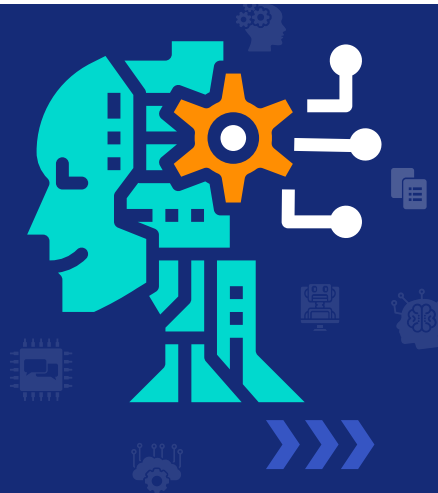
Revolutionising Intelligence
Edge AI Anywhere, Anytime

Ovidiu Vermesan, SINTEF
Giulio Urlini, STMicroelectronics
5 December 2024

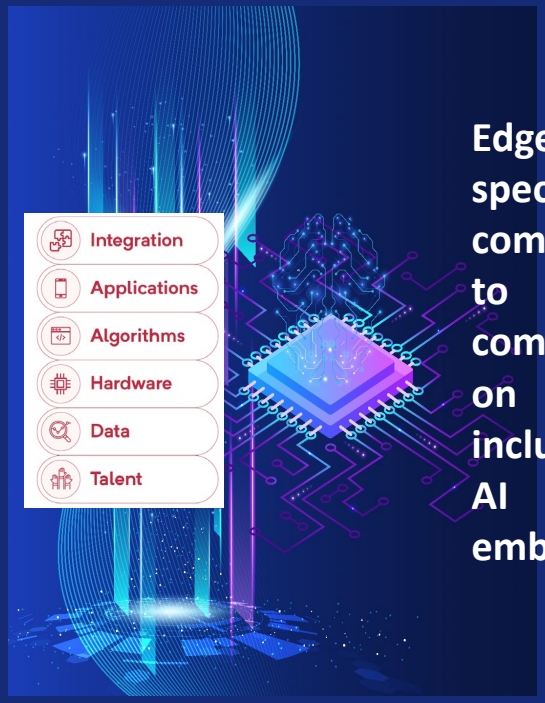
Edge AI



Edge Artificial Intelligence

- Edge AI symbolises the technology convergence of the **Internet of Things (IoT)**, **edge computing** and **AI**, which allows **processing data in real-time** at the edge and brings several benefits like **reduced latency**, **bandwidth requirements**, **power consumption** and **memory footprint** while **increasing security and data protection**.
- Edge AI needs specific hardware stacked up with software, AI algorithms, platforms and datasets.
- Edge processing redefines the interconnected device landscape reflected in the emergence of different edge layers, including **micro-edge**, **deep-edge**, and **meta-edge**.

Edge Intelligence Continuum

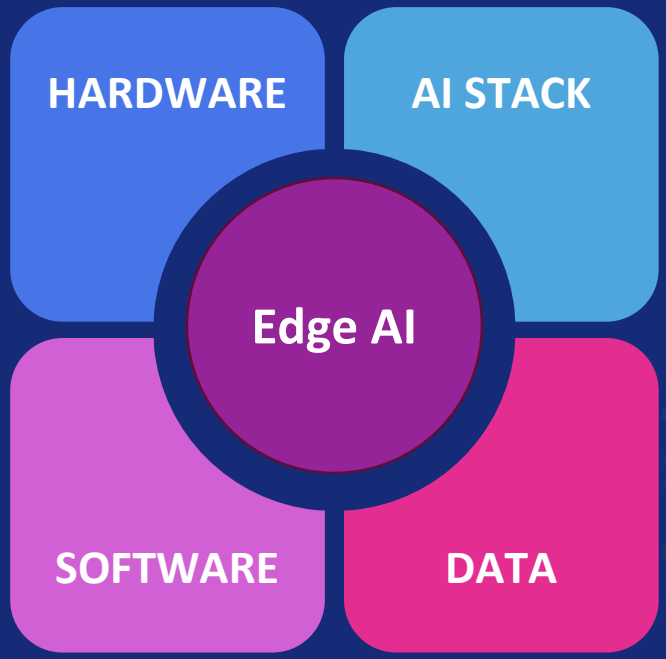


Edge AI chips, specialised hardware components designed to execute AI computations directly on edge devices, include AI processors, AI accelerators, AI-embedded sensors.

- HW platforms.
- CPUs, GPUs, TPUs
- ASICs, FPGAs
- Neuromorphic

- SW platforms
- Optimisation tools
- SDK tools
- Software engineering
- OSs

Holistic View



- AI Models
- Methods
- Algorithms
- Libraries
- Frameworks

- Data collection
- Features extraction
- Data types
- Data sets
- Training data
- Validation/Test data
- Inference data

Micro-edge

DSPs, FGPAs, CPUs, GPUs, ASICs Network Processing Unit (NPU), Intelligent Processing Unit (IPU). Tensor Processing Unit (TPU), Reduced Instr. Set Computer RISC-V, Neuromorphic.

Deep-edge

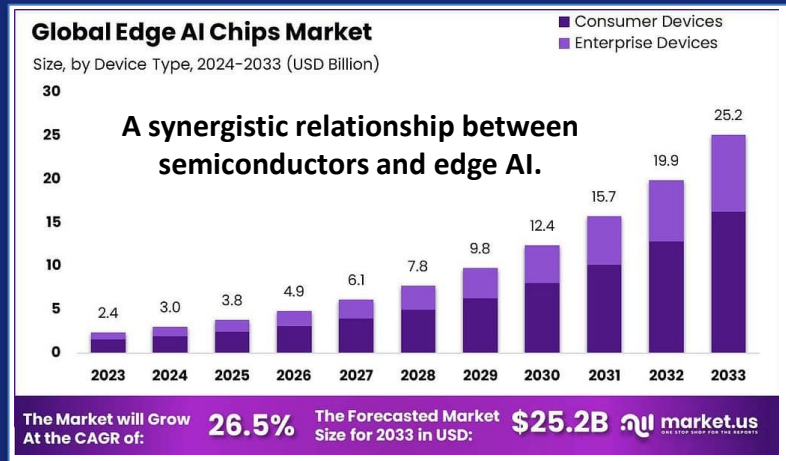
Computing units, network computing units (intelligent routers, switches, gateways and other communications hardware), intelligent controllers (PLCs, RTUs, DCS).

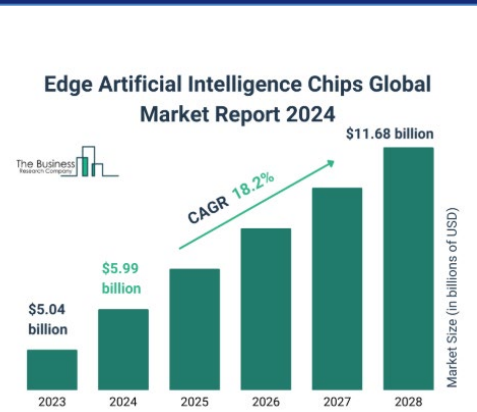
Meta-edge

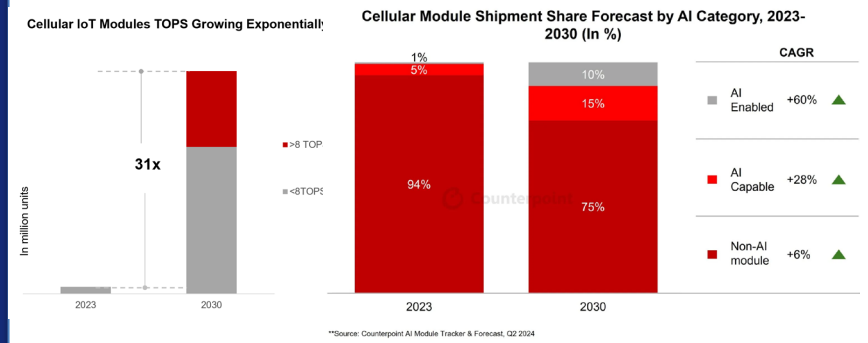
Micro and clustered servers to handle compute intensive tasks / workloads (e.g., high-end CPUs, GPUs, FGPAs, etc.), on premises edge computing, local edge.

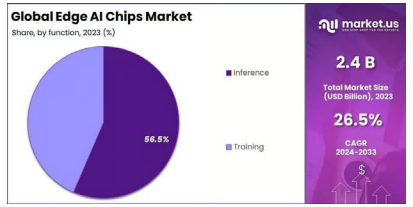
Edge AI Chips Market

- Growth, projected to reach USD 25.2 billion by 2033, at a CAGR of 26.5% (2024-2033).
- In **2023**, the **CPU segment** dominated the market with **36.7%** share.
- The **consumer device segment** in **2023**, accounting **64.5%** share.
- The **inference segment** led the market in **2023**, with a share of **56.5%**.
- **North America** was the leading region in **2023**, with **42.3%** share and revenues of **USD 1.01 billion**.
- Edge devices to process 18.2 zettabytes of data per minute by 2025.









Edge AI-embedded cellular modules are projected to comprise 25% of all IoT module shipments by 2030, up from 6% in 2023.

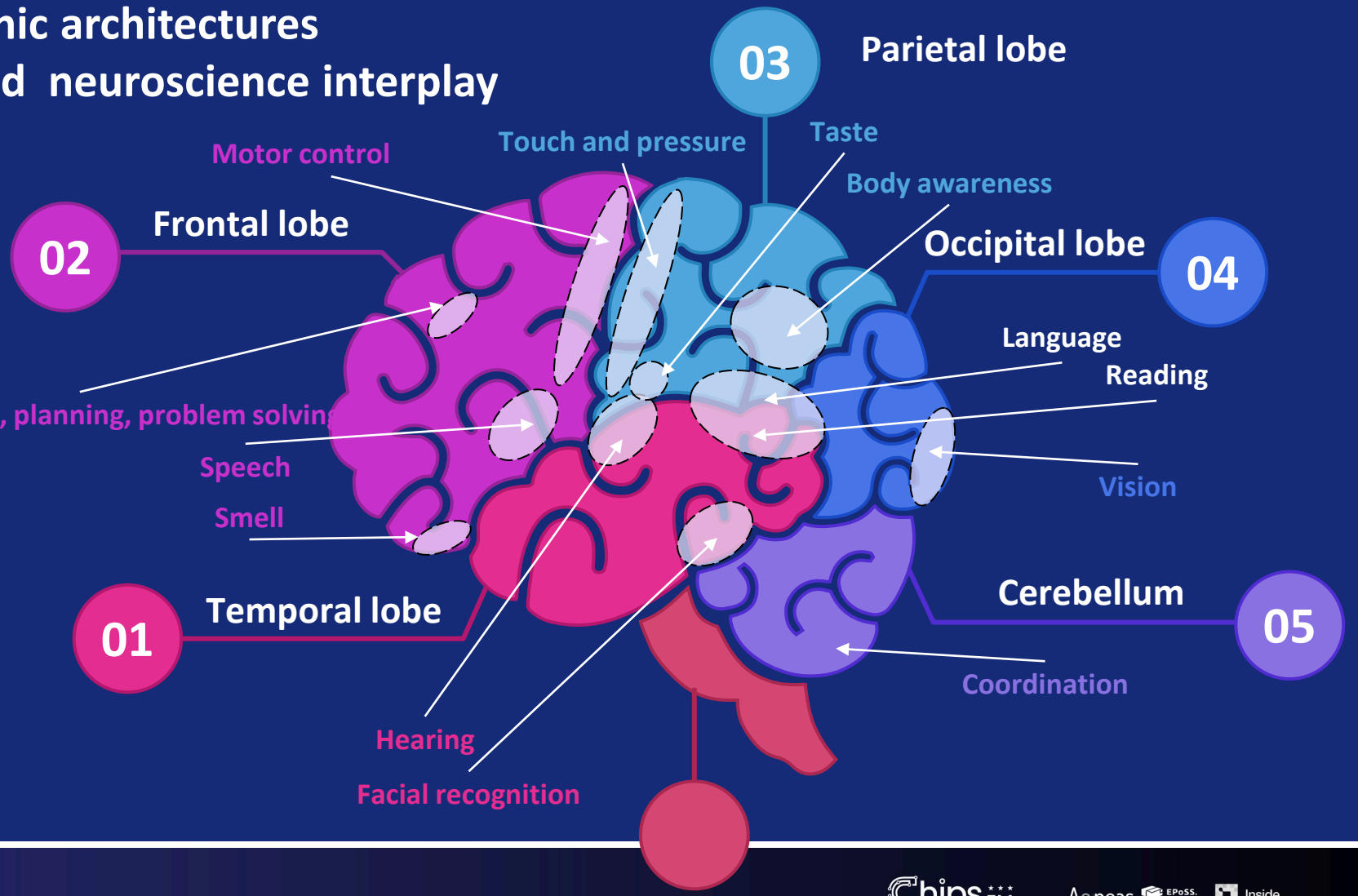
Source: Market.us **Source: The Business Research Company** **Source: Counterpoint Research**

Edge AI Synergetic Evolution - Bio Inspiration

Bio-Inspired edge AI – Neuromorphic architectures
Generative edge AI, biomimicry and neuroscience interplay

The human brain contains about 10^{11} neurons, has a computational power of about 1 Exaflops/s consuming about 20 W*.

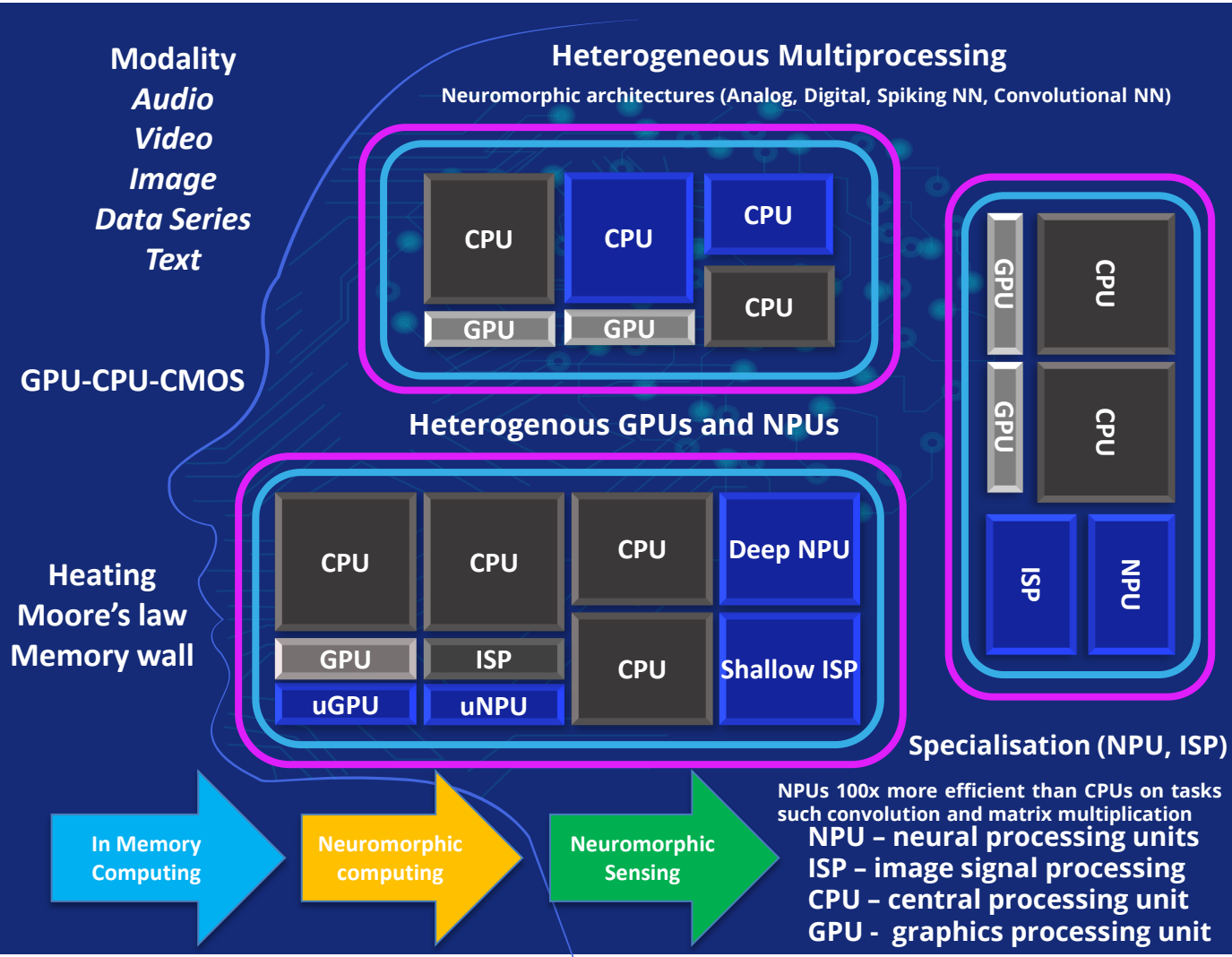
The human brain is a million times more energy-efficient than #1 supercomputer in the world.



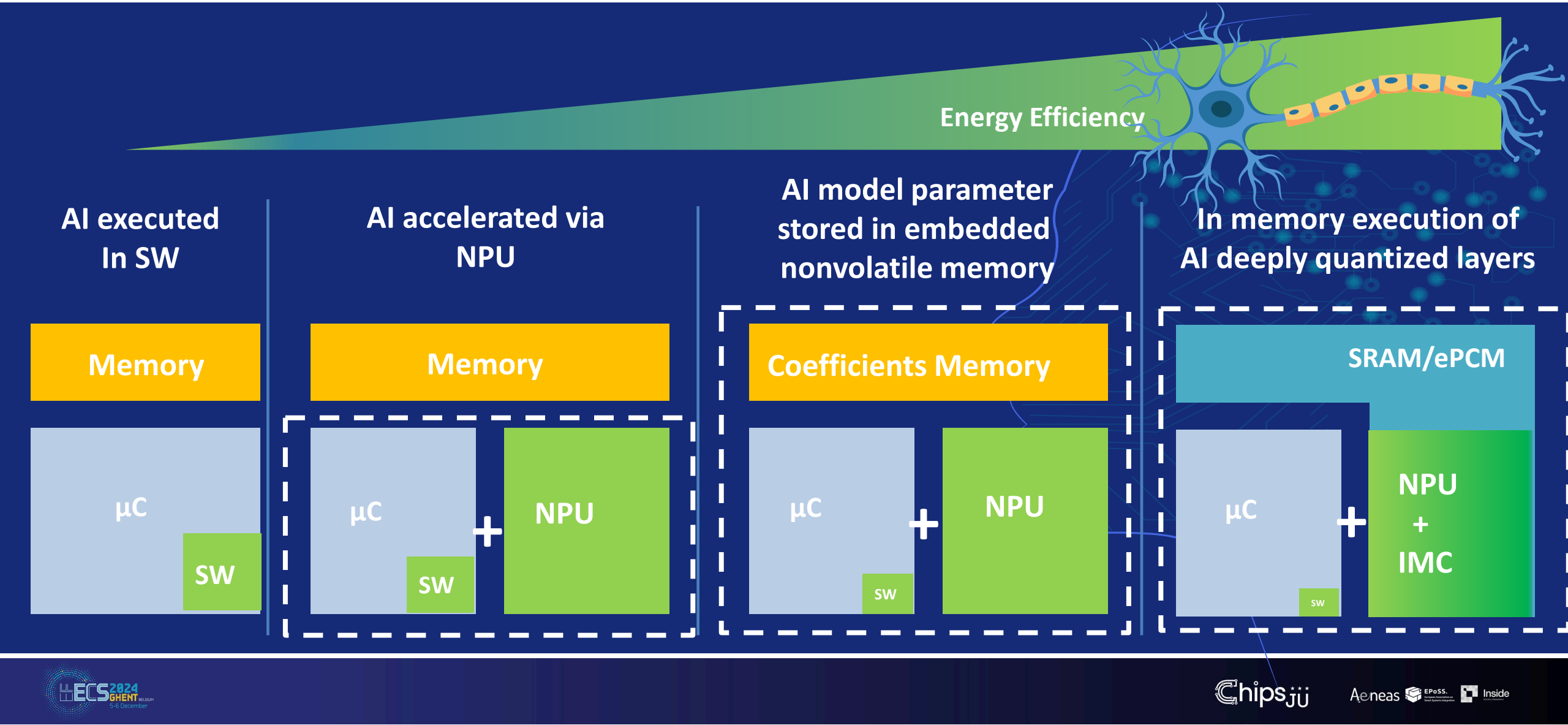
Edge AI Multi-modality and Heterogenity

Edge AI chips designed to perform efficiently process computations specific for AI functions including:

- **Compute cores:** Processing cores of an edge AI chip that include multiple compute cores to address parallel processing functions.
- **Tensor cores:** These specialised cores are designed to maximise the efficiency of DL operations.
- **Vector processors:** Perform vector operations, a key feature in NN operations and key task for edge AI chips.
- **Matrix Multiply Units:** Used in matrix multiplication computations, which is a vital NN feature.
- **Pooling units:** Used to perform pooling in convolutional neural network operations.
- **FPGAs**: Used with an edge AI chip to be programmed to accomplish a certain task and be reprogrammed for a different task or set of command inputs.



Edge AI Evolution of NPUs Towards Neuromorphic



Accelerating Edge AI with RISC-V

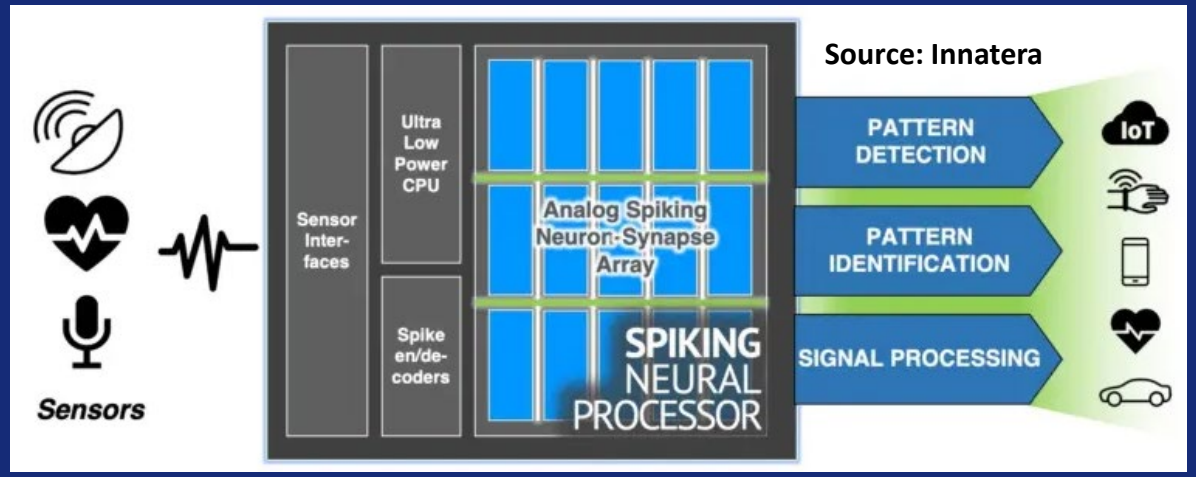
Reduced Instruction Set Computing (RISC)-V processor architectures address edge AI workloads. RISC-V's penetration into AI workloads is pushing RISC-V chip shipments in edge AI to 129 million by 2030.

Source: ABI Research

Challenges:

- Interoperability
- Rate of adoption
- Vendor “lock-in” with low-level software
- Custom extensions
- IP vendors

Spiking Neural Processor T1



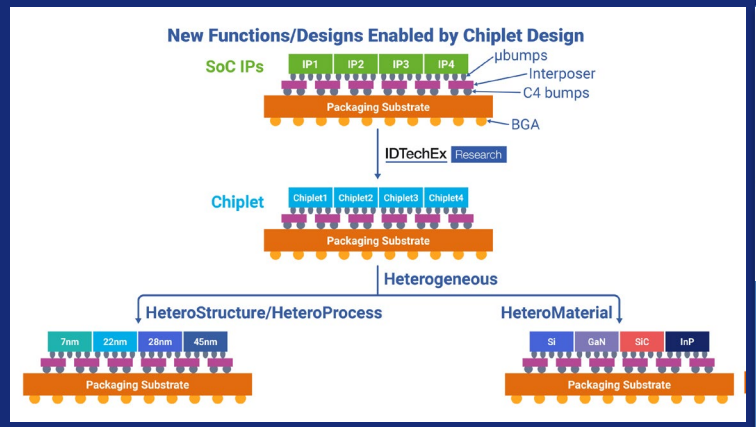
Neuromorphic microcontroller for edge AI sensor applications based on the **RISC-V architecture**. The chip can deliver **energy savings of up to 500x** with **100x shorter latency** across a range of applications compared to a traditional CPU, DSP or conventional AI accelerator.

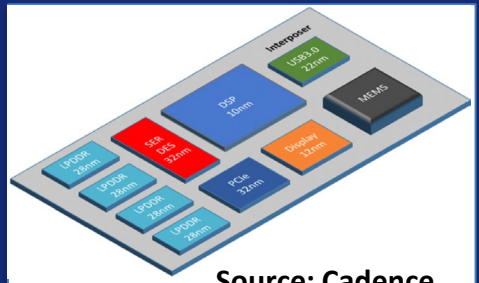
Edge AI – Chiplets and 3D Integration

Chiplets are logically partitioning an edge AI system and offer specialised functions optimised for the specific technology nodes.

- Multi-Die edge AI systems
- Advanced 3D packaging a path to keep Moore's Law alive and applied to edge AI.
- 3D packaging a way of physically integrating multi dies with specific functions.
- Chiplets and 3D stacking are loosely coupled and advance edge AI solutions.
- Chiplets and 3D integration, bring new opportunities to partition IP across dies.

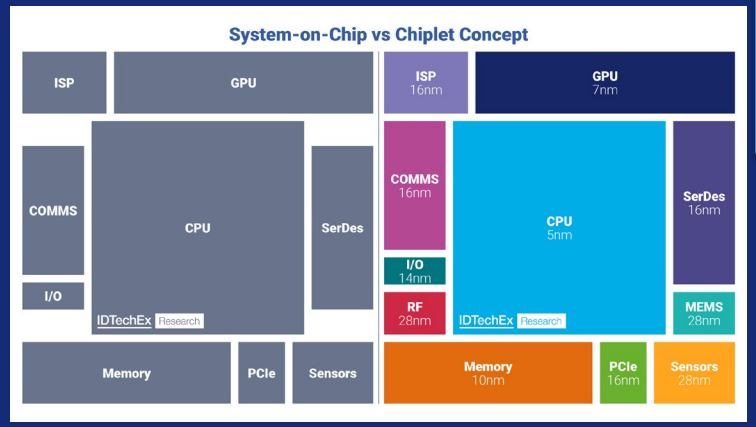
Heterogeneous integration and hyper-scalability



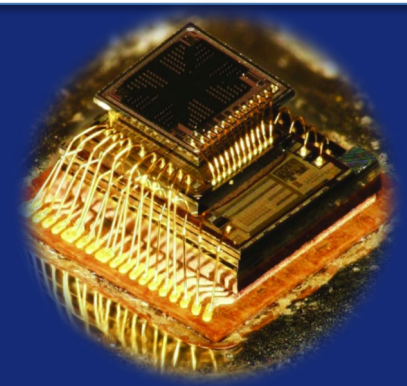


Source: Cadence

	Monolithic SoC	Chiplet-Based
Cost	High	
Effort	High	
Risk	High	
Power		Acceptable?
Performance		Acceptable?
Area		Acceptable?

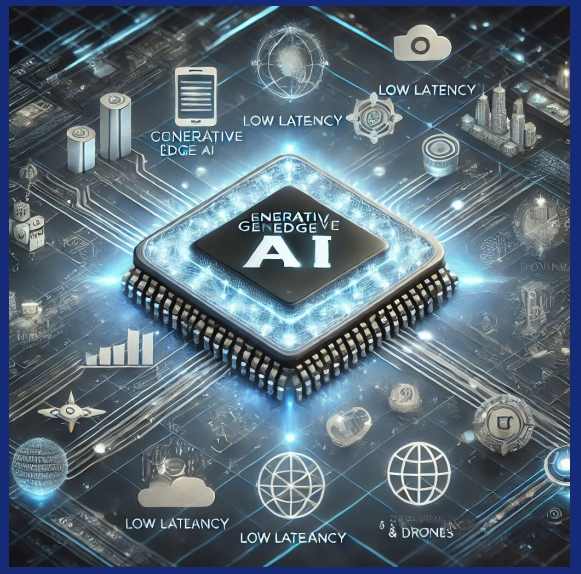


Source: IDTechEx report, "Chiplet Technology 2025-2035: Technology, Opportunities, Applications"



Generative Edge AI

- Deploying large language models (LLMs) on resource constrained devices with limited memory.
- Retrieval-augmented generation (RAG) is a potential solution as it is the process of optimizing the output of a LLM, so it references an authoritative knowledge base outside of its training data sources before generating a response.
- Multimodality embrace multimodal LLMs to use a combination of text, speech and images to deliver more contextually about tables, charts or schematics.

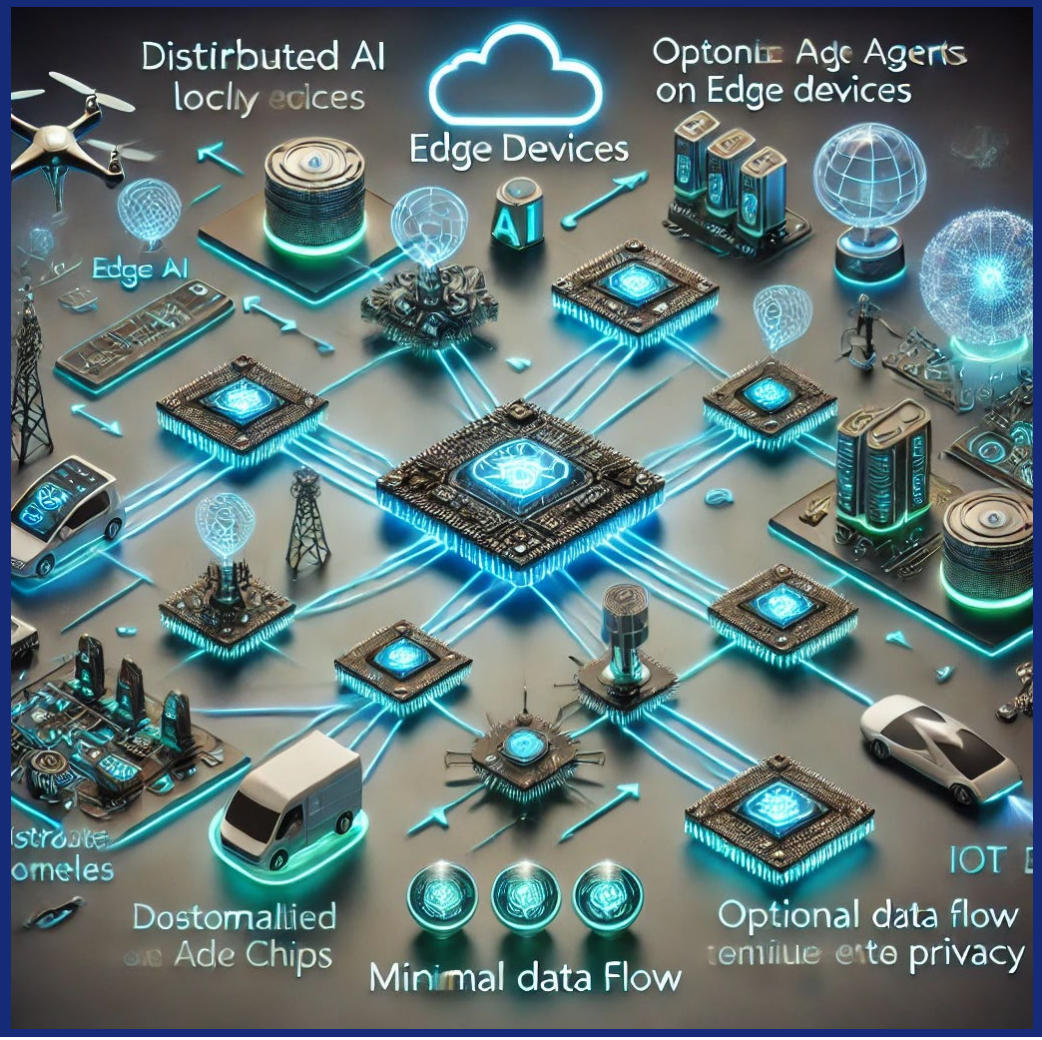


LLMs Challenges and the edge

- Presenting false information when it does not have the answer.
- Presenting out-of-date or generic information when the user expects a specific, current response.
- Creating a response from non-authoritative sources.
- Creating inaccurate responses due to terminology confusion, wherein different training sources use the same terminology to talk about different things.

Source:LLM in a flash: Efficient Large Language Model Inference with Limited Memory

Edge AI Agents

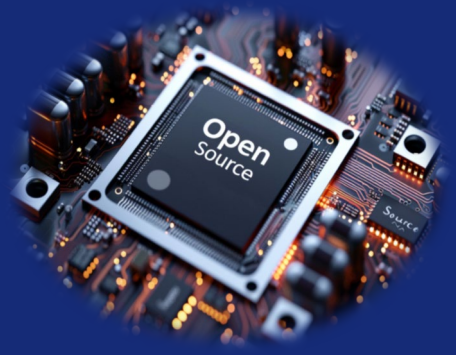


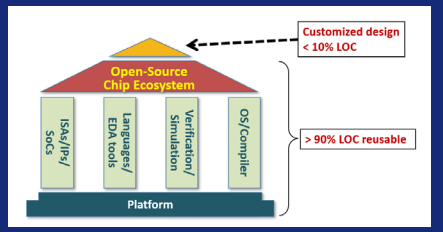
- **Edge AI agents** - autonomous entities powered by AI algorithms that operate directly on edge devices that perform tasks like decision-making, data analysis, and content generation locally, ensuring low latency, enhanced privacy, and reduced data transfer.
- Enable distributed, cooperative edge AI systems, using agents, to work in concert while leveraging on decentralization and collaboration.
- **Capabilities**: decentralised architecture, distributed collaboration, autonomous decision-making, generative capabilities, contextual awareness, learning on the edge, dynamic adaptation, task allocation and optimisation.

Edge AI – Open-Source

- Open source is a paradigm and practice that promotes the free access, use, and modification of software, hardware, or other resources. It emphasises collaboration, transparency, and community-driven development.
- Open source involves distributing the source code or design files, allowing users to study, modify, and improve the product.
- This approach fosters innovation and inclusivity, enabling diverse contributions and widespread adoption
- Open-source platforms and development tools for edge AI is democratising access to AI technologies, enabling a broader base of developers to create edge AI-powered solutions.

"The theory behind open source is simple. In the case of an operating system, the source code is free. Anyone can improve it, change it, exploit it. But those improvements, changes, and exploitations have to be made freely available."
Linus Torvalds





Source; https://www.sigarch.org/embracing-the-era-of-open-source-chips/

Open-source software - focuses on freely accessible source code and the ability to modify and redistribute software.
Open-source hardware - centres on making physical design files and documentation available for replication and modification.
Open-source AI - involves sharing code, models, and datasets for collaborative advancement, emphasising ethical considerations and transparency.

Open Source Initiative (OSI)
Open Source Hardware Association (OSHWA)

Edge AI Future Research

Generative edge AI

The integration of LLMs into edge AI technologies and new distributed agent-based solutions.

Edge AI Interoperability

Heterogeneity of edge AI systems require more efforts for new interoperability among various architectures and multi-modal data types.

Edge AI Tools and Platforms

Automated and generative edge AI design tools and methods.

Edge AI Autonomous Agents

Distributed edge AI agents to coordinate dynamically in real-time. Real-time functions, swarm agents to interact with APIs. Edge AI Defined X (EAIDX).

Energy and Resource Use Efficiency

Energy-efficient edge AI solutions to balance communication overhead with computational efficiency. Optimised resource usage.

Edge AI Trustworthiness

Edge AI dependable systems based on system engineering principles and integrating verification, validation, testing and benchmarking frameworks..

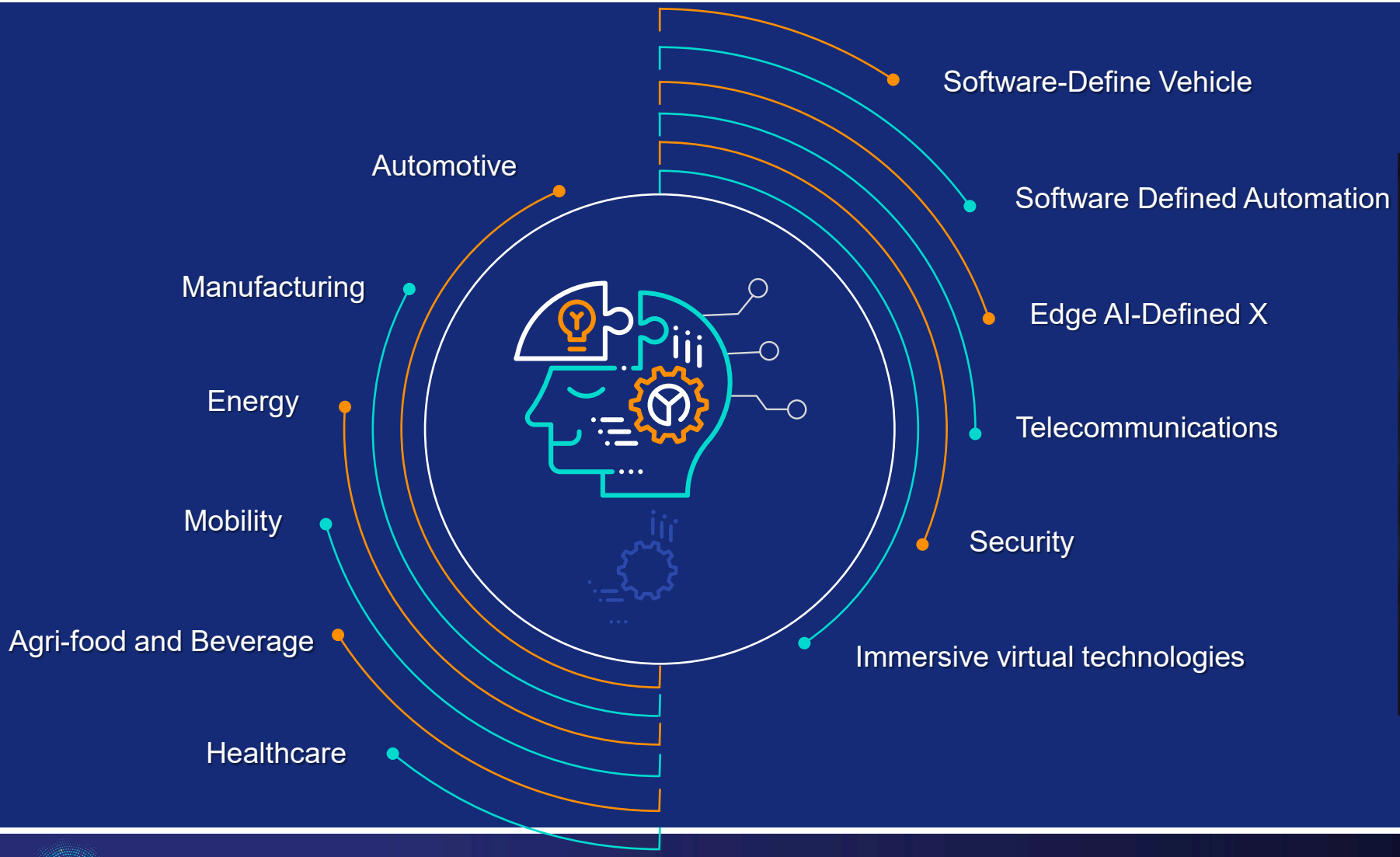
Explainability - Interpretability

Embedded efficient and balanced explainable and interpretable model techniques.

New Learning Paradigms

New resource efficient edge AI learning methods and solutions.

Edge AI Applications



Chips JU EdgeAI Project

EdgeAI (Edge AI Technologies for Optimised Performance Embedded Processing) develops **new electronic components and systems,** processing architectures, connectivity, software, algorithms, and middleware through the combination of **microelectronics, edge AI, embedded systems, and edge computing.**

Start	End
Dec **2022**	Dec **2025**
Duration	Budget
37M	**35.2M€**
Partners	Countries
48	**10**

Accelerate edge AI HW/SW design, innovation and development.

Objectives

Objective 1
Develop secure AI-based edge platforms for end-to-end hardware/software solutions addressing the AI design stack and middleware.

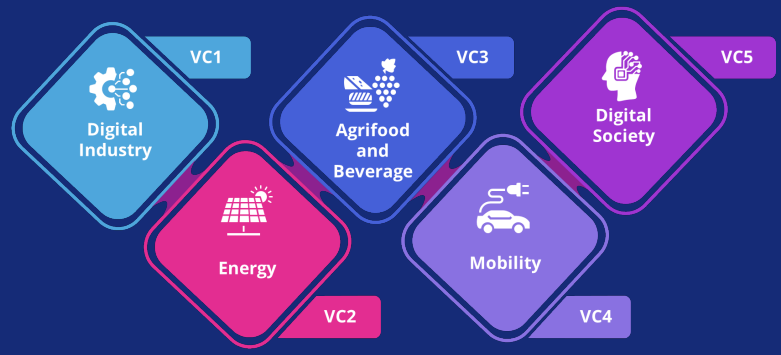
Objective 2
Provide scalable edge AI-based energy-efficient techniques, methods and frameworks supporting different OSs and hardware platforms.

Objective 3
Advance multi core SoC and SoM AI-based designs with embedded hybrid architectures, connectivity and IIoT devices for industrial environments.

Objective 4
Integration of scalable and modular AI Co-design: hardware/software, algorithms, topologies into novel AI open architecture platforms.

Objective 5
Implementation of reconfigurable AI-based architectures for increasing the re-use, updatability, upgradability and service life of AI.

Objective 6
Provide trustworthy and explainable edge-AI by design solutions with real-time operation capabilities and dynamic online learning.



Application Areas

EdgeAI developments implement applications across the edge continuum (micro, deep, meta-edge).

https://www.edge-ai-tech.eu

https://www.linkedin.com/company/edgeaiproject/

Ovidiu.Vermesan@sintef.no (Coordinator)

European Edge AI Ecosystem

EdgeAI
ANDANTE
SMARTY
AI4DI

AI4CSM
TRISTAN
CLEVER
REBECCA

dAIEDGE
SMARTEDGE
NEUROKIT2E
LoLiPoP IoT

European Edge AI Ecosystem

European Conference on EDGE AI Technologies and Applications - EEAI

Connecting the future and driving the next wave of technological advancements for a better world.

21-23 October 2024, Cagliari, Sardinia, Italy

Thank You!