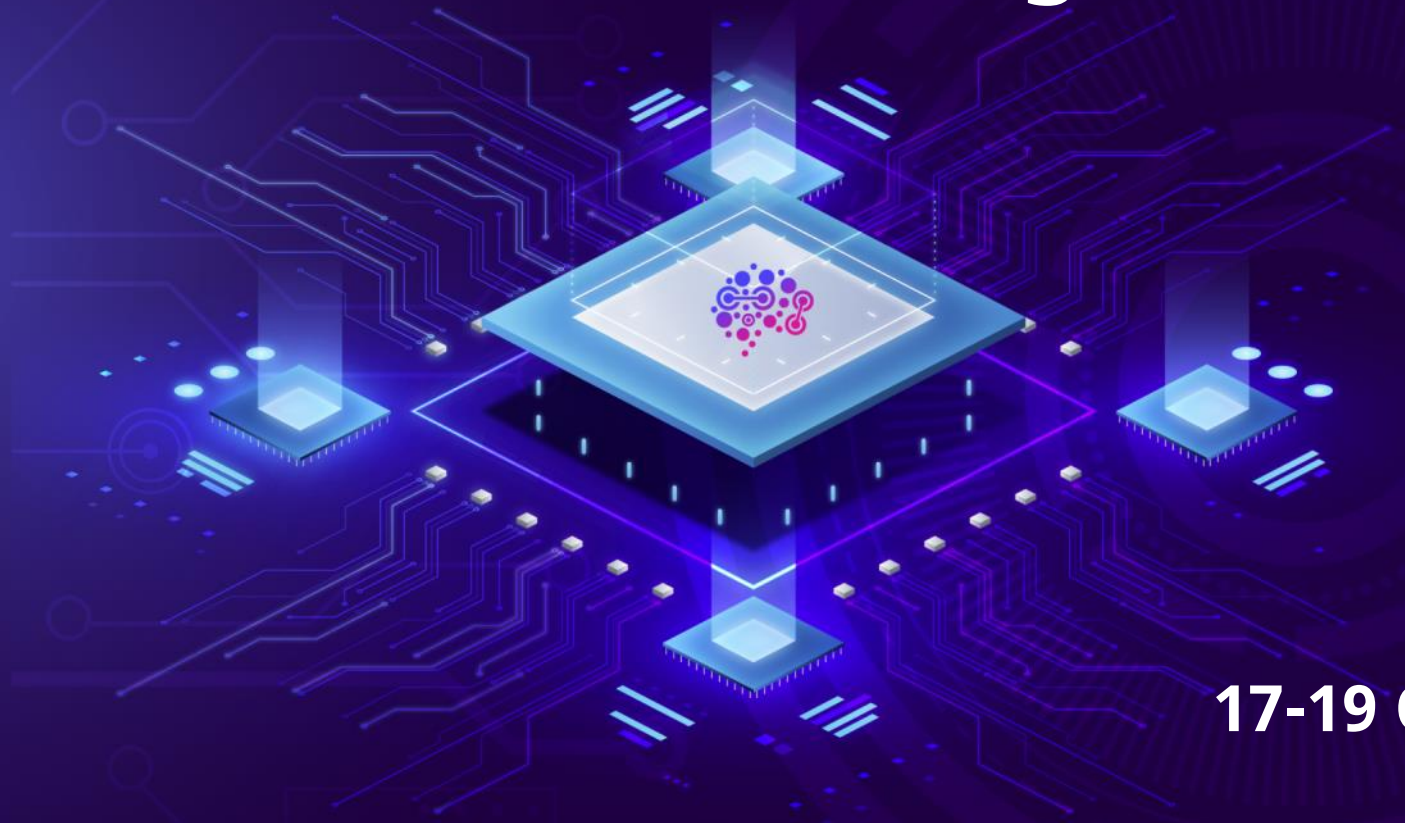
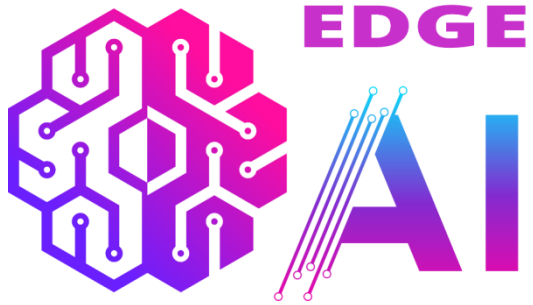
European Conference on EDGE AI Technologies and Applications - EEAI



Advancing emerging edge AI technologies and driving next generation intelligent applications.

17-19 October 2023, Athens, Greece

European Conference on EDGE AI Technologies and Applications - EEAI



AIDGE : a new framework for Deep Neural Network development, training and deployment on the Edge

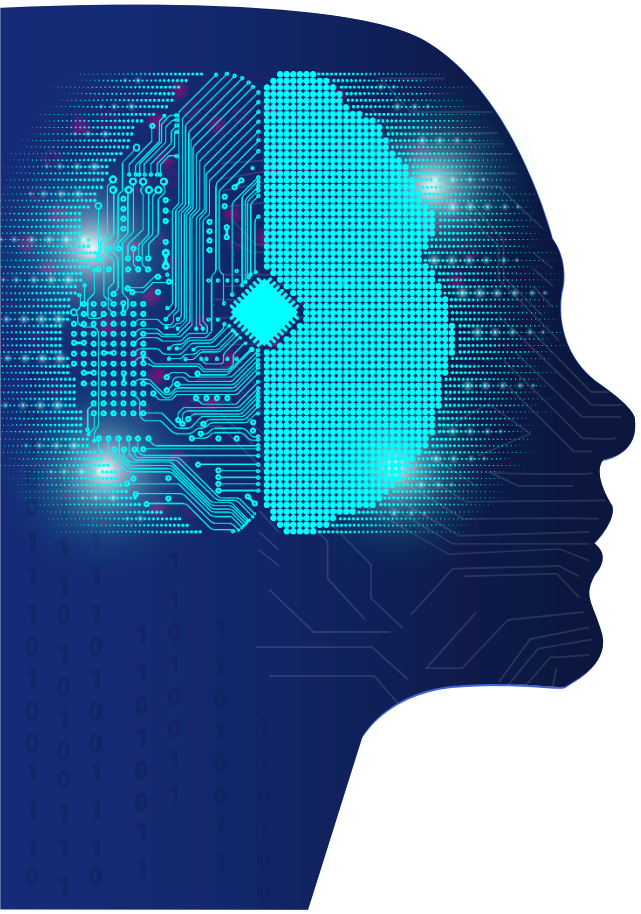
Fabian CHERSI, *CEA-List, France*

Olivier BICHLER, Cyril Moineau, Maxence Naud, Laurent Soutier, Vincent Templier, Thibault Allenet, Inna Kucher

17-19 October 2023 Athens, Greece

Copyright © 2023

Presentation Outline



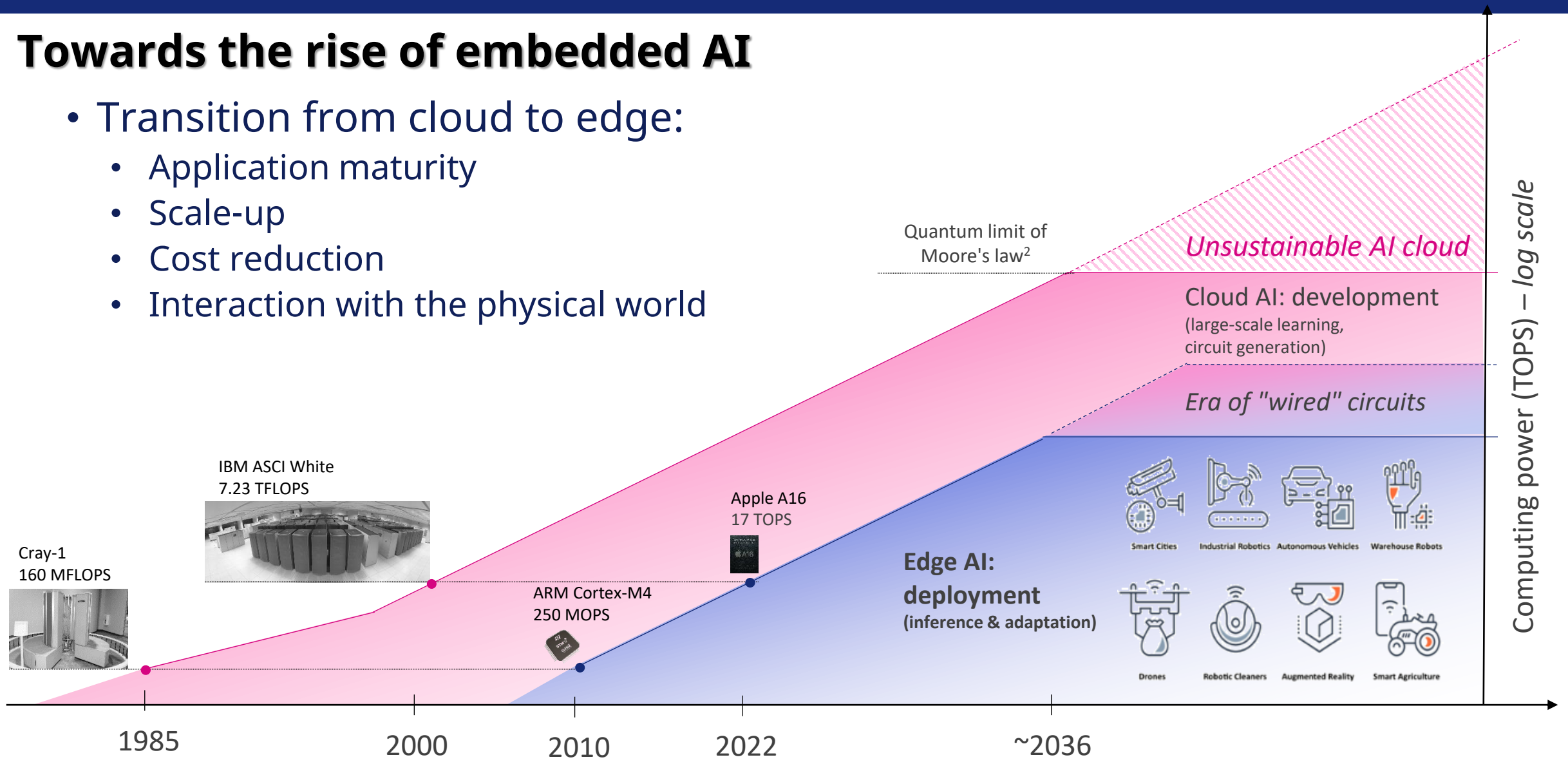
- Introduction
- State-of-the-art
- Research Goal
- Methodology
- Design and Implementation
- Experimental Results
- Future Research
- Conclusion

Copyright © 2023

Introduction

Towards the rise of embedded AI

- Transition from cloud to edge:
  - Application maturity
  - Scale-up
  - Cost reduction
  - Interaction with the physical world



Copyright © 2023

State-of-the-art

Tools for embedded AI (with strong orientations and dependencies)

Hexagon

Qualcomm snapdragon

android

NNAPI

SAMSUNG Exynos

iOS

TensorFlow Lite

Standard and for Microcontrollers

ARM Cortex-M

Adreno GPU

Qualcomm snapdragon

android

AMD XILINX

Vitis AI

TensorRT

VTA: Versatile Tensor Accelerator

tvm

Standard and Micro

- ***Operator level***
- ***Loop tiling***
- ***LLVM backend***

CoreML

android

NNAPI

intel

OpenVINO, oneDNN

TensorRT

AMD XILINX

Vitis AI

Qualcomm snapdragon

SNPE SDK

arm

ARM NN

AMD ROCm

ONNX RUNTIME

- ***Wrapper on proprietary libraries***

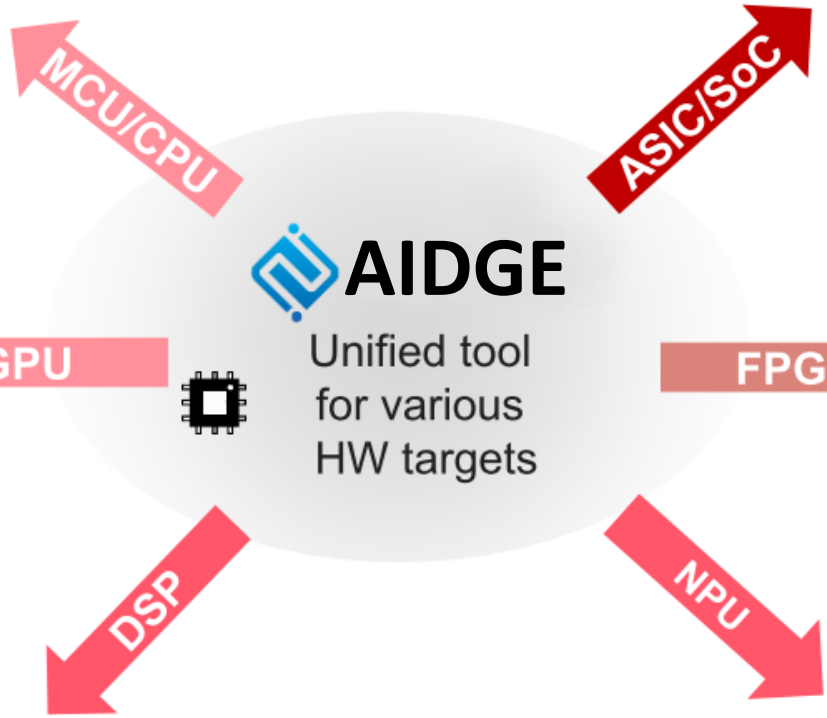
Copyright © 2023

Research Goal

Embedded HW code generation

General purpose architecture:
+ Complex tasks and large instruction set
+ Sequential computation
+ Large flexibility
- Memory handling
- Low core count

Fixed architecture:
+ Performance (low latency and power consumption)
+ Low cost on volumes
- Single specialized task
- High entry price (mask fabrication)



Parallel compute units:
+ High performances for AI
+ Highly parallel thanks to a large number of cores
- Large power consumption
- Large hardware footprint

Programmable architecture:
+ Flexible
+ Real time computation (low latency)
- Programming complexity
- Component cost

Specialized architecture for signal processing:
+ Mostly sequential
+ Large instruction set

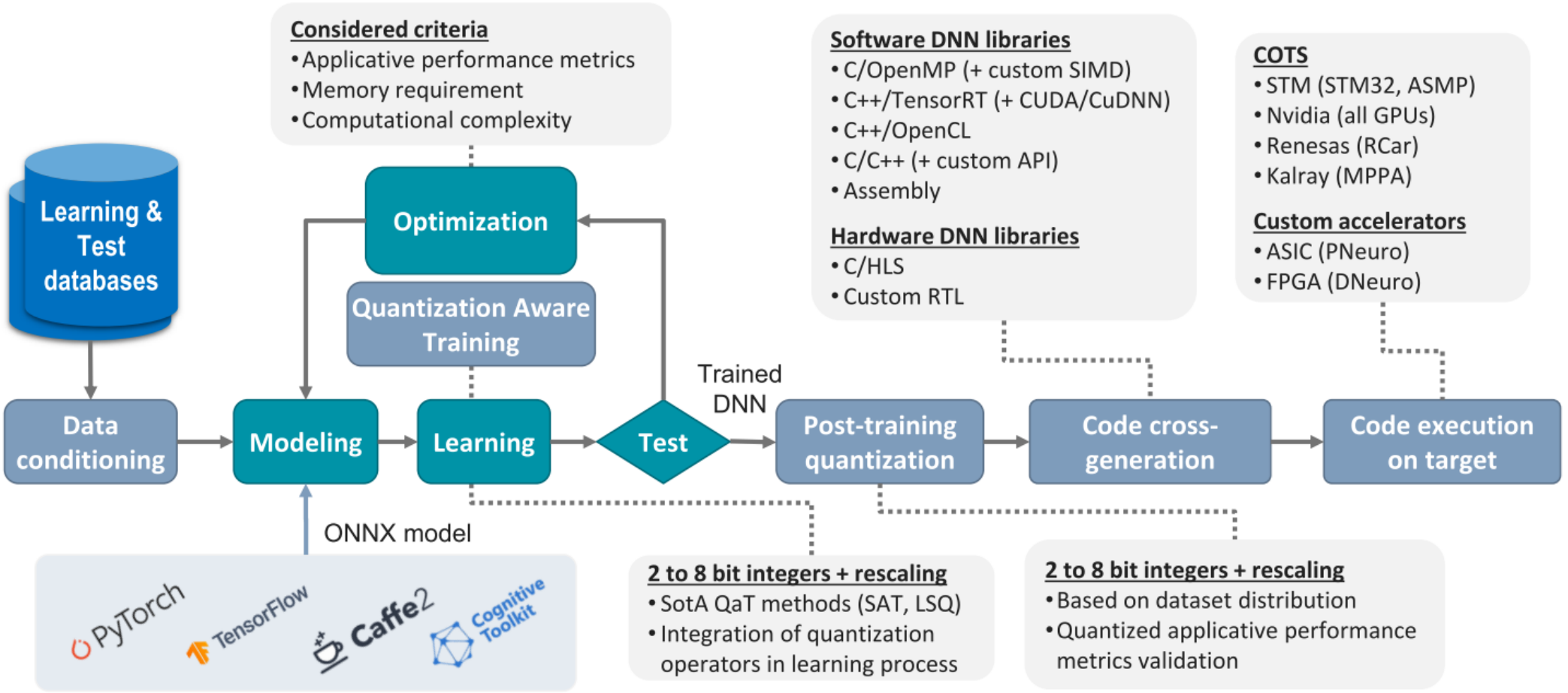
Specialized processor for neural networks:
+ Low energy consumption
+ Low hardware footprint
- Proprietary design
- Specific SDK

Copyright © 2023

Methodology

The AIDGE Platform

A unique platform for the design and exploration of DNN applications



Copyright © 2023

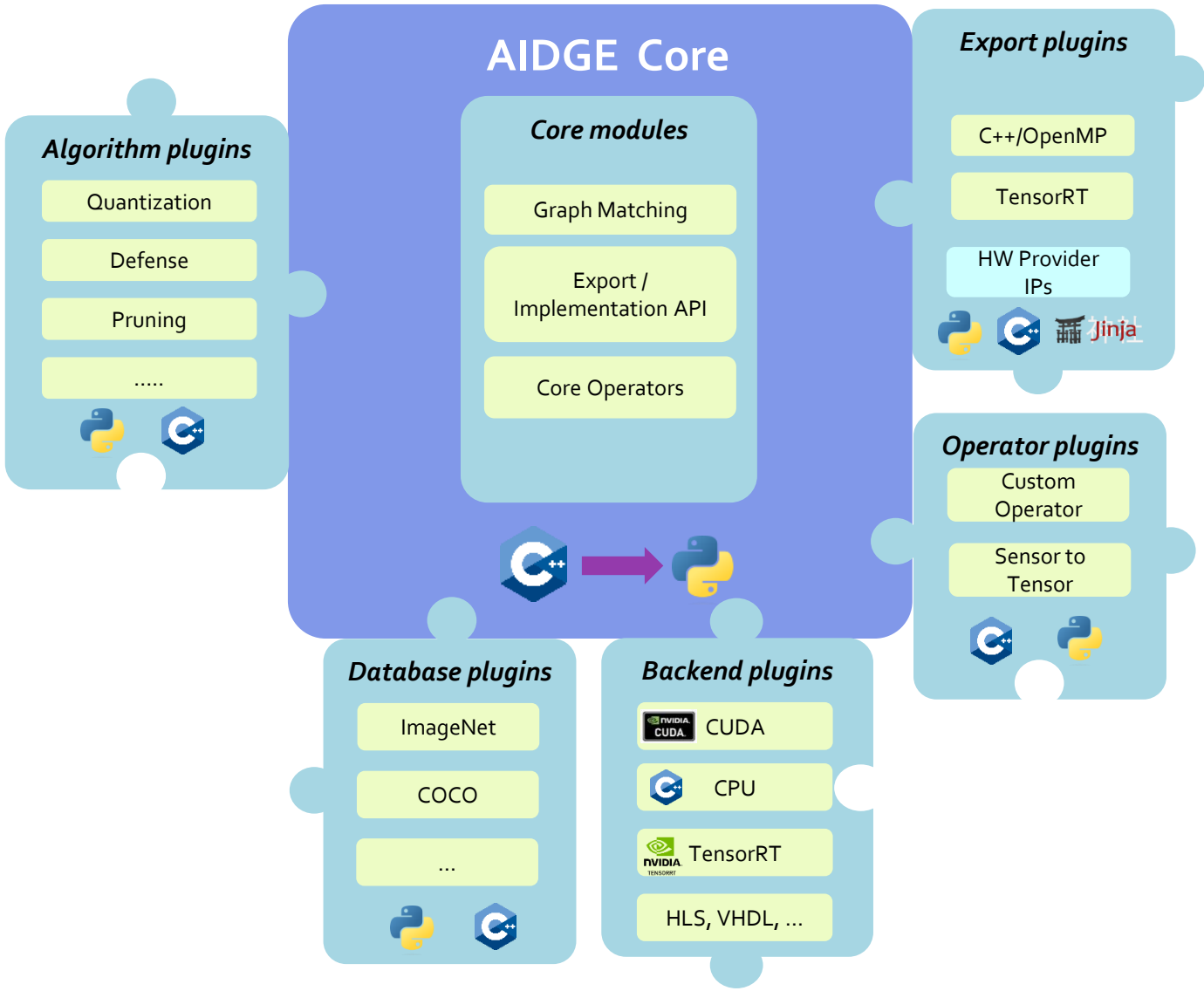
Design and Implementation

Modular Architecture

- **Easy maintenance**
- **Integration of new contributions**
- **AIDGE Core is very light**
  - Compilation only of necessary modules
  - The Core can be executed on multiple devices

Python and C++

- Optimized with C++
- All functions are utilizable in python thanks to bindings
- Rapid prototyping possible with python (low TRL)
- Interoperability for the final user

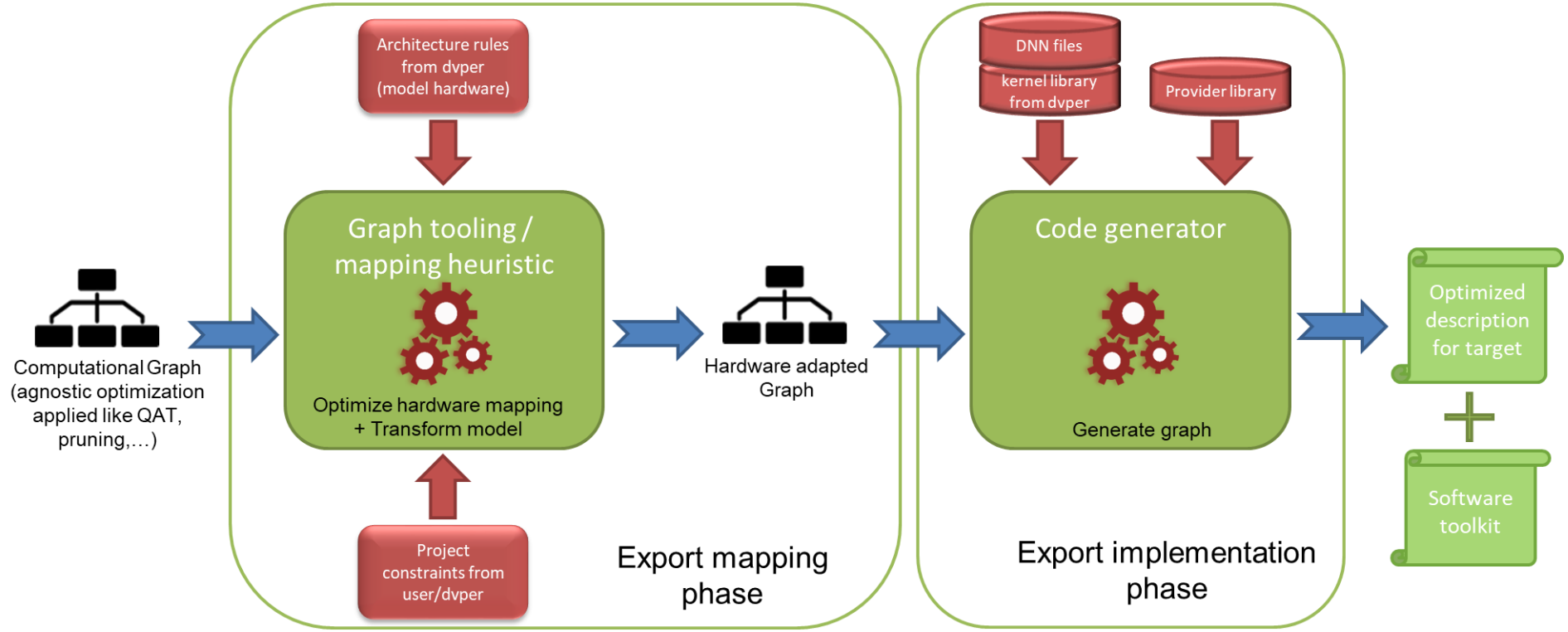


Copyright © 2023

Design and implementation

Important characteristics of this framework:

- It will produces a source code that is interpretable and verifiable by the user and the developer
- It will take into account the specifics of the target hardware already at the graph level
- It will provide a library of kernels and primitives for all the operators based on the low-level libraries provided by the hardware vendors



Copyright © 2023

Experimental Results: Python example (net import)

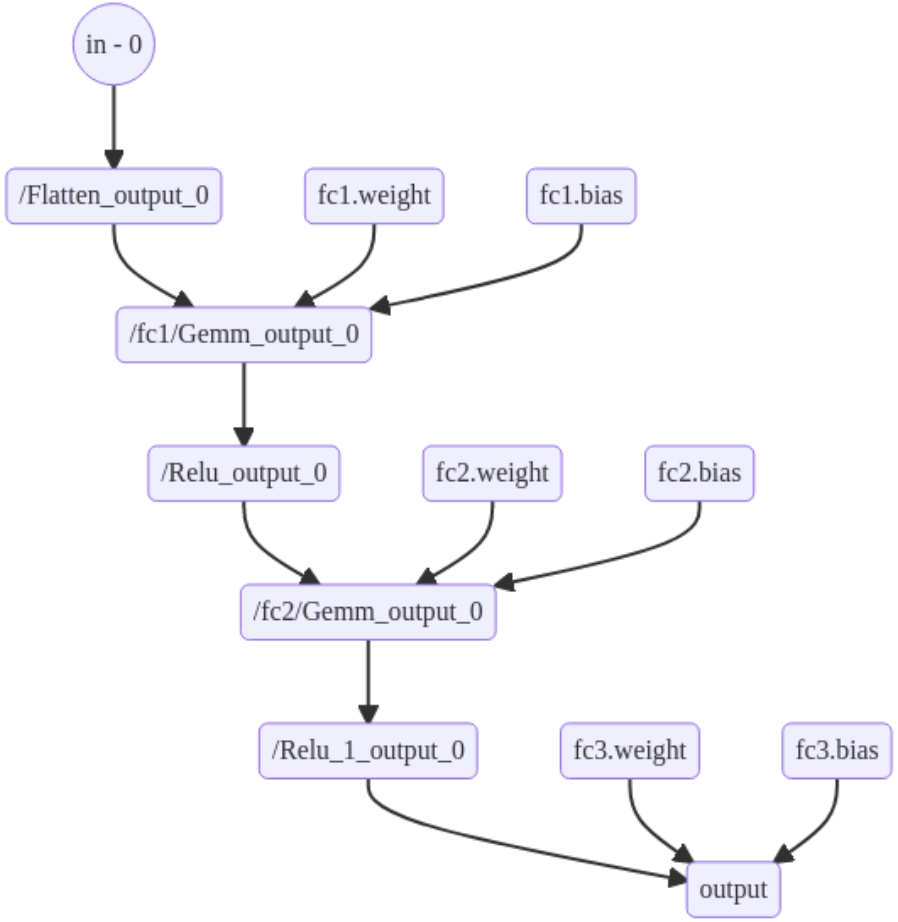
Aidge is built on a C++ Core Library that interfaces with multiple modules binded to python libraries.

aidge_core offers all the basic functionalities to create and manipulate the network/graph representation

aidge_backend_cpu is a module providing the linux cpu implem graph

Python example :

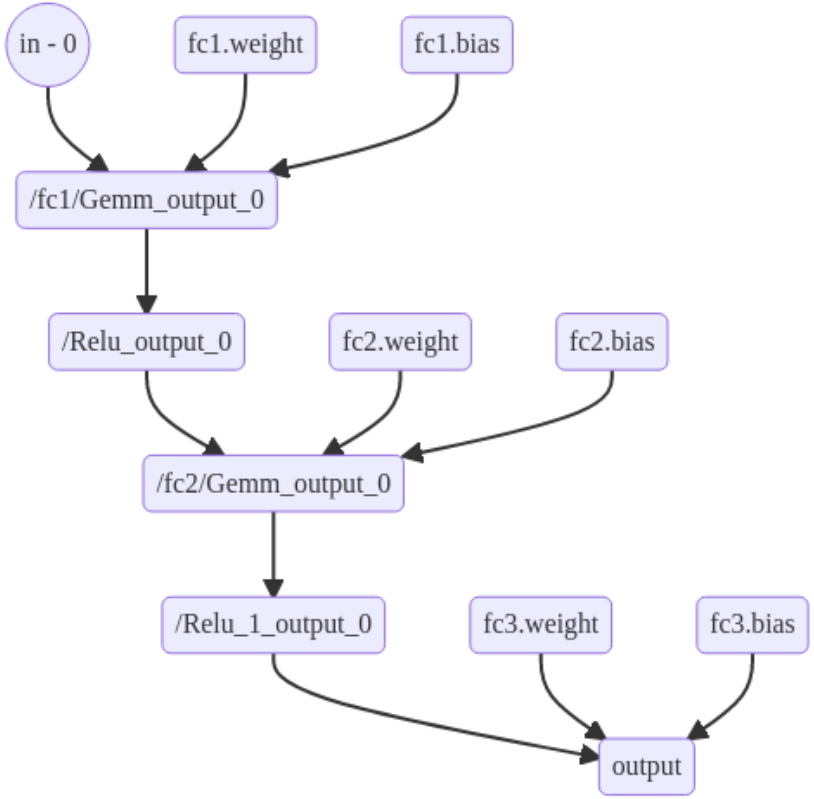
```
import aidge_core
import aidge_backend_cpu
import aidge_onnx
import numpy as np
import matplotlib.pyplot as plt


model = aidge_onnx.load_onnx("MLP_MNIST.onnx")


model.save("myModel")   # Save in mermaid format
```



Copyright © 2023

Experimental Results: Python example (net manipulation)

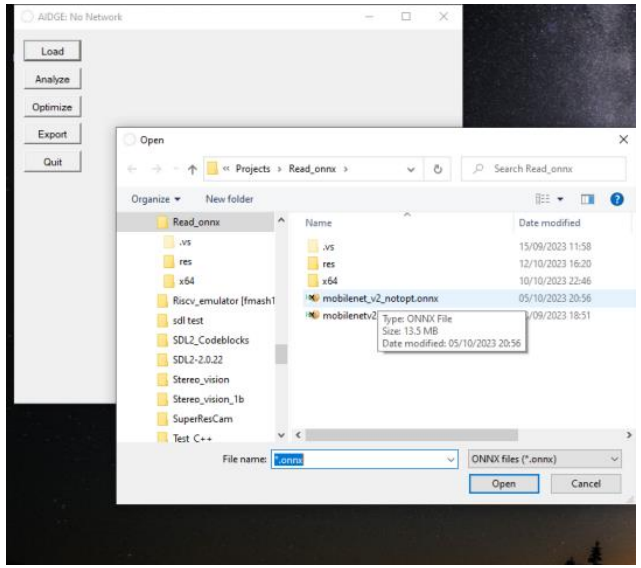
```
# Node Regex:  : Describe the conditions an operator has to fulfill
nodesRegexes = {}
nodesRegexes["Flatten"] = aidge_core.NodeRegex("Flatten")
# Graph Regex : Describe the graph topological pattern
graphRegex = ["Flatten;"]


graphMatching = aidge_core.GRegex(nodesRegexes, graphRegex)
all_match = graphMatching.match(model)


for mn in all_match.get_match_nodes():
  aidge_core.remove_flatten(mn)


model.set_datatype(aidge_core.DataType.Float32)
model.set_backend("cpu")
scheduler = aidge_core.SequentialScheduler(model)
scheduler.forward()    # Run inference !
```

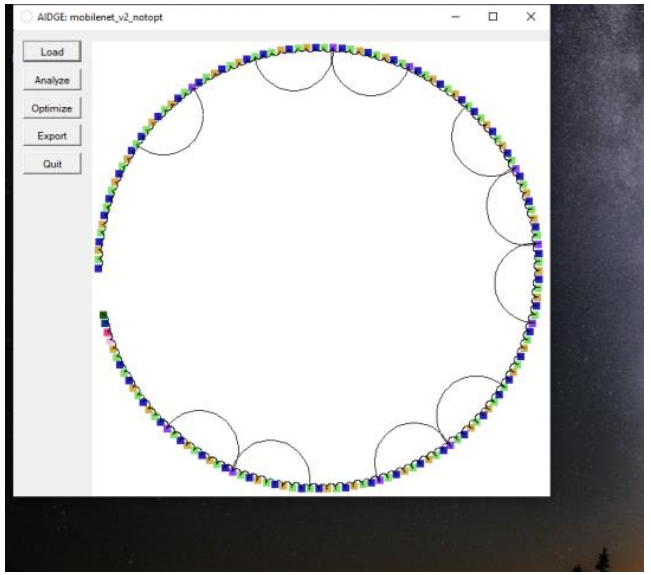


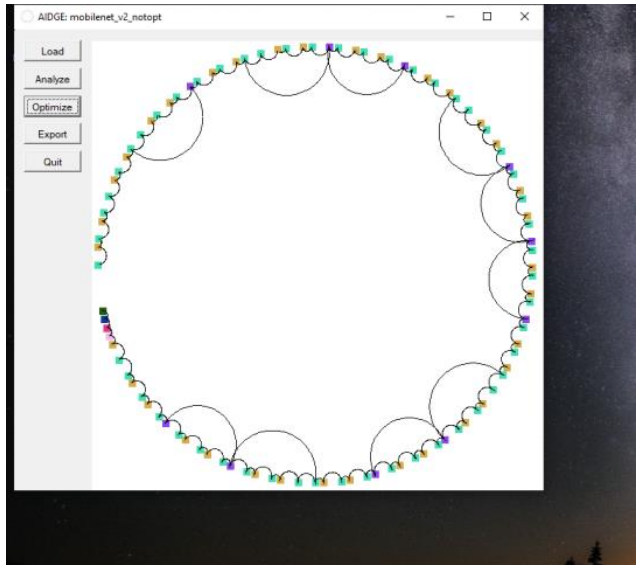
Copyright © 2023

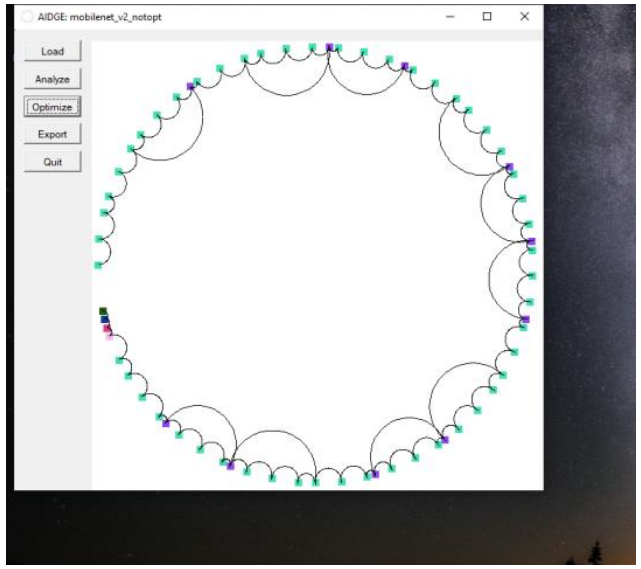
Experimental Results: AIDGE GUI

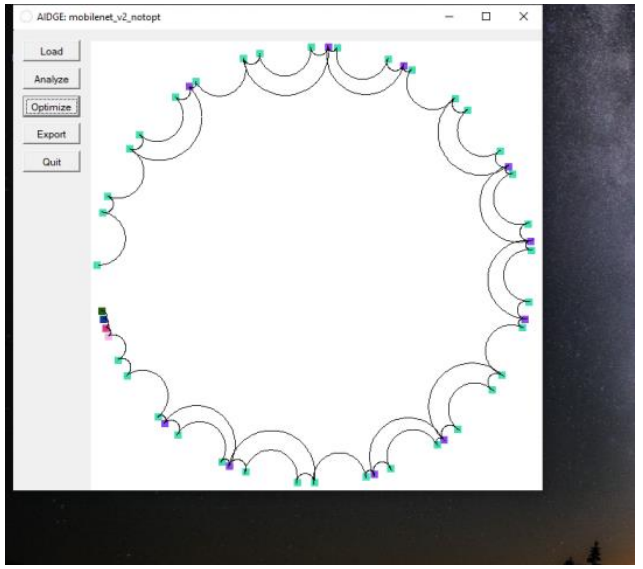
A GUI for AIDGE











Copyright © 2023

Future Research

- **Interoperability** :
  - Keras, PyTorch
- **Graph Export** :
  - Pneuro
  - SDK with guidelines
  - C++ STM32
- **Graph optimization** :
  - Tiling
  - Quantization Aware Training
- **Supported models (inference)**
  - Object detector
  - Semantic segmentation
  - Multimodal
  - Attention (Transformers)
- **Supported models** (learning)
  - CNN
- **Tutorials**
- **Graph Export** :
  - NeuroCorgi
  - Heterogeneous graph export
  - C++ RISC-V
- **Graph optimization**
  - Mixed precision quantization
  - Compression et Pruning
  - Spike coding
- **Benchmark KPI**
  - Perform Life Cycle Analysis
- **Supported models (learning)**
  - Attention (Transformers)
- **Ensure Robustness**
  - Ensemble hash defense
- **Learning on edge**

Copyright © 2023

Conclusions

Provide a DNN framework suitable also for embedded applications

Develop state-of-the-art hardware-aware optimization techniques

Reunite deep learning and neuromorphic

Provide generic and reusable HW building blocks and methodologies

Answer industrial challenges for usability of AI

Copyright © 2023

Thank You

For your attention

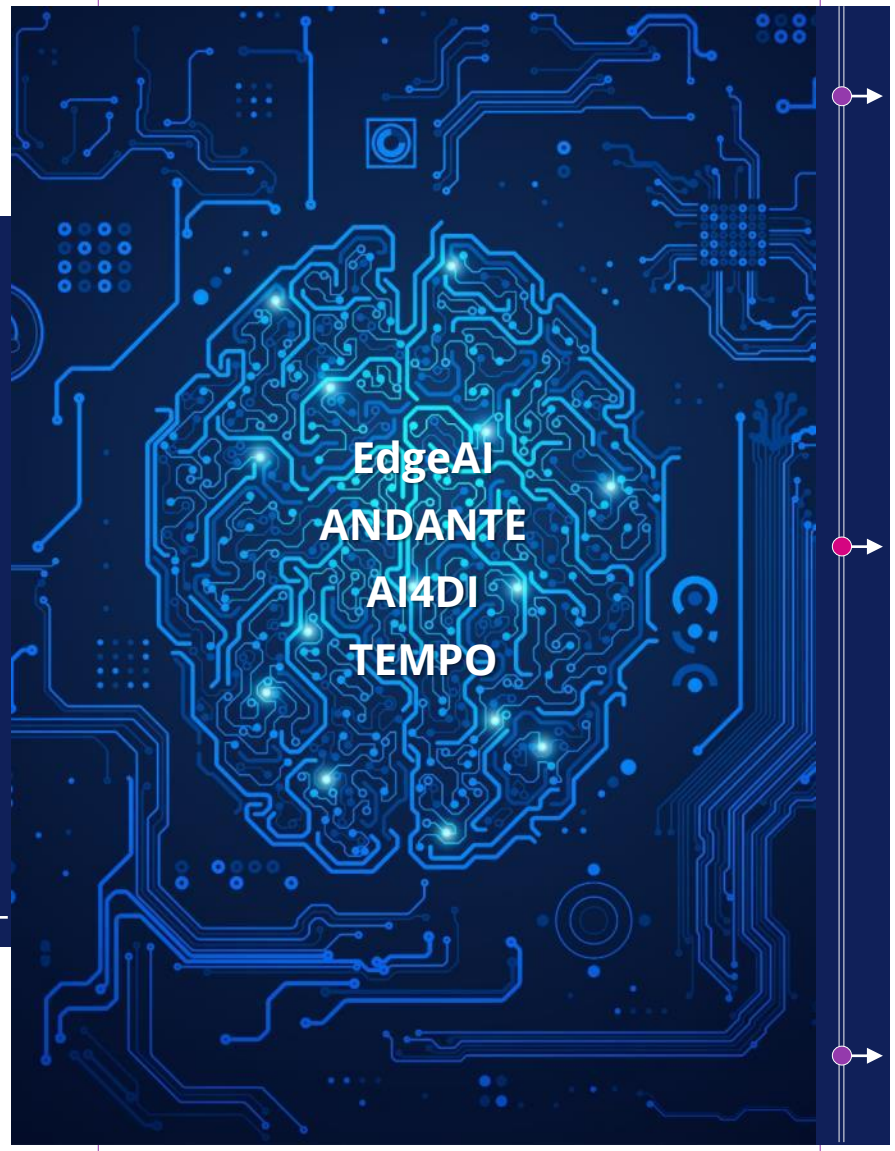
@ fabian.chersi@cea.fr

www.linkedin.com/in/neurokit2e-european-project/

https://projects.eclipse.org/projects/technology.edge

Copyright © 2023

Event Organisers



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. www.edge-ai-tech.eu

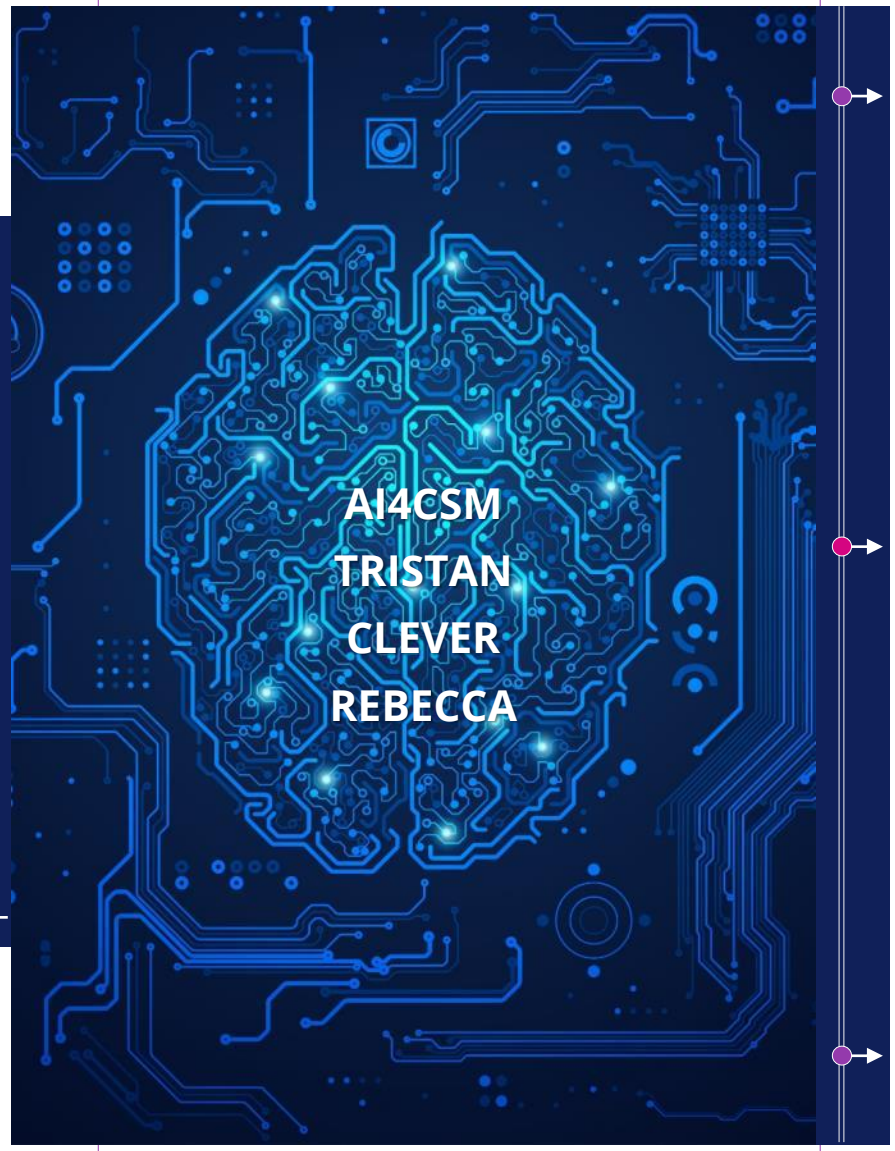
The ANDANTE project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 876925. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Belgium, France, Germany, The Netherlands, Portugal, Spain, Switzerland. www.andante-ai.eu

The TEMPO project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826655. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Belgium, France, Germany, The Netherlands, Switzerland. www.tempo-ecsel.eu

The AI4DI project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826060. The JU receives support from the European Union's Horizon 2020 research and innovation programme and the national authorities. www.ai4di.eu

Copyright © 2023

Event Organisers



AI4CSM (Automotive Intelligence for Connected Shared Mobility) develops advanced electronic components and systems (ECS) and architectures for future mass-market ECAS vehicles based on sensor fusion and perception platforms, efficient propulsion and energy modules, advanced connectivity and trustworthy AI techniques and methods. www.ai4csm.eu

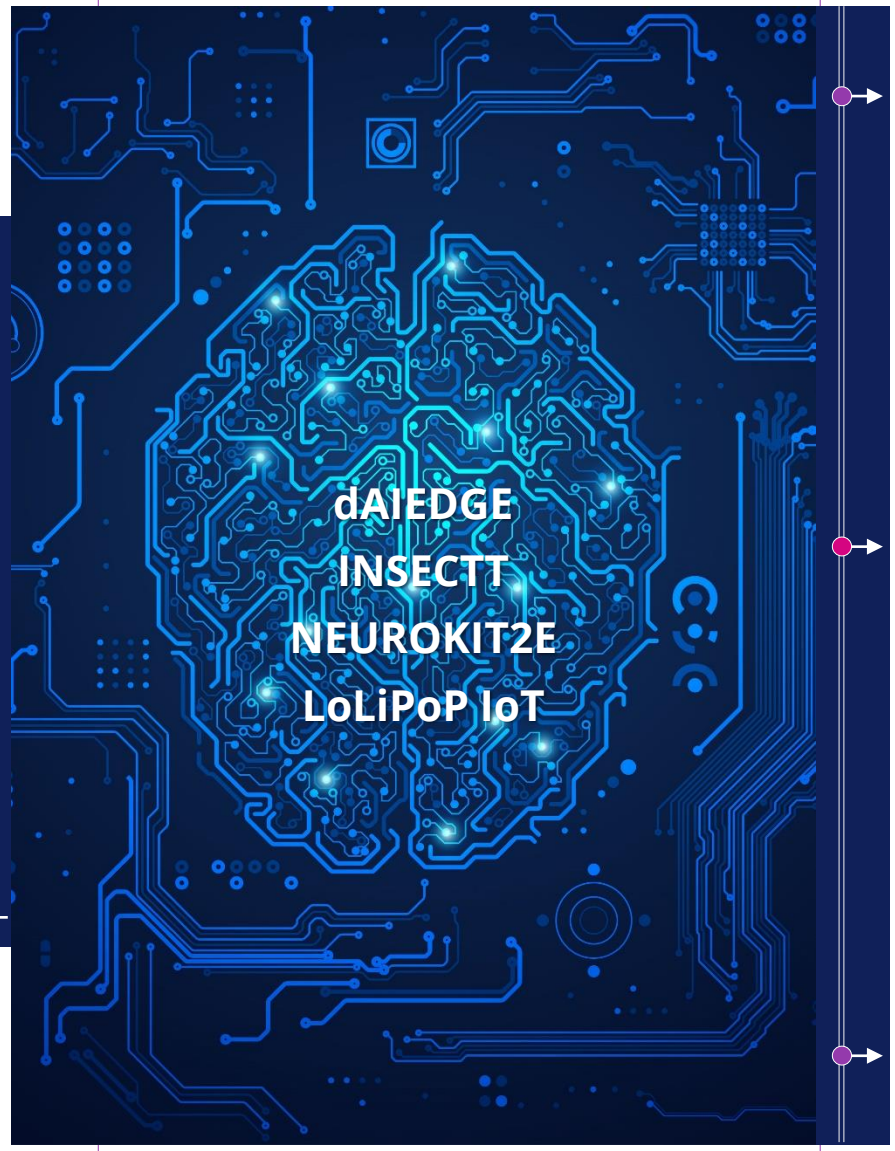
The ANDANTE project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 876925. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Belgium, France, Germany, The Netherlands, Portugal, Spain, Switzerland. www.tristan-project.eu

CLEVER (Collaborative edge-cLoud continuum and Embedded AI for a Visionary industry of thE futuRe) proposes innovations in hardware accelerators, design stack, and middleware software that revolutionize the ability of edge computing platforms to operate federated, leveraging sparse resources that are coordinated to create a powerful swarm of resources. www.cleverproject.eu

REBECCA (Reconfigurable Heterogeneous Highly Parallel Processing Platform for safe and secure AI) aims to democratize the development of edge AI systems and create a complete hardware and software stack centered around a RISC-V CPU, which offer higher performance, energy efficiency, safety, and security than existing systems. www.rebecca-chip.eu

Copyright © 2023

Event Organisers



dAIEDGE

dAIEDGE is the European Network of Excellence for distributed, trustworthy, efficient, and scalable AI at the Edge and promotes the application, development, and deployment of Artificial Intelligence (AI) on edge computing platforms.

InSecTT

INSECTT (Intelligent Secure Trustable Things) provide intelligent, secure, and trustworthy systems for industrial applications to provide comprehensive cost-efficient solutions of intelligent, end-to-end secure, trustworthy connectivity and interoperability to bring the Internet of Things and Artificial Intelligence together. www.insectt.eu

NEUROKIT2E

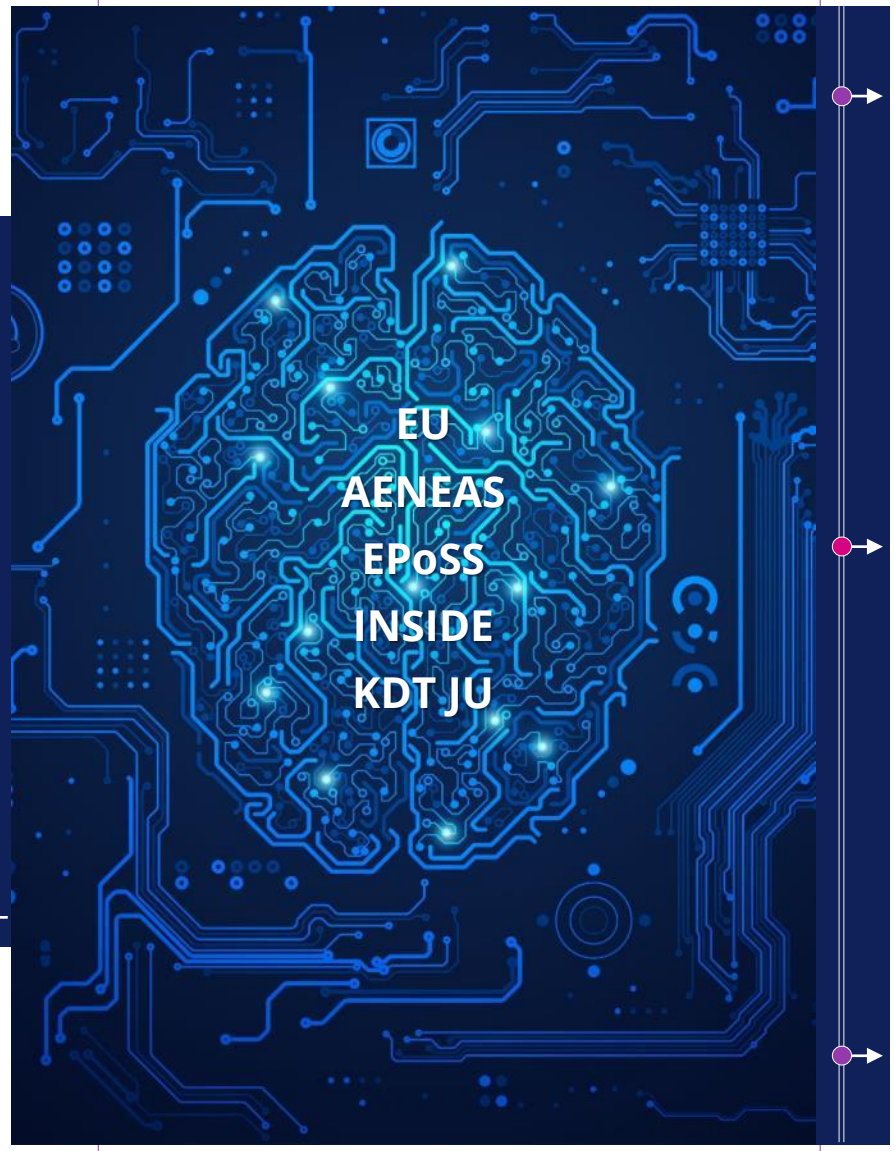
NEUROKIT2E (Open source deep learning platform dedicated to Embedded hardware and Europe) proposes a Deep Learning Platform for Embedded Hardware around an established European value chain (AI HW/SW). The solutions developed support neural network design, optimisation, and implementation on constrained HW.

LoLiPoP-IoT

The objectives of LoLiPoP IoT (Long Life Power Platforms for Internet of Things) are to develop energy harvesting-based innovative Long Life Power Platforms that enable retrofit of wireless sensor network edge devices for asset tracking, condition and performance monitoring. www.lolipop-iot.eu

Copyright © 2023

Supporting Organizations



AENEAS standing for Association for European NanoElectronics ActivitieS, is an industrial Association, established in 2006, providing unparalleled networking opportunities, policy influence & supported access to funding to all types RD&I participants in the field of micro and nanoelectronics enabled components and systems. https://aeneas-office.org/

The European Technology Platform on Smart Systems Integration is an industry-driven policy initiative, defining research, development and innovation needs as well as policy requirements related to Smart Systems Integration and integrated Micro- and Nanosystems. The main objective is to develop a vision and to set up a Strategic Research Agenda. www.smart-systems-integration.org

Inside Industry Association is the European Technology Platform for research, design and innovation on Intelligent Digital Systems and their applications. The Association is a membership organisation for the European research and innovation actors with more than 200 members and associates from all over Europe. www.inside-association.eu

The Key Digital Technologies Joint Undertaking - the Public-Private Partnership for research, development and innovation – funds projects for assuring world-class expertise in these key enabling technologies, essential for Europe's competitive leadership in the era of the digital economy. KDT JU is the successor to the ECSEL JU programme. www.kdt-ju.europa.eu

Copyright © 2023