Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing
www.supercomputing-icsc.it/en/icsc-home/

Architectures and Design Methodologies to Accelerate AI Workloads

Speaker
Prof. Stefania Perri
University of Calabria – Italy
s.perri@unical.it

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI INGEGNERIA MECCANICA, ENERGETICA E GESTIONALE DIMEG
DIPARTIMENTO DI INGEGNERIA INFORMATICA, MODELLISTICA, ELETTRONICA E SISTEMISTICA DIMES

The partnership

ICSC
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

total investment 319.938.979,26 €

The mission of the National Centre:
To maintain and upgrade the Italian HPC and Big Data infrastructure

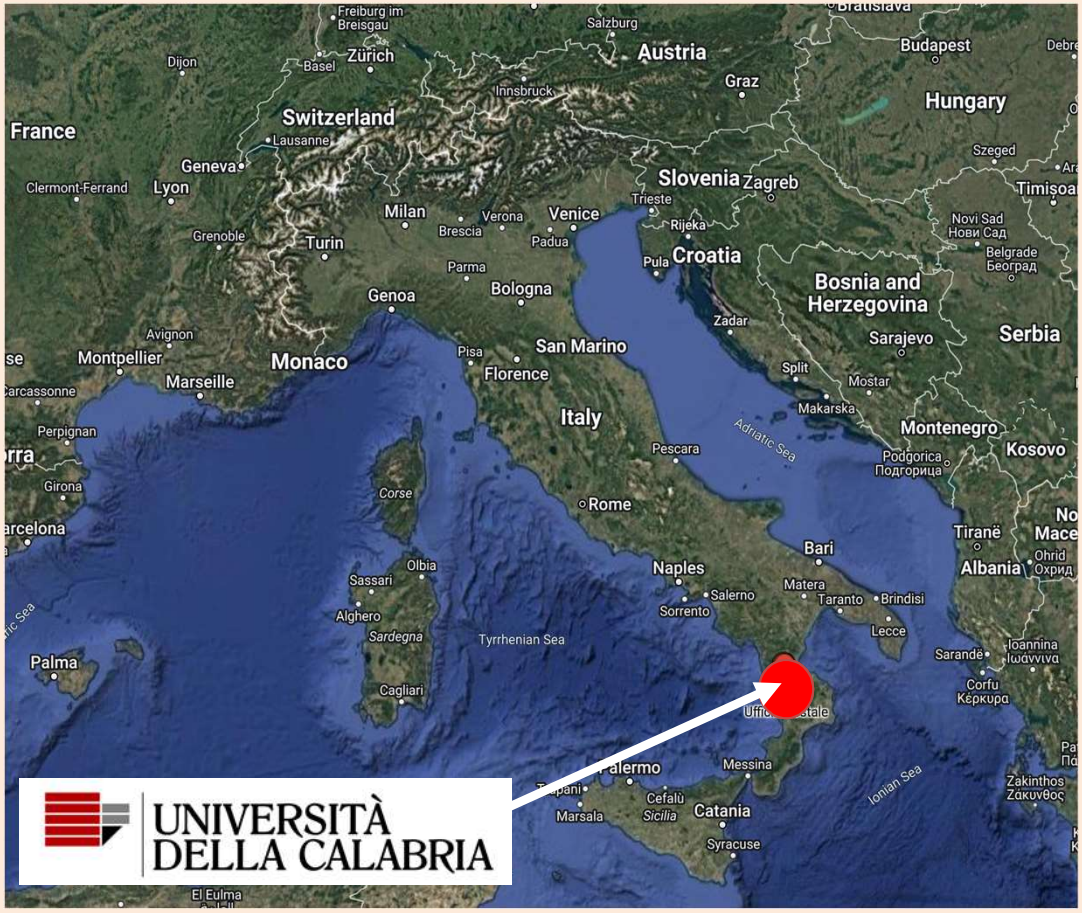
The organization: 51 partners - more than 200 researchers.

The Flagship2 of the Spoke 1

POLITECNICO MILANO 1863
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA
Università degli Studi di Ferrara
UNIVERSITÀ DI TORINO
UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II
Università di Catania
Politecnico di Torino
UNIVERSITÀ DELLA CALABRIA
ENEA Agenzia nazionale per le nuove tecnologie, l'energia e lo sviluppo economico sostenibile
TOR VERGATA UNIVERSITÀ DEGLI STUDI DI ROMA

My Institution

More than 800 Professors, 27000 students, 14 Departments, 200 fully-equipped laboratories, 73 EU Funded projects, more than 400 projects funded by the Italian Government

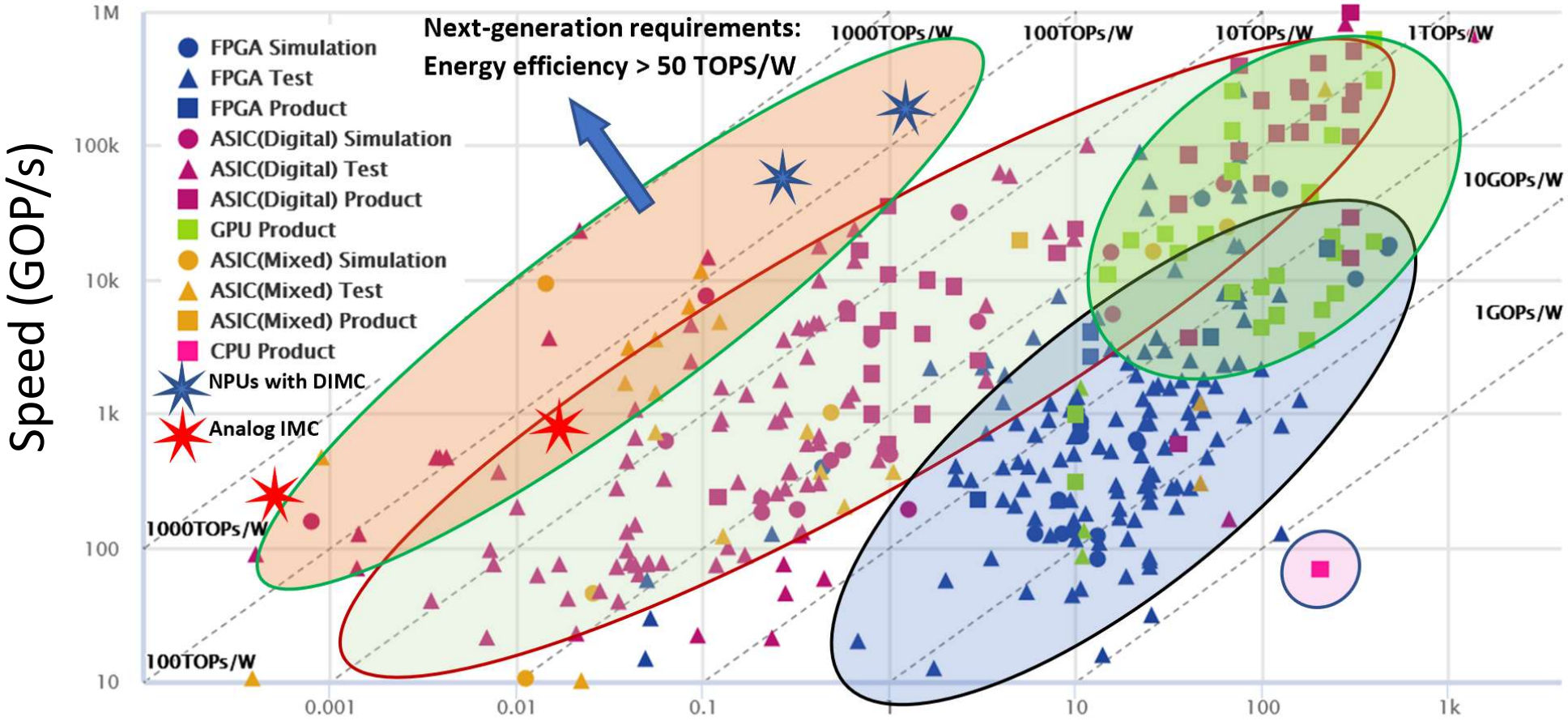


www.supercomputing-icsc.it/en/icsc-home/

www.unical.it

Energy Efficiency of AI accelerators: TOPs/W

ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing



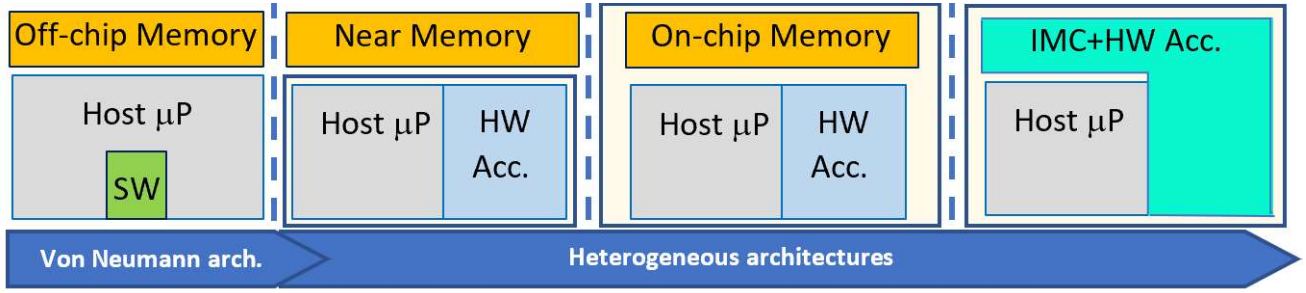
UNIVERSITÀ DELLA CALABRIA

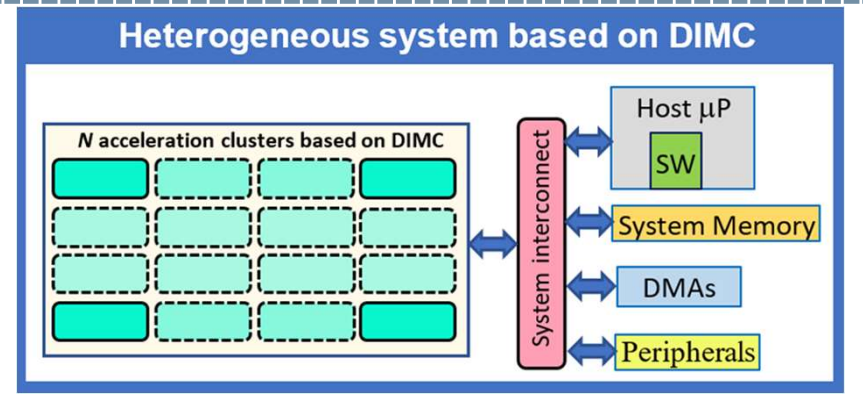
Sources: https://nicsefc.ee.tsinghua.edu.cn/project.html [Jhang et al.] "Challenges and Trends of SRAM-Based Computing-In-Memory for AI Edge Devices" IEEE Trans. On Circ. and Sys. I - 2021

Towards In-Memory Computing

UNIVERSITÀ DELLA CALABRIA

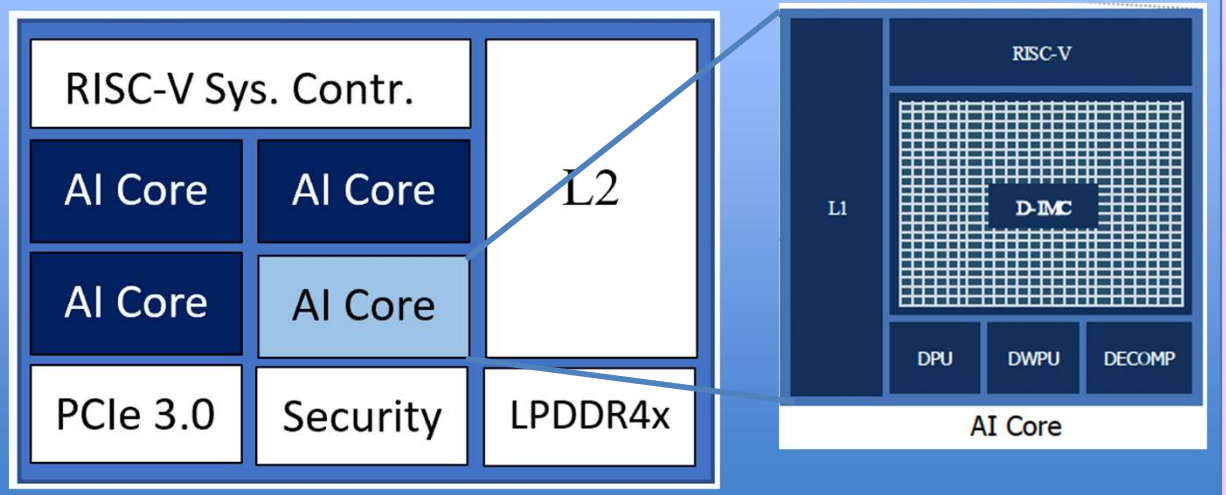
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing



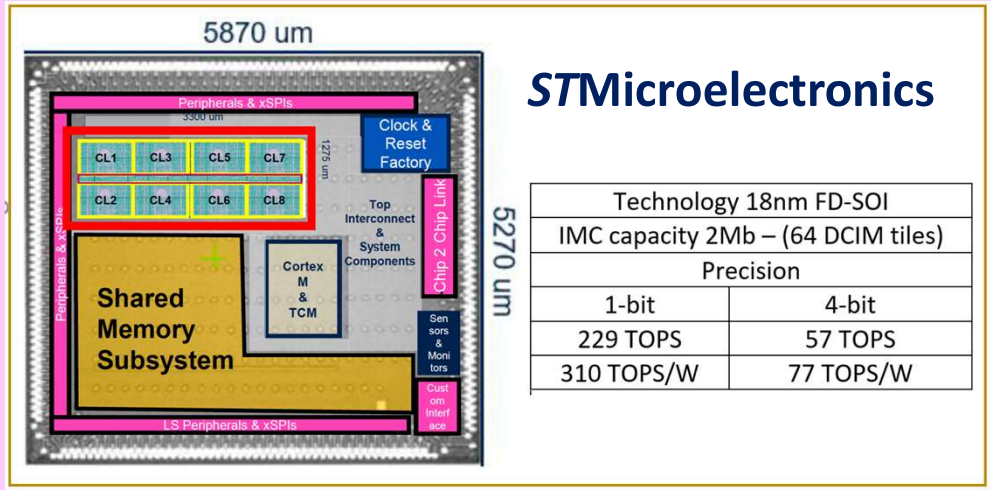


S. Perri, et al., "Digital In-Memory Computing to Accelerate Deep Learning Inference on the Edge". IPDPS 2024

Axelera Metis AIPU for edge AI inference



https://www.axelera.ai/blog/the-metis-ai-platform-in-detail



*ST*Microelectronics

Technology 18nm FD-SOI	
IMC capacity 2Mb – (64 DCIM tiles)	
Precision	
1-bit	4-bit
229 TOPS	57 TOPS
310 TOPS/W	77 TOPS/W

G. Desoli et al. "16.7 A 40-310TOPS/W SRAM-Based All-Digital Up to 4b In-Memory Computing Multi-Tiled NN Accelerator in FD-SOI 18nm for DL Edge Applications," IEEE ISSCC 2023

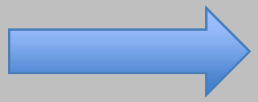
What do designers need?
Open Challenges

What are we doing?

C1. Specific design methodologies and EDA tools

R1. The new DIMCacti Framework
- 6T, 8T, 10T, 12T standard bit cells
- New memory cells optimized for IMC
- Different array sizes
- Energy, delay and area estimations

C2. Supporting different data precisions



R2. Introducing the SIMD paradigm
- 1b, 4b, 8b (16b?) operations
- Design of proper IPs for ASIC and FPGAs

C3. Optimizing energy and computational speed

R3. Introducing Approximate Computing
- Design of proper IPs for ASIC and FPGAs
- Adopting emerging technologies (e.g. RFETs)

CONTACT

stefania.perri@unical.it

www.unical.it

https://www.linkedin.com/in/stefania-perri-71a20673/

UNIVERSITÀ DELLA CALABRIA

DIPARTIMENTO DI **INGEGNERIA MECCANICA, ENERGETICA E GESTIONALE** DIMEG

DIPARTIMENTO DI **INGEGNERIA INFORMATICA, MODELLISTICA, ELETTRONICA E SISTEMISTICA** DIMES