and inference workloads both on-premises and at the deep edge,** is crucial. These should have a strong focus on deployment tools and low power chips alongside an unprecedented energy-efficiency envelope. Developing skills is foundational for supporting advanced technological research and innovation. Investments in **on-premises highly energy-efficient computing infrastructure and educational programmes** ensure that the EU has the necessary resources and talent to lead in fields such as AI and big data analytics.
- Encouraging **education and training** about AI usage and development involves creating teaching programmes that equip individuals with the skills to utilise and build AI technologies responsibly. This focus on applied AI ensures that the next generation workforce can meet the demands of the European digital economy and contribute to ethical AI development[42].
- **Reducing the brain drain** requires creating an environment that retains and attracts talent within the EU by federating the nations to act as a unified human resource organisation. This can be achieved by offering shared values, competitive opportunities, fostering innovation ecosystems, and providing support for research and entrepreneurship, thereby preventing the loss of skilled professionals to other regions.
- Supporting and simplifying the **creation of startups in the Edge AI domain** involves reducing bureaucratic hurdles, providing access to funding, and offering mentorship programmes. These measures encourage entrepreneurship, stimulate economic growth, and drive technological innovation within the EU. Facilitating the relocation of high-potential talent across the EU is essential for fostering innovation and sustaining knowledge growth. By enabling mobility, the EU will enhance knowledge retention and ensure a return on national investments in the education of students and young professionals. To better assess the impact of these efforts, data should be incorporated on the annual number of STEM graduates in the EU – an indicator of the region's potential intellectual capital generated through public educational investments.

Focusing efforts on areas not yet dominated by the US or Asia allows the EU to carve out niches in emerging technologies. By identifying and investing in underexplored sectors, the EU can establish leadership positions, diversify its technological portfolio, and reduce dependency on external technologies. There should be a unified and collaborative EU effort to achieve human brain energy-efficient chips (eg, 50 Peta Operations/W) focused on energy-efficient scalable on-premises and Edge AI computing. This would help bring about decarbonisation, CO_2 reduction, and water and energy savings, and avoid AI computing centres being supplied by nuclear reactors as happens in US for high performance and cloud GenAI computing. It should also be a strategic priority on techniques such as data cleaning, compression and augmentation, along with model optimisation methods including knowledge distillation, pruning and deep heterogenous quantisation.

42 *According to talentneuron, there will be a global shortage of over 85 million STEMs by 2030, with a potential loss of USD8.5 trillion in GDP (see https://www.talentneuron.com/blog/solutions-for-bridging-the-growing-stem-skills-gap). This means it is imperative that the EU develop the next generation of STEM professionals.*

Collectively, these strategies aim to bolster the EU's technological sovereignty, foster innovation, and ensure that European businesses and industries remain competitive on the global stage.

Implementing the following measures will help to achieve these goals. However, addressing the outlined objectives necessitates a comprehensive approach to advancing Edge AI within the EU.

- **Aligning strategy with global trends and EU initiatives:** To maintain competitiveness in Edge AI, it is crucial to synchronise strategies with global developments and EU initiatives. The EU has launched significant programmes, such as like the EU AI Champions Initiative and InvestAI, collectively mobilizing around €200 billion to accelerate AI innovation across the continent. Aligning with these initiatives ensures that efforts are cohesive, leveraging shared resources and knowledge to foster technological advancement.
- **Assessing the Current EU and Global State of the Art in Edge AI:** A thorough assessment of the existing landscape in Edge AI within the EU and globally is essential. This involves evaluating recent advancements, ongoing research, and emerging applications to identify strengths and areas needing improvement. For instance, the EU-funded dAIEDGE project unites leading research centers and industrial partners to develop new paradigms for distributed AI solutions, positioning Europe at the forefront of Edge AI innovation.
- **Reducing the Complexity of Edge AI Systems:** Simplifying Edge AI systems is vital for broader adoption and efficiency. Techniques such as data cleaning, compression, and augmentation, along with model optimization methods like pruning and quantization, can make AI models more suitable for deployment on resource-constrained edge devices. Additionally, system optimization strategies, including framework support and hardware acceleration, contribute to more efficient Edge AI workflows.
- **Organizing Application and Domain Consultations:** Engaging with diverse stakeholders through consultations is crucial for gathering insights and fostering collaboration. Initiatives like the EU's Internet of Things policy demonstrate the importance of cross-sector collaboration to boost industrial cooperation through open platforms and standards, thereby achieving European leadership across the entire edge ecosystem[43].
- **Setting Priorities, Topics, and Benchmarks for Future Chips Act JU Calls:** Establishing clear priorities and benchmarks is essential for guiding future research and funding. The European Chips Act, which came into force in September 2023, aims to double Europe's global semiconductor market share to 20% by 2030, providing €43 billion in public and private investment for chip research and development. Aligning future Joint Undertaking (JU) calls with this act ensures that resources are directed toward impactful areas in Edge AI.
- **Increasing Technology Readiness Level and Promoting Market Readiness:** Advancing the Technology Readiness Level (TRL) of Edge AI technologies involves moving innovations from the lab to real-world applications. The EU's investment of €180 million in breakthrough digital technologies, including AI, robotics, and new materials, underscores the commitment to bridging the gap between research and market deployment. Focusing on customer-centric research and development ensures that technologies meet market needs and are poised for successful adoption.

By addressing these facets, the EU can foster a robust Edge AI ecosystem that is innovative, competitive, and aligned with both regional and global technological advancements. The following objectives will provide guidance for overcoming the highlighted barrios and for maintaining and expanding the position of the European players.

43 https://digital-skills-jobs.europa.eu/en/actions/european-initiatives/europes-internet-things-policy

9.1 Objective 1: Create European ecosystem and enforce synergies between existing ecosystems for fast adoption of Edge AI solutions

The role of edge AI in the computing continuum is growing. The adoption of this technologies especially in safety-critical systems depends not only on resource efficient tiny ML approaches and models but also on the innovation in engineering and chip design processes including simulation and testing. The main technological milestones and R&I actions should address:

Advancements in Edge Artificial Intelligence (AI) necessitate a comprehensive understanding of various technological aspects to develop efficient, reliable, and user-centric systems. Here's an elaboration on the key areas:

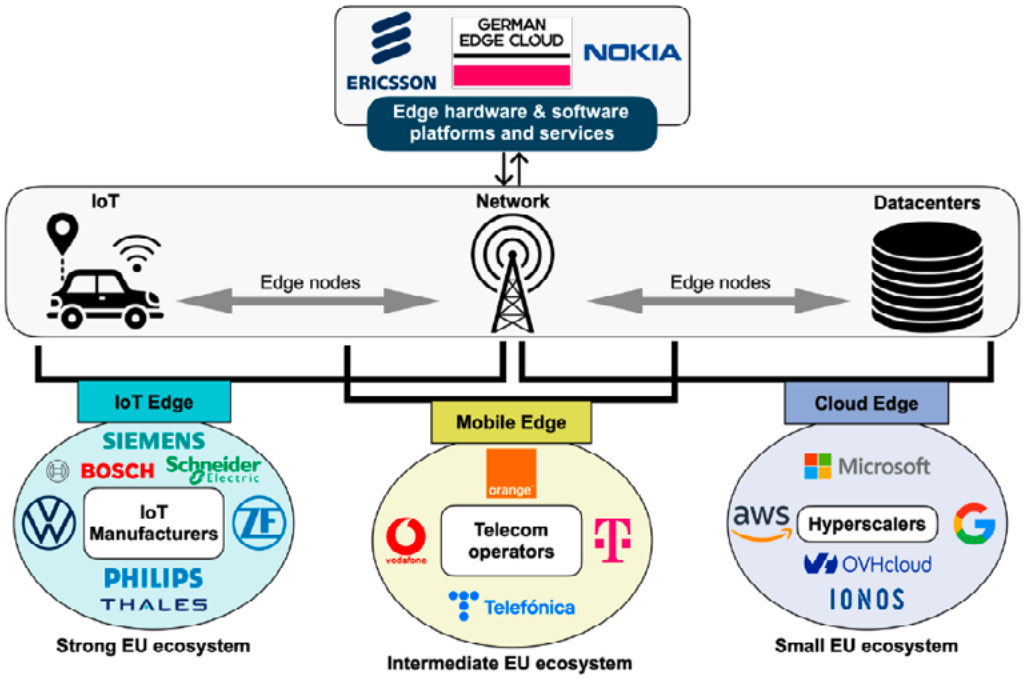
- **Migration of the Processing to the Edge:** Shifting computational tasks from centralised cloud servers to edge devices offers benefits like reduced latency and enhanced privacy. Techniques such as advanced memory management and in-memory computing accelerators are pivotal in this transition. In-memory computing reduces the energy consumption associated with data transfer between memory and processing units by performing computations directly within the memory hardware. This approach is particularly effective in low-power AI edge devices, addressing the memory-wall bottleneck inherent in traditional architectures.
- **Foundational Models, Data, and Learning Technologies:** Distributed Edge AI involves deploying AI models across multiple edge devices, enabling localised data processing and decision-making. This paradigm relies on foundational models tailored for edge environments, efficient data management strategies, and learning technologies that support decentralised training and inference. By distributing AI workloads, systems can achieve scalability and resilience, essential for applications such as autonomous vehicles and smart cities.
- **AI chips supporting multiple computing paradigms and multi-technology AI:** The development of AI chips capable of supporting various computing paradigms – such as classical computing, analogue, neuromorphic computing, and deep learning – is essential for versatile AI applications. For instance, BrainChip's Akida neural processor integrates event-based processing, mimicking neurological functions to enhance efficiency in Edge AI applications. Similarly, AMD's Instinct MI300 series combines traditional and AI-optimised cores to accelerate diverse workloads. The emerging novel spintronic hardware for Edge AI could also be a breakthrough in overcoming the current limitations of existing hardware architectures.
- **AI verification and certification:** Ensuring the reliability and safety of AI systems is critical, especially in sectors such as healthcare and autonomous driving. Verification and validation (V&V) processes systematically assess AI models to identify potential errors or biases, validating their performance against predefined criteria. Techniques include testing against representative datasets, conducting simulations, and analysing decision-making processes to ensure AI systems operate within acceptable bounds.
- **AI explainability, interpretability, verification and certification for building trust in AI systems:** Establishing trust in AI systems is essential for their widespread adoption and responsible deployment. This begins with explainability and interpretability, which aim to make AI decision-making processes understandable and transparent to humans – an increasingly important requirement for both user acceptance and regulatory compliance (eg, the AI Act). Equally important are verification and certification processes, which ensure that AI systems adhere to standards of safety, fairness and reliability. These practices help validate that AI behaves as intended, particularly in high-stakes applications. Trustworthy AI also encompasses model security, including the authentication of deployed models, monitoring their evolution over time, and verifying the quality and integrity of the data used during training. Together, these elements form the foundation for deploying AI systems that are not only powerful but also accountable and secure.

- **Interoperability, scalability and modularity:** Designing AI systems with these features ensures seamless integration across diverse platforms, as well as the ability to adapt to growing demands.
- **On-device training is the learning mechanism that powers Self-X functionalities:** Such training should include self-learning and self-adaptation, self-configuration, self-healing, and self-optimisation to enable AI systems to become robust, autonomous, context-aware and adaptive, all without cloud dependence, thus enhancing resilience and efficiency.
- **Engineering tools for designing, training, optimising, deploying, updating, and robustness against cyber-attacks and maintaining Edge AI:** Specialised engineering tools facilitate the lifecycle management of Edge AI applications. These tools assist in designing, training, updating and maintaining AI models, ensuring they remain effective and secure over time. For example, AI-powered verification tools enhance the efficiency of SoC design verification, reducing manual effort and improving accuracy.
- **Support for the entire lifecycle from requirement specification to end-of-life:** Comprehensive support throughout the AI system lifecycle – from requirement specification, design, development, deployment, operation, maintenance, evolution, to end-of-life – is vital for sustainability and compliance. This holistic approach ensures that AI systems are developed responsibly, maintained effectively and decommissioned safely, aligning with ethical and regulatory standards.
- **Human interaction with AI:** Optimising human–AI interaction focuses on creating natural interfaces and interactions, and ensuring AI systems understand and respond to human inputs effectively. This involves natural language processing, adaptive learning, and user-centric design principles to enhance user experience and trust in AI applications.
- **Intent-driven optimisation, machine-to-machine interaction, interaction with digital twins:** Intent-driven optimisation allows AI systems to anticipate and act upon user intentions, improving efficiency and personalisation. Machine-to-machine (M2M) interactions enable devices to communicate and collaborate without human intervention, essential for the IoT. Interaction with digital twins (metaverse and virtual worlds) – virtual replicas of physical systems – facilitates real-time monitoring, simulation and optimisation, enhancing decision-making and operational efficiency.

Understanding and integrating these aspects are crucial for advancing Edge AI technologies, leading to more efficient, reliable and user-friendly applications across various industries. **Education and professional training** should supplement the outlined R&I actions for skill and capacity building in Europe.

9.2 Objective 2: Foster collaboration along the AI value chain in Europe, from chip vendors to system integrators, along with collaboration across EU stakeholders in the ECS value chain, from chip designers to integrators to manufacturers

NVIDIA is currently the major market player with a growing ecosystem of hardware and software application providers in quickly evolving domains (such as robotics). It saves integration costs by providing complete solutions to consumers, while its easy-to-use software development kits (SDKs) for management, integration and deployment effectively create a vendor-lock and stronger customer retention. NVIDIA is in charge of the updates of the APIs, and maintains the value chain under its control.



Source: DECISION Etudes & Conseil

Figure 9.1: *Cloud-edge-IoT market structures and ecosystems*
(Source: https://op.europa.eu/en/publication-detail/-/publication/ff35c457-8f3b-11ee-8aa6-01aa75ed71a1/language-en)

The European Edge AI ecosystem is currently fragmented and lacks a dominant player. STMicroelectronics cooperates with NVIDIA, and provides APIs and tools for integration with NVIDIA management and deployment SDKs and hardware solutions. In addition, Infineon expanded its safe automated driving collaboration with the NVIDIA DRIVE™ Pegasus AI car computing platform in 2018.

European players are currently within the circle of suppliers for NVIDIA's solutions, making NVIDIA the fastest-growing and most valuable chip vendor company in the market. The latest Blackwell chip cost USD10 billion in R&D, according to public interviews by NVIDIA's CEO. No hardware industry in Europe can achieve such investment for a single chip, nor could one imagine EU taxpayer funds being used to achieve this level of investment. Therefore, on-premises and energy-efficient edge computing is the most effective EU alternative for investment. An embedded software, application and service ecosystem should be created to complement edge chips, as industries are already proving through a strong focus on SMEs to help them transition toward AI endorsement.

To challenge NVIDIA's dominance in the Edge AI sector, establishing an open ecosystem akin to Kubernetes[44] in cloud computing is essential. This ecosystem should encompass modular edge platforms and infrastructures, facilitating the integration of diverse European hardware and software solutions, thereby mitigating vendor lock-in.

Achieving seamless collaboration necessitates a unified vocabulary that bridges hardware and software domains, fostering effective communication and knowledge sharing. Leveraging large language models can assist in aligning disparate concepts and terminologies. Revisiting and updating existing reference architectures, such as RAMI 4.0[45], could further support this integration.

44 *https://kubernetes.io*

45 *Reference Architectural Model for Industry 4.0 (RAMI4.0): https://www.plattform-i40.de/IP/Redaktion/EN/Downloads/Publikation/rami40-an-introduction.html*

Such harmonisation would promote a balanced market landscape, enhancing the competitiveness of European enterprises. Virtualisation and simulation technologies are pivotal for ensuring seamless integration, trust and collaboration. Given the diverse stakeholders in the value chain, implementing certification mechanisms – such as automated and secure onboarding for systems of systems – and ensuring compliance with standards is crucial, especially in safety-critical applications.

Collaborating with global players, SMEs and startups is also vital. Large corporations offer generic components that can be customised, while SMEs and startups, closely aligned with end-users, act as system integrators, tailoring solutions to specific needs. A modular and integrated approach would streamline the development of customised solutions. However, challenges such as GDPR compliance and stringent security requirements may impede collaboration with SMEs.

The European Chips Act should emphasise the co-design of software and hardware, exploring AI methodologies beyond GenAI, including tinyML, federated learning, and reinforcement learning. Initiatives could focus on enhancing interoperability across the value chain through a shared vocabulary, potentially by funding platforms that support diverse approaches and modular architectures, revising frameworks such as RAMI4.0 or the Asset Administration Shell (AAS). Engaging open-source organisations, incubators and accelerators as partners is essential for ecosystem development. Identifying and collaborating with early adopters of Edge AI technologies – such as startups, researchers and SMEs – will be crucial in the coming years to scale novel solutions. Notably, countries such as South Korea and Singapore are more receptive to new technologies compared to Europe's conservative stance. European companies should strategise their market entry by considering global adoption dynamics.

To remain competitive globally, the technology stack must address both vertical and cross-domain aspects. Given the evolving requirements, a singular, generic technology stack is impractical. Instead, developing a flexible, modular stack composed of interoperable building blocks is imperative. This approach requires ongoing standardisation and interoperability efforts across various domains. Throughout this development process, identifying gaps will highlight opportunities for startups and SMEs, fostering innovation and growth within the European Edge AI ecosystem.

9.3 Objective 3: Create greater market impact along the AI value chain for Edge AI applications

The increasing demand for low-latency and energy-efficient solutions across diverse applications – such as autonomous driving, assistive systems and robotics – is creating new opportunities for Edge AI technologies. This trend is further reinforced by the increasing need for secure, AI-enabled manufacturing equipment, and the growing integration of AI into medical devices, enabling less invasive, more personalised treatments.

- **Delivering clear value through innovation:** To foster impactful innovation across Europe, we must begin with a clear vision of the end product and its real-world value. Academic research, particularly within universities, should be more closely aligned with industry needs and user-driven priorities. Strengthening the flow of information between research institutions, industrial partners and end-users will ensure that technological advancements are relevant, scalable, and contribute meaningfully to societal progress.
- **Balancing cost and value:** As semiconductor technologies advance, costs are rising – especially with the transition to 2nm and 3nm nodes, where transistor density increases significantly. The cost-reducing effects of Moore's Law are diminishing, particularly in the AI domain. However, by focusing on the value created for end-users, we can justify the necessary investments in infrastructure, supply chains and advanced manufacturing processes.

- **Aligning innovation with market need:** Semiconductor innovation must be market-driven. Europe cannot rely solely on advanced manufacturing capabilities; a strong high-end market is also essential. Strategic collaboration – especially among leading European players such as Infineon, NXP and STMicroelectronics – is crucial to develop demand and share investment burdens effectively.
- **Strengthening policy, research and industry collaboration:** Policymakers should shape funding and evaluation frameworks that prioritise long-term impact over short-term outputs. Researchers, in turn, should focus on application-oriented projects with clear pathways to market. Meanwhile, industry stakeholders must engage early in the research process to help guide innovation toward viable, high-impact solutions.
- **Encouraging knowledge and IP sharing within Europe:** Rigid intellectual property barriers can hinder early-stage innovation. A well-structured framework for shared innovation can amplify impact and reinforce Europe's industrial foundation. Cross-layer collaboration – spanning AI, hardware, and embedded systems – should be prioritised to avoid vendor lock-in and ensure platform portability across applications.
- **Transferring research to meet market demands:** Instead of measuring success by publication volume alone, we should emphasise innovations that have the potential to transform markets, address pressing societal challenges, and solidify Europe's leadership in critical technologies. Achieving this requires a long-term, collaborative approach across sectors, with a focus on strategic alignment and real-world applicability.
- **Enabling strategic, pre-competitive cooperation:** In sectors such as automotive, there are already tangible benefits from shared reference architectures for AI and semiconductors. For example, BMW is driving a collaborative ecosystem for software-defined vehicles with partners that include Bosch, Imec, Cadence, Synopsys, Siemens and Arm as part of the Automotive Chiplet Programme[46]. Standards and harmonisation will be key to building a competitive and open European technology landscape.
- **Fostering European and international collaboration:** To meet the challenges posed by GenAI and emerging technologies, Europe must intensify its collaborative efforts across the ecosystem. Initiatives such as the Edge AI Foundation – despite their North American origins – offer valuable platforms for European participation, knowledge sharing, and alignment with customer needs in an open, value-driven manner.
- **Achieving strategic technological autonomy:** Europe's continued reliance on foreign sources for key technologies, including semiconductors and AI, poses risks to its economy, democracy and technological sovereignty. Addressing this dependency is an urgent strategic imperative. By investing in our capabilities and reinforcing cross-border collaboration, we can secure Europe's leadership and autonomy in the global technology landscape.

Edge AI is at a pivotal moment. To unlock its full potential, Europe must accelerate innovation by improving development tools, reducing fragmentation, and fostering cross-sector collaboration. Unlike cloud AI, which benefits from standardised platforms, Edge AI faces complexity and heterogeneity, requiring tailored design approaches for everything from ultra-low power devices to high-performance chips. GenAI is a transformative force and key driver of current market momentum. Acting swiftly to develop European foundation models and AI tools is essential. Strategic cooperation and shared standards – as seen in efforts such as the Automotive Chiplet Programme – are vital to advancing software-defined mobility and Edge AI. Embracing system-level thinking, encouraging IP sharing under proper frameworks, and fostering cross-layer optimisation will be critical for Europe's leadership in next-generation AI.

46 https://www.imec-int.com/en/press/arm-ase-bmw-group-bosch-cadence-siemens-siliconauto-synopsys-tenstorrent-and-valeo-commit

10 Authors

Paolo Azzoni (INSIDE)

Kay Bierzynski (Infineon Technologies AG)

Gerardo Daaldero (NXP Semiconductors)

Philippe Dallemagne (CSEM)

Mario Diaznava (STMicroelectronics)

Marc Duranton (CEA-Leti [Commissariat à l'énergie atomique et aux énergies alternatives])

Wolfgang Ecker (Infineon Technologies AG)

Jacek Flak (VTT Technical Research Centre of Finland Ltd)

Andreas Hausrotter (esc Aerospace GmbH)

Deepak V Katkoria (Logiicdev GmbH)

Jan Langer (Fraunhofer ENAS)

Anders Lindgren (RISE Research Institutes of Sweden)

Michele Magno (ETH Zürich)

Harald Mathis (Fraunhofer FIT)

Danilo Pau (STMicroelectronics)

Bernhard Peischl (AVL List GmbH)

Pietro Perlo (I-FEVS Interactive Fully Electrical Vehicles S.r.l.)

Davis Sawyer (NXP Semiconductors)

Inessa Seifert (VDI/VDE-IT [EPoSS])

Petri Solanti (Siemens AG)

Markus Taube (Research Studio)

Salvatore Tedesco (Tyndall National Institute)

Hans-Jörg Vögel (BMW Group)

Lars Weimer (esc Aerospace GmbH)

ABBREVIATIONS

A

AAS – Asset Administration Shell
ADAS – Advanced driver-assistance system
ADC – Analogue-to-digital converter
AGI – Artificial General Intelligence
AI – Artificial Intelligence
AI4DI – Artificial Intelligence for Digitizing Industry
AIoT – Artificial intelligence of things
ANDANTE – AI for New Devices and Technologies at the Edge
ANN – Artificial neural network
API – Application programming interface
ASIC – Application-specific integrated circuit
ASRA – Advanced SoC Research for Automotive
AutoML – Automatic Machine Learning

B

BRICS – Brazil, Russia, India, China and South Africa

C

CAGR – Compound annual growth rate
CEA – Commissariat à l'énergie atomique et aux énergies alternatives
CEO – Chief executive officer
CES – Consumer Electronics Show
Chips JU – Chips Joint Undertaking
CIM – Compute-in-memory
CLI – Command-line interface
CMOS – Complementary metal-oxide semiconductor
CNN – Convolutional Neural Network
CoAP – Constrained application protocol
CPU – Central processing unit
CRISPR – Clustered regularly interspaced short palindromic repeats
CSRM – Cybersecurity risk management
CSS – Cybersecurity satellite

D

DL – Deep Learning
DNN – Deep Neural Network
DRAM – Dynamic random-access memory
DRE – Data Routing Engine

E

E/E – Electrical/electronic
ECHO – Edge CHip to clOud
ECS – Electronic components and systems
ECU – Electronic control unit
EDA – Electronic design automation
ePCM – Embedded phase change memory

F

FDSOI – Fully depleted silicon on insulator
FPGA – Field programmable gate array

G

GAN – Generative Adversarial Network
GDPR – General Data Protection Regulation
GenAI – Generative AI
GPGPU – General-purpose graphics processing unit
GPU – Graphics processing unit

H

HPC – High-performance computing

I

I/O – Input/output
IC – Integrated circuit
ICT – Information and communications technology
IoT – Internet of Things
IP – Internet protocol
ISP – Image signal processor
ISPU – Intelligent sensor processing unit

J

JU – Joint undertaking

L

LLM – Large Language Model

M

M2M – Machine-to-machine
M2TJ – Multi-level magnetic tunnel junction
MAS – Multi-agent system
MCU – Microcontroller
MEMS – Micro-electro-mechanical systems
ML – Machine Learning
MQTT – Message queuing telemetry transport
MRAM – Magnetic random-access memory

N

NLP – Natural language processing
NoE – Network of Excellence
NPU – Neural processing unit

O

OxRAM – Oxide-based resistive RAM

P

PCB – Printed circuit board
PCM – Phase-change memory
PCRAM – Phase-change RAM
PPU – Parallel Processing Unit
PUF – Physically unclonable function
PULP – Parallel Ultra-Low Power

R

R&I – Research and innovation
RAG – Retrieval-Augmented Training
RAM – Random-access memory
RAMI4.0 – Reference Architectural Model for Industry 4.0
ReRAM – Resistive RAM
RVV – RISC-V Vector Extension

S

SDK – Software development kit
SME – Small and medium-sized enterprise
SNN – Spiking Neural Network
SoC – System-on-a-chip
SOT – Spin-orbit torque
SPU – Signal Processing Unit
SRAM – Static random-access memory
ST – STMicroelectronics
STCO – System technology co-optimisation
STDP – Spike-Timing-Dependent Plasticity
STEM – Science, technology, engineering and mathematics

T

TinyML – Tiny Machine Learning
TOPS – Trillions of operations per second
TPU – Tensor processing unit
TRL – Technology readiness level
TSMC – Taiwan Semiconductor Manufacturing Co

V

V&V – Verification and validation
V2I – Vehicle-to-infrastructure
V2V – Vehicle-to-vehicle

LIST OF FIGURES AND TABLES

FIGURES

TABLES

Copyright © EPoSS e. V.
Permission to reproduce any text for non-commercial purposes is granted,
provided that it is credited as source.

Berlin, October 2025
smart-systems-integration.org