2.1

DRAFT

Cross-Sectional Technologies

EDGE COMPUTING AND EMBEDDED ARTIFICIAL INTELLIGENCE

1 Edge computing and embedded Artificial Intelligence

1 Scope

1.1 Introduction

Our world is drastically changing with the deployment of digital technologies that provide ever increasing performance and autonomy to existing and new applications at a constant or decreasing cost but with a big challenge concerning energy consumption. Especially cyber-physical systems (CPS) place high demands on efficiency and latency. Distributed computing systems have diverse architectures and in addition tend to form a continuum between extreme edge, fog, mobile edge[1] and cloud. Nowadays, many applications need computations to be carried out on spatially distributed devices, generally where it is most efficient. This trend includes edge computing, edge intelligence (e.g. Cognitive CPS, Intelligent Embedded Systems, Autonomous CPS) where raw data is processed close to the source to identify the insight data as early as possible bringing several benefits such as reduce latency, bandwidth, power consumption, memory footprint, and increase the security and data protection.

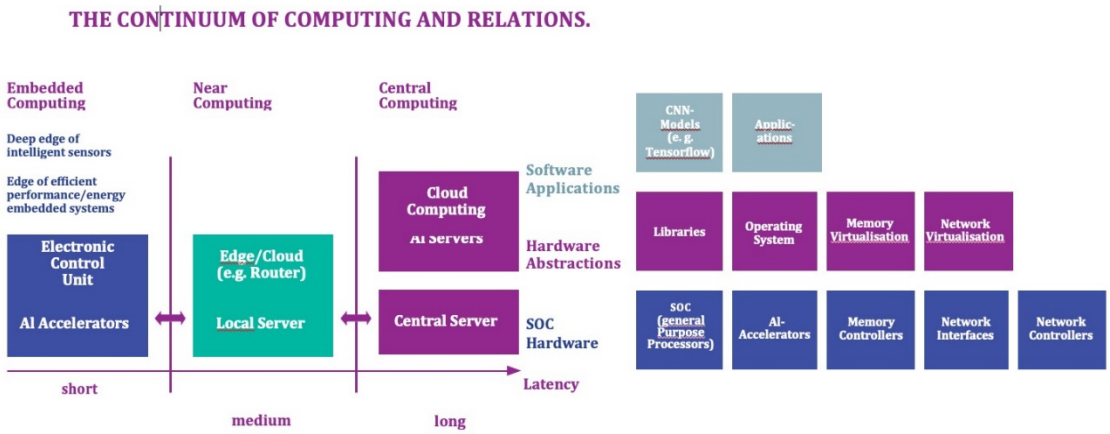


Figure 1 The continuum of computing and relations between the elements constituting an embedded AI system (figure from Gerd Teepe)

The introduction of Artificial Intelligence (AI) at the edge for data analytics brings important benefits for a multitude of applications. New advanced, efficient and specialized processing architectures (based on CPU, embedded GPU, accelerators, neuromorphic computing, FPGA and ASICs) are needed to increase, for several orders of magnitude, the edge computing performances and to drastically reduce the power consumption.

One of the mainstream uses of AI is to allow an easier and better interpretation of the data (unstructured data such as image files, audio files, or environmental data) coming from the physical world. Being able to interpret data from the environment locally triggers new applications such as autonomous vehicles. The use of AI in the edge will contribute to automate complex and advanced tasks, and represents one of the most important innovations being introduced by the digital transformation. Important examples are its contribution in the recovery

1 Multi-access Edge Computing standardization (ETSI/ISG)

from Covid-19 pandemic as well as its potential to ensure the required resilience in future crises[2].
This Chapter focuses on computing components, and more specifically Embedded architectures, Edge Computing devices and systems using Artificial Intelligence at the edge. These elements rely on process technology, embedded software, and have constraints on quality, reliability, safety, and security. They also rely on system composition (systems of systems) and design and tools techniques to fulfil the requirements of the various application domains.

Furthermore, this Chapter focuses on the trade-off between performances and power consumption reduction, and managing complexity (including security, safety and privacy[3]) for Embedded architectures to be used in different applications areas, which will spread Edge computing and Artificial Intelligence use and its contribution to the European sustainability.

1.2 Positioning edge and cloud solutions

The centralized cloud computing model, including data analysis and storage for the increasing number of devices in a network, is limiting the capabilities of many applications, creating problems regarding interoperability, latency and response time, connectivity, privacy and data processing.
Another issue is dependability that creates the risk of a lack of data availability for different applications, a large cost in energy consumption, and the solution's concentration in the hands of a few cloud providers that raise concerns related to data security and privacy.
The increased number of intelligent IoT devices provides new opportunities for enterprise data management, as the applications and services are moving the developments toward the edge and, therefore, from the IoT data generated and processed by enterprises, most of them could be processed at the edge rather than in the traditional centralized data centre in the cloud.

Edge Computing enhances the features and the capabilities (e.g. real-time) of IoT applications, embedded, and mobile processor landscape by performing data analytics through high-performance circuits using AI/ML techniques and embedded security. Edge computing allows the development of real-time applications, considering the processing is performed close to the data source. It can also reduce the amount of transmitted data by transforming an extensive amount of raw data into few insightful data with the benefits of decreasing communication bandwidth and data storage requirements, but also increasing security, privacy data protection, and reducing energy consumption. Moreover, edge computing provides mechanisms for distributing data and computing, making IoT applications more resilient to malicious events. Edge computing can also provide distributed deployment models to address more efficient connectivity and latency, solve bandwidth constraints, provide higher and more "specialized" processing power and storage embedded at the network's edge. Other benefits are scalability, ubiquity, flexibility, and lower cost.
In this Chapter, **Edge Computing** is described as a paradigm that can be implemented using different architectures built to support a distributed infrastructure of data processing (data, image, voice, etc.) as close as possible to the points of collection (data sources) and utilization. In this context, the edge computing distributed paradigm provides computing capabilities to the nodes and devices of the edge of the network (or edge domain) to improve the performance (energy efficiency, latency, etc.), operating cost, reliability of applications and services, and contribute significantly to the sustainability of the digitalization of the European society and

2 https://www.eenewsembedded.com/news/nxp-developing-neural-networks-identify-covid-19
3 Security, safety and privacy will be covered in the Chapter about “Quality, reliability, safety and security”

DRAFT

economy. Edge computing performs data analysis by minimizing the distance between nodes and devices and reducing the dependence on centralized resources that serve them while minimizing network hops. Edge computing capabilities include a consistent operating approach across diverse infrastructures, the ability to perform in a distributed environment, deliver computing services to remote locations, application integration, orchestration. It also adapts service delivery requirements to the hardware performance, and develops AI methods to address applications with low latency and varying data rates requirements – in systems typically subject to hardware limitations and cost constraints, limited or intermittent network connections.
For intelligent embedded systems, the edge computing concept is reflected in the development of edge computing levels (micro, deep, meta, explained in the next paragraphs) that covers the computing and intelligence continuum from the sensors/actuators, processing, units, controllers, gateways, on-premises servers to the interface with multi-access, fog, and cloud computing.

A description of the micro, deep and meta edge concepts is provided in the following paragraphs (as proposed by the AIoT community).
The **micro-edge** describes intelligent sensors, machine vision, and IIoT devices that generate insight data and are implemented using microcontrollers built around processors architectures such as ARM Cortex M4 or recently RISC-V which are focused on minimizing costs and power consumption. The distance from the data source measured by the sensors is minimized. The compute resources process this raw data in line and produce insight data with minimal latency. The hardware devices of the micro-edge physical sensors/actuators generate from raw data insight data and/or actuate based on physical objects by integrating AI-based elements into these devices and running AI-based techniques for inference and self-training.
Intelligent micro-edge allows IoT real-time applications to become ubiquitous and merged into the environment where various IoT devices can sense their environments and react fast and intelligently with an excellent energy-efficient gain. Integrating AI capabilities into IoT devices significantly enhances their functionality, both by introducing entirely new capabilities, and, for example, by replacing accurate algorithmic implementations of complex tasks with AI-based approximations that are better embeddable. Overall, this can improve performance, reduce latency and power consumption, and at the same time increase the devices' usefulness, especially when the full power of these networked devices is harnessed – a trend called AI on edge.

The **deep-edge** comprises intelligent controllers PLCs, SCADA elements, machine vision connected embedded systems, networking equipment, gateways and computing units that aggregate data from the sensors/actuators of the IoT devices that are generating data. Deep edge processing resources are implemented with performant processors and microcontrollers such as Intel i-series, Atom, ARM M7+, etc., including CPUs, GPUs, TPUs, and ASICs. The system architecture, including the deep edge, depends on the envisioned functionality and deployment options considering that these devices' cores are controllers: PLCs, gateways with cognitive capabilities that can acquire, aggregate, understand, react to data, exchange, and distribute information.
The **meta-edge** integrates processing units, typically located on-premises, implemented with high-performance embedded computing units, edge machine vision systems, edge servers (e.g., high-performance CPUs, GPUs, FPGAs, etc.) that are designed to handle compute-intensive tasks, such as processing, data analytics, AI-based functions, networking, and data storage.

This classification is closely related to the distance between the data source and the data processing, impacting overall latency. A high-level rough estimation of the communication latency

DRAFT

and the distance from the data sources are as follows. Micro edge the latency is below 1ms, and the distances from zero to max 15 m, deep edge with distances under 1 km and latency below 2-5 ms, meta edge latencies under 10 ms and distances under 50 km, beyond 50 km the fog computing, MEC concepts are combined with near edge 10-20 ms and 100 km, far edge 20-50ms 500 km and cloud and data centres more than 50 ms and 1000 km.

	Latency	Distance
Micro-edge	Below 1ms	From 0 cm to 15 m
Deep-edge	Below 2-5 ms	Below 1km
Meta-edge	Below 10 ms	Below 50 km
Fog	10-20 ms	Up to 50 km
MEC[4] + near edge	10-20 ms	100 km
Far edge	20-50 ms	200 km
Cloud/data centers/HPC	More than 50 -100 ms	1000 km and beyond

DRAFT

Deployments "at the edge" can contribute, thanks to its flexibility to be adapted to the specific needs, to provide more energy-efficient processing solutions by integrating various types of computing architectures at the edge (e.g., neuromorphic, energy-efficient microcontrollers, AI processing units), reduce data traffic, data storage and the carbon footprint (one way to reduce the energy consumption is to know which data and why it is collected , which targets are achieved and optimize all levels of processes, both at hardware and software levels, to achieve those targets, and finally to evaluate what is consumed to process the data). Furthermore, edge computing reduces the latency and bandwidth constraints of the communication network by processing locally and distributing computing resources, intelligence, and software stacks among the computing network nodes and between the centralized cloud and data centres.
In general, the edge (in the peripheral of a global network as the Internet) includes compute, storage, and networking resources, at different levels as described above, that may be shared by several users and applications using various forms of virtualization and abstraction of the resources, including standard APIs to support interoperability.
More specifically, an edge node covers the edge computing, communication, and data analytics capabilities that make it smart/intelligent. An edge node is built around the computing units (CPUs, GPUs/FPGAs, ASICs platforms, AI accelerators/processing), communication network, storage infrastructure and the applications (workloads) that run on it.
The edge can scale to several nodes, distributed in distinct locations and the location and the identity of the access links is essential. In edge computing, all the nodes can be dynamic. They are physically separated and connected to each other by using wireless/wired connections in topologies such as mesh. The edge nodes can be functioning at remote locations and operates semi-autonomously using remote management administration tools.
The edge nodes are optimized based on the energy, connectivity, size, cost and their computing resources are constrained by these parameters. In different application cases, it is required to provide isolation of edge computing from data centres in the cloud to limit the cloud's domain interference and its impact on edge services.

4 Multi-access Edge Computing (ETSI/ISG)

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

1.3 Positioning Embedded Artificial Intelligence

Thanks to the fast development in Machine Learning during the last decade, Artificial Intelligence is nowadays widely used. However, it demands huge quantity of data, especially for supervised learning using Deep Learning techniques, to get accurate results level. According to the application complexity, neuronal deep learning architectures are becoming more and more complex and demanding in terms of calculations time. As a result, the huge AI success, its perversive deployment and its computing costs, the worldwide energy consumed will be increased dramatically to levels that will be unsustainable in the near future. However, for a similar performances, due to increase of the efficiency of the algorithm and various quantization and pruning techniques, the computing and storage need tends to decrease over time. Complex tasks such as voice recognition which required models of 100 GB in the cloud are now reduced to less than half a gigabyte and can be run on local devices, such as smartphones.

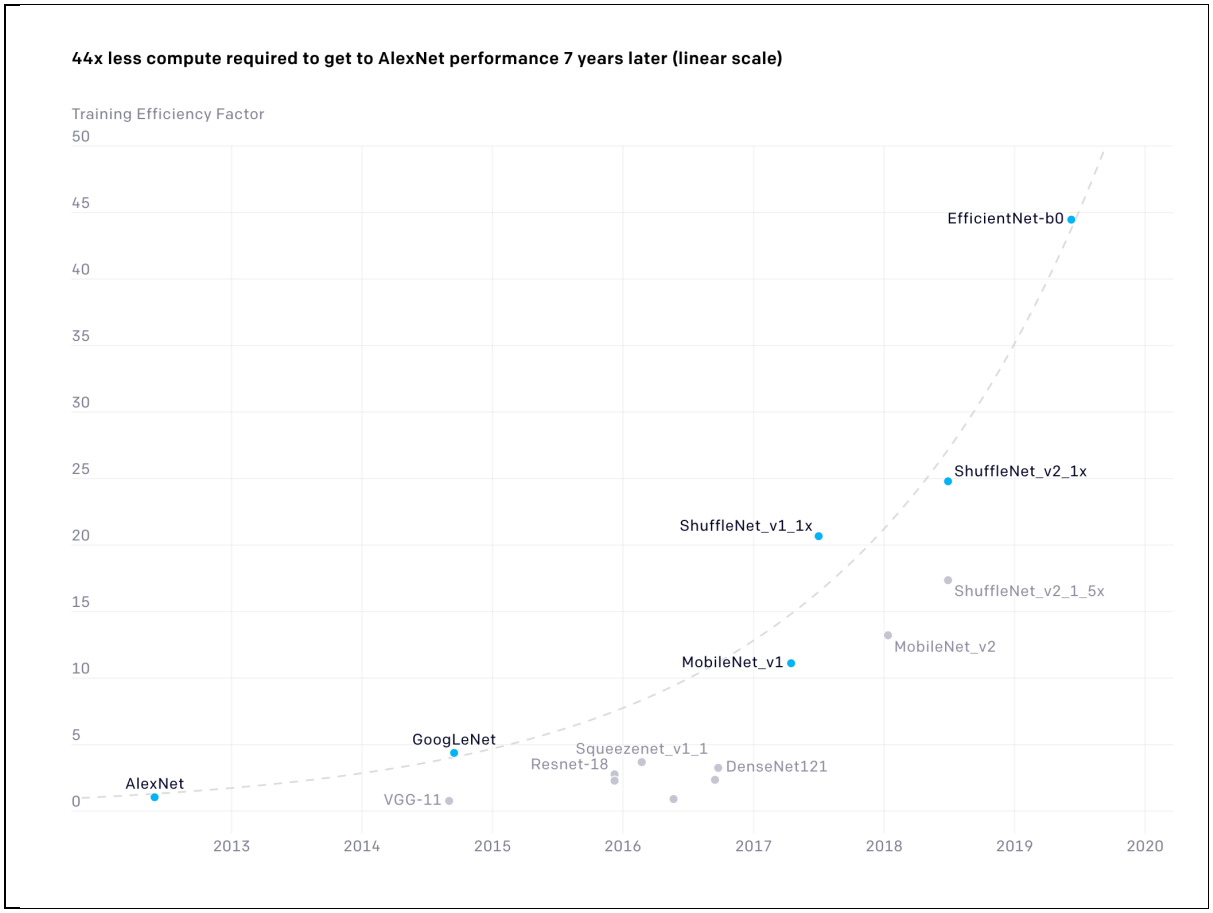


Figure 1: Increase of efficiency used to train to AlexNet level performance (from https://openai.com/blog/ai-and-efficiency/).

Artificial Intelligence is a very efficient tool for several applications (e.g., image recognition and classifications, natural language understanding, complex manufacturing optimization, supply chain improvements, etc.) where pattern detection and process optimization can be done.

As a side effect, data collection is exploding with high heterogeneity levels, coming from numerous and very various sensors. On top, the bandwidth connecting Data Centers is limited and not all data need to be processed in the Cloud.
Naturally, systems are evolving from a centralized to a distributed architecture. Artificial Intelligence is, then, a crucial element that allows soft and optimized operation of distributed systems. Therefore, it is increasingly more embedded in the various network nodes even down to the very edge.
Such powerful tool allows Edge Computing to be more efficient in treating the data locally, while also minimizing the necessary data transmission to the upper network nodes. Another advantage of Embedded Artificial Intelligence is its capacity to self-learn and adapt to the environment through the data collected. Today's learning techniques are still mostly based on supervised learning, but semi-supervised, self-supervised, unsupervised or federative learning techniques are being developed.
At the same time, semiconductor technologies, hardware architectures, algorithms and software are being developed and industrialized to reduce memory size, time for data treatment and energy consumption, thus making Embedded AI an important pillar for Edge Computing. Tools for Embedded AI are also rapidly evolving leading to faster and easier implementation at all levels of the network.

1.4 Scope of the Chapter

DRAFT

The scope of this Chapter is to cover the hardware architectures and their realizations (Systems of Chip, Embedded architectures), mainly for edge and "near the user" devices such as IoT devices, cars, ICT for factories and local processing and servers. Data centres and electronic components for data centres are not the focus of the Chapter, except when the components can be used in local processing units or local servers (local clouds, swarm, fog computing, …). We therefore also cover this "edge" side of the "continuum of computing" and the synergies with the cloud. Hardware for HPC centres is also not the focus, even if the technologies developed for HPC systems are often found in high end embedded systems a few years (decades?) after.
Each Section of this Chapter is split into 2 sub-Sections, from the generic to the more specific:

- Generic technologies for compute, storage and communication (generic Embedded architectures technologies) and technologies that are more focused towards edge computing.
- Technologies focused for devices using Artificial Intelligence techniques (at the edge).

The technological aspects, at system level (PCB, assembly, system architecture, etc.), and embedded and application software are not part of this Chapter as they are covered in other Chapters.

Therefore, this Chapter shall cover mainly the elements foreseen to be used to compose AI or Edge systems:

- **processors** with high energy efficiency,
- **accelerators (for AI** and for other tasks, such as security),
- DPU (Data processing Unit, e.g. logging and collecting information for automotive and other systems) and processing data early (decreasing the load on processors/accelerators),

- **memories and associated controllers,** specialized for low power and/or for processing data locally (e.g. using non-volatile memories such as PCRAM, CBRAM, MRAM for synaptic functions, and In/Near Memory Computing), …
- **power management.**

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

- **Security infrastructure** (e.g. Secure Enclave) with placeholder for customer-specific secure elements (PUF, cryptographic IPs…). Security requirements are dealt with details in the corresponding Chapter.
- **Field connectivity IPs** (see connectivity Chapter, but the focus here is on field connectivity) (all kinds, wired, wireless, optical), ensuring interoperability.
- Integration using chiplet and interposer interfacing units will be detailed in the technology Chapter.
- And all other elements such as coherent cache infrastructure for many-cores, scratchpad memories, smart DMA, NoC with on-chip interfaces at router level to connect cores (coherent), memory (cache or not) and IOs (IO coherent or not), SerDes, high speed peripherals (PCIe controllers and switches, ...), trace and debug hardware and low/medium speed peripherals (I2C, UART, SPI…).

However, the Chapter will not detail the challenges for each of these elements, but only the generic challenges that will be grouped in 1) Edge computing and 2) Embedded Artificial Intelligence domains.
In a nutshell the main recommendation is a paradigm shift towards distributed low power architectures/topologies:

- Distributed computing,
- and AI using distributed computing, leading to distributed intelligence.

1.5 State of the Art

This paragraph gives an overview of the importance that AI and embedded intelligence is playing in the sustainable development, the market perspectives for the AI components and the indication of some semiconductor companies providing components and key IPs.

Impact of AI and embedded intelligence in sustainable development
AI and particularly embedded intelligence, with its ubiquity and its high integration level having the capability "to disappear" in the environment (ambient intelligence), is significantly influencing many aspects of our daily life, our society, the environment, the organizations in which we work, etc. AI is already impacting several heterogeneous and disparate sectors, such as companies' productivity[5], environmental areas like nature resources and biodiversity preservation[6], society in terms gender discrimination and inclusion[7], [8], smarter transportation

5 Acemoglu, D. & Restrepo, P. Artificial Intelligence, Automation, and Work. NBER Working Paper No. 24196 (National Bereau of Economic Research, 2018).
6 Norouzzadeh, M. S. et al. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proc. Natl Acad. Sci. USA 115, E5716–E5725 (2018).
7 Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V. & Kalai, A. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Adv. Neural Inf. Process. Syst. 29, 4349–4357 (2016).
8 Tegmark, M. Life 3.0: Being Human in the Age of Artificial Intelligence (Random House Audio Publishing Group, 2017)

DRAFT

systems[9], etc. just to mention a few examples. The adoption of AI in these sectors is expected to generate both positive and negative effects on the sustainability of AI itself, of the solutions based on AI and on their users[10 11]. It is difficult to extensively assess these effects and there is not, to date, a comprehensive analysis of their impact on sustainability. A recent study[12] has tried to fill this gap, analysing AI from the perspective of 17 Sustainable Development Goals (SDGs) and 169 targets internationally agreed in the 2030 Agenda for Sustainable Development[13]. From the study it emerges that AI can enable the accomplishment of 134 targets, but it may also inhibit 59 targets in the areas of society, education, health care, green energy production, sustainable cities and communities.

From a technological perspective AI sustainability depends, at first instance, on the availability of new hardware[14] and software technologies. From the application perspective, automotive, computing and healthcare are propelling the large demand of AI semiconductor components and, depending on the application domains, of components for embedded intelligence and edge AI. This is well illustrated by car factories on hold because of the current shortage of electronic components. Research and industry organizations are trying to provide new technologies that lead to sustainable solutions redefining traditional processor architectures and memory structure. We already saw that computing near, or in-memory, can lead to parallel and high-efficient processing to ensure sustainability.
The second important component of AI that impacts sustainability concerns software and involves the engineering tools adopted to design and develop AI algorithms, frameworks and applications. The majority of AI software and engineering tools adopts an open-source approach to ensure performance, lower development costs, time to market, more innovative solutions, higher design quality and software engineering sustainability. However, the entire European community should contribute and share the engineering efforts at reducing costs, improving the quality and variety of the results, increasing the security and robustness of the designs, supporting certification, etc .
The report on "Recommendations and roadmap for European sovereignty on open source hardware, software and RISC-V Technologies[15]" details these aspects in more details.

Sustainability through open technologies extends also to open data, rules engines[16] and libraries. The publication of open data and datasets is facilitating the work of researchers and developers for ML and DL, with the existence of numerous images, audio and text databases that are used to train the models and become benchmarks[17]. Reusable open- source libraries[18] allow to solve recurrent development problems, hiding the technical details and simplifying the

9 Adeli, H. & Jiang, X. Intelligent Infrastructure: Neural Networks, Wavelets, and Chaos Theory for Intelligent Transportation Systems and Smart Structures (CRC Press, 2008).
10 Jean, N. et al. Combining satellite imagery and machine learning to predict poverty. Science (80-.) 353, 790–794 (2016).
11 Courtland, R. Bias detectives: the researchers striving to make algorithms fair. Nature 558, 357–360 (2018).
12 Vinuesa, R., Azizpour, H., Leite, I. et al. The role of artificial intelligence in achieving the Sustainable Development Goals. Nat Commun 11, 233 (2020).
13 UN General Assembly (UNGA). A/RES/70/1Transforming our world: the 2030 Agenda for Sustainable Development. Resolut 25, 1–35 (2015).
14 AI is boosting the semiconductor industry with a market of $68.5 billion already by the mid-2020s, according to IHS Markit. The boom of this market is due to the availability of emerging processor architectures for GPUs, FPGAs, ASICs and CPUs that enables applications based on deep learning and vector processing.
15 https://digital-strategy.ec.europa.eu/en/library/recommendations-and-roadmap-european-sovereignty-open-source-hardware-software-and-risc-v
16 E.g. Clips, Drools distributed by red Hat, DTRules by Java, Gandalf on PHP
17 A few examples are ImageNet (14 million images in open data), MNIST or WordNet (English linguistic basis)
18 E.g. Nvidia Rapids, Amazon Comprehend, Google NLU Libraries

DRAFT

access to AI technologies for developers and SMEs, maintaining a high-quality results, reducing time to market and costs.

Eventually, open-source initiatives (being so numerous, heterogeneous, and adopting different technologies) provide a rich set of potential solutions, allowing to select the most sustainable one depending on the vertical application. At the same time, open-source is a strong attractor for applications developers as it gathers their efforts around the same kind of solutions for given use cases, democratizes those solutions and speeds up their development. However, some initiatives should be developed, at European level, to create a common framework to easily develop different types of AI architectures (CNN, ANN, SNN, etc). This initiative should follow the examples of GAMAM (Google, Amazon, Meta, Apple, Microsoft). GAMAM have greatly understood its value and elaborated business models in line with open source, representing a sustainable development approach to support their frameworks[19]. It should be noted that Open-Source hardware should not only cover the processors and accelerators, but also all the required infrastructure IPs to create Embedded architectures and to ensure that all IPs are interoperable and well documented, are delivered with a verification suite and remain maintained constantly to keep up with errata from the field and to incorporate newer requirements. The availability of automated SoC composition solutions, allowing to build Embedded architectures design from IP libraries in a turnkey fashion, is also a desired feature to quickly transform innovation into PoC (Proof of Concept) and to bring productivity gains and shorter time-to-market for industrial projects.

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

It will be a challenge for Europe to excel in this race, but the emergence of AI at the edge, and its know-how in embedded systems, might be winning factors. However, the competition is fierce and the big names are in with big budgets and Europe must act quickly, because US and Chinese companies are already also moving in this "intelligence at the edge" direction (e.g. with Intel Compute Stick, Google's Edge TPU, Nvidia's Jetson Nano and Xavier, and multiples start-ups both in US and China, etc.).

Recently, the attention to the identification of sustainable computing solutions in modern digitalization processes has significantly increased. Climate changes and initiative like the European Green Deal[20] are generating more sensitivity to sustainability topics, highlighting the need to always consider the technology impact on our planet, which has a delicate equilibrium with limited natural resources[21]. The computing approaches available today, as cloud computing, are in the list of the technologies that could potentially lead to unsustainable impacts. A recent study[22] has clearly confirmed the importance that edge computing plays for

19 See e.g. DL networks with Tensorflow at Google, PyTorch / Caffe at Facebook, CNTK at Microsoft, Watson at IBM, DSSTNE at Amazon

20 https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal_en

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

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

DRAFT

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

Market perspectives

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

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

From another side, PWC expects that the market for AI-related semiconductors to grow to more than US$30bn by 2022. The market for semiconductors powering inference systems will likely remain fragmented because potential use cases (e.g. facial recognition, robotics, factory automation, autonomous driving, and surveillance) will require tailored solutions. In comparison, training systems will be primarily based on traditional CPUs, GPUs, FPGAs infrastructures and ASICs.

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

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

AI components vendors

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

Qualcomm has launched the fifth generation Qualcomm AI Engine, which is composed of Qualcomm Kyro Central Processing Unit (CPU), Adreno Graphics Processing Unit (GPU), and Hexagon Tensor Accelerator (HTA). Developers can use either CPU, GPU, or HTA in the AI Engine to carry out their AI workloads. Qualcomm launched also Qualcomm Neural Processing Software Development Kit (SDK) and Hexagon NN Direct to facilitate the quantization and deployment of AI models directly on Hexagon 698 Processor.

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

DRAFT

ARM is developing the new Cortex-M55 core ~~designed~~ for machine learning applications and used in combination with the Ethos-U55 AI accelerator. Both are designed for resource-constrained environments. The new ARM's cores are designed for customized extensions and for ultra-low power machine learning.

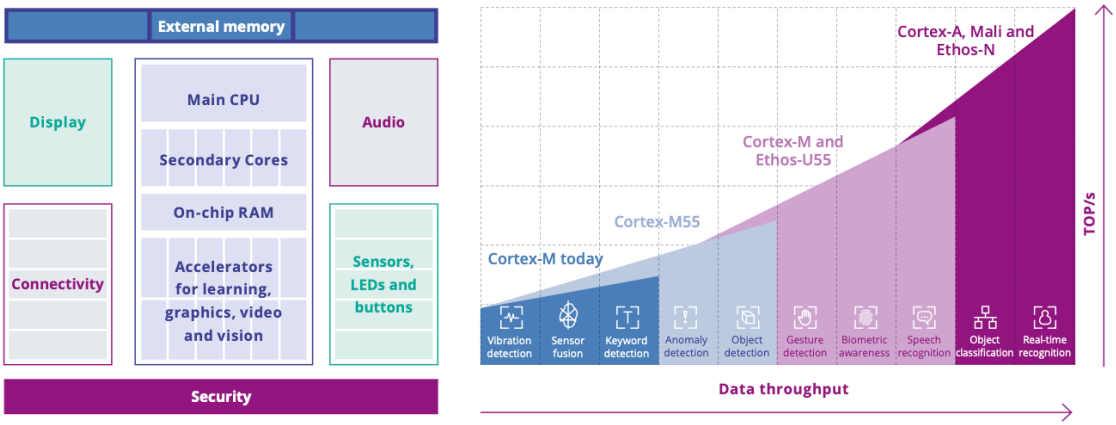


Figure 2 Example of architecture of a modern SoC (from Paolo Azzoni, see also Chapter 1.3) / Arm's Cortex-M55 and Ethos-U55 Tandem. Provide processing power for gesture recognition, biometrics, and speech recognition applications (Source: Arm).

DRAFT

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

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

2 Technology-enabled societal benefits

Driven by Moore's Law over the last 40 years[23], computing and communication brought important benefits to society. Complex computations in the hands of users and hyper-connectivity have been at the source of significant innovations and improvements in productivity, with a significant cost reduction for consumer products at a global level, including products with a high electronic content, traditional products (e.g. medical and machinery products) and added value services.
Computing is at the heart of a wide range of fields by controlling most of the systems with which humans interact. It enables transformational science (Climate, Combustion, Biology,

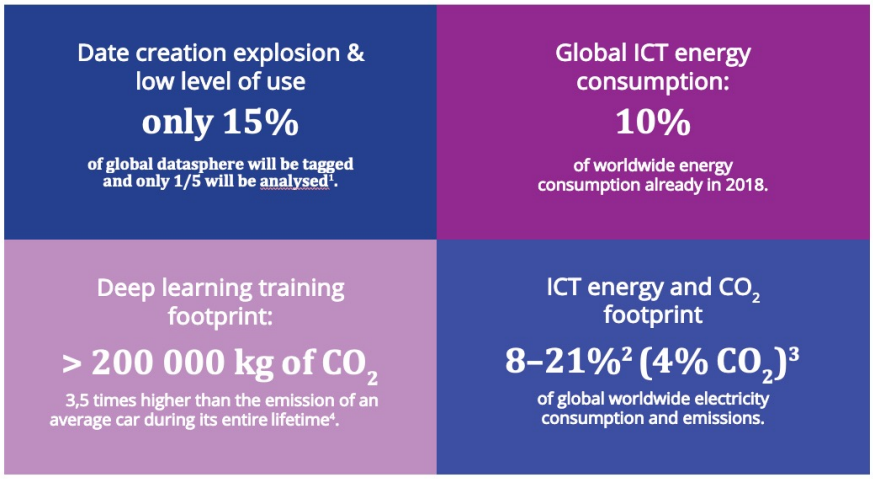
[23] Moore's law is diminishing, however including Ai and accelerator at the edge might increase Moore's law duration, see https://www.synopsys.com/glossary/what-is-sysmoore.html

Astrophysics, etc.), scientific discovery and data analytics. But the advent of Edge Computing and of AI on the edge, enabling complete or partially autonomous cyber-physical systems, requires tremendous improvements in terms of semantics and use case knowledge understanding, and of new computing solutions to manage it. Even if deeply hidden, these computing solutions directly or indirectly impact our ways of life: consider, for example, their key role in solving the societal challenges listed in the application Chapters, in optimizing industrial processes costs, in enabling the creation of cheaper products (e.g. delocalized healthcare).
They will also enable synergies between domains: e.g., self-driving vehicles with higher reliability and predictability will directly benefit medical systems, consumer smart bracelets or smart watches for lifestyle monitoring reduce the impact of health problems[24] with a positive impact on the healthcare system costs, first-aid and insurance services are simplified and more effective thanks to cars location and remote control functionalities.
These computing solutions introduce new security improvements and threats. Edge Computing allows a better protection of personal data, being stored and processed only locally, and this ensures the privacy rights required by GDPR. But at the same time, the easy accessibility to the devices and new techniques, like AI, generates a unique opportunity for hackers to develop new attacks. It is, then, paramount to find interdisciplinary trusted computing solutions and develop appropriate counter measures to protect them in case of attacks. For example, Industry 4.0 and forthcoming Industry 5.0[25] requires new architectures that are more decentralized, new infrastructures and new computational models that satisfy high level of synchronization and cooperation of manufacturing processes, with a demand of resources optimization and determinism that cannot be provided by solutions that rely on "distant" cloud platforms or data centres[26], ensure low-latency data analysis, that are extremely important for industrial application[27].
These computing solutions have also to consider the man in the loop: especially with AI, solutions ensuring a seamless connection between man and machine will be a key factor. Eventually, a key challenge is to keep the environmental impact of these computing solutions under control, to ensure the European industry sustainability and competitiveness.
The following figure illustrates an extract of the challenges and expected market trend of Edge Computing and AI at the edge:

DRAFT

24 https://indianexpress.com/article/technology/gadgets/apple-watch-panic-attack-detection-feature-watchos7-6404470/
25 https://research-and-innovation.ec.europa.eu/research-area/industry/industry-50_en

26 Chen, B., Wan, J., Shu, L., Li, P., Mukherjee, M., Yin, B.: Smart Factory of Industry 4.0: Key Technologies, Application Case, and Challenges. IEEE Access. 6, 6505–6519 (2018).
27 Jeschke, S., Brecher, C., Meisen, T., Özdemir, D., Eschert, T.: Industrial Internet of Things and Cyber Manufacturing Systems. In: Jeschke, S., Brecher, C., Song, H., and Rawat, D.B.(eds.) Industrial Internet of Things. pp. 3–19. Springer International Publishing, Cham (2017).



[1] IDC Data Age 2025 study, sponsored by Seagate, April 2017 | [2] Challenges 2015, 6, 117-157; doi:10.3390/challe6010117, projection from Anders Andrae, https://www.nature.com/articles/d41586-018-06610-y | [3] Internation Energy Agency | [4] https://lejournal.cnrs.fr/articles/numerique-le-grand-gachis-energetique

Figure 3 challenges and expected market evolution.

AI introduces a radical improvement to the intelligence brought to the products through micro-electronics and could unlock a completely new spectrum of applications and business models. The technological progress in microelectronics has increased the complexity of microelectronic circuits by a factor of 1000 over the last 10 years alone, with the integration of billions of transistors on a single microchip. AI is therefore a logical step forward from the actual micro-electronics control units and its introduction will significantly shape and transform all vertical applications in the next decade.

AI and Edge Computing have become core technologies for the digital transformation and to drive a sustainable economy. AI will allow to analyse data on the level of cognitive reasoning to take decisions locally on the edge (embedded artificial intelligence), transforming the Internet of Things (IoT) into the Artificial Intelligence of Things (AIoT). Likewise, control and automation tasks, which are traditionally carried out on centralized computer platforms will be shifted to distributed computing devices, making use of e.g. decentralized control algorithms. Edge computing and embedded intelligence will allow to significantly reduce the energy consumption for data transmissions, will save resources in key domains of Europe's industrial systems, will improve the efficient use of natural resources and will also contribute to improve the sustainability of companies.

DRAFT

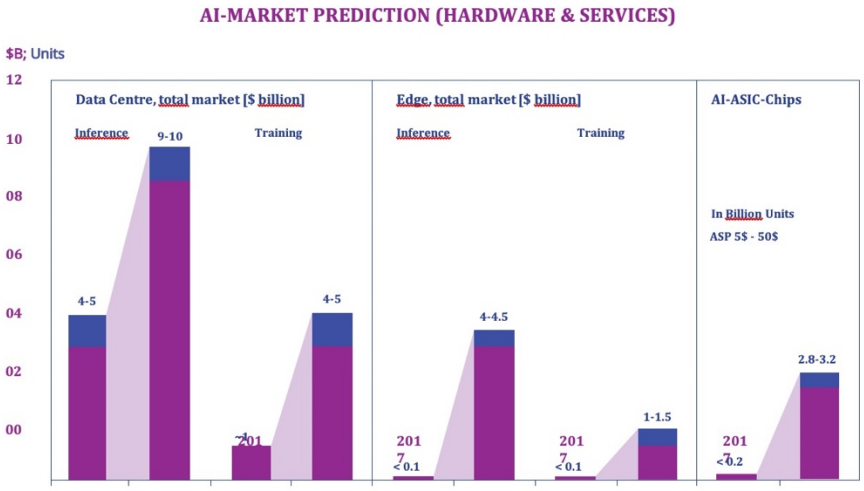


Figure 4 Illustration of an extract of the challenges and the expected market trend for AI and edge computing AI-Market prediction (Hardware & Services) (Source: Tractica, May 2019, McKinsey & Company)

3 Applications breakthroughs

Technologies allowing low power solutions are almost here. What is now key is to integrate these solutions as close as possible to the production of data and sensors.
The key issues to the digital world are the availability of affordable computing resources and transfer of data to the computing node with an acceptable power budget. Computing systems are morphing from classical computers with a screen and a keyboard to smart phones and to deeply embedded systems in the fabric of things. This revolution on how we now interact with machines is mainly due to the advance in AI, more precisely of machine learning (ML) that allows machines to comprehend the world not only on the basis of various signal analysis but also on the level of cognitive sensing (vision and audio). Each computing device should be as efficient as possible and decrease the amount of energy used.
Low-power neural network accelerators will enable sensors to perform online, continuous learning and build complex information models of the world they perceive. Neuromorphic technologies such as spiking neural networks and compute-in-memory architectures are compelling choices to efficiently process and fuse streaming sensory data, especially when combined with event-based sensors. Event based sensors, like the so-called retinomorphic cameras, are becoming extremely important especially in the case of edge computing where energy could be a very limited resource. Major issues for edge systems, and even more for AI-embedded systems, is energy efficiency and energy management. Implementation of intelligent power/energy management policies are key for systems where AI techniques are part of processing sensor data and power management policies are needed to extend the battery life of the entire system.
As extracting useful information should happen on the (extreme) edge device, personal data protection must be achieved by design, and the amount of data traffic towards the cloud and the edge-cloud can be reduced to a minimum. Such intelligent sensors not only recognize low-level features but will be able to form higher level concepts as well as require only very little (or no) training. For example, whereas digital twins currently need to be hand-crafted and built bit-for-bit, so to speak, tomorrow's smart sensor systems will build digital twins autonomously by aggregating the sensory input that flows into them.

To achieve intelligent sensors with online learning capabilities, semiconductor technologies alone will not suffice. Neuroscience and information theory will continue to discover new

DRAFT

ways[28] of transforming sensory data into knowledge. These theoretical frameworks help model the cortical code and will play an important role towards achieving real intelligence at the extreme edge.
AI systems use the training and inference for providing the proper functions of the system, and they have significant differences in terms of computing resources provided by the AI chips. Training is based on past data using datasets that are analysed, and the findings/patterns are built into the AI algorithm. Current hardware used for training needs to provide computation accuracy, support sufficient representation accuracy, e.g., floating-point or fixed-point with long word-length, large memory bandwidth, memory management, synchronization techniques to achieve high computational efficiency and fast write time and memory access to a large amount of data[29]. However, recent research points to increasing training potential for complex CNN models even on constrained edge devices.[30]

Reinforcement learning (RL) is a booming area of machine learning and is based on how agents ought to take actions in an environment in order to maximize the notion of cumulative reward. Recent work[31] develops systems that were able to discover their own reward function from scratch. Similarly, Auto-ML allows to determine a "good" structure for a DL system to be efficient in a task. But all those approaches are also very compute demanding.

New deep learning models are introduced at an increasing rate and one of the recent one, with large applications potential, are transformers. Based on the attention model[32] , it is a "sequence-to-sequence architecture" that transforms a given sequence of elements into another sequence. Initially used for NLP (Natural Language Processing) where it can translate one sequence in a first language into another one, or complement the beginning of a text with potential follow-up, it is now extended to other domains such a video processing. It is also a self-supervised approach: for learning it don't need labelled examples, but only part of the sequence, the remaining part being the "ground truth". The biggest models, such a GPT3 are based on this architecture. GPT3 was in the highlight in may 2020 because of its potential use in many different application (the context being given by the beginning sequence) such as generating new text, summarizing text, translating language, answering to questions and even generating code from specifications. Even if today transformers are mainly used for cloud applications, this kind of architecture will certainly ripple down in embedded systems in the future. For example, Meta OPT, released in may 2022, has 1/7 of the CO_2 footprint of GPT3 with similar performances. The new
GPUs of Nvidia, supports float8 in order to efficiently implement transformers.

The inference is the application of the learned algorithm to the real devices to solve specific problems based on present data. The AI hardware used for inference needs to provide high speed, energy efficiency, low cost, fixed-point representation, efficient reading memory access and efficient network interfaces for the whole hardware architecture. The development of AI-based devices with increased performance, and energy efficiency allows the AI inference "at

28 Even though our understanding of how the brain computes is still in its infancy, important breakthroughs in cortical (column) theory have been achieved in the last decade.
29 GPT-3 175B from OpenAI is trained with 499 Billion tokens (https://lambdalabs.com/blog/demystifying-gpt-3/) and required 3.14E23 FLOPS of computing for training.
30) https://proceedings.mlr.press/v162/patil22b.html
31 https://arxiv.org/pdf/2007.08794.pdf
32 https://arxiv.org/abs/1706.03762

DRAFT

the edge" (embedded intelligence) and accelerates the development of middleware allowing a broader range of applications to run seamlessly on a wider variety of AI-based circuits. Companies like Google, Gyrfalcon, Mythic, NXP, STMicroelectronics and Syntiant are developing custom silicon for the edge. As an example, Google was releasing Edge TPU, a custom processor to run TensorFlow Lite models on edge devices. NVidia is releasing the Jetson Orin Nano range of products, allowing to perform up to 40 TOPS of sparce neural networks within a 15W power range[33].

The Tiny ML community (https://www.tinyml.org/) is bringing Deep Learning to microcontrollers with limited resources and at ultra-low energy budget. The MLPerf allows to benchmark devices on similar applications (https://github.com/mlcommons/tiny), because it is nearly impossible to compare performances on figures given by chips providers.

In summary we see the following disruptions on the horizon, once embedded AI enters the application space broadly:

- Various processing, especially concerning AI functionalities, are moved to local devices, such as voice and environment recognition, allowing privacy preserving functionalities.
- The latent intelligence of things will be enabled by AI.
- Federated functionalities will emerge (increasing the functionality of a device by using capabilities, resources, or neighbouring devices).
- Connected functionalities will also show up: this will extend the control and automation of a single system (e.g. a truck, a car) to a network of systems (e.g. a truck platoon), resulting in networked control of a cyber-physical system. The benefit of this is generally better performance and safety. It will also set the foundation for autonomous machines (including vehicles).
- The detection of events by camera and other long-range sensors (radar, lidar, etc.) is coming into action. Retina sensors will ensure low power operation of the system. Portable devices for blind people will be developed.
- The possibilities for disabled people to move their arms and legs comes into reach, as AI-conditioned sensors will directly be connected to the brain.
- The use of conversational interfaces will be drastically increased, improving the human machine interface with reliable understanding of natural language.

DRAFT

4 Strategic advantage for EU

Edge computing and Embedded Artificial Intelligence are key enablers for the future, and Europe should act quickly to play a global role and have a certain level of control of the assets we use in Europe. Further development of AI can be a strategic advantage for Europe, but we are not in a leading position.

Already today AI is being used as a strategic competitive advantage. Tesla is the first car company which is marketing a driving-assistance-system as “auto-pilot”. Although it is not qualified to operate without human intervention, it is a significant step forward towards autonomous

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

driving. Behind this feature is one of the strongest AI-processors, which can be found in driver assistance systems. However, the chips employed are not freely available on the market, but are exclusive for Tesla and they are developed internally now to train their self-learning capabilities. This example shows clearly the importance of system ownership in AI, which must be secured for Europe, if its companies want to be able to sell competitive products when AI is becoming pervasive.

In this context, Europe must secure the knowledge to build AI-systems, design AI-chips, procure the AI-software ecosystem, and master the integration task into its products, and particularly into those products where Europe has a lead today.

Adapted to the European industry structure, which is marked by a vibrant and versatile ecosystem of SMEs together with larger firms, we need to build and enhance the AI-ecosystem for the particular strengths but also weaknesses of Europe.

A potential approach could be to:

- To rely on existing application domains where we are strong (e.g. automotive, machinery, chemistry, energy, …).
- Promoting to keep, catch-up and get all expertise in Europe that are required to build competitive Edge Computing systems and Embedded Intelligence, allowing us to develop solutions that are adapted to the European market and beyond. All the knowledge is already present in Europe, but not structured and focused and often the target of non-European companies. The European ecosystem is rich and composed of many SMEs, but with little focus on common goals and cooperation.
- Open-source Hardware can be an enabler or facilitator of this evolution, allowing this swarm of SMEs to develop solutions more adapted to the diversity of the market.
- Data-based and knowledge-based modelling combined into hybrid modelling is an important enabler.
- Particular advantage will be cross-domain and cross-technology cooperation between various European vendors combining the best hardware and software know-how and technologies.
- Cooperation along and across value chains for both hardware and software experts will be crucial in the field of smart systems and the AI and IoT community.

DRAFT

While Europe is recognized for its know-how in embedded systems architecture and software, it should continue to invest in this domain to remain at the state of the art, despite fierce competition from countries like USA, China, India, etc. From this perspective, the convergence between AI and Edge Computing, what we call *embedded intelligence*, should be a top priority. Europe should take benefit of its specificities, such as the drive of the "European Green Deal" to make its industry sustainable AND competitive.

European companies are also in the lead for embedded microcontrollers. Automotive, IoT, medical applications and all embedded systems utilize many low-cost microcontrollers, integrating a complete system, computing, memory, and various peripherals in a single die. Here, pro-active innovation is necessary to upgrade the existing systems with the new possibilities from AI, Cyber-Physical Systems and Edge computing, with a focus on local AI. Those new applications will require more processing power to remain competitive. In addition, old applications will require AI-components to remain competitive. But power dissipation must not increase accordingly, in fact a reduction would be required. Europe has lost some ground in the processor domain, but AI is also an opportunity to regain parts of its sovereignty in the domain of computing, as completely new applications emerge. Mastering key technologies for the

future is mandatory to enforce Europe, and for example, to attract young talents and to enables innovations for the applications.

Europe no longer has a presence in "classical" computing such as processors for laptops and desktop, servers (cloud) and HPC, but the drive towards Edge Computing, part of a computing continuum, might be an opportunity to use the solid know how in embedded systems and extend it with high performance technology to create Embedded (or Edge) High Performance Computers (eHPC) that can be used in European meta-edge devices. The initiative of the European Commission, "for the design and development of European low-power processors and related technologies for extreme-scale, high-performance big-data and emerging applications, in the automotive sector" could reactivate an active presence of Europe in that field and has led to the launch of the "European Processor Initiative – EPI ". New initiatives around RISC-V and Open source hardware are also key ingredients to keep Europe in the race.

AI optimized hardware components such as CPUs, GPUs, DPUs, FPGAs, ASICs accelerators and neuromorphic processors are becoming more and more important. European solutions exist, and the knowledge on how to build AI-systems is available mainly in academia. However, more EU action is needed to bring this knowledge into real products in view to enhance the European industry with its strong incumbent products. Focused action is required to extend the technological capabilities and to secure Europe's industrial competitiveness. A promising approach to prevent the dependence on closed processing technologies, relies on Open Hardware initiatives (Open Compute Project, RISC-V, OpenCores, OpenCAPI, etc.). The adoption of an open ecosystem approach, with a globally and incrementally built know-how by multiple actors, prevents that a single entity can monopolize the market or cease to exist for other reasons. The very low up-front cost of open hardware/silicon IP lowers the barrier of innovation for small players to create, customize, integrate, or improve Open IP to their specific needs. Thanks to Open Hardware freely shared, and to existing manufacturing capabilities that still exist in Europe, prototyping facilities and the related know-how, a new wave of European start-ups could come to existence, building on top of existing designs and creating significant value by adding the customization needed for industries such as automotive, energy, manufacturing or health/medical. Another advantage of Open-Source hardware is that the source code is auditable and therefore inspected to ensure quality (and less prone to attack if correctly analysed and corrected).

In a world, in which some countries are more and more protectionist, not having high-end processing capabilities, (i.e. relying on buying them from countries out of Europe) might become a weakness (leaving for example the learning/training capabilities of AI systems to foreign companies/countries). China, Japan, India and Russia are starting to develop their own processing capabilities in order to prevent potential shortage or political embargo.

It is also very important for Europe to master the new key technologies for the future, such as AI, the drive for more local computing, not only because it will allow to sustain the industry, but also master the complete ecosystem of education, job creation and attraction of young talents into this field while implementing rapidly new measures as presented in Major Challenge 4.

5 Major challenges

DRAFT

5.1 For Edge Computing

Four Major Challenges have been identified for the further development of computing systems, especially in the field of embedded architectures and Edge Computing:

1 Increasing the energy efficiency of computing systems:
   1.a Processing data where it is created.
   1.b Co-design: algorithms, HW, SW and topologies.
2 Managing the increasing complexity of systems:
   2.a Balanced mechanisms between performance and interoperability.
   2.b Realizing self-X, self-optimize, reconfiguration and self-management.
   2.c Using AI techniques to help in complexity management.
3 Supporting the increasing lifespan of devices and systems:
   3.a HW supporting software upgradability.
   3.b Improving interoperability (with the same class of application) and between classes, modularity and complementarity between generations of devices.
   3.c Developing the concept of 2nd life for components.
   3.d Implementation on the smallest devices, high quality data, meta-learning, neuromorphic computing and other novel hardware-architectures.
4 Ensuring European sustainability in Embedded architectures design:
   4.a Open-source HW.
   4.b Energy efficiency improvement.
   4.c Engineering support to improve sustainable AI, edge computing and Embedded architectures.

DRAFT

5.2 For Embedded Intelligence

The world is more and more connected. Data collection is exploding. Heterogeneity of data and solutions, needs of flexibility in calculation between basic sensors and multiple sensors with data fusion, protection of data and systems, extreme variety of use cases with different data format, connectivity, bandwidth, real time or not, etc … increase the complexity of systems and their interactions. This leads to systems of systems solutions, distributed between deep edge to cloud and possibly creating a continuum in this connected world.

Ultimately, energy efficiency becomes the key criteria as digital world is taking a more and more significant percentage of produced electricity.

Embedded Intelligence is then foreseen as a crucial element to allow a soft and optimized operation of distributed systems. It is a powerful tool to achieve goals such as:

- Power energy efficiency by treating data locally and minimizing the necessary data sent to the upper node of network.
- Securing the data (including privacy) keeping them local.
- Allowing different systems to communicate to each other and adapt over the time (increasing their lifetime).
- Increasing resilience by learning and becoming more secured, more reliable.
- Keeping systems always on and accessible towards a network continuum.

On top, Embedded Intelligence can be installed to all levels of the chain. However, many challenges have to be solved to achieve those goals.

First priority is the energy efficiency. The balance between Embedded AI energy consumption and overall energy savings must be carefully reviewed. New innovative architectures and technologies (Near-Memory-Computing, In-Memory-Computing, Neuromorphic, …) need to be developed as well as sparsity of coding and of the algorithm topology (e.g. for Deep Neural Network). It also means to carefully choose which data is collected and for which purposes. Avoiding data transfers is also key for low power: Neural Networks, where storage (the synaptic weights) and computing (the neurons) are closely coupled lead to architectures which may differ from the Von Neumann model where storage and computation are clearly separated. Computing In or Near memory are efficient potential architectures for some AI algorithms.
Second, Embedded AI must be scalable and modular all along the distributed chain, increasing flexibility, resilience and compatibility. Stability between systems must be achieved and tested. Thus, benchmark and validation tools for Embedded AI and related techniques have to be developed.
Third, self-learning techniques (Federative learning, unsupervised learning, ...) will be necessary for fast and automatic adaptation.
Last but not least, trust in AI is key for societal acceptance. Explainability and Interpretability of AI decisions for critical systems is an important factor for AI adoption, together with certifications processes.

Algorithms for Artificial Intelligence can be realized in stand-alone, distributed (federated, swarm, …) or centralized solution (of course, not all algorithms can be efficiently implemented in the 3 solutions). For energy, privacy and all the reasons explained above, it is preferable to have stand-alone or distributed solutions (hence the name "Intelligence at the edge"). The short term might be more oriented towards stand-alone AI (e.g. self-driving car) and then distributed (or connected, like car2car or car2infrastructure).

Summarizing, four Major Challenges have been identified:

- Increasing the energy efficiency:
    - Development of innovative (and heterogeneous) hardware architectures: e.g. Neuromorphic.
    - Avoiding to move large quantities of data: processing at the source of data, sparse data coding, etc. Only processing when it is required (sparse topology, algorithms, etc.).
    - Interoperability (with the same class of application) and between classes
    - Scalable and Modular AI.
- Managing the increasing complexity of systems:
    - Development of trustable AI.
    - Easy adaptation of models.
    - Standardized APIs for hardware and software tool chains, and common descriptions to describe the hardware capabilities.
- Supporting the increasing lifespan of devices and systems:
    - Realizing self-X (unsupervised learning, transfer learning, etc.).
    - Update mechanisms (adaptation, learning, etc.).
- Ensuring European sustainability in AI:

DRAFT

- Developing solutions that correspond to European needs and ethical principles.
- Transforming European innovations into commercial successes.
- Cultivating diverse skillsets and expertise to address all parts of the European embedded AI ecosystem

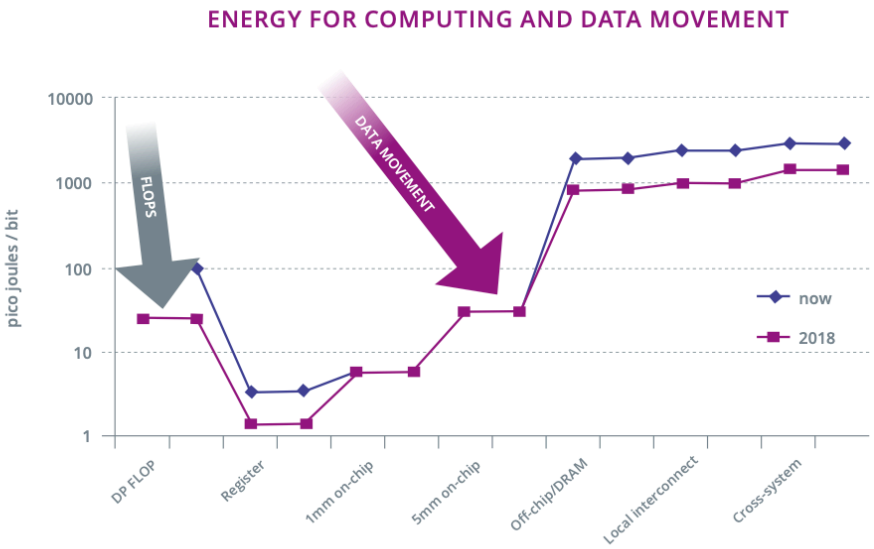
Of course, as seen above, all the generic challenges found in Embedded architectures are also important for Embedded AI based systems, but we will describe more precisely which is specific for each subsection (Embedded architectures/Edge computing and Embedded Intelligence).

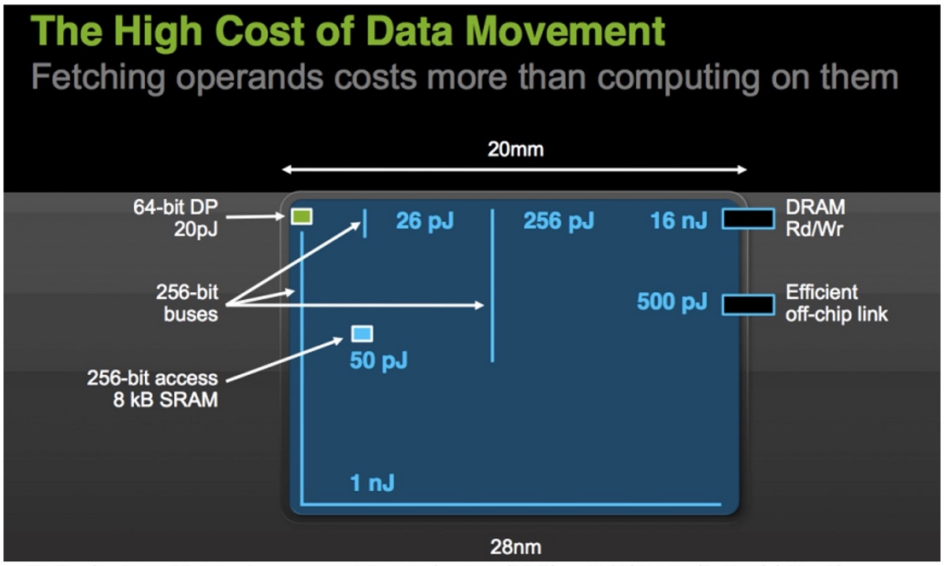
5.3 Major Challenge 1: increasing the energy efficiency of computing systems

5.3.1 For Edge Computing

The advantages of using digital systems should not be hampered by their cost in terms of energy. For HPC or data centres, it is clear that the main challenge is not only to reach the "exaflops", but to reach "exaflops" at reasonable energy cost, which impacts the cooling infrastructure, the size of the "power plug" and globally the cost of ownership. At the other extremity of the spectrum, micro-edge devices should work for months on a small battery, or even by scavenging their energy from the environment (energy harvesting). Reducing the energy footprint of devices is the main charter for fulfilling sustainability and the "European Green deal". Multimode energy harvesting (e.g. solar/wind, regenerative braking, dampers/shock absorbers, thermoelectric, etc.) offers huge potential for electrical vehicles and other battery-, fuel cells -operated vehicles in addition to energy efficiency design, real-time sensing of integrity, energy storage and other functions.
Power consumption should not be only seen just at the level of the device, but at the level of the aggregation of functions that are required to fulfil a task.
The new semiconductor technology nodes don't really bring improvement on the power per device, Dennard's scaling is ending and going to a smaller node does not anymore lead to a large increase of the operating frequency or a decrease of the operating voltage. Therefore, dissipated energy per surface, the power density of devices is increasing rather than decreasing. Transistor architectures, such as FinFet, FDSOI, GAA, nanosheets mainly reduce the leakage current (i.e. the energy spent by an inactive device). However, transistors made on FDSOI substrates achieve the same performance than FinFet transistors at a lower operating voltage, reducing dynamic power consumption.
In addition, comes the memory wall. Today's limitation is not coming from the pure processing power of systems but more from the capacity to bring data to the computing nodes within a reasonable power budget fast enough.

DRAFT





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

Figure 5 Energy for compute and data movement[34] This explain the order of magnitude of the problem of data movement, and this problem is still relevant in all technology nodes

[35]

Furthermore, the system memory is only part of a broader Data Movement challenge which requires significant progress in the data access/storage hierarchy from registers, main memory (e.g. progress of NVM technology, such as the Intel's 3D-xpoint, etc.), to external mass storage devices (e.g. progress in 3D-nand flash, SCM derived from NVM, etc.). In a modern system, large parts of the energy is dissipated in moving data from one place to another. For this reason, new architectures are required, such as **computing in or near memory**, neuromorphic architectures (also where the physics of the NVM - PCM, CBRAM, MRAM, OXRAM, ReRAM, FeFET, etc. - technology can be used to compute) and lower bit count processing are of primary importance.

[34] Source: S. Horst, Optical Interconnect Conference, 2013
[35] Source: S. Horst, Optical Interconnect Conference, 2013

Power consumption can be reduced by local treatment of collected data, not only at circuit level, but also at system level or at least at the nearest from the sensors in the chain of data transfer towards the data centre (for example: in the gateway). Whereas the traditional approach was to have sensors generate as much data as possible and then leave the interpretation and action to a central unit, future sensors will evolve from mere data-generating devices to devices that generate semantic information at the appropriate conceptual level. This will obviate the need for high bit rates and thus power consumption between the sensors and the central unit. In summary, raw data should be transformed into relevant information (what is really useful) as early as possible in the processing continuum to improve the global energy efficiency:

- Only end or middle points equipment are working, potentially with low or sleeping consumption modes.
- Data transfer through network infrastructures is reduced. Only necessary data is sent to the upper level.
- Usage of computing time in data centres is also minimized.
- The development of benchmarks and standardization for HW/SW and data sets could be an appropriate measure to reduce power consumption. Hence, energy consumption evaluation will be easy and include the complete view from micro-edge to cloud.

5.3.1.1 Key focus areas

DRAFT

To increase the energy efficiency of computing systems, especially in the field of systems for AI and Edge Computing requires the development of innovative hardware architectures at all levels with their associated software architectures and algorithms:

- At technology level (FinFet, FDSOI, silicon nanowires or nanosheets), technologies are pushing the limits to be Ultra-low power. On top, advanced architectures are moving from Near-Memory computing to In-Memory computing with potential gains of 10 to 100 times. Technologies related to advanced integration and packaging have also recently emerged (2.5D, chiplets, active interposers, ...) that open up innovative design possibilities, particularly for what concerns tighter sensor-compute and memory-compute integration.
- At device level, several type of circuit architectures are currently running, tested, or developed worldwide. The list is moving from the well-known CPU to some more and more dedicated accelerators integrated in Embedded architectures (GPU, DPU, TPU, NPU, DPU, …) providing accelerated data processing and management capabilities, which are implemented very variously going from fully digital to mixed or full analog solutions:
    - Fully digital solutions have addressed the needs of emerging application loads such as AI/DL workloads using a combination of parallel computing (e.g., SMP and GPU) and accelerated hardware primitives (such as systolic arrays), often combined in heterogeneous Embedded architectures. Low-bit-precision (8 bit integer or less) computation as well as sparsity-aware acceleration have been shown as effective strategies to minimize the energy consumption per each elementary operation in regular AI/DL inference workloads; on the other hand, there remain many challenges in terms of hardware capable of opportunistically exploiting the characteristics of more irregular mixed-precision networks. Applications, including AI/DL also require further development due to their need for more flexibility and precision in numerical

representation (32 or 16 bit floating point), which puts a limit to the amount of hardware efficiency that can be achieved on the compute side.
    - Avoiding to move data: this is crucial because the access energy of any off-chip memory is currently 10-100x more expensive than access to on-chip memory. Emerging non-volatile memory technologies such as MRAM, with asymmetric read/write energy cost, could provide a potential solution to relieve this issue, by means of their greater density at the same technology node. Near-Memory Computing (NMC) and In-Memory Computing (IMC) techniques move part of the computation near or inside memory, respectively, further offsetting this problem. While IMC in particular is extremely promising, careful optimization at the system level is required to really take advantage of the theoretical peak efficiency potential.
    - Another way is also to perform invariant perceptive processing and produce semantic representation with any type of sensory inputs.
- At system level, micro-edge computing near sensors (i.e integrating processing inside or very close to the sensors or into local control) will allow Embedded architectures to operate in the range of 10 mW to 100 mW with an estimated energy efficiency in the order of 100s of GOPs/Watt up to a few TOPs/Watt in the next 5 years. This could be negligible compared to the consumption of the sensor (for example, a Mems microphone can consume a few mA). On top, the device itself can go in standby or in sleep mode when not used, and the connectivity must not be permanent. Devices currently deployed on the edge rarely process data 24/7 like data centres: to minimize global energy, a key requirement for future edge Embedded architectures is to combine high performance "nominal" operating modes with lower-voltage high compute efficiency modes and, most importantly, with ultra-low-power sleep states, consuming well below 1 mW in fully state-retentive sleep, and less than 1-10 uW in deep sleep. The possibility to leave Embedded architectures in an ultra-low power state for most of the time has a significant impact on the global energy consumed. The possibility to orchestrate and manage edge devices becomes fundamental from this perspective and should be supported by design. On the contrary, data servers are currently always on even if they are loaded only at 60% of their computing capability.
- At data level, memory hierarchies will have to be designed taking into account the data reuse characteristics and access patterns of algorithms, which strongly impact load and store access rate and hence, the energy necessary to access each memory in the hierarchy. For example (but not only), weights and activations in a Deep Neural Network have very different access patterns and can be deployed to entirely separate hierarchies exploiting different combinations of external Flash, DRAM, non-volatile on-chip memory (MRAM, FRAM…) and SRAM.
- At tools level, HW/SW co-design of system and their associated algorithms are mandatory to minimize the data moves and optimally exploit hardware resources, particularly if accelerators are available, and thus optimize the power consumption.

The challenge is not only at the component level, but also at the system and even infrastructure level: for example, the Open Compute Project was started by Facebook with the idea of delivering the most efficient designs for scalable computing through an open-source hardware community.

DRAFT

5.3.2 For Embedded Intelligence

Training AI models can be very energy-demanding. As an example, according to a recent study[36], the model training process for natural-language processing (NLP, that is, the sub-field of AI focused on teaching machines to handle human language) could end emitting as much carbon as five cars in their lifetimes[37]. However, if the inference of that trained model is executed billions of times (e.g., by billion users' smartphones), its carbon footprint could even offset the training one. Another analysis[38], published by the OpenAI association, unveils a dangerous trend: "since 2012, the amount of compute used in the largest AI training runs has been increasing exponentially with a 3.5 month-doubling time (by comparison, Moore's law had a 2-years doubling period)". These studies reveal that the need for computing power (and associated power consumption) for training AI models is dramatically widening. Consequently, the AI training processes need to turn greener and more energy-efficient.

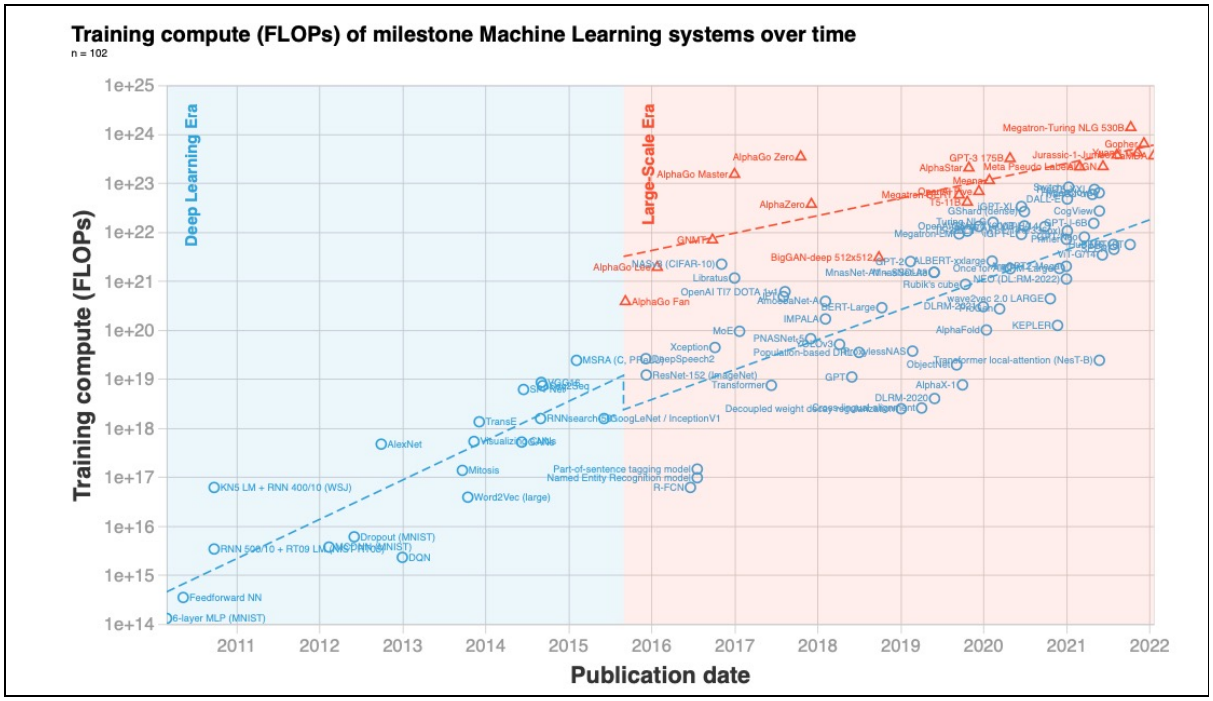


Figure 6 Evolution of the size of the most advanced deep learning networks (from https://arxiv.org/abs/2202.05924)

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

In addition, novel learning paradigms such as transfer learning, federated learning, self-supervised learning, online/continual/incremental learning, local and context adaptation, etc., should

36 https://arxiv.org/pdf/1906.02243.pdf
37 https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/
38 https://openai.com/blog/ai-and-compute/

be preferred not only to increase the effectiveness of the inference models but also as an attempt to decrease the energy cost of the learning scheme. Indeed, these schemes avoid retraining models from scratch all the times or reduce the number and size of the model parameters to transmit back and forth during the distributed training phase.

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

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

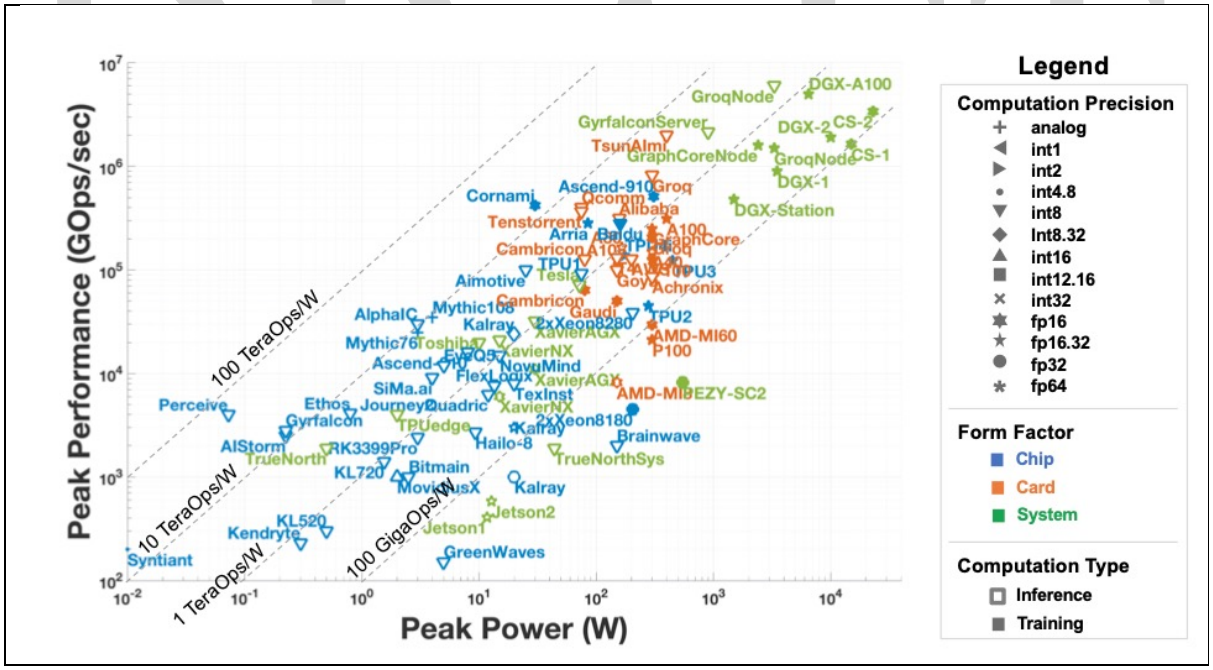


Figure 7 Landscape of AI chips according to their peak power consumption and peak performance[39].

To efficiently support new AI-related applications, for both, the server and the client on the edge side, new accelerators need to be developed. For example, DL does not usually need a 32/64/128-bit floating point for its learning phase, but rather variable precision including dedicated formats such as bfloats. However, a close connection between the compute and storage parts are required (Neural Networks are an ideal "compute in memory" approach). Storage also

39 Source: **AI Accelerator Survey and Trends, Albert Reuther, Peter Michaleas, Michael Jones, Vijay Gadepally, Siddharth Samsi, Jeremy Kepner, 2021 https://arxiv.org/abs/2109.08957**

needs to be adapted to support AI requirements (specific data accesses, co-location compute and storage), memory hierarchy, local vs. cloud storage.

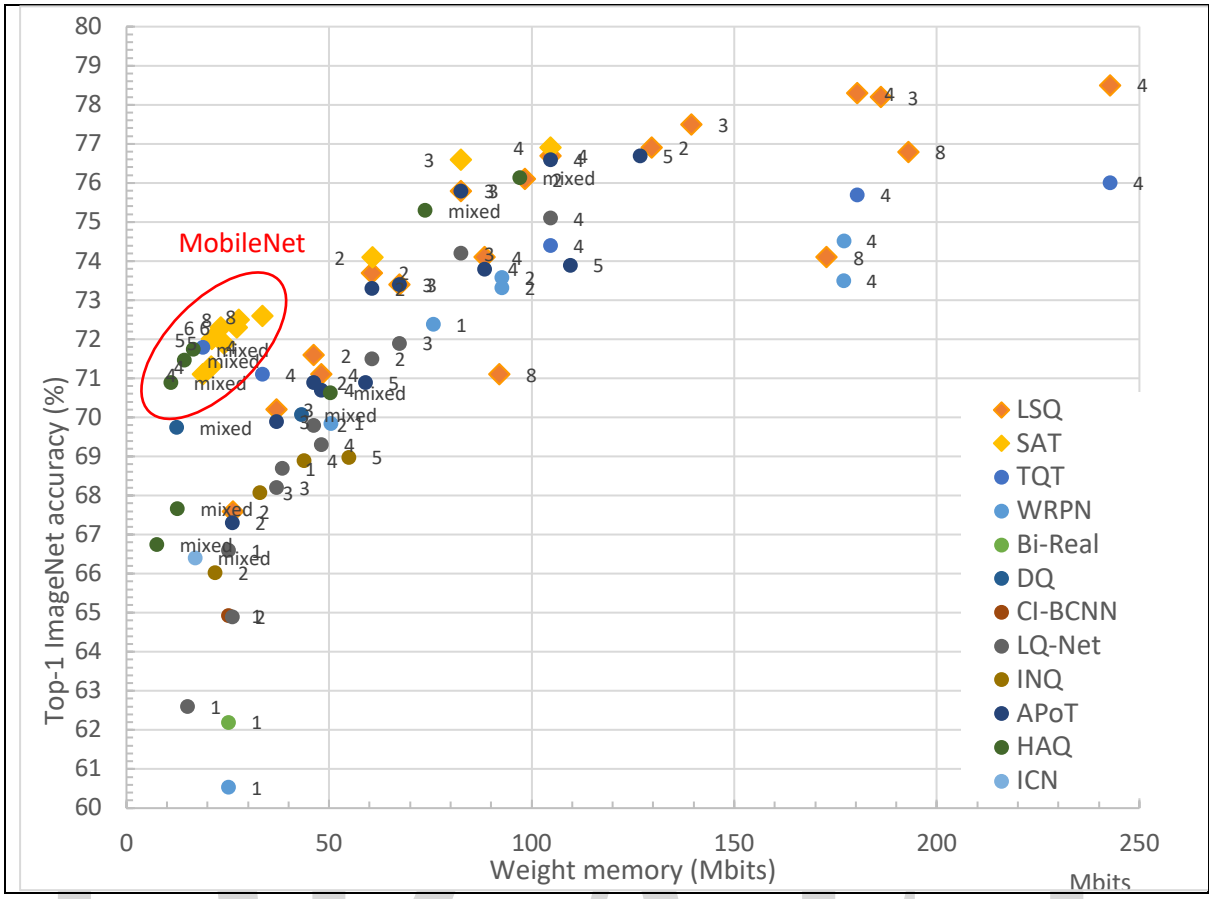
Similarly, at the edge side, accelerators for AI applications will particularly require real-time inference, in view to reduce the power consumption. For DL applications, arithmetic operations are simple (mainly multiply-accumulate) but they are done on data sets with a very large set of data and the data access is therefore challenging. In addition, clever data processing schemes are required to reuse data in the case of convolutional neural networks or in systems with shared weights. Computing and storage are deeply intertwined. And of course, all the accelerators should fit efficiently with more conventional systems.

Reducing the size of the neural networks and the precisions of computation is key to allow complex deep neural networks to run on embedded devices. This can be achieved either by pruning the topology of the networks, and/or by reducing the number of bits storing values of weight and neuron values. These processes can be done during the learning phase, or just after a full precision learning phase, or can be done (with less performances) independently of the learning phase (example: post-training quantization). The pruning principle is to eliminate nodes that have a low contribution to the final result. Quantization consists either in decreasing the precision of the representation (from float 32 to float 16 or even float8, as supported by the Nvidia GPUs mainly for Transformers networks), or to change the representation from float to integers. For the inference phase, current techniques allow to use 8-bit representations with a minimal loss of performances, and sometimes to reduce further the number of bits, with an acceptable reduction of performance or small increase of the size of the network. Most major developments environments (TensorFlow Lite[41], N2D2[42], etc) support post-training quantization, and the Tiny ML community is actively using it. Supporting better tools and algorithms to reduce size and computational complexity of Deep Neural Networks is of paramount importance for allowing efficient AI applications to be executed at the edge.

DRAFT

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

[42] https://github.com/CEA-LIST/N2D2



Results of various quantization methods versus Top-1 ImageNet accuracy

DRAFT

Finally, new approaches can be used for computing Neural-Networks, such as analogue computing, or using the properties of specific materials to perform the computations (although with low precision and high dispersion, but the Neural Networks approach is able to cope with these limitations).

Besides DL, the "Human Brain Project", a H2020 FET Flagship Project which targets the fields of neuroscience, computing, and brain-related medicine, including, in its SP9, the Neuromorphic Computing platform SpiNNaker and BrainScaleS. This Platform enable experiments with configurable neuromorphic computing systems.

5.3.2.1 Key focus areas

The focus areas rely on Europe maintaining a leadership role in embedded systems, CPS, components for the edge (e.g., sensors, actuators, embedded microcontrollers), and applications in automotive, electric, connected, autonomous, and shared (ECAS) vehicles, railway, avionics, and production systems. Leveraging AI in these sectors will improve the efficient use of energy resources and increase productivity.

However, running computation-intensive ML/DL models locally on edge devices can be very resource-intensive, requiring, in the worst-case, high-end processing units to be equipped in the end devices. Such stringent requirement not only increases the cost of edge intelligence but can also become either unfriendly or incompatible with legacy, non-upgradeable devices endowed with limited computing and memory capabilities. Fortunately, inferring in the edge with

the most accurate DL model is not a standard requirement. It means that, depending on the use case, different trade-offs among inference accuracy, power consumption, efficiency, security, safety, and privacy can be met. This awareness can potentially create a permanently accessible AI continuum. Indeed, the real game-changer is to shift from a local view (the device) to the "continuum" (the whole technology stack) and find the right balance between edge computation (preferable whenever possible, because it does not require data transfer) and data transmission towards cloud servers (more expensive in terms of energy). The problem is complex and multi-objective, meaning that the optimal solution may change over time, needing to consider changing cost variables and constraints. Interoperability/compatibility among devices and platforms is essential to guarantee efficient search strategies in this search space.

AI accelerators are crucial elements to improve efficiency and performances of existing systems (to the cost of more software complexity, as described in the next challenge, but one goal will be to automatize this process). For the training phase, the large amount of variable precision computations requires accelerators with efficient memory access and large multi-computer engine structures. In this phase, it is necessary to access large storage areas containing training instances. However, the inference phase requires low-power efficient implementation with closely interconnected computation and memory. In this phase, efficient communication between storage (i.e., the synapses for a neuromorphic architecture) and computing elements (the neurons for neuromorphic) are paramount to ensure good performances. Again, it will be essential to balance the need and the cost of the associated solution. For edge/power-efficient devices, perhaps not ultra-dense technologies are required; cost and power efficiency matter perhaps more than raw computational performances. It is also important to develop better tools and algorithms to reduce size and computational complexity of Deep Neural Networks for allowing efficient AI applications to be executed at the edge.

DRAFT

Other architectures (neuromorphic) need to be further investigated and to find the sweet use case spot. One key element is the necessity to save the neuronal network state after the training phase as reinitializing after switch-off will increase the global consumption. The human brain never stops.

It is also crucial to have a co-optimization of the software and hardware to explore more advanced trade-offs. Indeed, AI, and especially DL, require optimized hardware support for efficient realization. New emerging computing paradigms such as mimicking the synapses, using unsupervised learning like STDP (Spike-timing- dependent plasticity) might change the game by offering learning capabilities at relatively low hardware cost and without needing to access large databases. Instead of being realized by ALU and digital operators, STDP can be realized by the physics of some materials, such as those used in Non-Volatile Memories. These novel approaches need to be supported by appropriate SW tools to become viable alternatives to existing approaches.

Developing solutions for AI at the edge (e.g., for self-driving vehicles, personal assistants, and robots) is more in line with European requirements (privacy, safety) and know-how (embedded systems). Solutions at the extreme edge (small sensors, etc.) will require even more efficient computing systems because of their low cost and ultra-low power requirements.

The Deep learning approach is based on the neural networks paradigm coming initially from the work of Mc Culloch and Pitts, where a neuron is a small computing element connected to its pairs by weights called synapses. It is a structure where computing and storage are naturally

closely mixed. It is therefore important to address memories and topologies in such AI architectures. Sparsity of coding and of the neural network topology are important to reduce energy consumption, both by decreasing data communication and taking benefit of the sparsity of coding and of the topology.

5.4 Major Challenge 2: managing the increasing complexity of systems:

5.4.1 For Edge Computing

The increasing complexity of electronic embedded systems, hardware and software algorithms has a significant impact on the design of applications, engineering lifecycle and the ecosystems involved in the product and service development value chain.

The complexity is the result of the incorporation of hardware, software and connectivity into systems, and their design to process and exchange data and information without addressing the architectural aspects. As such, architectural aspects such as optimizing the use of resources, distributing the tasks, dynamically allocating the functions, providing interoperability, common interfaces and modular concepts that allow for scalability are typically not sufficiently considered. Today's complexity to achieve higher automation levels in vehicles and industrial systems is best viewed by the different challenges which need to be addressed when increasing the number of sensors and actuators offering a variety of modalities and higher resolutions. These sensors and actuators are complemented by ever more complex processing algorithms to handle the large volume of rich sensor data. The trend is reflected in the value of semiconductors across different vehicle types. While a conventional automobile contains roughly $330 value of semiconductor content, a hybrid electric vehicle with a full sensor platform can contain up to $1000 and 3,500 semiconductors. Over the past decade, the cost contribution for electronics in vehicles has increased from 18% to 20% to about 40% to 45%, according to Lam Research. The numbers will further increase with the introduction of autonomous, connected, and electric vehicles which make use of AI-based HW/SW components.

This approach necessitates the use of multiple high-performance computing systems to support the cognition functions. Moreover, the current Electrical and Electronic (E/E) architectures impose that the functional domains are spread over separated and dedicated Electronic Control Units (ECU's). This approach hampers upscaling of the automation functionality and obstructs effective reasoning and decision making.

5.4.1.1 Key focus areas

The major recommendations at the Embedded architectures infrastructure level are:

- Improving interoperability of systems: this is mainly covered by design methodology, where tools should be able to build a system from IPs coming from various sources. That means also that the description of the IPs, even if they are proprietary (black box), should contain all the view required to smoothly integrate them together. This is also a requirement for Open-Source Hardware. This can be extended at the level of integration in 2.5D systems based on interposers and chiplets: an ecosystem will only proliferate and flourish if a large catalogue of chiplets (in this case) are available and easily connected. As infrastructure for Embedded architectures, the "common

DRAFT

platform" initiated by the European Processor Initiative (EPI) is an example of a template that allows to build different ICs with minimum efforts.

- Facilitating the easy addition of modules to a system: what is done at the Embedded architectures level can also be promoted at the system levels, where reuse of existing core could simplify the design, but perhaps at a cost of more complex software.
- Developing common interfaces and standards: this a basic element if we want to increase the productivity by reuse and the efficiency by using interoperability.
- Using AI techniques to help complexity management: existing Embedded architectures are so complex that humans cannot understand all the interactions and corner cases. Tools and techniques using Operational Research or Artificial Intelligence can be used to explore the space of conception, and recommend optimum combinations and architectures. Automated Design Space Exploration is an emerging field, and AI is already used in backend tools by the major CAD tools providers (and by Google to design their TPUs).

Similarly, the solutions and recommendations for Edge devices are similar of those for embedded computing:

- Improving interoperability of systems
- Facilitating the easy addition of modules to a system
- Developing common interfaces and standards, standardized APIs for hardware and software tool chains
- Using AI techniques to help complexity management

5.4.2 For Embedded Intelligence

To still achieve the required increased level of automation in automotive, transportation and manufacturing, disruptive frameworks are being considered offering a higher order of intelligence. Several initiatives to deliver hardware and software solutions for increased automation are ongoing. Companies like Renesas, NVIDIA, Intel/Mobileye, and NXP build platforms to enable Tier1s and OEMs to integrate and validate automated drive functions. Still, the "vertical" distribution of AI functionality is difficult to manage across the traditional OEM/Tier-1/Tier-2 value chain. Due to the long innovation cycle associated with this chain, vertically integrated companies such as Tesla/Waymo currently seem to hold an advantage in the space of autonomous driving. Closed AI component ecosystems represent a risk as transparency in decision making could prove hard to achieve and sensor level innovation may be stifled if interfaces are not standardized. Baidu (Apollo), Lyft, Voyage and Comma.ai take a different approach as they develop software platforms which are open and allow external partners to develop their own autonomous driving systems through on-vehicle and hardware platforms. Such open and collaborative approach might be the key to accelerate development and market adoption.

Next generation energy- and resource-efficient electronic components and systems that are connected, autonomous and interactive will require AI-enabled solutions that can simplify the complexity and implement functions such as self-configure to adapt the parameters and the resource usage based on context and real time requirements. The design of such components and systems will require a holistic design strategy based on new architectural concepts and optimized HW/SW platforms. Such architectures and platforms will need to be integrated into new design operational models that consider hardware, software, connectivity and sharing of information (1) upstream from external sources like sensors to fusion computing/decision

DRAFT

processes, (2) downstream for virtualization of functions, actuation, software updates and new functions, and (3) mid-stream information used to improve the active user experience and functionalities.

Still, it is observed that the strategical backbone technologies to realize such new architectures are not available. These strategical backbone technologies include smart and scalable electronic, components and systems (controllers, sensors, and actuators), the AI accelerator hardware and software, the security engines, and the connectivity technologies. A holistic end-to-end approach is required to manage the increasing complexity of systems, to remain competitive and to continuously innovate the European electronic components and systems ecosystem. This end-to-end approach should provide new architecture concepts, HW/SW platforms that allow for the implementation of new design techniques, system engineering methods and leverage AI to drive efficiencies in the processes.

Based on the European's semiconductor expertise and in view of its strategic autonomy, we see an incentive for Europe to build an ecosystem on electronic components, connectivity and software AI, especially when considering that the global innovation landscape is changing rapidly due to the growing importance of digitalization, intangible investment and the emergence of new countries and regions. As such, a holistic end-to-end AI technology development approach enables the advances in other industrial sectors by expanding the automation levels in vehicles and industrial systems while increasing the efficiency of power consumption, integration, modularity, scalability, and functional performance.
The new strategy should be anchored into a new bold digitalization transformation as digital firms perform better and are more dynamic: they have higher labour productivity, grow faster, and have better management practices.
The reference architectures for future AI-based systems need to provide modular and scalable solutions that support interoperability and interfaces among platforms that can exchange information and share computing resources to allow the functional evolution of the silicon-born embedded systems.
The evolution of the AI-based components and embedded systems is no longer expected to be linear and will depend on the efficiency and the features provided by AI-based algorithms, techniques and methods applied to solve specific problems. This allows to enhance the capabilities of the AI-based embedded systems using open architecture concepts to develop HW/SW platforms enabling continuous innovation instead of patching the existing designs with new features that ultimately will block the further development of specific components and systems.

Europe has an opportunity to develop and use open reference architecture concepts for accelerating the research and innovation of AI-based components and embedded systems at the edge and deep-edge that can be applied across industrial sectors. The use of reference open architecture will support the increase of stakeholder diversity and AI-based embedded systems, IoT/IIoT ecosystems. This will result in a positive impact on market adoption, system cost, quality and innovation, and will support to ensure the development of interoperable and secure embedded systems supported by a strong European R&I&D ecosystem.

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

- Infineon is well positioned to fully realize AI's potential in different tech domains. By adding AI to its sensors, e.g., utilizing its PSOC microcontrollers and its Modus toolbox,

DRAFT

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

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

- STMicroelectronics integrated Edge AI as one of the main pillars of its product strategy plan. By combining AI-ready features in its hardware products to a comprehensive ecosystem of software and tools, ST ambitions to overcome the uphill challenge of AI : opening technology access to all and for a broad range of applications. For the smart building domain, the STM32 microcontrollers embed optimized machine learning algorithms to determine room occupancy, count people in a corridor or automatically read water meters. The AI code compression is performed by users through the low-code STM32Cube.ai optimizer tool which enables a drastic reduction of the power consumption while maintaining the accuracy of the prediction. In Anomaly detection for industry 4.0, NanoEdge AI studio, an Auto-ML software for edge-AI, automatically finds and configure the best AI library for STM32 microcontroller or smart MEMS that

DRAFT

contain ST's embedded Intelligent Sensor Processing Unit (ISPU) while being able to do learning on device. It results in the early detection of arc-fault or technical equipment failure and extend the lifetime of industrial machines. Designers can now use NanoEdge AI Studio to distribute inference workloads across multiple devices including microcontrollers (MCUs) and sensors with ISPUs in their systems, significantly reducing application power consumption. Always-on sensors that contain the ISPU can perform event detection at very low power, only waking the MCU when the sensor detects anomalies.

Europe can drive the development of scalable and connected HW/SW AI-based platforms. Such platforms will efficiently share resources across platforms and optimize the computation based on the needs and functions. As such, the processing resource will dynamically adjust the type, speed and energy consumption of processing resource depending on the instantaneous required functionality.
This can be extended at the different layers of the architecture by providing scalable concepts for hardware, software, connectivity, AI algorithms (inference, learning) and the design of flexible heterogenous architectures that optimize the use of computing resources.

Optimizing the performance parameters of AI-based components, embedded systems within the envelope based on energy efficiency, cost, heat dissipation, size, weight using reference architecture that can scale across the information continuum from end point deep-edge to edge, cloud and data centre.

DRAFT

5.4.2.1 Key focus areas

- Evolving the architecture, design and semiconductor technologies of AI-based components and systems, integration into IoT/IIoT semiconductor devices with applications in automation, mobility, intelligent connectivity, enabling seamless interactions and optimized decision-making for semi-autonomous and autonomous systems.
- New AI-based HW/SW architectures and platforms with increased dependability, optimized for increased energy efficiency, low cost, compactness and providing balanced mechanisms between performance and interoperability to support the integration into various applications across the industrial sectors.
- Edge and deep-edge components, architectures, and interoperability concepts for AI edge-based platforms for data tagging, training, deployment, and analysis. Use and development of standardized APIs for hardware and software tool chains.
- Deterministic behaviours, low latency and reliable communications are also important for other vertical applications, such as connected cars, where edge computing and AI represent "the" enabling technology, independently from the sustainability aspects. The evolution of 5G is strongly dependent on edge computing and multi-access edge computing (MEC) developments.
- Developing new design concepts for AI born embedded systems to facilitate trust by providing the dependable design techniques, that enable the end-to-end AI systems to be scalable, make correct decisions in repetitive manner, provide mechanisms to

be transparent, explainable, interpretable and able to achieve repeatable results and embed features for AI model's and interfaces' interpretability.

- Distributed edge computing architecture with AI models running on distributed devices, servers, or gateways away from data centres or cloud servers.
- Scalable hardware agnostics AI models capable of delivering comparable performance on different computing platforms, (e.g. Intel, AMD or ARM architectures).
- Seamless and secure integration at HW/SW embedded systems with the AI models integrated in the SW/HW and APIs to support configurable data integrated with enterprise authentication technologies through standards-based methods.
- Development of AI based HW/SW for multi-tasking and provide techniques to adapt the trained model to produce close or expected outputs when provided with a different but related set of data. The new solutions must provide dynamic transfer learning, by assuring the transfer of training instance, feature representation, parameters, and relational knowledge from the existing trained AI model to a new one that addresses the new target task.
- HW/SW techniques and architectures for self-optimize, reconfiguration and to self-manage the resource demands (e.g. memory management, power consumption, model selection, hyperparameter tuning for automated machine learning scenarios, etc.).
- Edge-based robust energy efficient AI-based HW/SW for processing incomplete information with incomplete data, in real time.
- End-to-end AI architecture including the continuum of AI-based techniques, methods and interoperability across sensor-based system, device-connected system gateway-connected system, edge processing units, on-premises servers, etc.
- Developing tools and techniques helping in the management of complexity, e.g. using AI methods.

DRAFT

5.5 Major Challenge 3: supporting the increasing lifespan of devices and systems

5.5.1 For Edge Computing

Increasing lifetime of an electronic object is very complex and has multiple facets. It covers the life extension of the object itself up to the move of some of its critical parts in other objects and ultimately in the recycling of raw material in new objects. This domain of lifetime extension is very error prone as it is extremely easy to confuse some very different concepts such as upgradability, reuse up to recycling.

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

The second aspect of increasing lifetime is to reuse a system in an application framework less demanding in term of performance, power consumption, safety, etc.

5.5.1.1 Key focus areas

For re-using something in an environment for which it was not initially designed, it is key to be able to qualify the part in its new environment. To achieve this very challenging goal the

main question is “what are the objective parameters to take into account to guarantee that the degraded part is compatible with its new working environment?”

- Intelligent reconfigurable concepts are an essential key technology for increasing the re-use and service life of hardware and software components. Such modular solutions on system level require the consideration of different quality or development stages of sensors, software, or AI solutions. If the resulting uncertainties (measurements, predictions, estimates by virtual sensors, etc.) are considered in networked control concepts, the interoperability of agents/objects of different generations can be designed in an optimal way.
- Distributed monitoring: continuous monitoring and diagnosis also play a crucial role for the optimization of product lifetime. Where a large amount of data is collected during daily life operation (e.g. usage, environment, sensor data), big data analysis techniques can be used to predictively manipulate the operational strategy, e.g. to extend service life. Similarly, an increase in power efficiency can be achieved by adjusting the calibration in individual agents. For example, consider a fuel cell electric vehicle where the operation strategy decisively determines durability and service life. Distributed monitoring collects data from various interconnected agents in real-time (e.g. a truck platoon, an aircraft swarm, a smart electricity distribution network, a fleet of electric vehicles) and uses these data to draw conclusions about the state of the overall system (e.g. the state of health or state of function). On the one hand, this allows to detect shifting behaviour or faulty conditions in the systems and to even isolate them by attributing causes to changes in individual agents in the network or even ageing of individual objects and components. Such detection should be accomplished by analysing the continuous data stream that is available in the network of agents. A statistical or model-based comparison of the individual objects with each other provides additional insights. Thus, for example, early failures of individual systems could be predicted in advance. This monitoring should also cover the performance of the semiconductor devices themselves, especially to characterize and adjust to aging and environmental effects and adjust operations accordingly.
- Another essential factor for increasing the lifespan of products is the intelligent use and handling of real-world data from products that are already in use and from previous generations of these. On the one hand, this allows for an optimal adaptation of the operating strategy to, for example, regionally, seasonally, or even individually varying use patterns. On the other hand, the monitoring of all agents (e.g. fleet of vehicles) also enables very precise estimates and predictions of certain conditions. This enables the detection of early failures of individual objects but also the timely implementation of countermeasures. Such approaches can be referred to as distributed monitoring.
- Distributed predictive optimization is possible, whenever information about future events in a complex system is available. Examples are load predictions in networked traffic control or demand forecasts smart energy supply networks. In automation, a concept dual to control is monitoring and state observation, leading to safety-aware and reconfigurable automation systems. Naturally, all these concepts, as they concern complex distributed systems must rely on the availability of vast data, which is commonly associated with the term big data. Note that in distributed systems the information content of big data is mostly processed, condensed, and evaluated locally thus relieving both communication and computational infrastructure.

DRAFT

5.5.2 For Embedded Artificial Intelligence

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

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

Whatever the solution used to extend lifetime of systems, this cannot be achieved without a strong framework regarding standards and, even more important, for AI qualification framework of solutions. AI systems are new and show little standardization currently. Therefore, it is of high importance to devote effort to this aspect of AI-hardware and -software developments. Europe has a very diverse industrial structure, and this is a strength if all those players have early access to the standards frameworks for AI and its development vectors. Open access is therefore as important for the European AI ecosystem as the ability to upgrade and participate in the development of AI-interfaces. Another very important point is how we qualify an AI solution. Comparing to computing systems based on algorithm, where it exists a lot of tools and environment to detect and certify that a system has a given property thanks to static code analysis, formal proof, worst case execution time, … In case of AI, most of these solutions are not applicable as the performance of the system depends on the quality of datasets used for training and quality of data used during the inference phases.

For this reason, we suggest a strong and dedicated focus on upcoming AI-standards. Nevertheless, we need to keep in mind the strong business lever of standard and make sure that European companies will be able to build on top of standards and generate value at European level. For

DRAFT

instance, android is open source but no way to make a competitive smartphone without a Google android license.

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

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

At the application level, edge AI has a potential positive impact on ecologic sustainability: consider e.g. the application of AI to optimize and reduce the power consumption in manufacturing plants, buildings, households, etc. The potential impact is evident but, to ensure a real sustainable development and a real benefit, edge AI solutions will have to ensure that the costs savings are significantly larger than the costs required to design, implement and train AI.

More generally, the implementation, deployment and management of large-scale solutions based on edge AI could be problematic and unsustainable, if proper engineering support, automation, integration platforms and remote management solutions will not be provided. At this level, the problem of sustainability includes business models, organizational aspects, companies' strategies, partnerships, and it extends to the entire value chain proposing edge AI-based solutions.

DRAFT

5.5.2.1 Key focus areas

- Developing HW/SW architectures and hardware that support software upgradability and extension of software useful life. Secure software upgradability is necessary in nearly all systems now and hardware should be able to support future updates. AI introduces additional constraints compared to previous systems. Multiplicity of AI approaches (Machine learning, DL, semantic, symbolic, etc.), multiplicity of neural network architectures based on a huge diversity of neuron types (CNN, RNN, etc.), potential complete reconfiguration of neural networks for a same system (linked to a same use case) with a retraining phase based on an adapted set of data make upgradability much more complex. This this why HW/SW, related stacks, tools, data sets compatible with the Edge AI system must be developed in synergy. HW/SW plasticity is necessary whatever is the AI background principle of each system to make them as much as possible upgradable and interoperable and to extend the system lifetime. HW virtualization will help to achieve it as well as standardization. The key point is that lifespan extensions, like power management, are requirements which

must be considered from day one of the design of the system. It is impossible to introduce them near the end without a strong rework.

- Standardization: standards are very difficult to define as they shouldn't be too restrictive to avoid limitation to innovation but not too open also to avoid plenty of objects compliant to the standard but not really interoperable because not supporting the same options of the same standard. For this reason, the concept of introducing standards early in the innovation process, must be complemented with a visionary perspective in view to expand the prospective standards for future expansions in function, feature, form, and performance.
- Re-use: One concept called the "2nd life" is actually the re-use parts of systems. Such re-use could be adapted to edge AI as far as some basic rules are followed. First, it is possible to extract the edge AI HW/SW module which is performing a set of functions. For example, this module performs classification for images, movements detection, sounds recognition, etc… Second, the edge AI module can be requalified and recertified downgrading its quality level. A module implemented in aeronautic systems could be reused in automotive or industrial applications. A module used in industrial could be reused in consumer applications. Third, an AI system may be re-trained to fit the "2nd life" similar use case, going for example from smart manufacturing to smart home. Last, business model will be affordable only if such "2nd life" use is on a significant volume scale. A specific edge AI embedded module integrated in tens of thousands of cars could be removed and transferred in a new consumer product being sold on the market.
- Prediction and improvements: prediction / improvements with pure analytics techniques is always difficult. Very often the analytic behaviours of some system parts is not known and then either approximate models are build-up or it is just ignored. Thanks to AI, the system will be able to evolve based on data collected during its running phase. AI techniques will allow better prediction method based on real data allowing the creation of aggregated and more pertinent indicators not possible with pure analytic approach.
- Realizing self-X (adaptation, reconfiguration, etc.): for embedded systems self-adaptation, self-reconfiguration has an enormous potential in many applications. Usually in self-reorganizing systems the major issue is how to self-reorganize while preserving the key parameters of a system (performance, power consumption, real time constraints, etc.). For any system, there is an operating area which is defined in the multi-dimensional operating parameter space and coherent with the requirements. Of course, very often the real operating conditions are not always covering the whole operating domain for which the system was initially designed. Thanks to AI, when some malfunctioning parts are identified it could then be possible to decide, relying on AI and the data accumulated during system operation, if it affects the behaviours of the system regarding its real operating conditions. If it is not the case, it could be considered that the system can continue to work, with maybe some limitations, but which are not vital regarding normal operation. It would then extend its lifetime "in place". The second case is to better understand the degraded part of a system and then its new operating space. This can be used to decide how it could be integrated in another application making sure that the new operating space of the new part is compatible with the operating requirements of the new hosting system.

DRAFT

- Self-learning techniques are promising. Prediction on Natural Language Understanding (NLU) on the fly or keyboard typing, predictive maintenance on mechanical systems (e.g. motors) are more and more studied. Many domains can benefit of the AI in mobility, smart building, communication infrastructure.
- Dynamic reconfiguration: a critical feature of the AI circuits is to dynamically change their functions in real-time to match the computing needs of the software, AI algorithms and the data available and create software-defined AI circuits and virtualize AI functions on different computing platforms. The use of reconfigurable computing technology for IoT devices with AI capabilities allows hardware architecture and functions to change with software providing scalability, flexibility, high performance, and low power consumption for the hardware. The reconfigurable computing architectures, integrated into AI-based circuits can support several AI algorithms (e.g. convolutional neural network (CNN), fully connected neural network, recursive neural network (RNN), etc.) and increase the accuracy, performance and energy efficiency of the algorithms are integrated as part of software define functions.
- From the engineering perspective, leveraging open source will help developing European advanced solutions for edge AI (open-source hardware, software, training datasets, open standards, etc.).

As a summary, intelligence at the edge sustainable engineering will have to face many challenges:

- Supply chain integrity for development capability, development tools, production, and software ecosystems, with support for the entire lifecycle of edge AI based solutions.
- Security for AI systems by design, oriented also to certify edge AI based solutions. European regulations and certification processes would lead to a global compelling advantage.
- Europe needs to establish and maintain a complete R&D ecosystem around AI.
- Europe need to address the end to end value chain and supports its SMEs.
- Identification of a roadmap for standardization that does not hinder innovation: the right balance that ensure European leadership in edge AI.
- Europe must strive for driving a leading and vibrant ecosystem for AI, with respect to R&D, development and production, security mechanisms, certifications, and standards.

5.6 Major Challenge 4: ensuring European sustainability

5.6.1 For Edge Computing

One of the major challenges that need to be accounted for in the next few years is related to the design of progressively more complex electronic systems to support advanced functionalities such as AI and cognitive functionality. This is particularly challenging in the European landscape, which is dominated by small and medium enterprises (SMEs) with only some large actors that can fund and support larger-scale projects. To ensure European competitiveness and sustainability in advanced Embedded architectures it is therefore crucial to create an ecosystem, and the means, in which SMEs can cooperate and increase their level of innovation and productivity. This ecosystem needs to cover at the best all part of the value chain from concept to design till production. The definition of open industrial standards and a market of Intellectual Properties (IPs) are required to accelerate the design, competitivity and create a larger

DRAFT

market. Open-source on Software, Hardware and tools can play an extremely important role in this regard, open-source solutions significantly allow to reduce engineering costs for licensing and verification, lowering the entry barrier to design innovative products.

5.6.1.1 Key focus areas

- Energy efficiency improvement:
  - New materials, new embedded non-volatile memories with high density and ultra-low power consumption, substrates and electronic components oriented to low and ultra-low power solutions.
  - 3D-based device scaling for low power consumption and high level of integration.
  - Strategies for self-powering nodes/systems on the edge.
  - Low and ultra-low power and interoperable communications components
  - Efficient cooling solutions.
- Improving sustainability edge computing:
  - Efficient and secure code mobility.
  - Open edge computing platforms, providing remote monitoring and control, security and privacy protection.
  - Solutions for the inclusion/integration of existing embedded computers on the edge.
  - Policies and operational algorithms for power consumption at edge computing level.
  - New benchmarking approach considering sustainability.
- Leveraging open source to help developing European advanced solutions on the edge:
  - Open-source hardware (and its complete ecosystem of Ips and tools).
  - Open-source software.
  - Europe must address the end to end value chain.
- Engineering support to improve sustainable edge computing:
  - Engineering process automation for full lifecycle support.
  - Edge devices security by design.
  - Engineering support for edge computing, verification and certification, addressing end-to-end edge solutions.

5.6.2 For Embedded Artificial Intelligence

First, as Embedded Artificial Intelligence is developing quickly and in many different directions for new solutions, it is crucial that an European ecosystem emerge gathering all steps of the value chain. It has then to include the hardware, the software, the tools chain for AI development and the data sets in an trustable and certifiable environment. Both Edge Computing and Embedded Artificial Intelligence ecosystems are tied together.

Next, technology is strongly affected by sustainability that, very often, tips the scale between the ones that are promising, but not practically usable, and the ones making the difference. e.g. cloud computing, based on data centres, plays a fundamental element for the digitalization process. However, data centres consume a lot of resources (energy[43] , water, etc.) and they are

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

DRAFT

responsible for significant carbon emissions, during their entire lifecycle, and generate a lot of electronic and chemical waste.

Today, the percentage of worldwide electricity consumed by data centres is estimated to exceed the 3%, while the CO2 emissions are estimated to reach the 2% of worldwide emissions[44] [45], with cloud computing that is responsible for half of these emissions. A recent study predicts that, without energy efficient solutions, by 2025 eight data centres will consume 20% percent of the world's energy, with a carbon footprint rising to 5.5% of the global emissions. Data centres are progressively becoming more efficient, but shifting the computing on the edge, for example, allows to temporally reduce data traffic, data centres storage and processing. However, only a new computing paradigm could significantly reduce their environmental footprint and ensure sustainability. Edge Computing could contribute to reach this goal by the introduction of ultra-low and efficient computing solutions.

Indeed, from a wider perspective, digital transformation relies largely on other technologies that could significantly impact sustainability, including edge and fog computing, AI, IoT hyper connectivity, etc. In recent years, artificial intelligence and cloud computing have been the focus of the scientific community, environmental entities and public opinion for the increasing levels of energy consumption, questioning the sustainability of these technologies and, indirectly, their impact on corporate, vertical applications and societal sustainability. For example, devices are already producing enormous amounts of data and a recent study[46] estimates that by 2025 communications will consume 20% of all the world's electricity. This situation has been worsening with COVID-19 pandemic that generated a worldwide reduction of power consumption as a result of global lockdown restrictions but, at the same time, caused a huge spike in Internet usage: NETSCOUT measured an increase of 25-35% of worldwide Internet traffic in March 2020, just due to remote work, online learning and entertainment. This spike in Internet use provides a flavour of the implications of digitalization on sustainability. Reducing energy of computing and storage devices is a major challenge (see Major Challenge 1 on "Increasing the energy efficiency of computing systems").

Shifting to green energy is certainly a complementary approach to ensure sustainability, but the conjunction of AI and edge computing, the Edge AI, has the potential to provide sustainable solutions with a wider and more consolidated impact. Indeed, a more effective and longer-term approach to sustainable digitalization implies reconsidering the current models adopted for data storage, filtering, analysis, processing, and communication. By embracing edge computing, for example, it is possible to significantly reduce the amount of useless and wasteful data flowing to and from the cloud and data centres, with an architectural and structural more efficient solutions that permanently reduces the overall power consumption and bring other important benefits such as real-time data analysis reducing the amount of data to be stored and then a better data protection. The Edge Computing paradigm also makes AI more sustainable: it is evident that cloud-based machine learning inference is characterized by a huge network load, with a serious impact on power consumption and huge costs for organizations. Transferring machine learning inference and data pruning to the edge, for example, could exponentially decrease the digitization costs and enable sustainable businesses. To avoid this type of drawbacks, new AI

[44] Koronen, C., Åhman, M. & Nilsson, L.J. Data centres in future European energy systems—energy efficiency, integration and policy. Energy Efficiency 13, 129–144 (2020).
[45] https://datacentrereview.com/content-library/490-how-to-reduce-data-centre-energy-waste-without-sinking-it-into-the-sea
[46] Andrae, A., & Edler, T. (2015). On global electricity usage of communication technology: trends to 2030. Challenges, 6, 117–157.

DRAFT

components should be developed based on neuromorphic architectures and considering the application areas, in some cases, this could bring to a more specialised and very efficient solutions.

Sustainability of Edge Computing and AI is affected by many technological factors, on which Europe should invest, and, at the same time, they have a positive impact on the sustainability of future digitalization solutions and related applications.

GAMAM already master these technologies and are progressively controlling the complete value chain associated with them. To follow this trend and aim at strategic autonomy, Europe has therefore to fill the technology gaps and address the value chain end to end, with a particular attention to SMEs (which generate a large part of European revenues) and leveraging on the cooperation between the European stakeholders in the value chain to develop successful products and solutions. From this perspective, European coordination to develop AI, edge computing and edge AI technologies is fundamental to create a sustainable value chain based on alliances and capable to support the European key vertical applications.

It will be a challenge for Europe to be in this race, but the emergence of AI at the edge, and its know-how in embedded systems, might be winning factors. However, the competition is fierce and the big names are in with big budgets and Europe must act quickly, because US and Chinese companies are already also moving in this "intelligence at the edge" direction.

DRAFT

5.6.2.1 Key focus areas

On top of the key focus area for Edge computing, Embedded Artificial Intelligence also requires:

- Energy-efficiency improvement:
  - New memories used to mimic synapses.
  - Advanced Neuromorphic components.
- Improving sustainability of AI:
  - Re-use and share of knowledge and models generated by embedded intelligence.
  - Energy- and cost-efficient AI training.
  - New benchmarking AI approach considering sustainability.
- Leveraging open source to help developing European AI advanced solutions on the edge:
  - Open-source training datasets.
  - Open Frameworks including AI tools.
  - Europe must address the end-to-end Embedded Intelligence value chain.
- Engineering support to improve sustainable AI:
  - Edge AI security by design.
  - Engineering support for AI verification and certification.
  - Education and support to deploy Edge AI.

6 Timeline

Short term: 2022-2026 - Medium 2027-2031 - Long term: beyond 2031
Legend:

- (EC): concern Edge Computing
- (eAI:) concern Embedded Artificial Intelligence

MAJOR CHALLENGE	TOPIC	SHORT TERM	MEDIUM TERM	LONG TERM
		2023-2027	**2028-2032**	**BEYOND 2032**
Major Challenge 1: increasing the energy efficiency of computing systems	**Processing data where it is created** **(EC and eAI)**	Development of algorithms and applications where processing is performed. Moving processing towards edge when it is possible New memory management	Development of hybrid architectures, with smooth integration of various processing paradigms (classical, neuromorphic, deep learning), including new OSs supporting multiple computing paradigms Advanced memory management	Dynamic instantiation of r paradigm computing reso according to the specifica of the task to be performe Automatic interfacing, dis ery, and configuration of r sources
	Development of innovative hardware architectures (EC)	Development of computing paradigms (e.g. using physics to perform computing). Use of other technologies than silicon (e.g. photonics) Use of 2.5D, interposers and chiplets, with efficient interconnection network, e.g. using photonics) New In-memory computing accelerators	Supporting tools integrating multiple computing paradigms. Advanced In-memory computing accelerators	Integration in the same pa of multiple computing pa digms (classical, Deep Lea neuromorphic, photonic, Complete 2.5D (interpose chiplets) ecosystem, with increasing productivity an use of chiplets in differen signs
	Development of innovative hardware architectures: e.g. neuromorphic (eAI)	Development of neuromorphic based chips and support of this new computing model. New In-memory computing accelerators for AI New architecture to support new DNN models in the embedded space, such as transformers. Development tools allowing to prune/quantize big networks in order to map them onto embedded devices	Integration of neuromorphic and other computing within classical systems Supporting tools integrating multiple AI computing paradigms. Automatic adaptation of complex networks to embedded systems with a minimum loss of performances	Integration in the same pa of multiple computing pa digms (classical, Deep Lea neuromorphic, photonic, Exploring potential use of tum computing in Artificia telligence?

DRAFT

Developing distributed edge computing systems (EC)	Development of Edge (ex: fog) type of computing (peer to peer)	Edge computing demonstrating high performance for selected applications	
Developing distributed edge AI systems (eAI)	Development of efficient and automated transfer learning: only partial relearning required to adapt to a new application (Ex: Federative learning) Support of recent Neural networks models such as Transformers, architectures for state of the art Neural Networks algorithms.	Federated learning or similar approach demonstrating high performance for selected applications	
Interoperability (with the same class of application) and between classes (EC and eAI)	Create gateways between various solutions, beyond ONNX (for eAI) Developing open architectures (for fast development) with maximum reuse of tools and frameworks Interfaces standards (more than solutions) (could help explainability, with a move from black to grey boxes)	Common interface architecture, with dynamic binding: publishing of capabilities for each device/block, flexible data structure and data converters, dynamic interconnect. Promoting European standard for interoperability cross application silos. Interfaces publishing non-functional properties (latency, bandwidth, energy, …)	At all levels (from chips to tems), automatic interope ity, adaptation to the data structure and physical inte face, considering the com cation characteristics. (Mi term? Automatic translate data and data format) Global reconfiguration of resources to satisfy the fu tional and non-functional quirements (latency, ener etc.)
Scalable and Modular AI (eAI)	Using the same software development infrastructure from deep edge to edge and possibly HPC applications for AI developments Use of similar building blocks from deep edge to edge AI devices	Scalable architecture (in 3 dimensions). Use of interposer and chiplets to build chips for various applications (for edge and for HPC applications) with the same AI hardware building blocks Complete 2.5D (interposers and chiplets) ecosystem, with tools increasing productivity and reuse of chiplets in different designs of AI systems	Linear and/or functional s bility of AI systems
Scalable and Modular systems (EC)	Using the same software development infrastructure from deep edge to edge and possibly HPC applications. Use of similar building blocks from deep edge to edge devices	Scalable architecture (in 3 dimensions). Use of interposer and chiplets to build chips for various applications (for edge and for HPC applications) with the same hardware building blocks Complete 2.5D (interposers and	Linear and/or functional s bility Digital twin (Functionaliti simulation)

DRAFT

			chiplets) ecosystem, with tools increasing productivity and re-use of chiplets in different designs	
	Co-design: algorithms, HW, SW and topologies (EC)	Quick implementation and optimization of HW for the new emerging algorithms	Tools allowing semi-automatic design exploration of the space of configurations, including variants of algorithms, computing paradigms, hardware performances, etc.	Auto-configuration of a di uted set of resources to s the application requireme (functional and non-funct

Major Challenge 2: managing the increasing complexity of systems	**Balanced mechanisms between performance and interoperability** **(EC)**	Exposing the non-functional characteristic of devices/blocks and off-line optimization when combining the devices/blocks	On-line (dynamic) reconfiguration of the system to fulfil the requirements that can dynamically change (Self-x)	Drive partitioning throu standards
	Development of trustable AI (eAI)	Move to security Chapter	Move to security Chapter	Move to security Chapt
	Developing distributed edge computing systems (EC)	See items above in *Increasing the energy efficiency of computing systems*	See items above in *Increasing the energy efficiency of computing systems*	See items above in *Incr ing the energy efficienc computing systems*
	Scalable and Modular AI (eAI)	See items above in *Increasing the energy efficiency of computing systems*	See also items above in *Increasing the energy efficiency of computing systems* Data and learning driven circuits design	See items above in *Incr ing the energy efficienc computing systems*
	Easy adaptation of models (eAI)	Development of efficient and automated transfer learning: only partial re-learning required to adapt to a new application (Ex: Federative learning) Create a European training reference database for same class of applications/use cases network learning	Optimization of the Neural Network topology from a generically learned networks to an application specific one.	Generic model based d AI development system

DRAFT

	Easy adaptation of modules (EC)	Easy migration of application on different computing platforms (different CPU – x86, ARM, Risc V- different accelerators)	Use of HW virtualization Automatic transcoding of application for a particular hardware instance (à la Rosetta 2)	Generic model based d development system
	Realizing self-X Self-optimize, reconfiguration and self-management (EC)	Add self-assessment feature to edge devices	Automatic reconfiguration of operational resources following the self-assessment to fulfil the goal in the most efficient way	Modelling simulation to for scalable digital twin
	Using AI techniques to help in complexity management (EC and eAI)	Using AI techniques for the assessment of solutions and decrease the design space exploration	Automatic generation of architecture according to a certain set of requirements (in a specific domain)	Modelling simulation to for scalable digital twin

DRAFT

Major Challenge 3: supporting the increasing lifespan of devices and systems	**HW supporting software upgradability** **(eAI)**	Create a European training reference database for same class of applications/use cases network learning Develop European training benchmarks (Methods and methodologies) Build framework tools for HW/SW for fast validation and qualification Establish interfaces standards compatible with most of AI approaches	HW virtualization based on AI algorithms Generic AI functions virtualization European training standards (Compliance/Certification) Certifiable AI (and paths towards explanability and interpretability)	Explainable AI
	Realizing self-X Also partially in ***Managing the increasing complexity of systems (eAI)***	Unsupervised learning technics Development of efficient and automated transfer learning: only partial relearning required to adapt to a new application (Ex: Federative learning)	HW virtualization based on AI algorithms Generic AI functions virtualization Certifiable AI (and paths towards explanability and interpretability)	Explainable AI

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

Major Challenge 4: ensuring	**Energy efficiency improvement**	Materials and electronic components oriented to low and ultralow power solutions Low and ultra-low power	3D-based device scaling for low energy consumption	

DRAFT

European sustainability in Edge computing and embedded Artificial Intelligence	**(EC)**	communications Strategies for self-powering nodes/systems on the edge Efficient cooling solutions		
	Improving sustainability of Edge computing (EC)	Inclusion of existing embedded systems on the edge (huge market opportunity)	Efficient and secure code mobility	
	Improving sustainability of embedded Artificial Intelligence (eAI)	Energy and cost-efficient AI training	Reuse of knowledge and models generated by embedded intelligence	
	Leveraging open source to help developing European AI advanced solutions on the edge (eAI)	Open-source software Open-source training datasets Open edge computing platforms	Open-source hardware	
	Engineering support to improve sustainable Edge computing (EC)	Sustainability through engineering process automation Continuous engineering across the product life cycle	Holistic development environment Engineering support for verification and certification	
	Engineering support to improve sustainable embedded Artificial Intelligence (eAI)	Sustainability through engineering process automation Continuous engineering across the product life cycle	Holistic development environment Engineering support for AI verification and certification Edge AI security by design	

DRAFT

7 Synergy with other themes

The scope of this Chapter is to focus on computing components, and more specifically towards Embedded architectures / Edge Computing and Intelligence at the edge. These elements rely heavily on Process Technologies, Equipment, Materials and Manufacturing, Embedded Software and Beyond, limits on Quality, Reliability, Safety and Cybersecurity, and are composing

systems (System of Systems) that use Architecture and Design techniques to fulfil the requirements of the various application domains. Please refer to all these Chapters in this SRIA for more details.

For example, there are close links with the Chapter on Quality, Reliability, Safety and Cybersecurity on the topics of increasing "trustworthiness" of computing systems, including those using AI techniques:

- Making AI systems "accepted" by people, as a certain level of explainability is required to build trust with their users.
- Developing approaches to verify, certify, audit and trace computing systems.
- Making systems correct by construction, and stable and robust by design.
- Systems with predictable behaviour, including those using deep learning techniques.
- Supporting European principles, such as privacy and having "unbiased" databases for learning, for example.

Embedded Software is also important, and the link to this is explained in the corresponding Chapter. Systems and circuits used for AI are of course developed applying Architecture and Design, and tools techniques and manufactured based on technologies developed in Process Technologies (e.g. use of non-volatile memories, 3D stacking, etc). Artificial intelligence techniques can be also used to improve efficiency in several application.

DRAFT