RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

cea cnrs Inria

FRANCE 2030
PROGRAMME DE RECHERCHE
INTELLIGENCE ARTIFICIELLE

AdaptING :

Adaptive architectures for embedded artificial INtelliGence

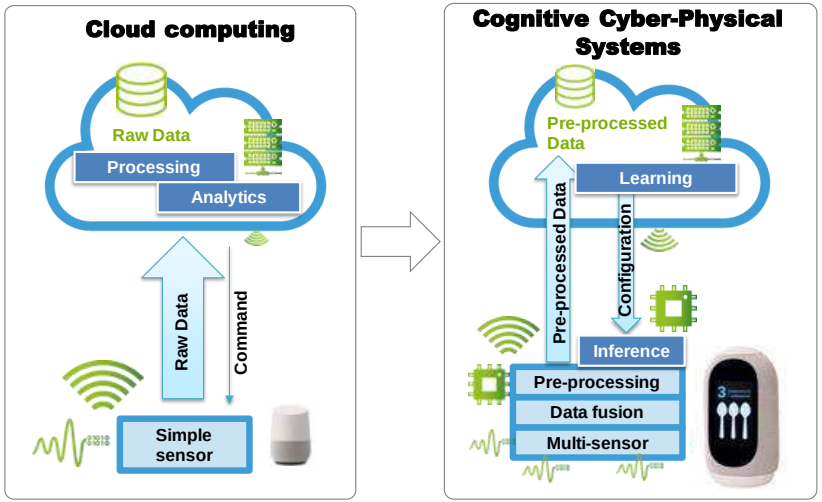
Alberto Bosio (ECL-INL)
Ivan Miro-Panades (CEA-List)
Sébastien Pillement (Nantes Université – IETR)
Philippe Coussy (Université Bretagne Sud / Lab-STICC)
Elisa Fromont (Université Rennes, IRISA /INRIA)
Andrea Pinna (Sorbonne Université LIP6, SU-Lip6)

anr agence nationale de la recherche

Motivation

Current AI applications are based on cloud servers.

- Data privacy issues
- Bandwidth limitations
- Excessive power consumption
- High latency response



Current model

Edge AI

Source: Denis Dutoit (CEA)

Versatile and intelligent edge devices will address the *new era*

Frugal embedded AI architectures with local learning are key to overcome current limitations

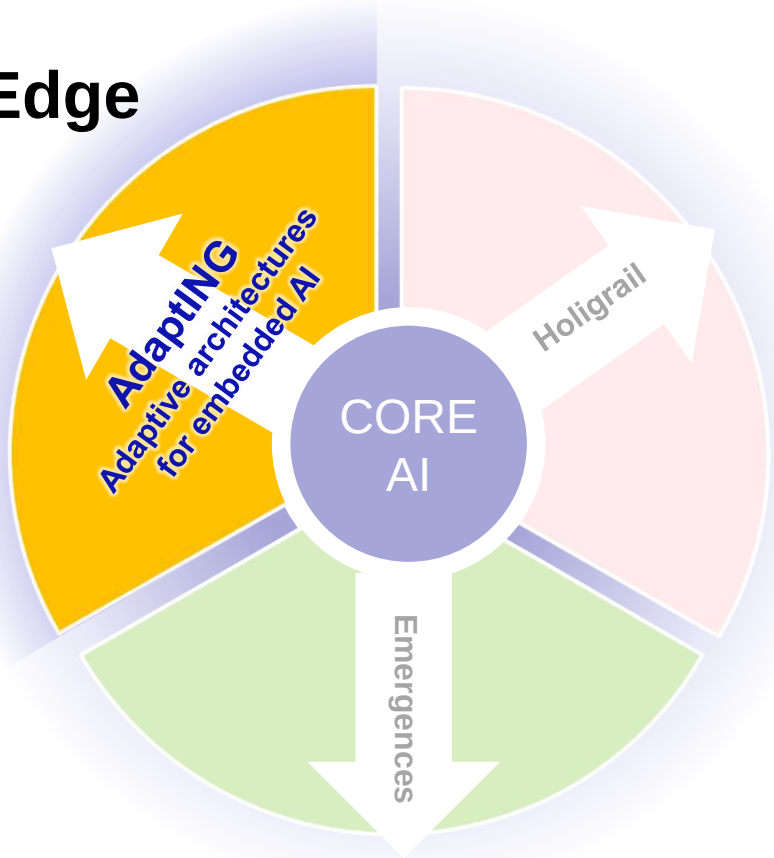
AI reliability (i.e., capability to provide correct results) is a must have

AdaptING targets on developing these new architectures

From Cloud to Edge

Challenges:

1. Flexibility
2. Learning-on-Chip
3. Energy efficiency
4. Reliability



Consortium members:

Challenges: Flexibility

A flexible architecture able to run different AI algorithms

Taking into account the application constraints in terms of accuracy, energy, latency and reliability

- For the same result, the architecture can be:
  - Executed faster with high power consumption and low latency

  - Executes slowly with lower power consumption but higher latency

This flexibility contributes towards making the hardware future-proof, leading to higher sustainability

Challenges: Learning-on-Chip

Evolves the AI model retrain paradigm from server to edge device

Learning-on-chip:

- Guarantee privacy
- Energy reduction since no data has to be systematically transferred to and from the cloud

Training at device level is not trivial 

- Not possible to embed huge databases in the device
- Catastrophic forgetting issues when retraining an AI model
- Energy budged is reduced

Challenges: Energy efficiency

High-throughput architectures (GPU, TPU)

- Good efficiency but high power consumption ⇒ Not for embedded solutions

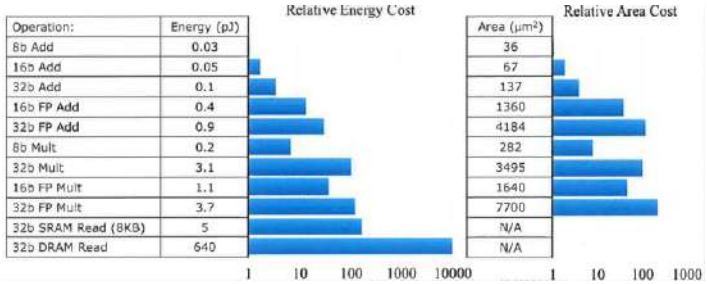
Specialized AI architectures

- High energy efficiency but reduced flexibility

Achieving at the same time flexibility and energy efficiency is a major challenge !

The main energy efficient killer is the data movement between device memory and processing elements.

⇒Reducing the data movement is key to minimize the power consumption



Operation:	Energy (pJ)	Area (μm²)
8b Add	0.03	36
16b Add	0.05	67
32b Add	0.1	137
16b FP Add	0.4	1360
32b FP Add	0.9	4184
8b Mult	0.2	282
32b Mult	3.1	3495
16b FP Mult	1.1	1640
32b FP Mult	3.7	7700
32b SRAM Read (8KB)	5	N/A
32b DRAM Read	640	N/A

Source: Mark Horowitz, “Computing's energy problem (and what we can do about it)”, ISSCC 2019

Challenges: Reliability

Similar to traditional computing hardware, is subject to errors that can have several sources: variability in fabrication process parameters, latent defects or even environmental stress

One of the overlooked aspects is the role that hardware errors can have in AI decision

There is a common belief that AI applications have an intrinsic high-level or resilience w.r.t. errors and noise

However, Hardware for AI is not always immune to HW errors

Traditional online fault-testing is not straightforward on AI-HW

Safety techniques like Dual-Core Lock Step (DCLS), doubles the area and the power consumption

Consortium members

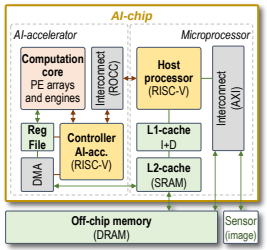
École Centrale de Lyon, Institut des Nanotechnologies de Lyon (ECL-INL)

Project coordinator

Previous works in the field

- Hardware accelerators in deep-learning applications through Approximate Computing
- Design and implement a precision re-configurable systolic array-based hardware accelerator for DNNs
- Analyze the impact of hardware faults on the reliability, safety, trust, and explainability of AI decisions

Alberto BOSIO



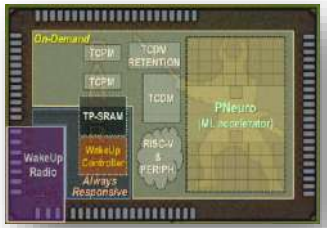
Commissariat à l'énergie atomique et aux énergies alternatives (CEA)

Project coordinator

Previous works in the field

- Open-source AI training tool (N2D2)
- ASIC design with low power techniques for High Performance Computing
- Design of AI architectures optimized for embedded platforms

Ivan
MIRO PANADES



SamurAI: Edge AI architecture

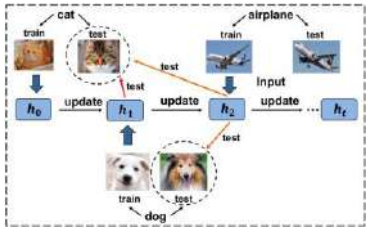
Consortium members

Université Rennes, IRISA /INRIA

Previous works in the field

- Algorithms related to federated, incremental and few shot learning
- Algorithms related to “explainable AI” in a adapting system ⇒ learning shift monitoring
- Strong expertise in fault-tolerant architectures and fault mitigation in existing architectures

Elisa Fromont



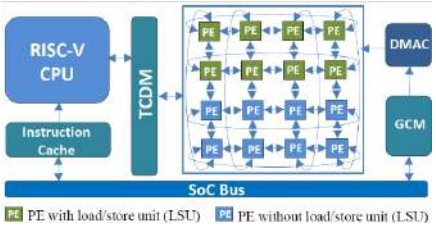
Incremental learning

Université Bretagne Sud / Lab-STICC

Previous works in the field

- AI accelerators targeting Content Addressable Memories (CAM) and prototyped on FPGA
- Coarse-grained reconfigurable architecture and associated compilation tool have been proposed for low power design
- Dataflow and near-memory computing

Philippe COUSSY



CGRA SoC

Consortium members

Sorbonne Université LIP6 (SU-Lip6)

Previous works in the field

- Embedded smart system design for medical device and diagnosis support
- Frugal AI and incremental learning as well as lightweight deep network design
- Precision Tuning Using Stochastic Arithmetic
- High-coverage compact functional test set for neuromorphic processors

Andrea PINNA

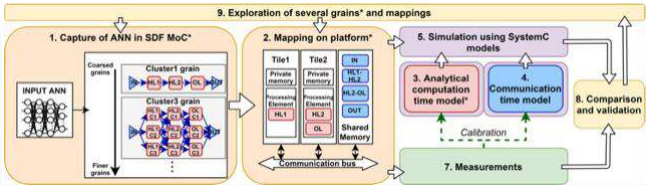
Cyclope: polyps detect capsule

Nantes Université - IETR

Previous works in the field

- High Level models for performance evaluation of embedding AI and related design space exploration methodology
- Design of reliable architecture using reconfiguration integrating fault-tolerant mechanisms
- Specific methodology to evaluate different fault-tolerant strategies in the early-stage of the design process for critical applications

Sebastien PILLEMENT



AI performance estimation

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

FRANCE 2030
PROGRAMME DE RECHERCHE
INTELLIGENCE ARTIFICIELLE

WP0 - Project management

WP1 - Machine Learning and Embedded Systems

AI Algorithms [Learning & Inference]

AI Hardware

CPU SoC FPGA

Profiling & Trustworthiness

Kernels

Design Space

Mapped Kernels [Systolic, In-Memory, MAC]

WP3 - Contexte & Hardware Aware

KPIs from Real Data
Latency - Precision - Energy
Reliability

In Field Validation

Adaptable AI

Rules

On-Line Monitoring

CPU SoC FPGA ASIC

Mapped Kernels

HW Monitors

Interconnect

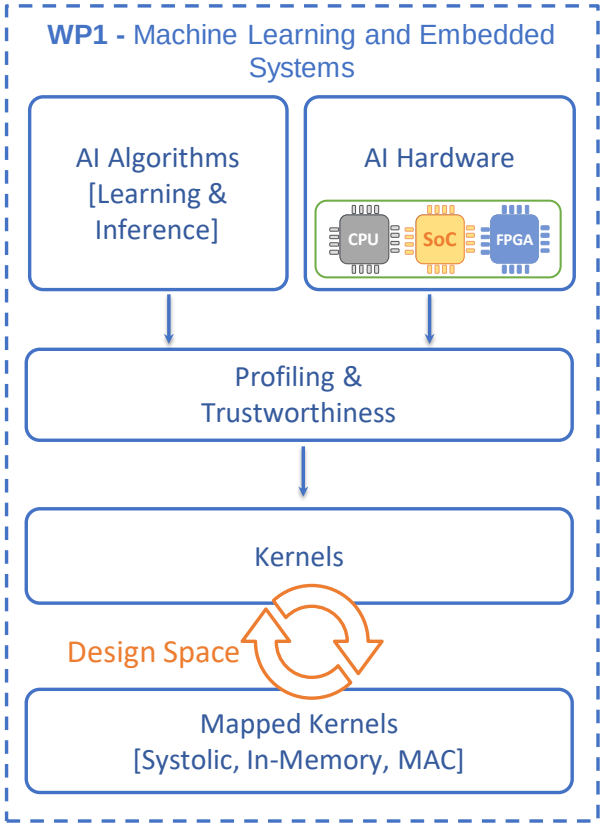
Embedded System Mapping

HW-SW CoDesign

Synthesis Flow

Reconfigurable Architectures

WP2 - Reconfigurable computing and architecture (adaptation/optimization/acceleration)

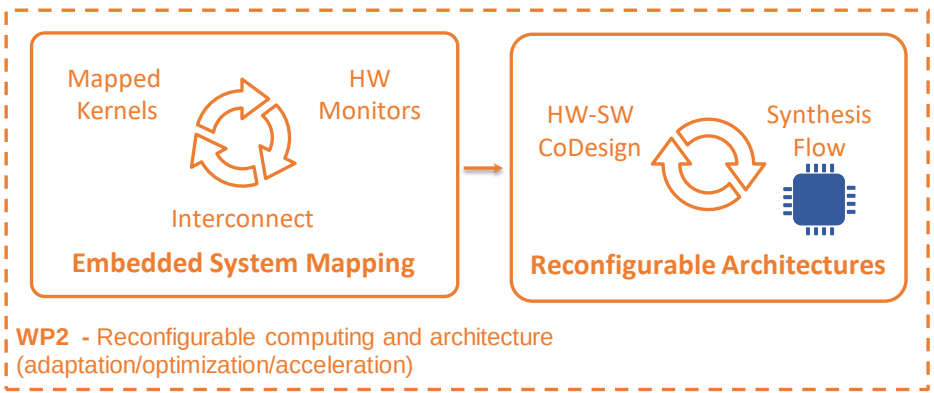


- WP1 **explore AI** (learning and inference models) at the algorithmic level
- **Identify AI algorithms** and use cases that will be used during the project.
- **Profiled AI models** on HW architectures (bottleneck, memory requirements and energy consumption)
- *Emergences* and *HOLIGRAIL* computing kernels are also explored
- Outcomes: mapped kernels used by WP2 and the use case definition that will be used in WP3.
- 9 Ph.D. thesis (6 of them will be co-supervised by two partners)
- 1 Post-Doc

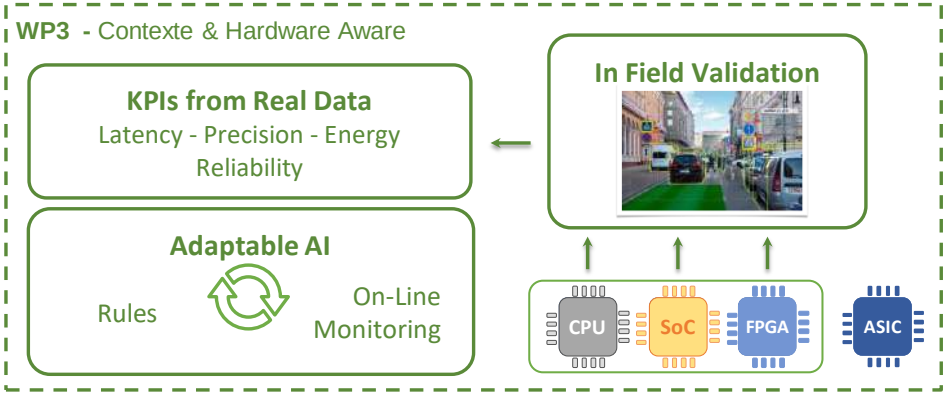
- Research of **methodologies and tools** for designing AI reconfigurable architectures.
- **Explored mapped kernels from WP1**: inference, training and monitoring kernels.
- Different HW/SW implementations are considered, varying in energy efficiency, latency, precision, and trustworthiness..
- The goal is to **select the kernel to be used and how they have to be interconnected** in terms of technology and topology to generate the AI reconfigurable architectures for a given use case.
- The resulting architectures will be heterogeneous, including components like CPUs, SoCs, FPGAs, and ASICs.
- HW/SW co-design and synthesis flow to estimate the energy efficiency and the latency of the reconfigurable architecture

- Synthetized reconfigurable architectures will be used in WP3 for the context aware monitoring and in-field validation

- 9 Ph.D. thesis (8 co-supervised by two partners)
- 2 Post-Docs



- WP3 focuses on **online monitoring** and validation of AI architectures.
- It builds upon the AI architectures provided by WP2, incorporating **reconfigurability mechanisms and hardware (HW) monitors**
- WP3 designs and implements online monitoring solutions for these architectures.
- Using AI architectures in the field with real data
- Online monitoring determines the status of running AI algorithms using information from embedded HW monitors
- **Status is evaluated** in terms of Latency, Accuracy, Energy, and Trustworthiness (KPIs).
- In mission mode, external context and data can change, triggering reconfiguration based on status
- Reconfiguration can **involve training AI with new data, modifying AI structure** to adapt to new data or adjust KPIs, or signaling inadequate reconfiguration mechanisms.



Nantes Université

- Solutions for inadequate mechanisms include modifying AI architecture or identifying new AI algorithms
- 7 Ph.D. theses (2 co-supervised by two partners)

AdaptING community

- 27 permanent researchers
- 3 research engineers
- 25 PhD students
- 2 PostDoc students
- 10 internships students

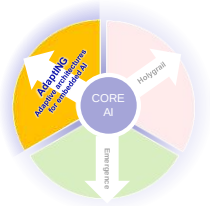
KPI evaluation

AI, especially deep neural networks, presents challenges for hardware due to its complexity, including energy consumption, memory requirements, computation speed, and scalability.

There is a need for AI hardware accelerators capable of handling high-dimensional and computationally-intensive workloads.

AdaptING aims to design adaptable AI-HW architectures that are flexible like general-purpose processors and more efficient than current AI accelerators.

KPI	Examples
Accuracy	Number of correct inferences/total inferences
Energy	Joule/Inference; Joule/training
Latency	Second/inference; second/training
Adaptability	Highest variation of conditions (internal/external) that can be tolerated thanks to the reconfigurability of the architecture
Trustworthiness	Inference/training confidence level



Project outcomes

Algorithmic

- Identification and characterization of applications and AI models (including real-time, mission- and safety-critical applications)

Architecture

- Development of flexible architectures

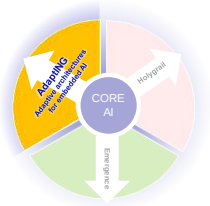
System

- Virtual prototype of the AI-HW architectures as an open-hardware platform for validation through simulation/emulation.
- Deployment of different AI models

Comparison of the real prototype (physical implementation) with state-of-the-art open AI-HW architectures such as the Pulp platform, Gemini, and Eyeriss, based on Key Performance Indicators.

Federation of French Hardware AI community

⇒ GDR SoC2, GDR IASIS, GDR BioComp, GDR RADIA, Summer School



Exploitation

The results of AdaptING can be further exploited through **technology transfers** with the objective to be integrated in **industrial products** and **open-source platforms** for embedded AI.

Examples are the N2D2 platform, PNeuro and NeuroCorgi platforms from CEA, PULP ecosystem from ETHZ and industrial solutions provided by companies (i.e. Greenwaves and Asygn).

Maturation of architectures: from low TRL (4/5) on AdaptING to high TRL (6/7) on DeepGreen

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

cea

cnrs

Inria

FRANCE 2030

PROGRAMME DE RECHERCHE

INTELLIGENCE ARTIFICIELLE

Retrouvez toutes nos actualités

anr agence nationale de la recherche