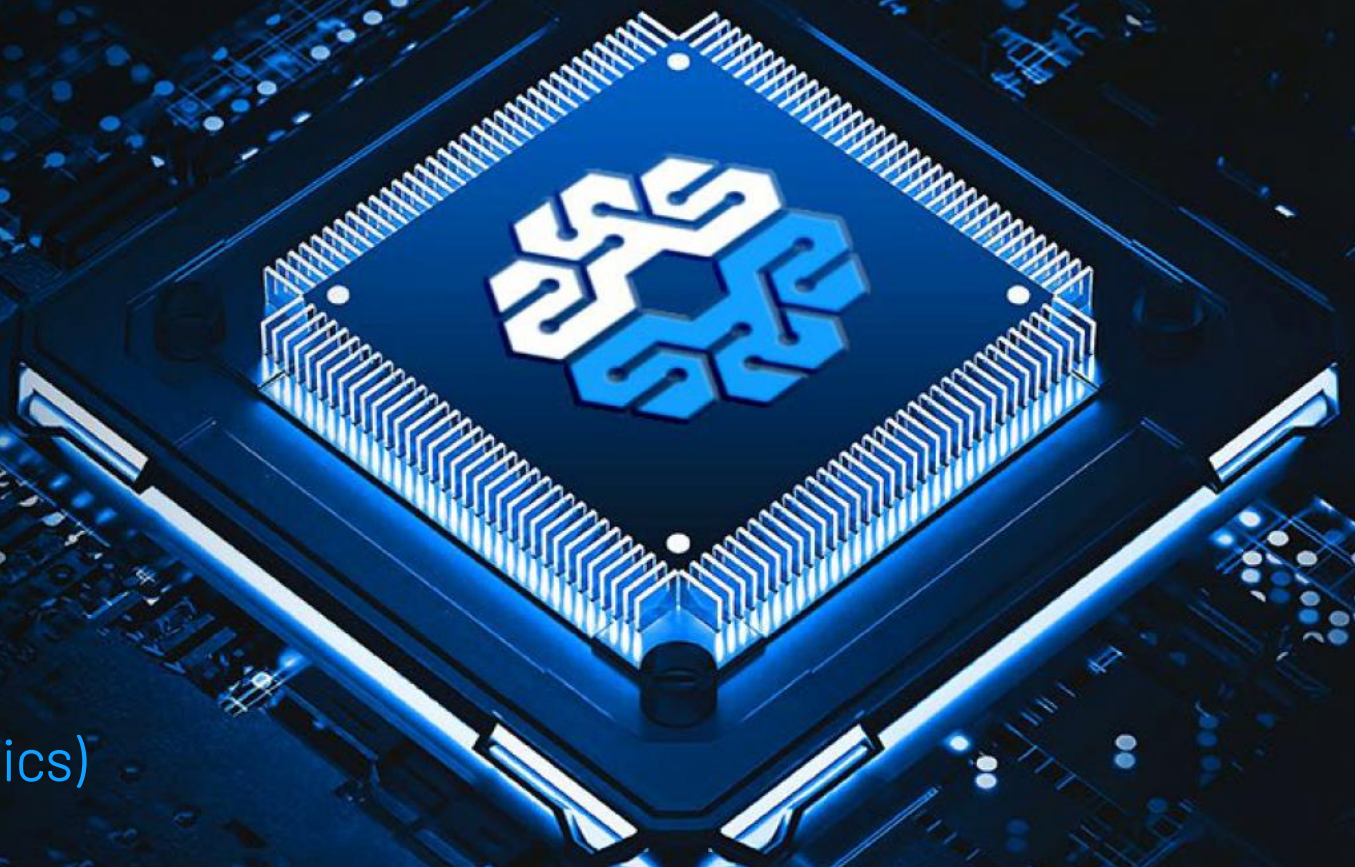
neuroSoC

NeuroSoC project

Use of RISC-V as multiprocessor host subsystem with In Memory computing based on NVM memories for AI inference

Carmine CAPPETTA (STMicroelectronics)



Funded by the European Union

Innovate UK

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

This work was supported by European Union (Horizon Europe Grant Agreement n°101070634), Swiss State Secretariat for Education, Research and Innovation (SERI) under contracts number SBFI 22.00202 and 23.00205 and UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10040829]

Summary

- NeuroSoC project
- NeuroSoC IMNPU architecture
- Enhancement of the NeuroSoC architecture with RISC-V support
- Architecture of the RISC-V core
- RISC-V co-processor unit
- Emulation platforms
- Future works

NeuroSoC project

The NeuroSoC project is supported by European Union (Horizon Europe Grant Agreement n°101070634), Swiss State Secretariat for Education, Research and Innovation (SERI) under contracts number SBFI 22.00202 and 23.00205 and UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (grant number 10040829).

Stands for

A multiprocessor System-on-Chip with In-Memory neural processing unit

It is a 42-month EU/UKRI/Switzerland funded project aiming at using Phase Change Memory and FD-SOI 28 nm technologies to develop an advanced multiprocessor System-on-Chip

NeuroSoC Rationale

The explosive growth of artificial intelligence

Its movement to the edge and end devices

Significant research on highly energy efficient and low-latency non-von Neumann computing paradigms such as in-memory computing (IMC)

neurSoC answer

Develop a **flexible computing system where an analog IMC-based neural processing unit** is integrated into a **multi-processor functional safe and secure system-on-chip**

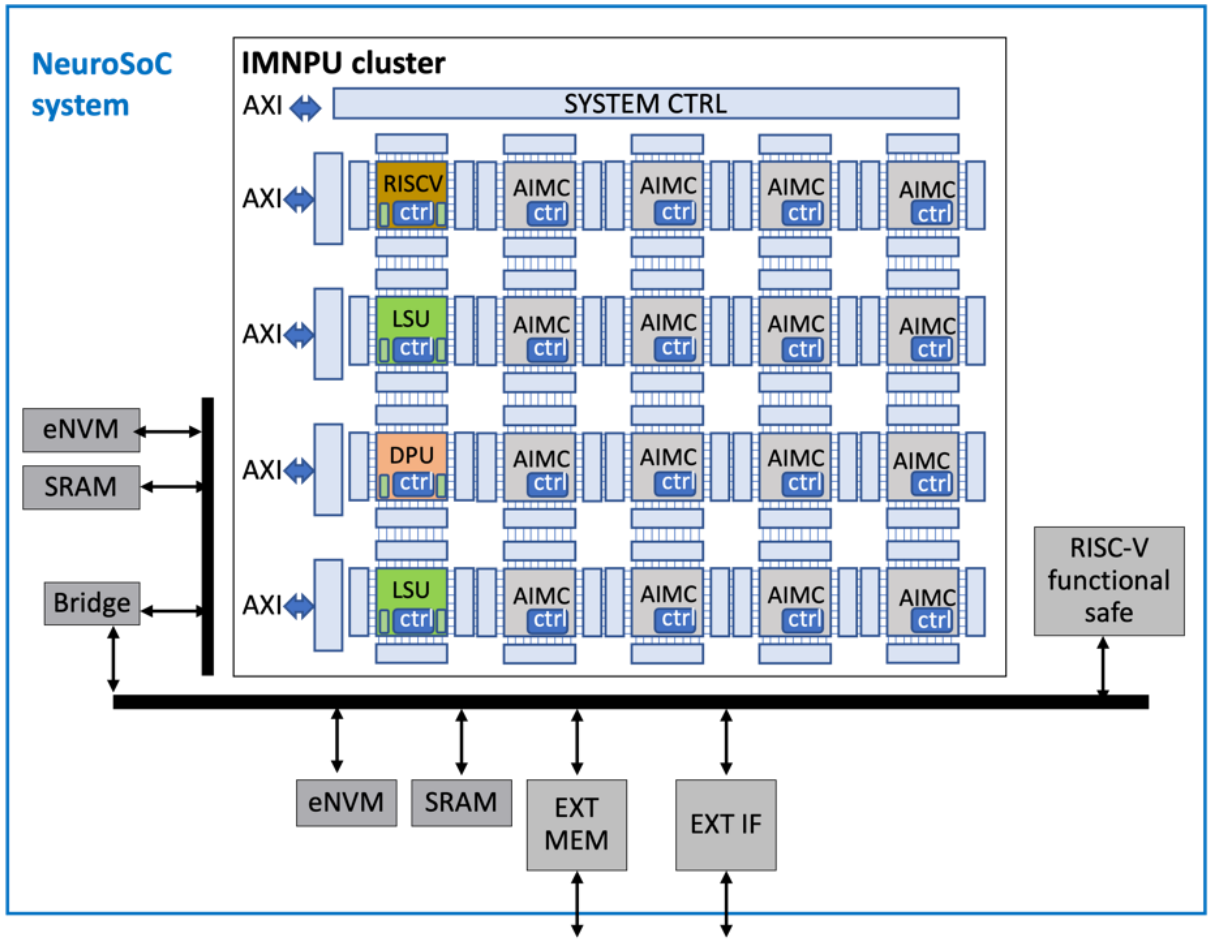
To tackle the requirements of a wide set of edge-AI applications.

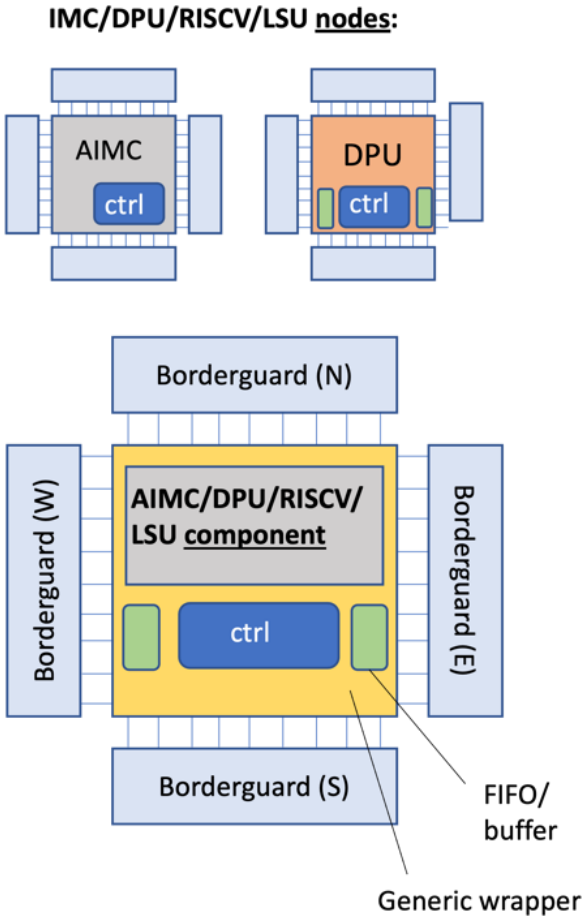
Relying on a solid, mature, and qualified reliable Phase Change Memory technology

Will enable the creation of an industrially proven path answering to the level of maturity need compatible with a mass volume production and cost

NeuroSoC IMNPU architecture

2D mesh-based IMNPU cluster and NeuroSoC system





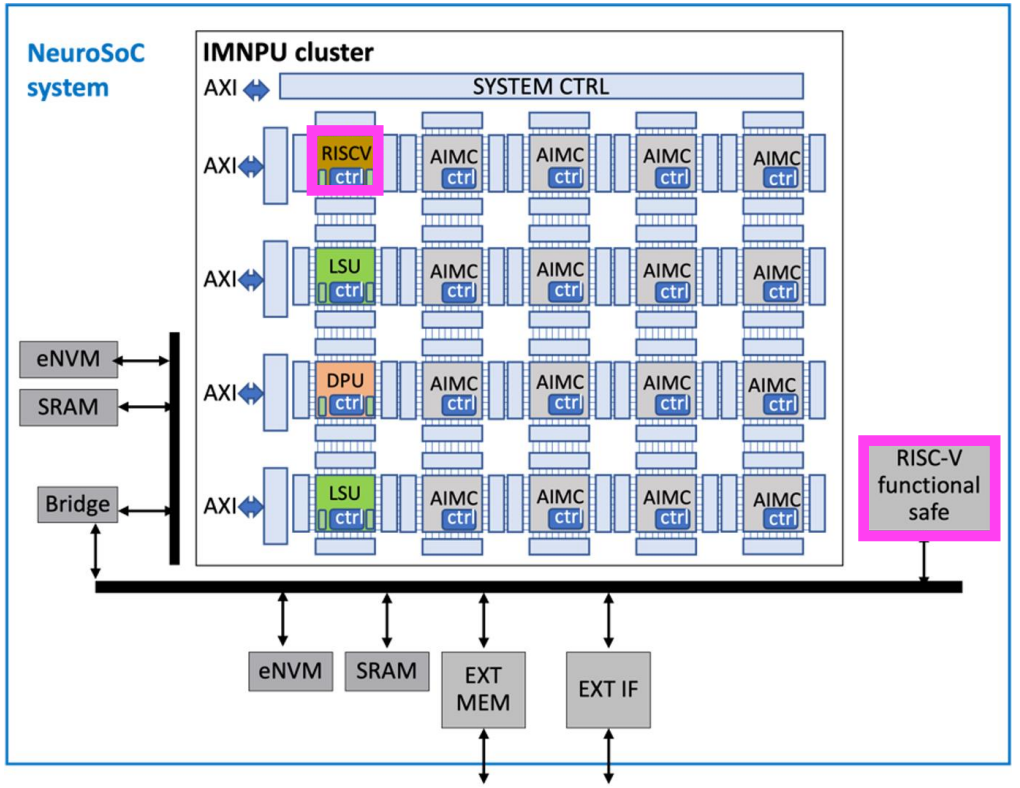
NeuroSoC System on Chip system level architecture comprises of:

- Cluster of PCM analog in-memory computing tiles
- Non-volatile memory and SRAM memory support
- Functional safe host processor
- Specialized digital processing units
- RISC-V co-processor

Enhancement of the NeuroSoC architecture with RISC-V support

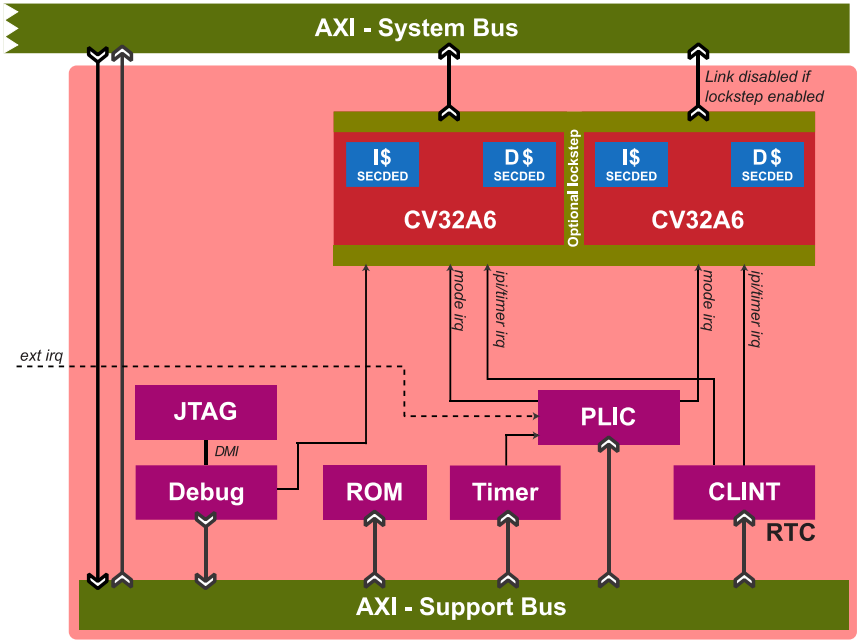
- The RISC-V co-processor will provide a programmable inference accelerator directly embedded in the NPU cluster along the project developed Analog In-Memory Compute (AIMC) components. It is highlighted with the pink shadow inside the Neural Processing Unit

- The RISC-V host multicore component will provide a RISC-V multi-core enhanced with functional safety features as main control element of the NeuroSoC system. It is highlighted with the pink shadow in the bottom right corner in the picture.

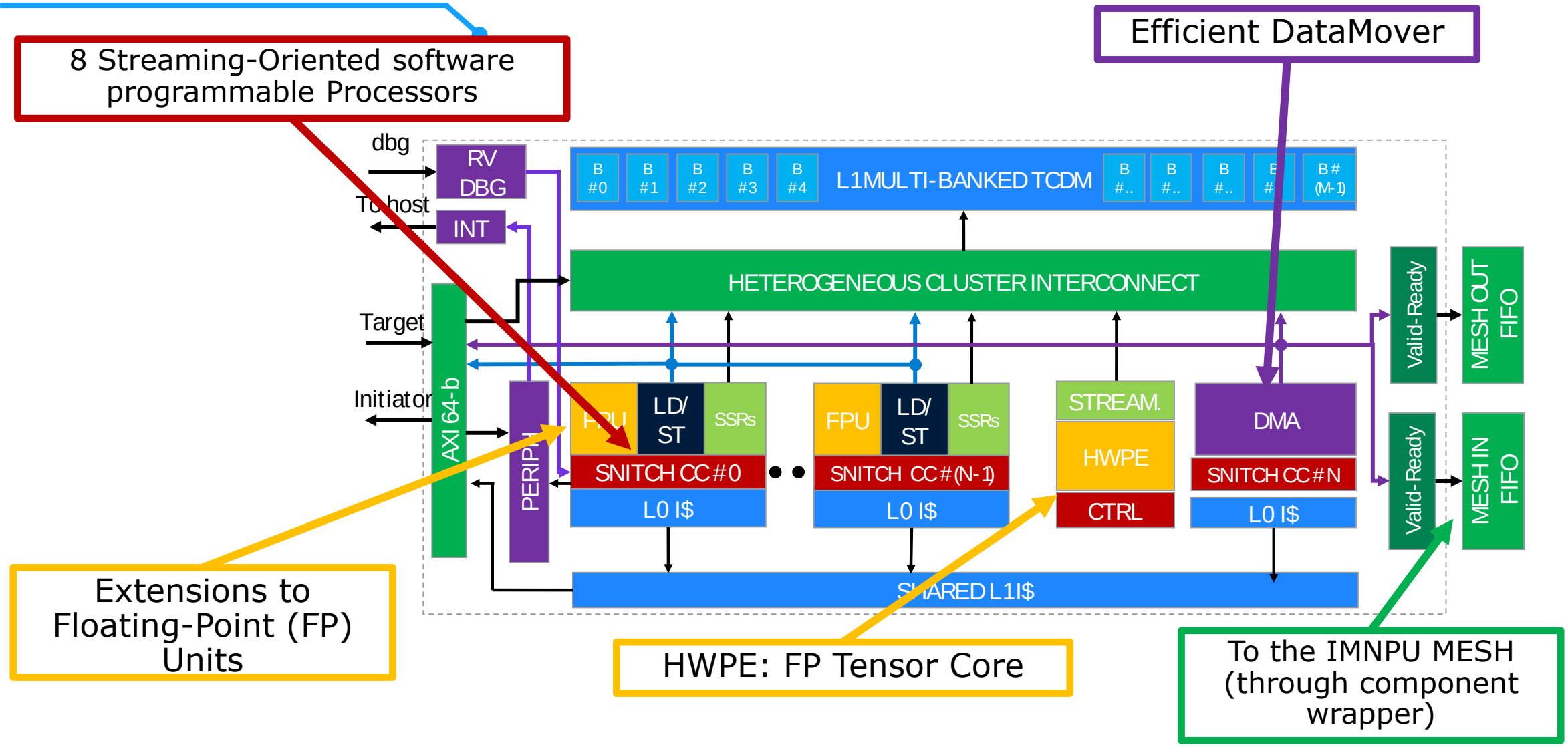


Functional-safe host RISC-V processor

- Based on **CVA6**
  - An open-source RISC-V application core, able to boot Linux
  - Curated at the OpenHW Group
  - Available in 32- and 64-bit: CV32A6, CV64A6
- Dual-core CV32A6 host developed in NeuroSoC
- Safety features:
  - Dual-core configured as AMP (asymmetric multi-processor) or DCLS (dual-core lock step)
  - Internal memory hardening with error correction

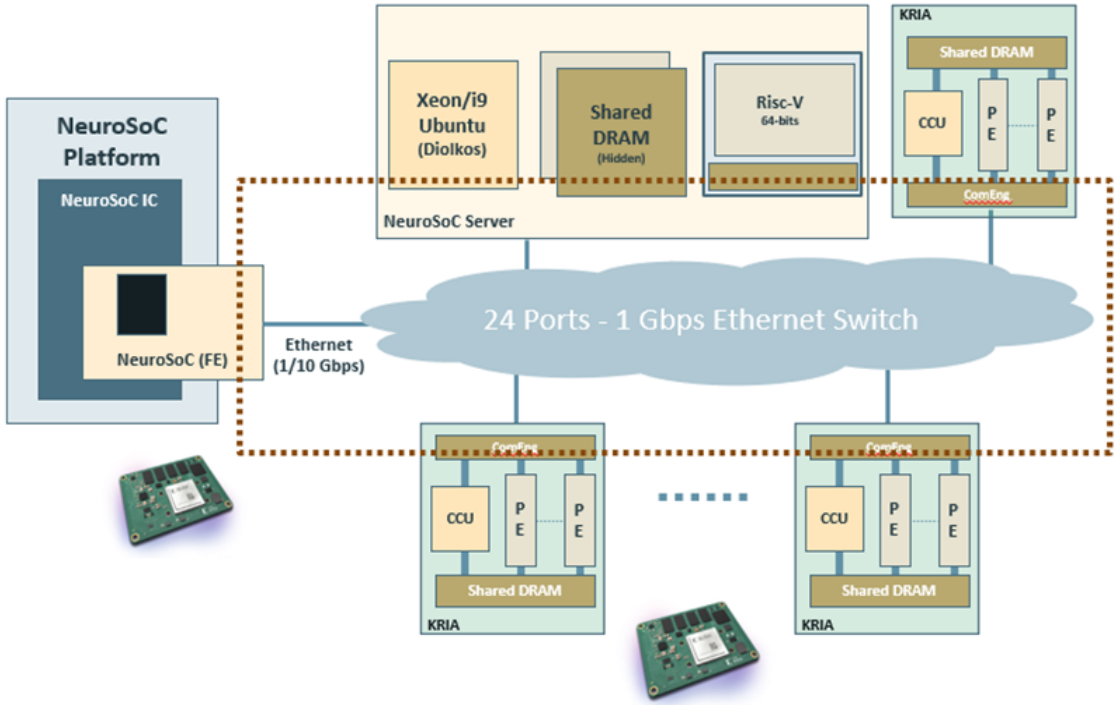


RISC-V co-processor unit



Emulation platforms (I)

- The first version of the NeuroSoC Emulator is shown in the figure
  - It is based on multiple FPGAs (Xilinx Kria boards) organized in a distributed computing environment
  - The NeuroSoC FE board communicates with the NeuroSoC Server and a network of Kria boards via Gbps Ethernet links.
  - Each Kria may support several PEs, probably of different types, while there is a number of reliable links (based on TCP/IP connections) between the PEs in order to exchange data reliably

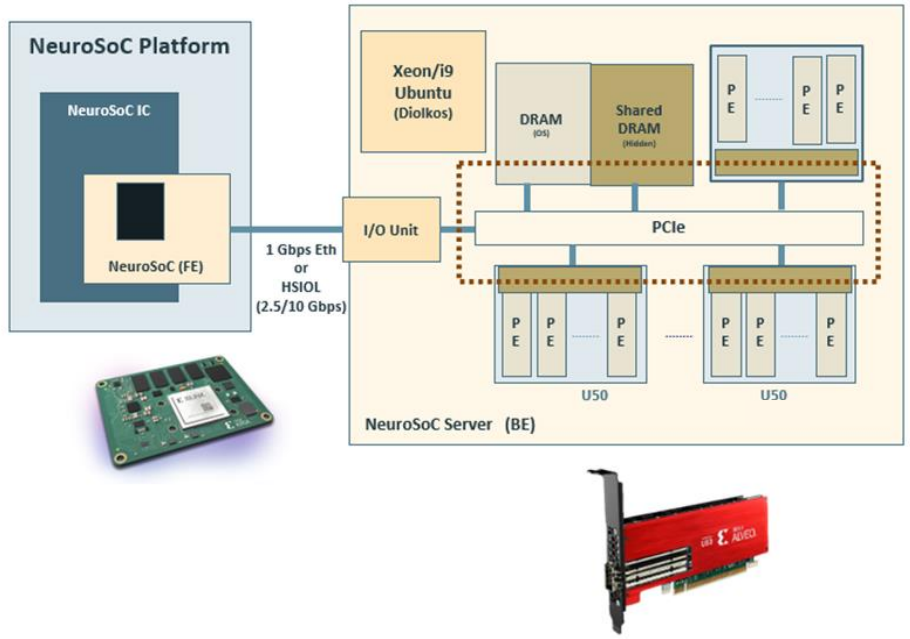


PROS of the solution

- Rapid NeuroSoC prototyping
- Easy scalability by adding more Kria boards to the network in a plug-and-play fashion, utilizing the TCP/IP protocol
- Built-in detailed system debugging and timestamping mechanisms

Emulation platforms (II)

- The second version consists of several Alveo U50 boards plugged in the PCIe slots of the NeuroSoC Server interfacing with each other over PCIe
- This version will have similar functionality as the one described previously, providing a much faster but less flexible environment



PROS of the solution

- Alveo U50 offers eight times more resources than the Kria, allowing for more PEs on the same board.
- PCIe Gen 3 with 16 lanes offers up to 128 Gbps transfer rate for fast data transfers
- Achieving much higher processing rates

Future Work

- The design of the architecture has been completed and the porting on the emulator is on going
- The next step will be to validate the prototype in terms of
  - Quality of detection
  - Efficiency
  - Security enhancement