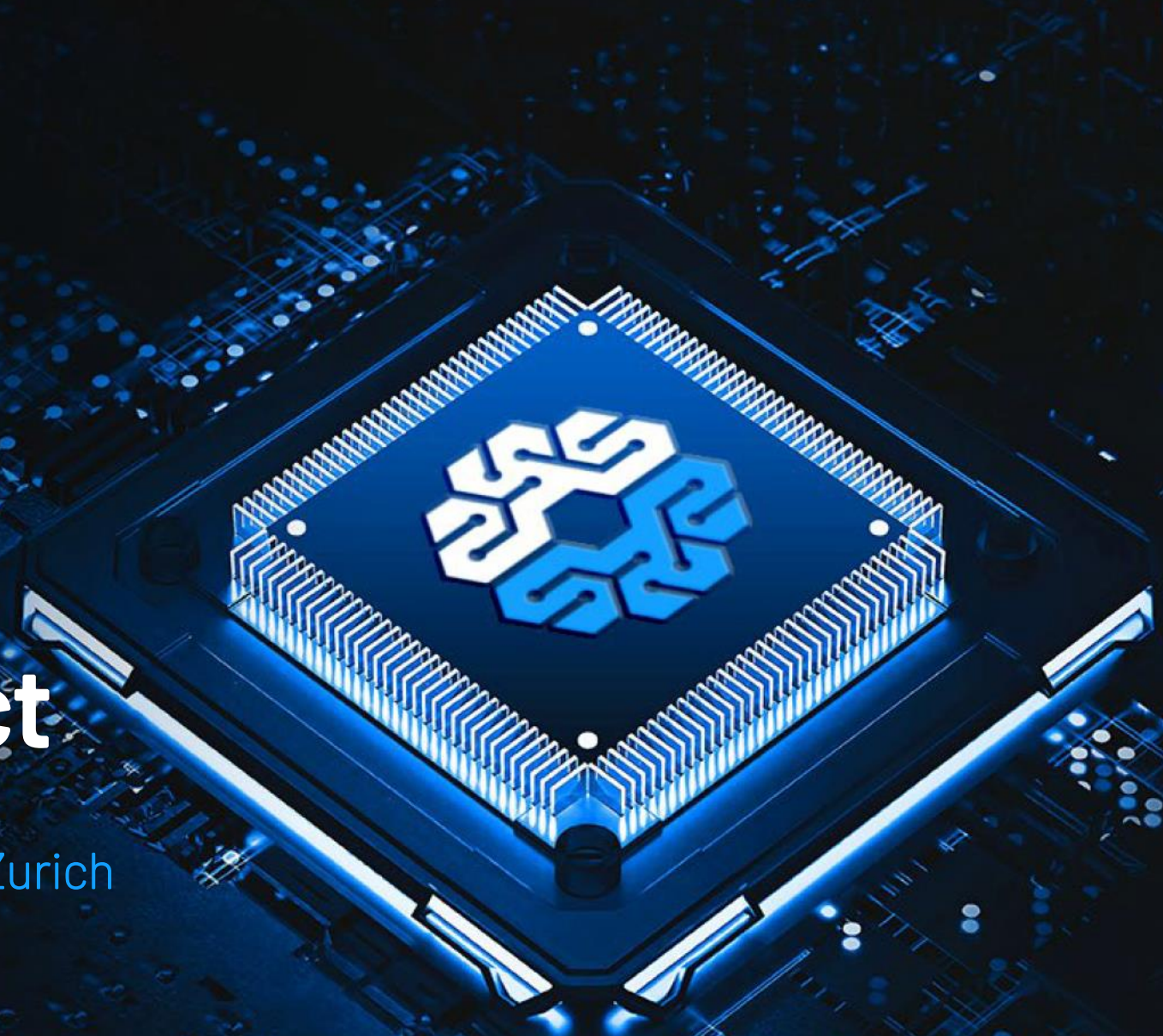


neur  SoC

NeuroSoC Project

Irem Boybat, IBM Research Europe - Zurich

NeuroEdge, 18th January 2024



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 SBF1 22.00202 and 23.00205 and UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10040829]



Presentation overview

- ❁ NeuroSoC project overview and rationale
- ❁ AI at the edge and promise of in-memory computing
- ❁ Building blocks of the edge SoC
 - Computational phase-change memory technology
 - Analog in-memory computing tiles based on phase-change memory devices
 - NeuroSoC SoC architecture
 - Algorithms and software tools
 - Applications requirements, integration, and use- cases demonstrations



About NeuroSoC

 NeuroSoC stands for:

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

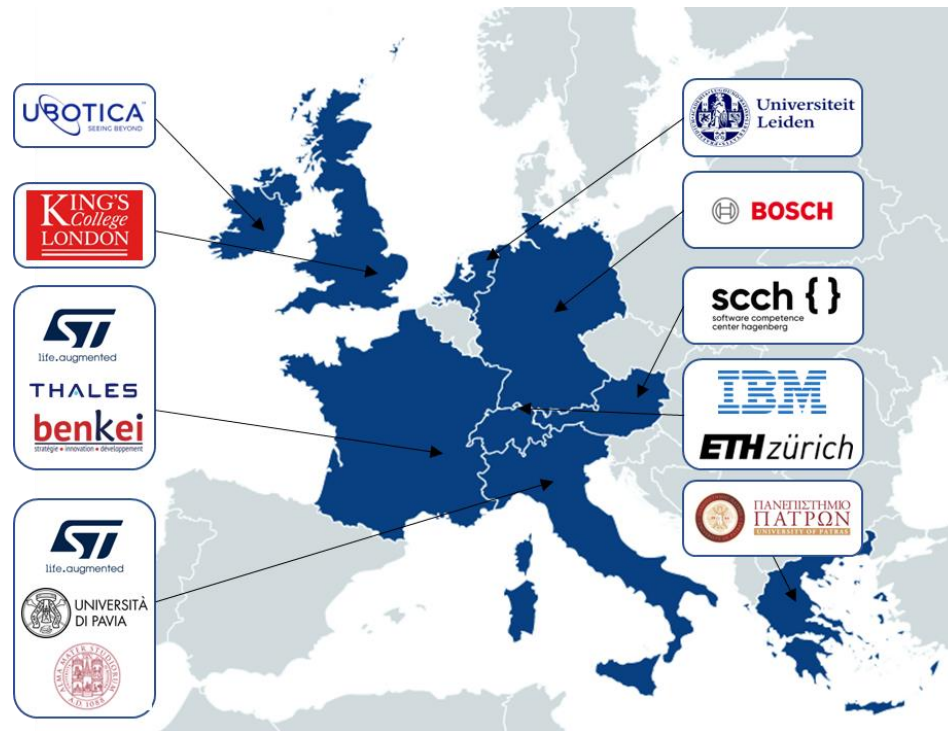
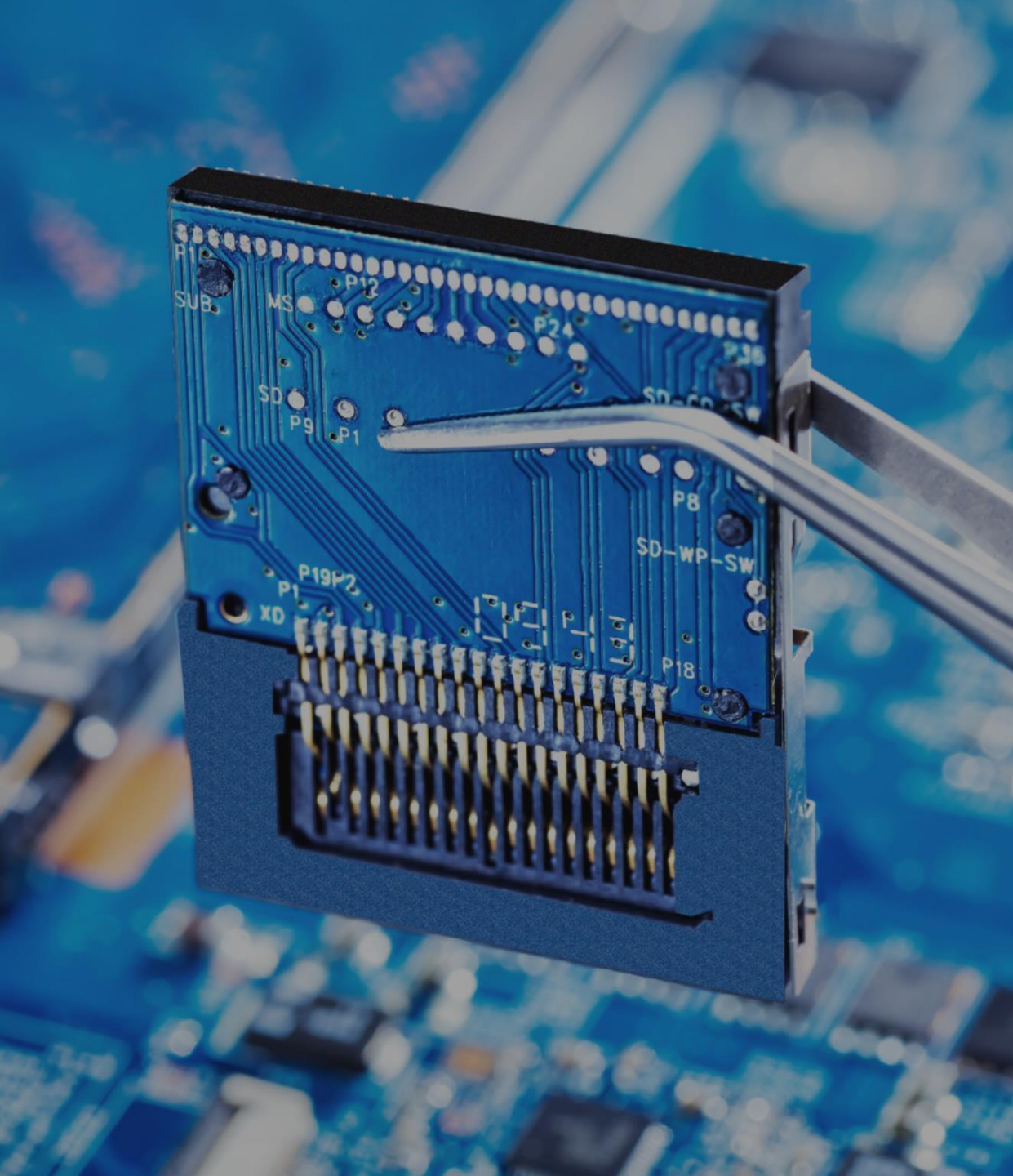
A 42-month EU/UKRI/Switzerland funded project aiming at using Phase Change Memory and FD-SOI 28 nm technologies to develop an advanced multiprocessor System-on-Chip



NeuroSoC at a glance

 Call and Topic/Activity:	HORIZON-CL4-2021-DIGITAL-EMERGING-01-01 – Ultra-low-power, secure processors for edge computing (RIA)
 GA number:	101070634
 Type of action:	RIA (Research & Innovation Action)
 Project cost:	7 952 677 EUR (only beneficiaries)
 Duration:	42 months; start 1 September 2022
 Website:	www.neurosoc.eu

An European strong value chain





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)

neuroSoC answer

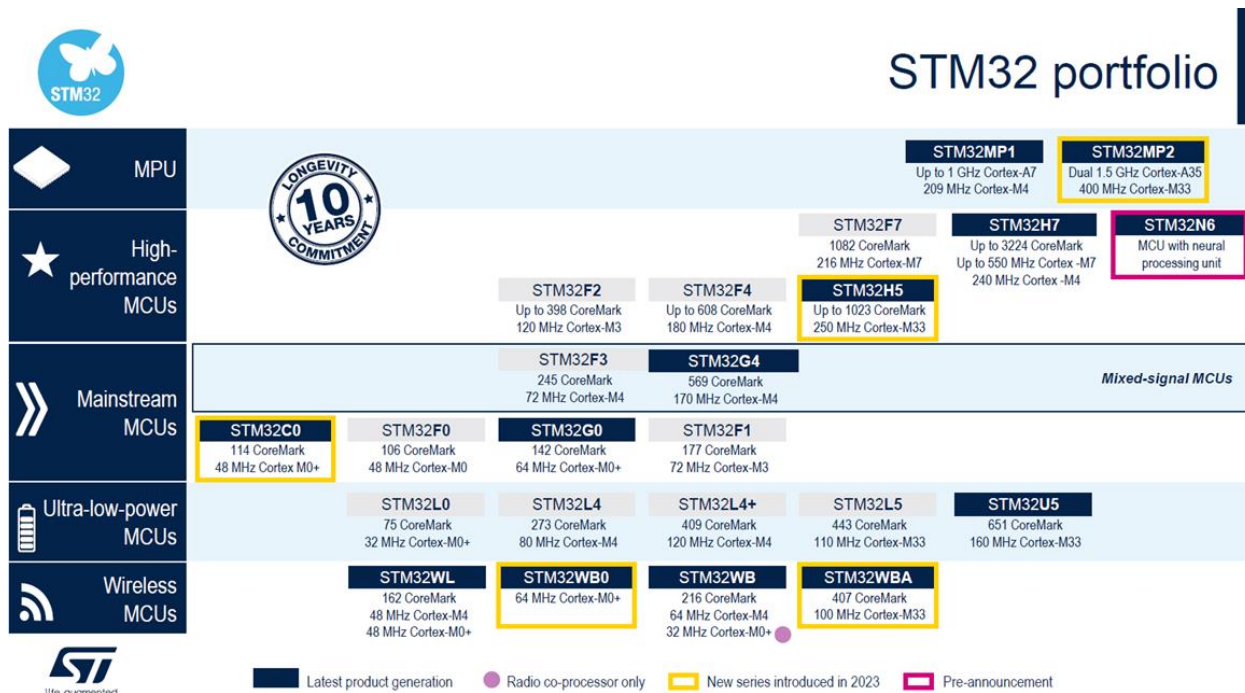
Develop a flexible computing system where an analog IMC-based neural processing unit is integrated into a multi-processor functional safe and secure system-on-chip

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

ST roadmap towards the AI at the edge



STM32N6 upcoming general-purpose microcontroller with ST Neural-Art Accelerator™, a Neural Processing Unit





Evolution of Neural Processing Units (NPU)

Energy Efficiency

AI executed
In SW

Memory

μ C

SW

AI accelerated via
Neural Processing Unit

Memory

μ C

SW

NPU

AI model parameter
stored in embedded
non volatile memory

Coefficients Memory

μ C

SW

NPU

In memory execution of
AI deeply quantized layers

SRAM/ePCM

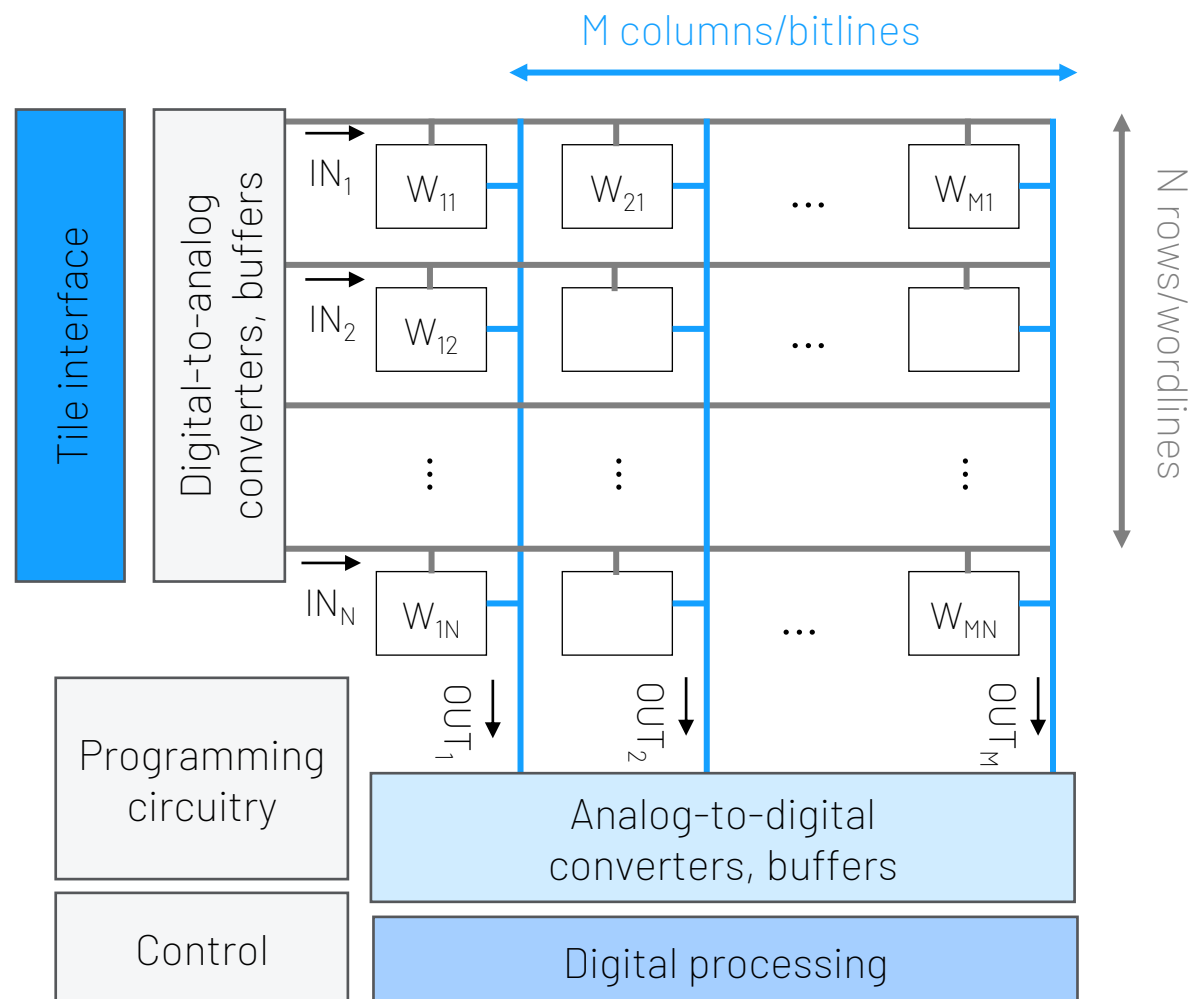
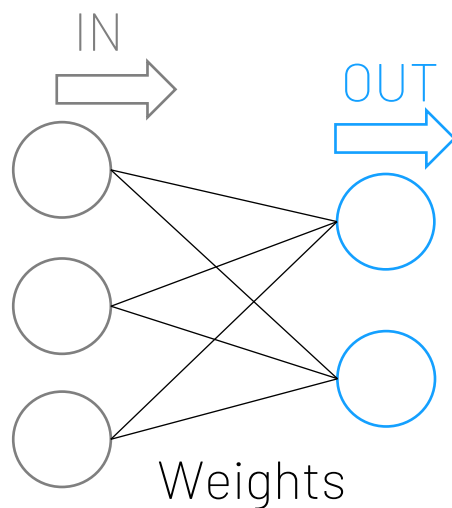
μ C

SW

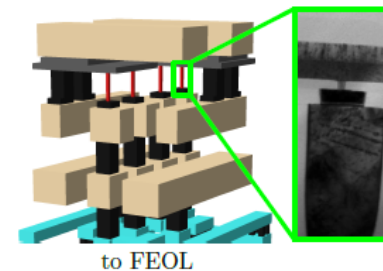
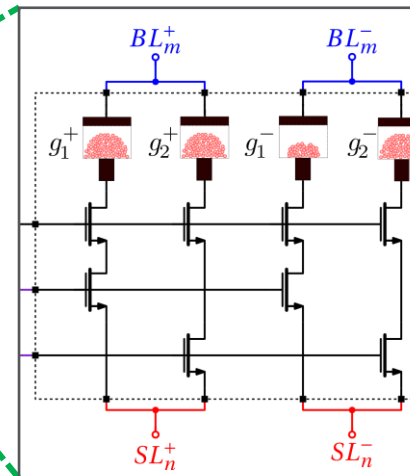
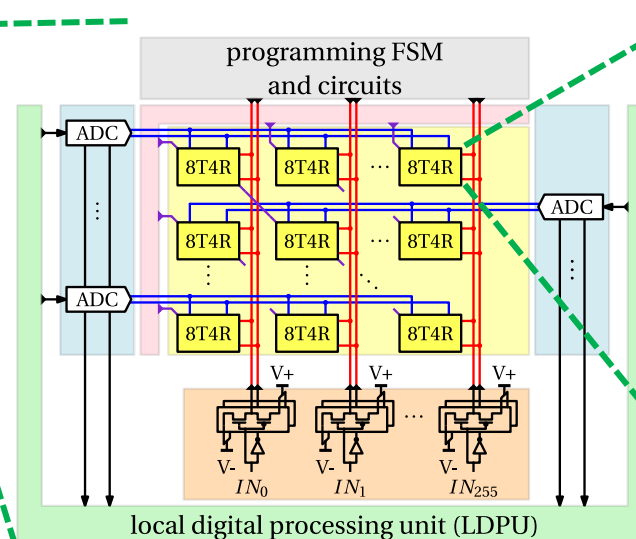
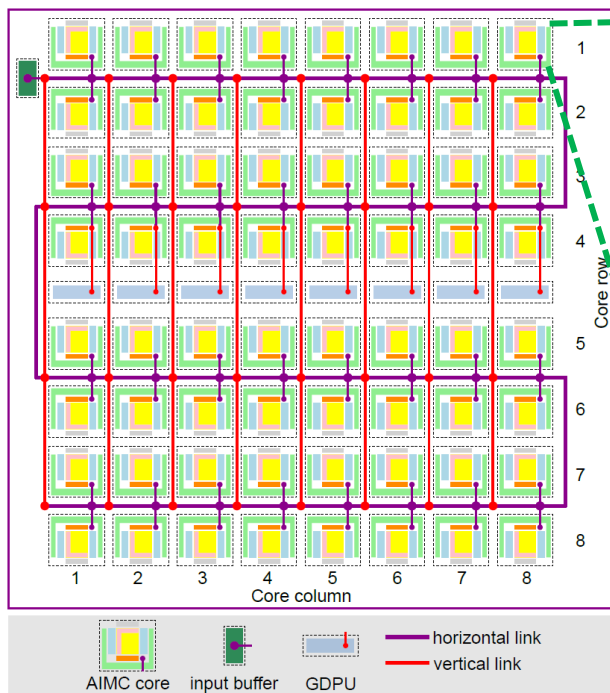
NPU
+
IMC

Analog in-memory computing basics

$$\begin{bmatrix} W_{11} & \dots & W_{1N} \\ \vdots & \ddots & \vdots \\ W_{M1} & \dots & W_{MN} \end{bmatrix} \begin{bmatrix} IN_1 \\ \vdots \\ IN_N \end{bmatrix} = \begin{bmatrix} OUT_1 \\ \vdots \\ OUT_M \end{bmatrix}$$



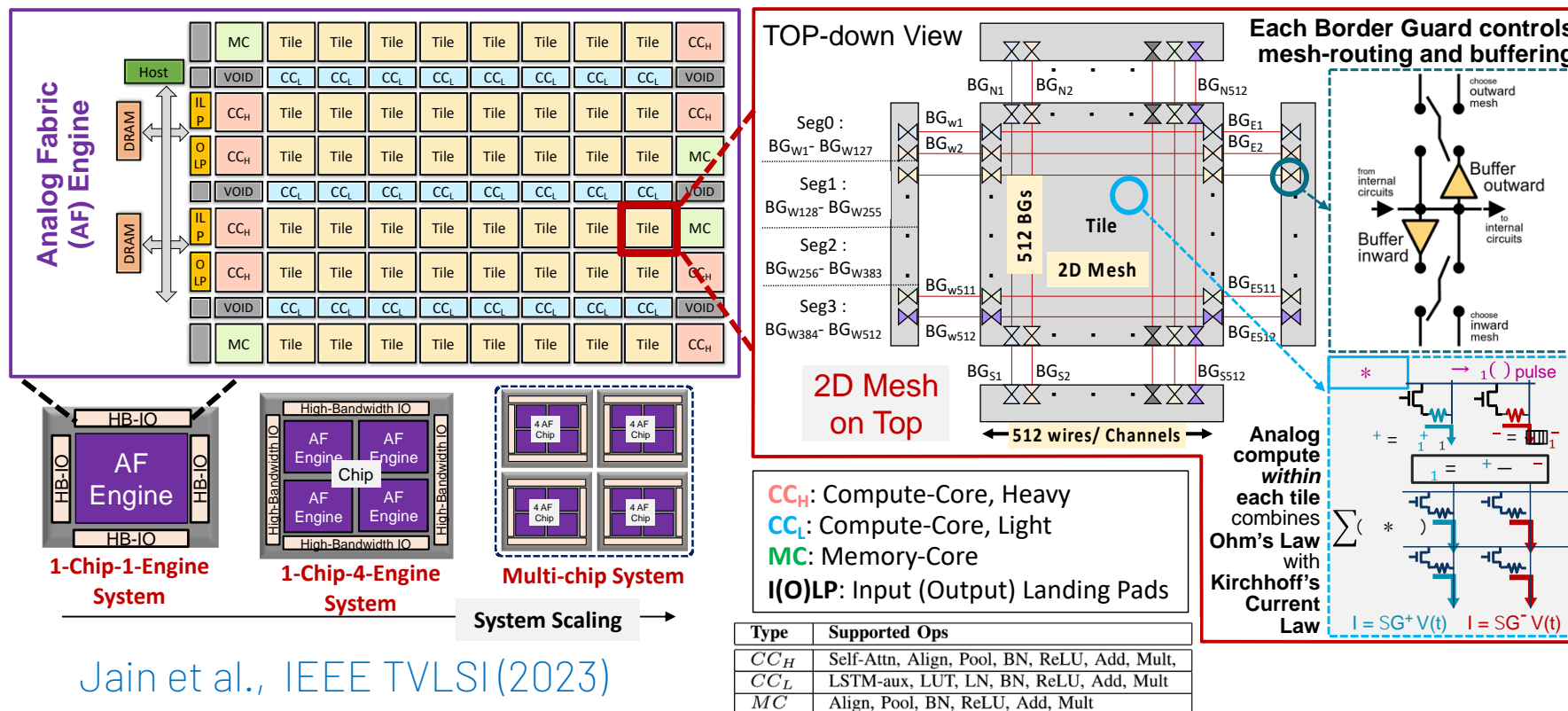
IBM HERMES project chip



Le Gallo et al., Nature Electronics (2023)

- Each of the 64 cores comprises 256x256 crossbar arrays of unit cells with peripheral circuitry (4M unit-cells)
- On-chip local and global digital processing as well as a communication fabric
- Each unit cell comprises of four phase-change memory devices (16M PCM devices)

IBM Heterogeneous architecture with 2D-mesh



- A heterogeneous architecture that combines AIMC compute cores with special-function compute cores for auxiliary digital computation
- A dense and efficient circuit-switched 2D mesh serves as the communication fabric

IBM AI HW Kit

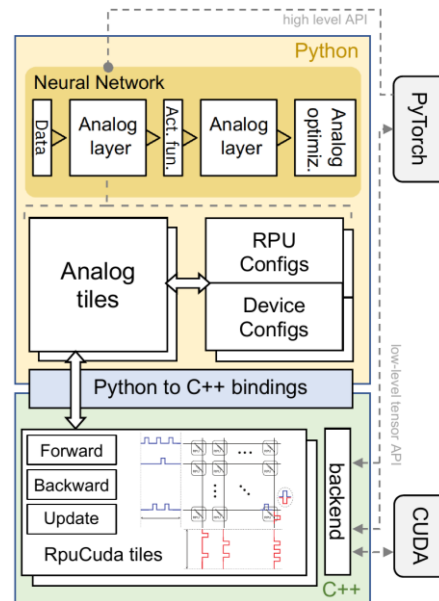
Rasch et al., Proc. AICAS (2021)

Le Gallo et al., APL Machine Learning (2023)

Overview

<https://github.com/IBM/aihwkit>

- Simulator that focuses on the algorithmic level and algorithmic advances of Analog in-memory computing
- AIMC training and inference simulations
- Bring your own models and datasets to evaluate the impact of emerging AIMC hardware on your DL workloads using the flexibility of PyTorch



Roadmap

- Additional neural network layers
- Algorithmic advances to improve training and inference accuracy
- Premium hardware demonstrations
- Real hardware demonstrations

- The IBM Analog HW Acceleration Kit is an excellent tool for developing and testing algorithms for hardware-aware training
- Equipped with an inference simulator with drift and statistical (programming) noise models calibrated on hardware, direct HW access will be enabled in the near future
- Full GPU support and substantial online documentation



Presentation overview

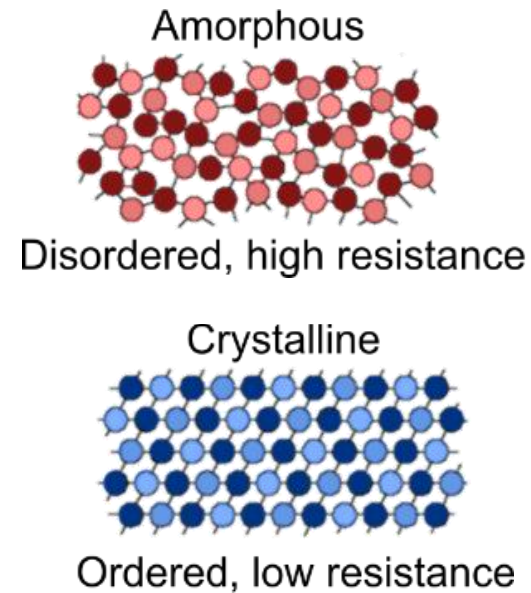
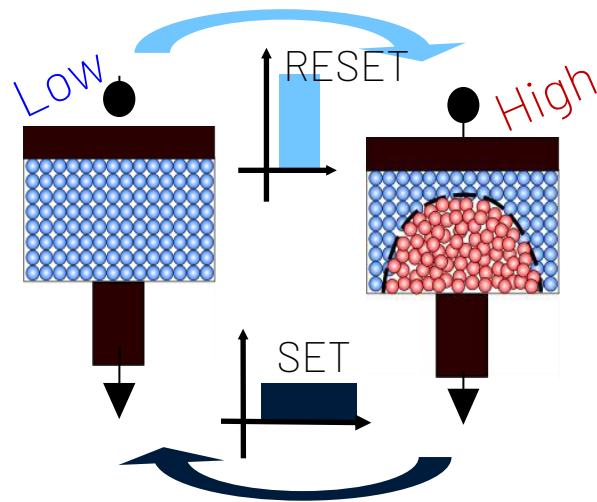
- ❁ NeuroSoC project overview and rationale
- ❁ AI at the edge and promise of in-memory computing
- ❁ Building blocks of the edge SoC
 - Computational phase-change memory technology
 - Analog in-memory computing tiles based on phase-change memory devices
 - NeuroSoC SoC architecture
 - Algorithms and software tools
 - Applications requirements, integration, and use- cases demonstrations



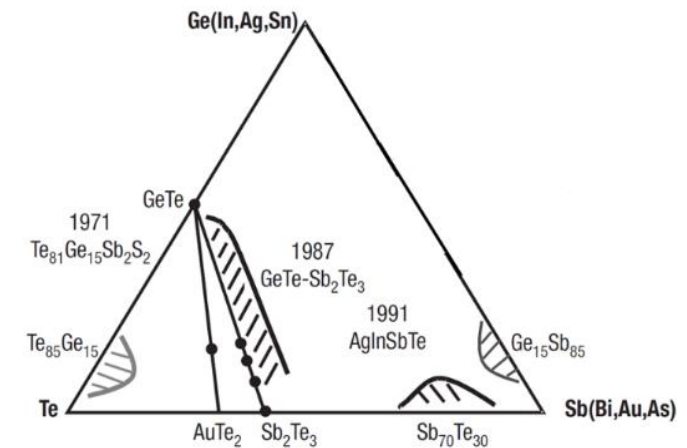
Focus on the PCM memory

- ❁ Characterization and modelling of a Phase Change Memory (PCM) device developed by ST-I in FD-SOI 28nm technology as building block of the In-Memory Computing (IMC) tile
- ❁ Optimization vs temporal drift and noise
- ❁ Statistical evaluation of programming algorithms, current distributions, and reliability of the analog IMC PCM cell
- ❁ Characterization of the computational precision and compensation (drift/read noise/temperature dependence)
- ❁ Development the analog IMC tile

Phase-change memory



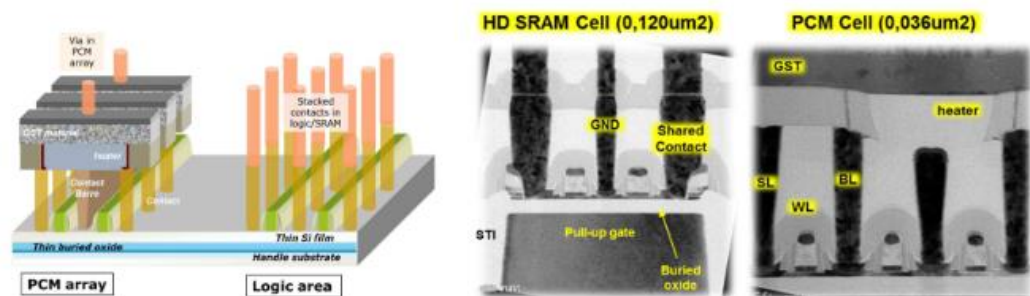
Commonly used phase change materials



Wuttig & Yamada, *Nature Materials*, 2007
 Le Gallo et al., *J. Phys. D*, 2020

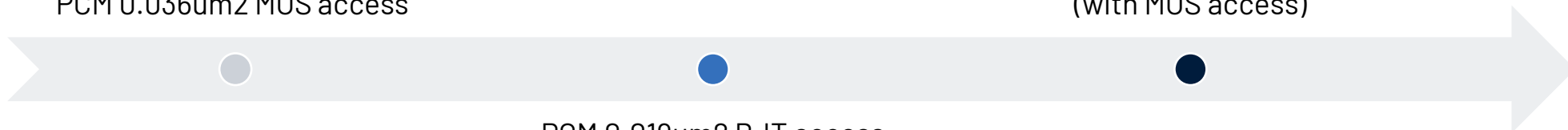
- A nanometric volume of phase change material between two electrodes
- A reversible phase transition is induced via Joule heating between crystalline (SET) and amorphous phases (RESET)
- Continuum of conductance levels can be achieved via intermediate phase configurations

ST High Density Embedded PCM Cell in 28nm FDSOI



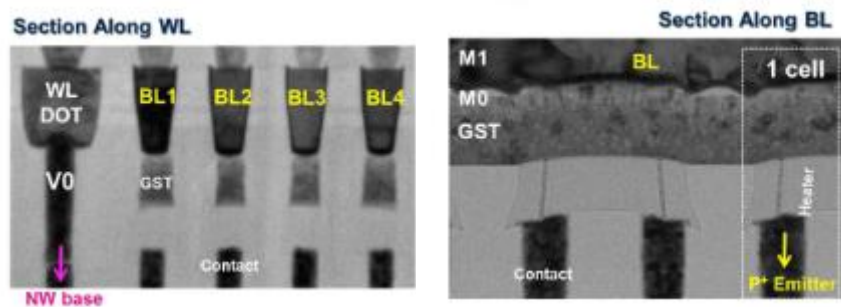
Arnaud et al., IEDM 2018
PCM 0.036um² MOS access

Computational PCM in NeuroSoC
(with MOS access)

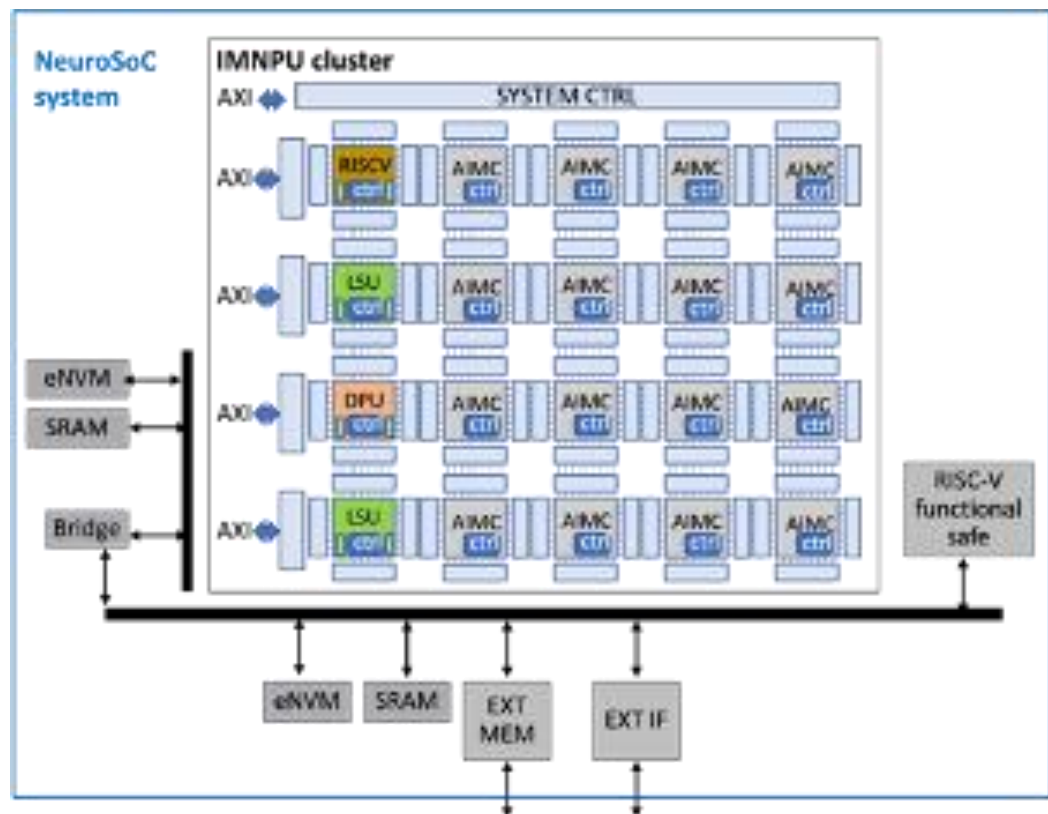


PCM 0.019um² BJT access

Arnaud et al., IEDM 2020



NeuroSoC SoC Architecture



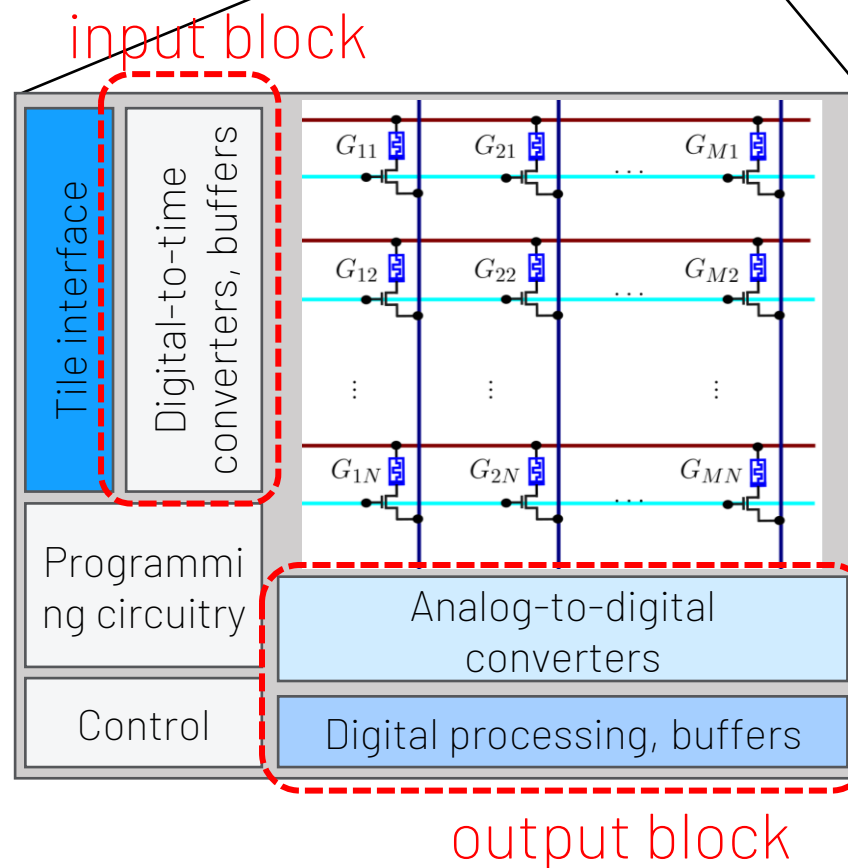
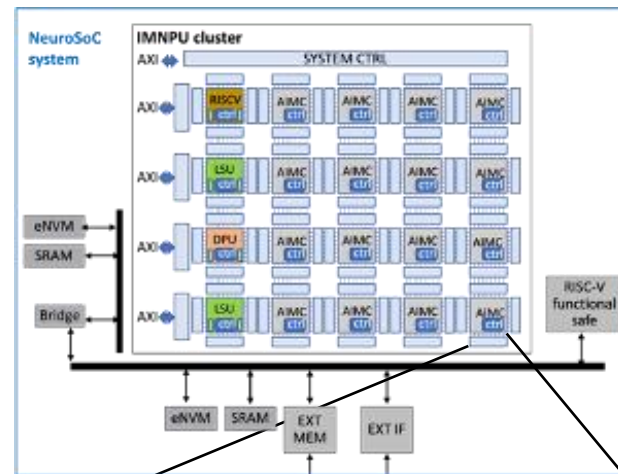
NeuroSoC System on Chip system level architecture comprises of:

- 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

IMC PCM tile

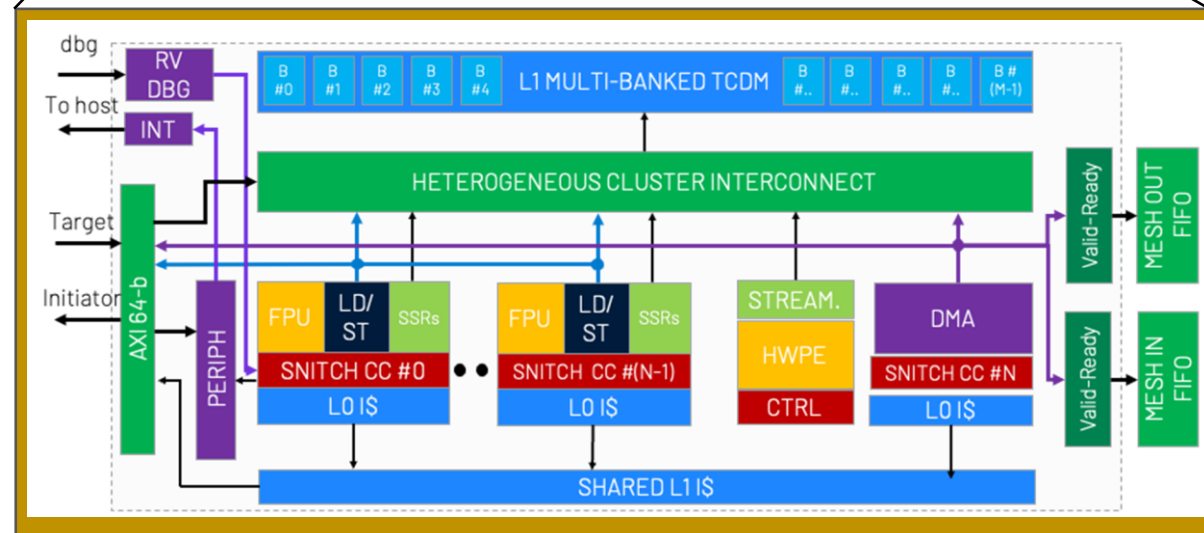
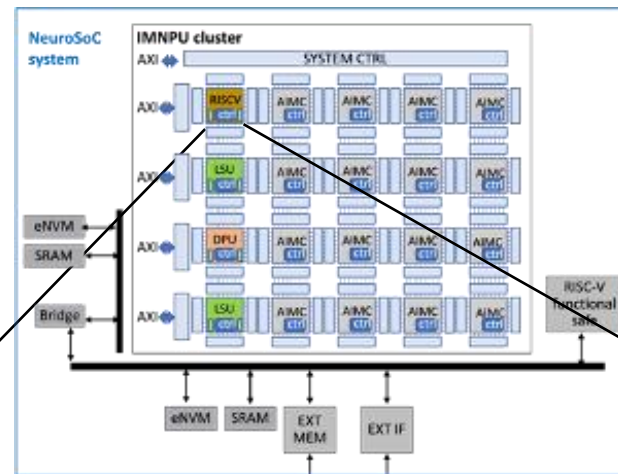
❁ Leveraging the multilevel PCM device to design an analog IMC tile

- ❁ Definition of the unit cell and a suitable array structure
- ❁ Design of the associated digital and analog circuits
- ❁ Anticipate inputs from security analysis to make the resulting IMC tile more robust against side channels attacks and for improved security



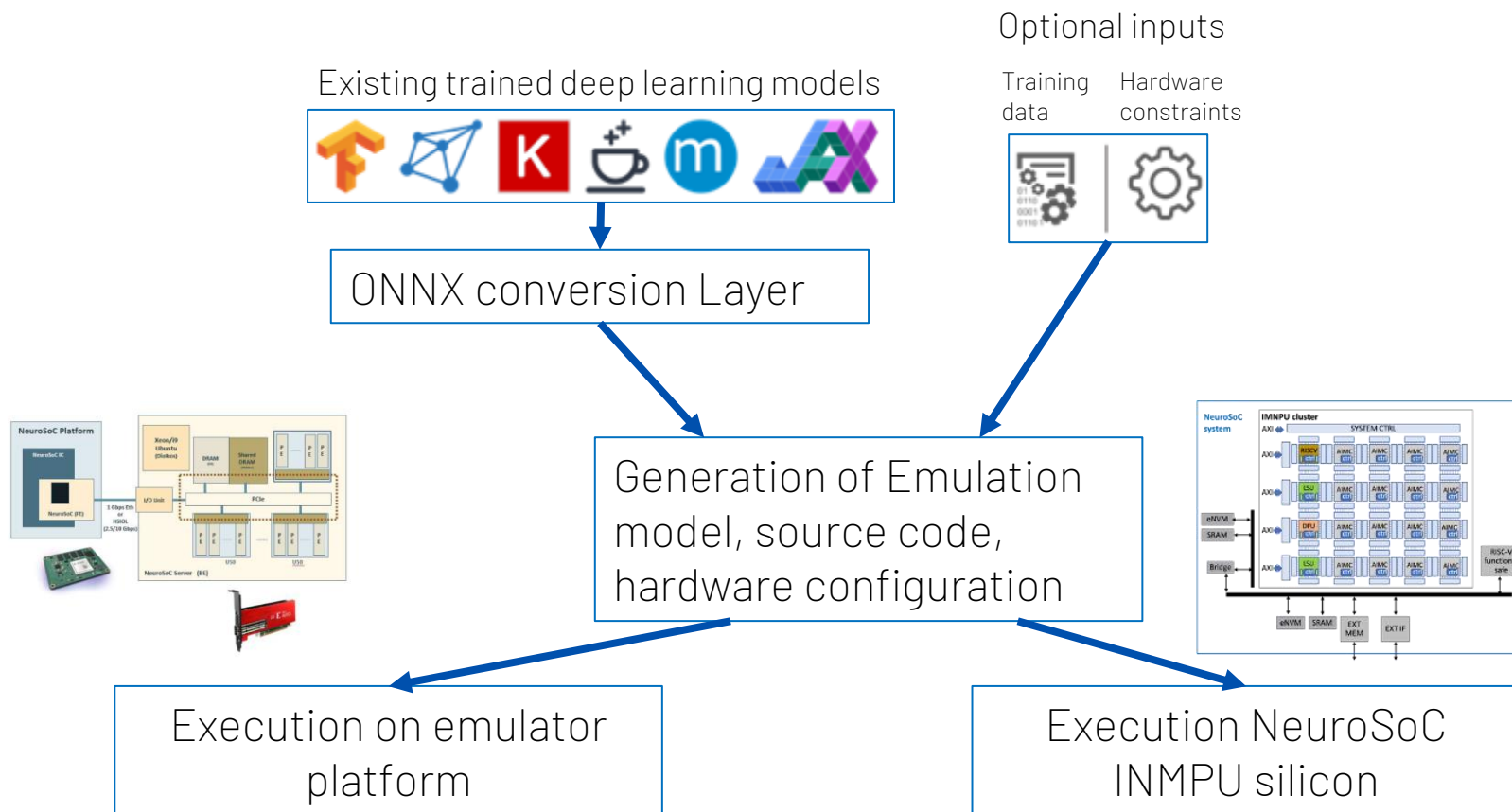
RISC-V Co processor features

- Complement the computing capabilities of the analog in-memory computing (AIMC) tiles and other specialized digital processing units (DPUs) present in the IMNPU
- Handles the execution of Deep Neural Network (DNN)-related workloads that must be executed at higher dynamic range for accuracy concerns, exploiting the floating-point arithmetic
- Supports various activation functions (ReLU, sigmoid, tahn), complex layers such as upsampling, depth-wise, softmax



Zaruba et al., IEEE Transactions on Computers (2020)

NeuroSoC Toolchain









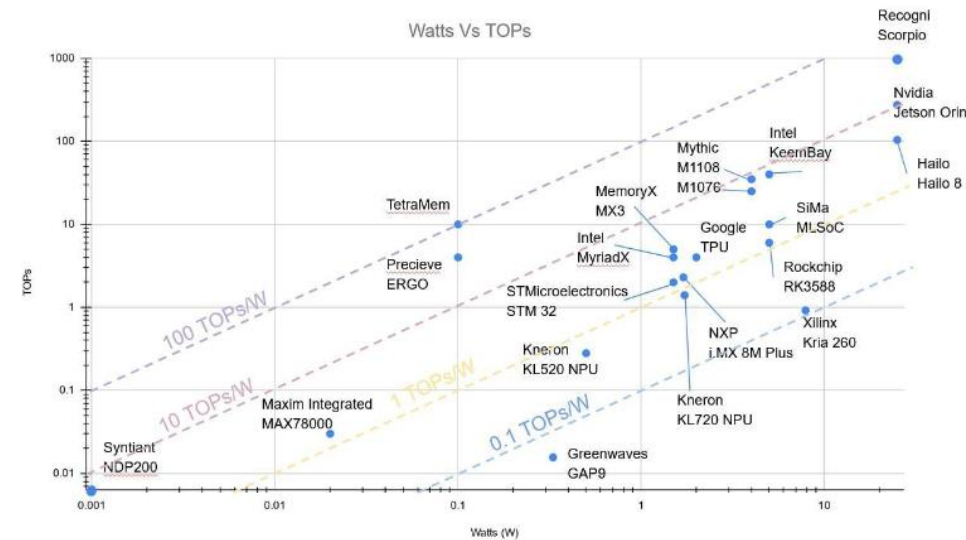
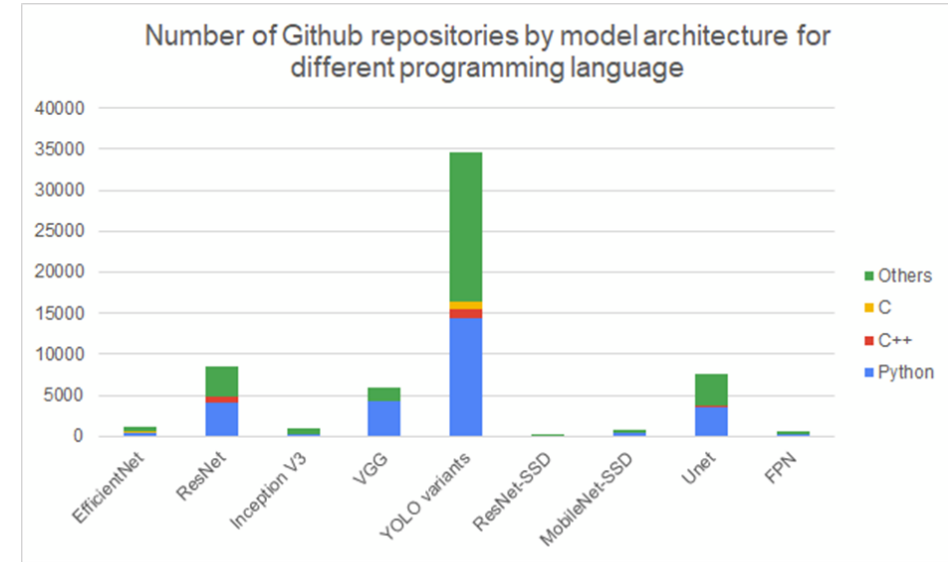
Software toolchain

- From high level network description
- Converted in an intermediate format
- Optimized for the specific platform (reconfigurable)
- Performance and functional emulation
- Execution on hardware once available

Applications requirements, integration, and use- cases demonstrations



-  Investigate edge-applications where NeuroSoC can offer a compelling advantage.
-  Selection and qualification of applications.
-  Benchmarking of SoA and emerging solutions.
-  Proposition of an evaluation framework.
 -  Toolchains, Accuracy, Power, Size, Throughput
-  Assessment of performances vs requirements.



Contact details

STMicroelectronics

 M. Giulio Urlini, Project Coordinator

 giulio[dot]urlini[at]stmicroelectronics[dot]com

Benkei

 Mrs Fabienne Brutin, Project administrative manager,

 fabienne[at]benkei[dot]fr

Acknowledgments

❁ 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 SBF1 22.00202 and 23.00205 and UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10040829]



**Funded by
the European Union**



**Innovate
UK**



Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

❁ Stay tuned: www.neurosoc.eu, <https://www.linkedin.com/company/neurosoc>