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/

Exploration of Digital In-Memory Computing to Accelerate Deep Learning on the Edge

Cristina Silvano, Politecnico di Milano

Stefania Perri, Università della Calabria

POLITECNICO MILANO 1863
UNIVERSITÀ DELLA CALABRIA

AI Accelerators: Motivations

- The proliferation of AI applications is pushing on accelerating AI on hardware.
- Nvidia's Blackwell is part of this trend to meet the computational demands of AI.
- As AI applications move beyond HPC and data centers to edge devices, GPUs and TPUs are not enough anymore.
- AI accelerators on the edge pose stringent constraints on power consumption, latency, memory footprint and cost.
- Neural processing units (NPUs) are becoming a standard in heterogeneous platforms to enable AI inference on the edge.

D. Garisto, "Accelerating AI: The cutting-edge chips powering the computing revolution", Nature, News Feature, Vol.630, June 2024.

Computer Science > Hardware Architecture

[Submitted on 27 Jun 2023 (v1), last revised 12 Jul 2024 (this version, v2)]

A Survey on Deep Learning Hardware Accelerators for Heterogeneous HPC Platforms

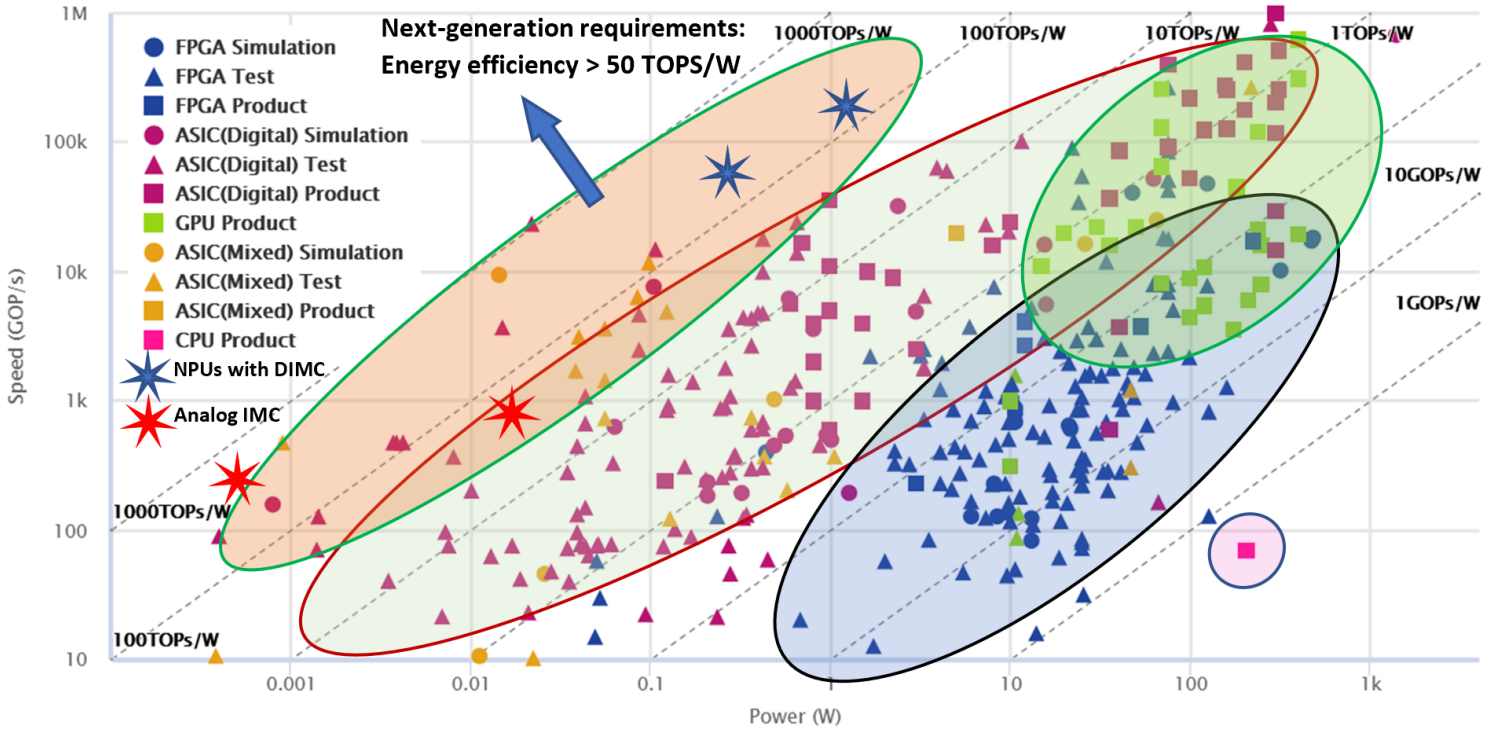
Cristina Silvano, Daniele Ielmini, Fabrizio Ferrandi, Leandro Fiorin, Serena Curzel, Luca Benini, Francesco Conti, Angelo Garofalo, Cristian Zambelli, Enrico Calore, Sebastiano Fabio Schifano, Maurizio Palesi, Giuseppe Ascia, Davide Patti, Nicola Petra, Davide De Caro, Luciano Lavagno, Teodoro Urso, Valeria Cardellini, Gian Carlo Cardarilli, Robert Birke, Stefania Perri

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

Survey Organization

https://arxiv.org/abs/2306.15552

Energy Efficiency



✱ Memory cell array + WL drivers

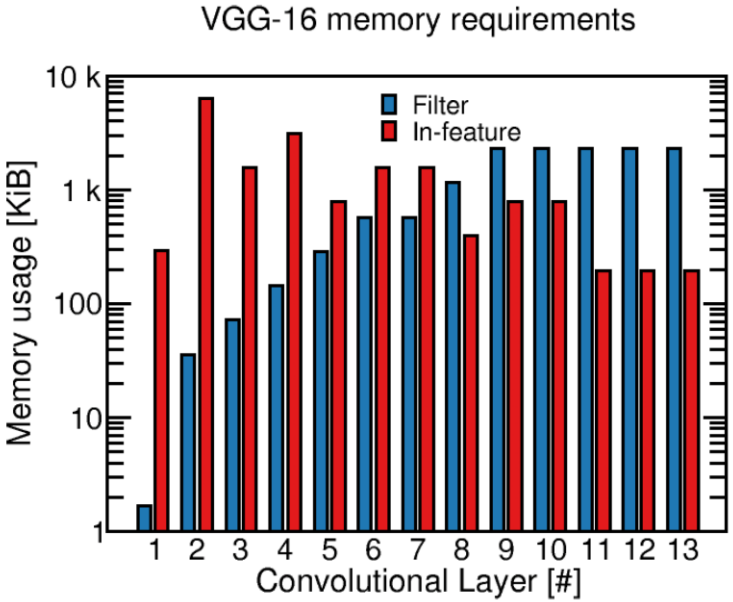
Sources: https://nicsefc.ee.tsinghua.edu.cn/project.html

[Jhang et al.] "Challenges and Trends of SRAM-Based Computing-In-Memory for AI Edge Devices" IEEE Trans. On Circ. and Sys. I - 2021

Memory usage per layer & Energy cost of RAM accesses

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

Memory usage changes significantly per layer of DNN models [A. Erdem, TACO 2020]



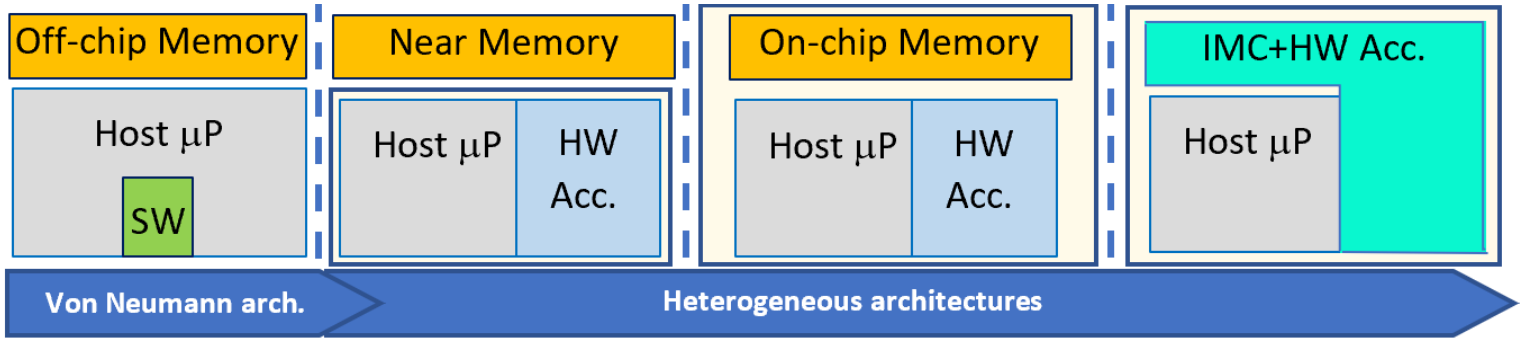
Data transfers from/to off-chip memory dominate the energy consumption

Energy per operation values from [Horowitz, ISSCC 2014]

Operation	E [pJ]	Relative Energy Cost
8b Add	0.03	
16b Add	0.05	
32b Add	0.1	
16b FP Add	0.4	
32b FP Add	0.9	
8b Multiply	0.2	
32b Multiply	3.1	
16b FP Multiply	1.1	
32b FP Multiply	3.7	
32b SRAM Read (8KB)	5	
32b DRAM Read	640	

Towards In-Memory Computing

Trend of HW accelerators to integrate processing and memory resources



Improved compute density and energy efficiency

Analog vs. Digital In-Memory Computing

To optimize the energy efficiency in terms of TOPs/Watt, **In-Memory Computing** represents an effective path towards the next generation of HW accelerators for data-intensive DNN tasks on the edge.

Analog IMC

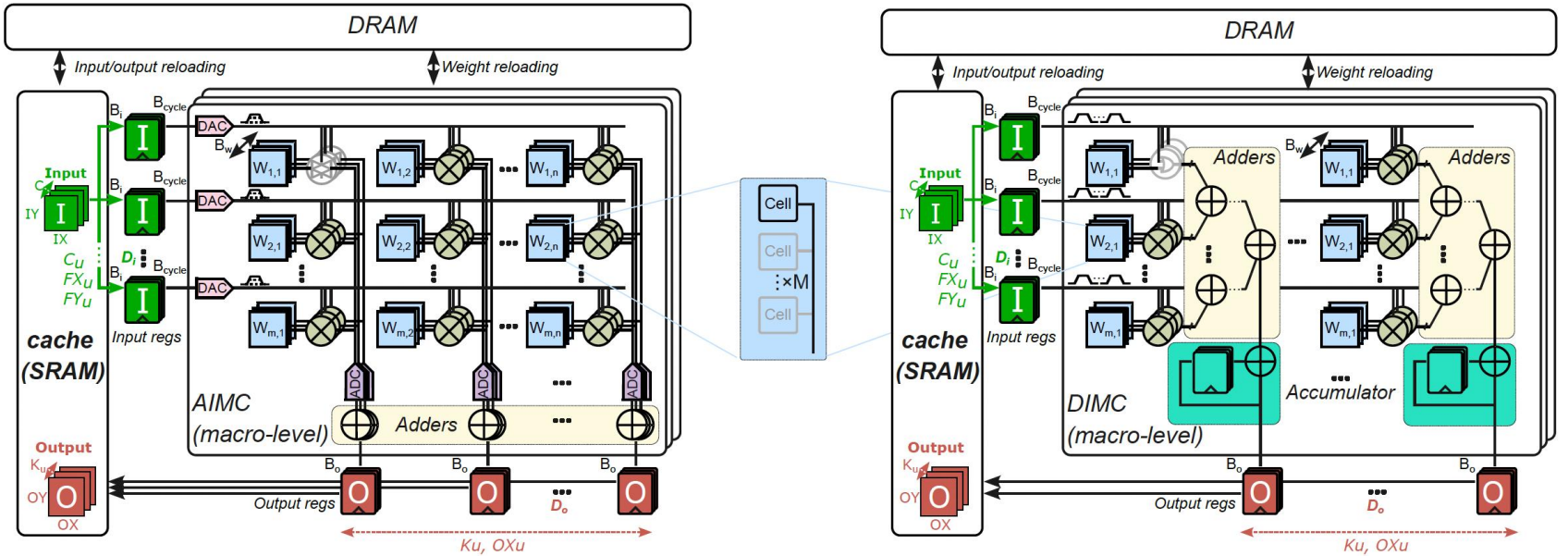
- Based on emerging NVM technologies such as Resistive RAM (ReRAM) and Phase Change Memory (PCM).
- Better energy efficiency
- Sensitivity to process and temperature variations
- Circuits nonidealities leads to low density and computing inaccuracies

Digital IMC

- Fully digital SRAM-based technologies
- Compatible for integration in SoC nanotechnologies.
- Deterministic behavior
- Robusteness
- Accuracy
- Higher computation density
- More flexibility

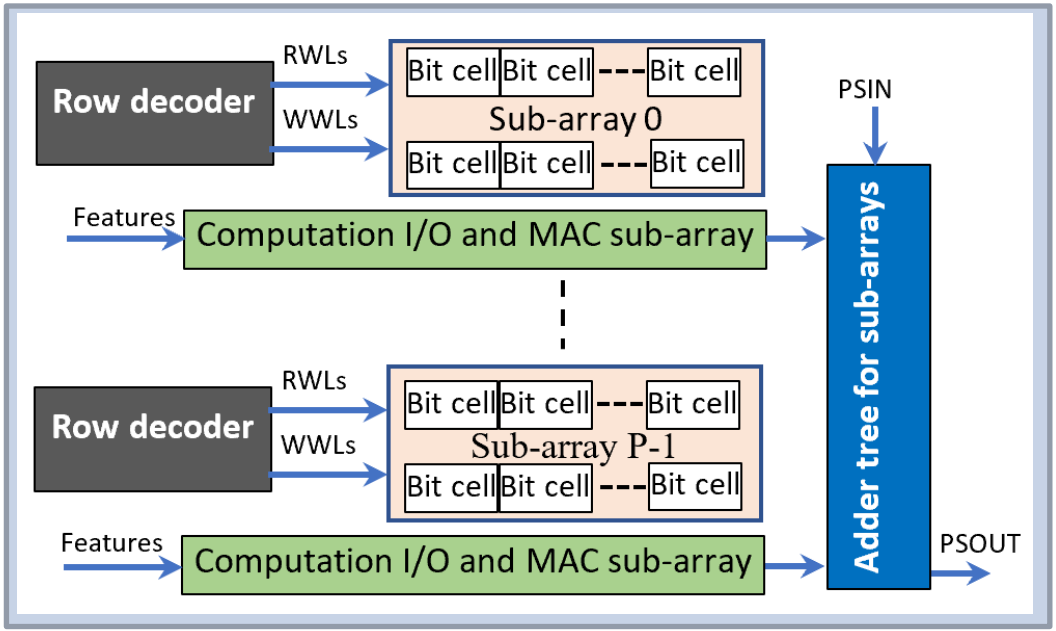
S. Perri, C. Zambelli, D. Ielmini, C. Silvano, "Digital In-Memory Computing to Accelerate Deep Learning Inference on the Edge". IPDPS (RAW Workshops) 2024

Analog vs. Digital In-Memory Computing: A system-level perspective



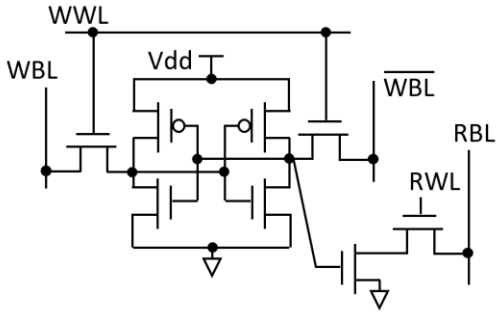
J. Sun et al., “Analog or Digital In-Memory Computing? Benchmarking Through Quantitative Modeling”, ICCAD 2023, 10.1109/iccad57390.2023.10323763

Digital IMC architecture - ISSCC 2023



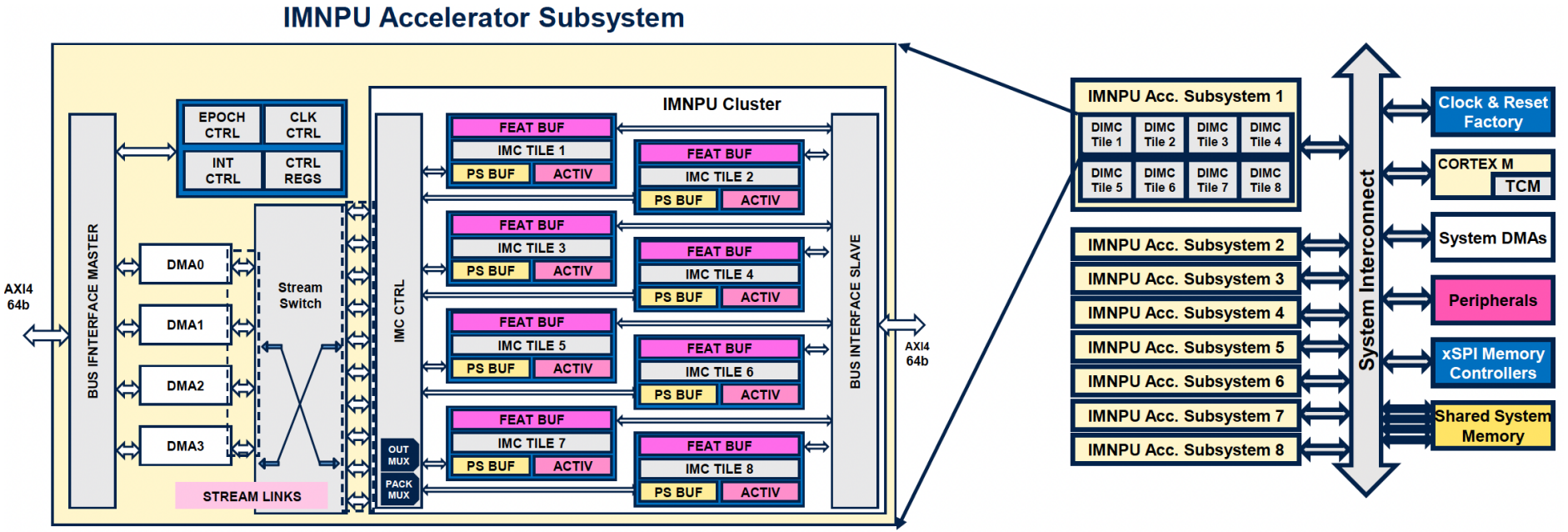
- 32Kb SRAM-based IMC supporting 1, 2, and 4b operations;
- 256-bit word-length;
- 4 sub-arrays, each 32x256 bit;
- Each sub-array has a local row decoder, sense amplifiers, output buffers, I/O and computing circuitry;

8T SRAM cells with decoupled read and write ports



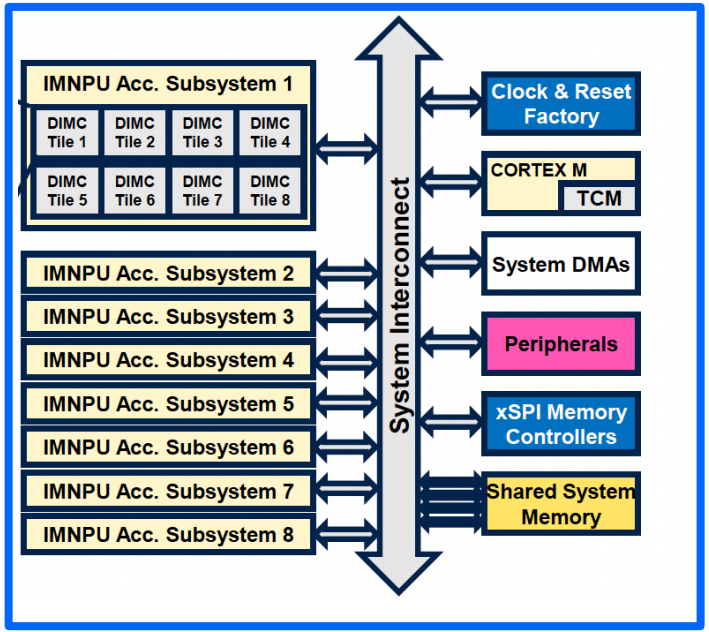
[G. Desoli et al.] "16.7 A 40-310TOPS/W SRAM-Based All-Digital Up to 4b In-Memory Computing Multi-Tiled NN Accelerator in FD-SOI 18nm for DL Edge Applications," IEEE Int. Solid-State Circ. Conf. (ISSCC), 2023

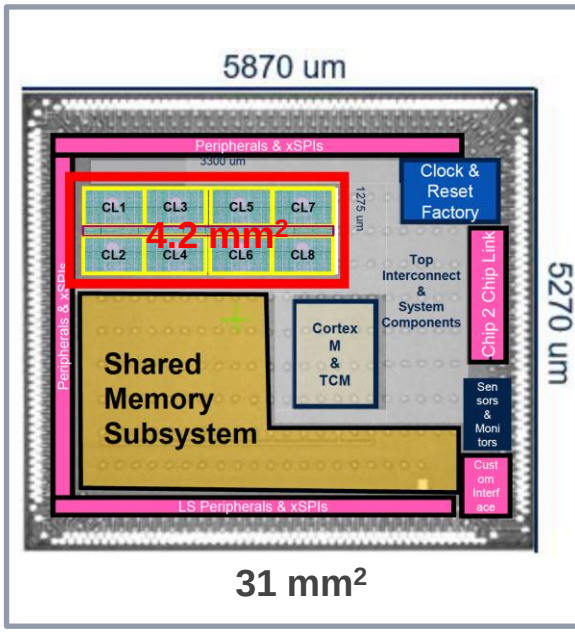
Digital IMC-based NPU System-on-Chip in 18nm



[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 Int. Solid-State Circ. Conf. (ISSCC), 2023

Digital IMC-based NPU System-on-Chip in 18nm (cont'd)

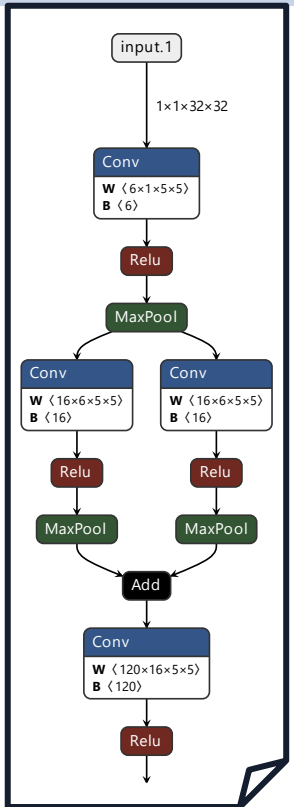


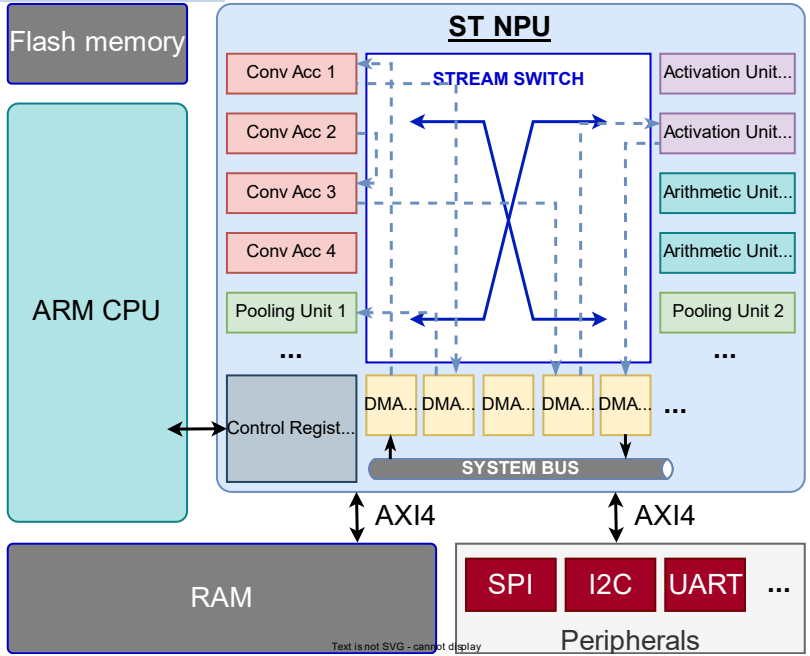


Technology 18nm FD-SOI	
IMC capacity 2Mb – (64 DCIM tiles)	
Precision	
1-bit	4-bit
229 TOPS	57 TOPS
310 TOPS/W	77 TOPS/W

[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 Int. Solid-State Circ. Conf. (ISSCC), 2023

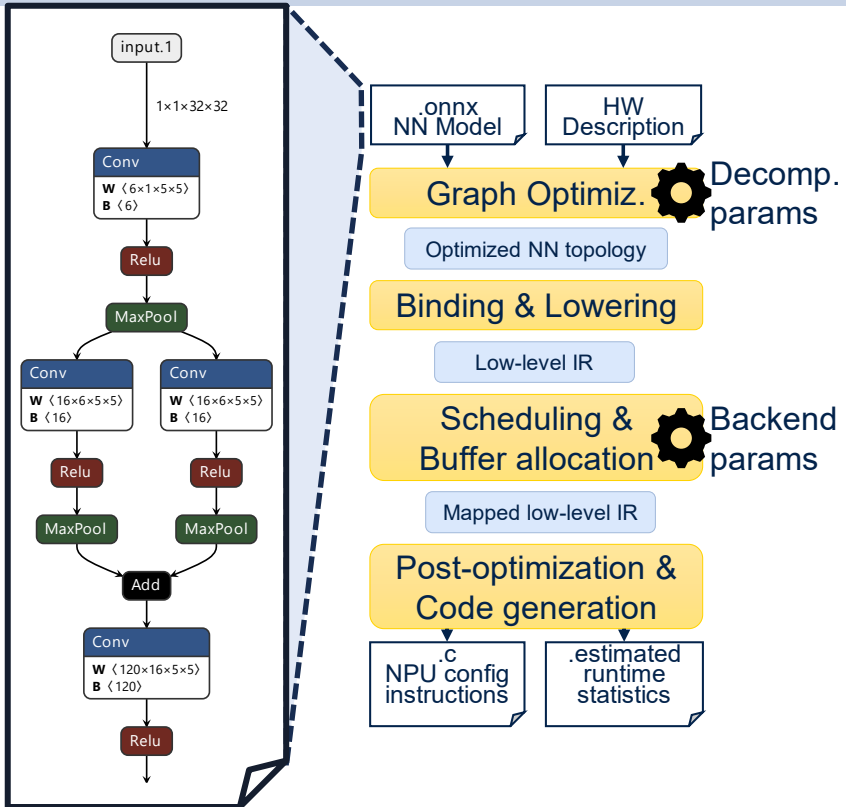
Layer-wise Exploration of NPU Compiler's Optimization Space

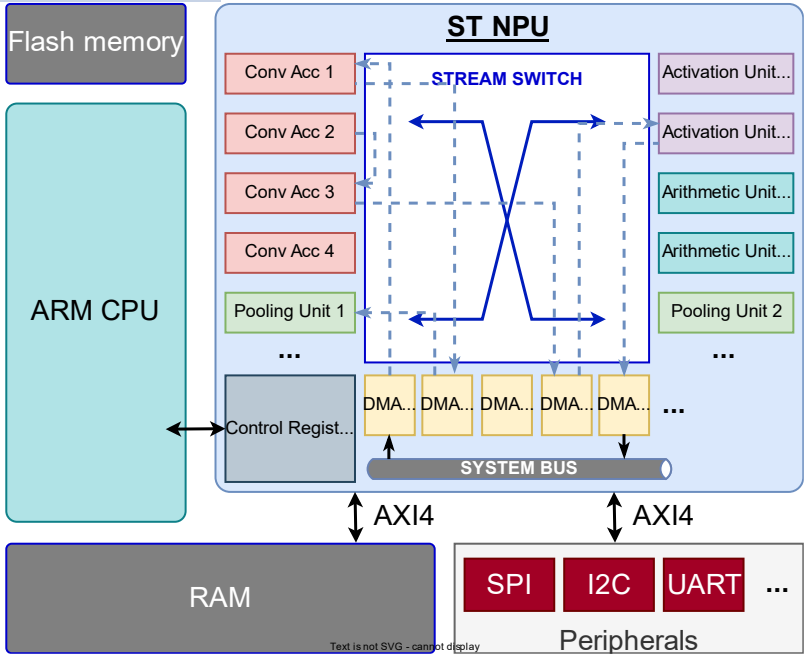




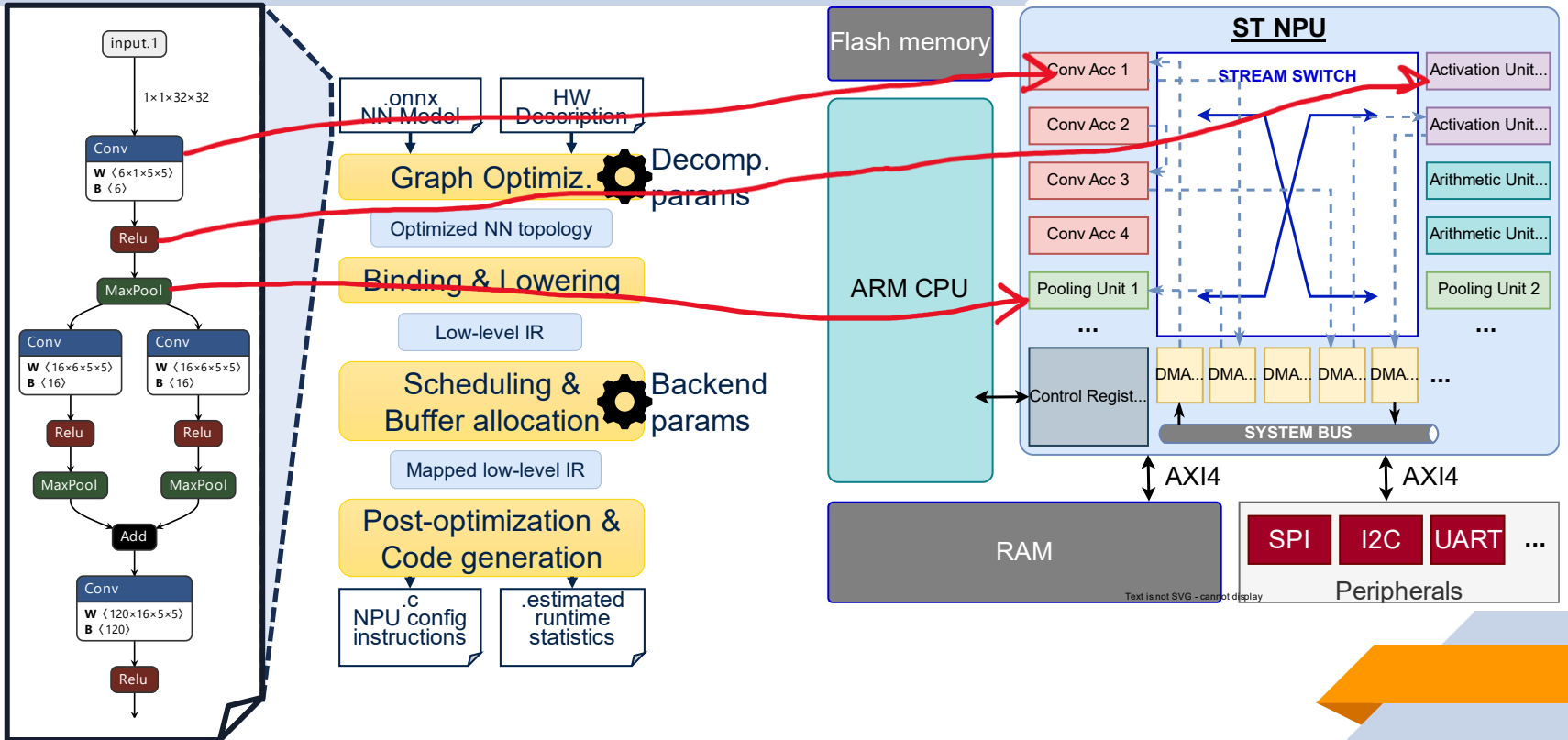
F. Indirli et al., "Layer-wise Exploration of a NPU Compiler's Optimization Space, ICCTA 2024

Layer-wise Exploration of NPU Compiler's Optimization Space

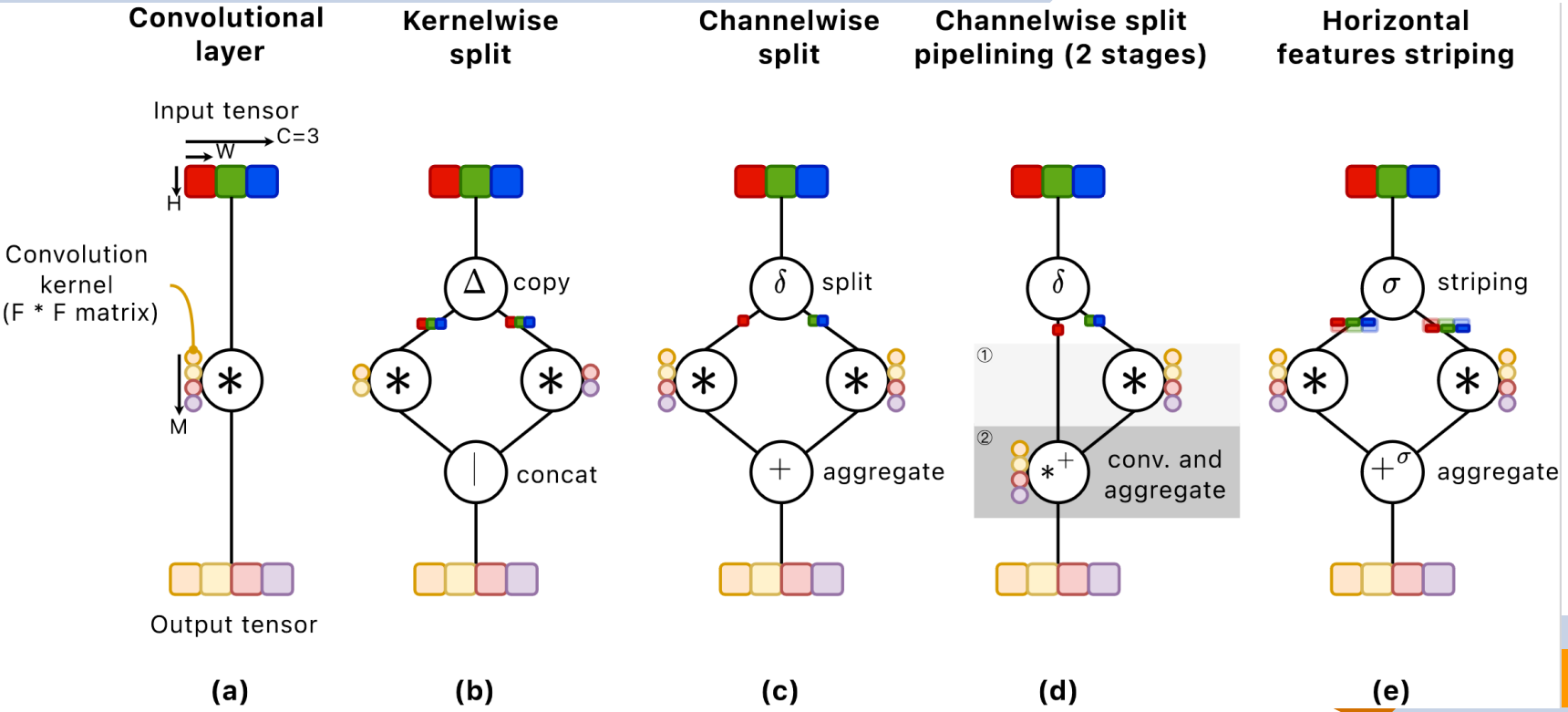




Layer-wise Exploration of NPU Compiler's Optimization Space

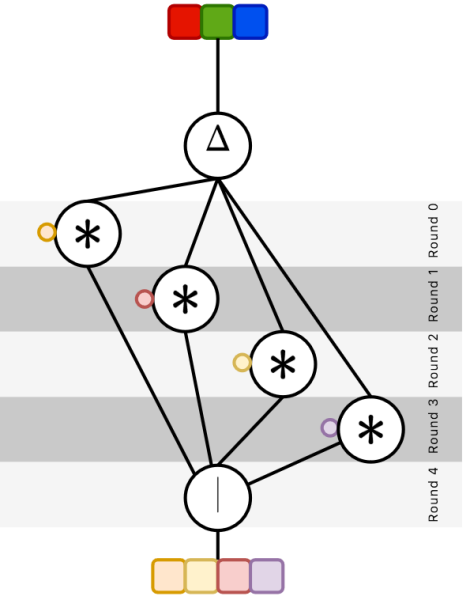


Graph transformations and mappings on PEs

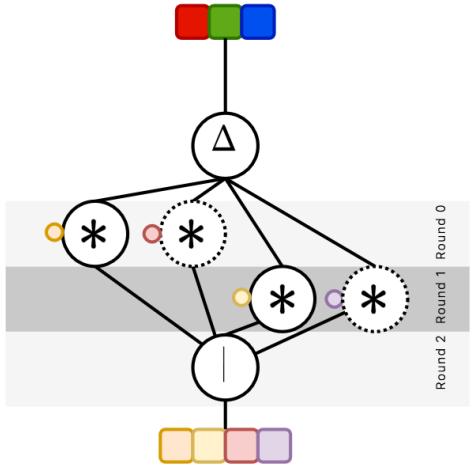


Graph transformations and mappings on PEs

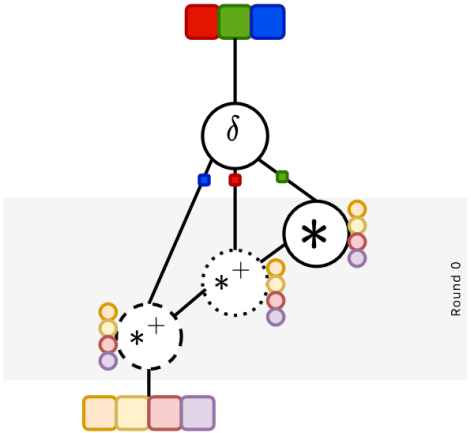
Kernelwise
(mapping 1PE/round)

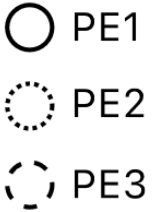


Kernelwise
(mapping 2PE/round)



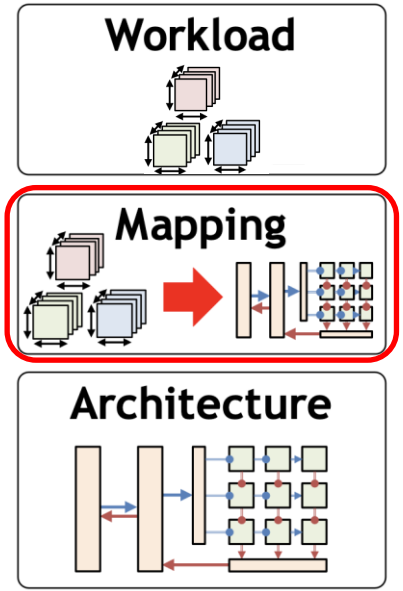
Channelwise
pipelining (3 PE)





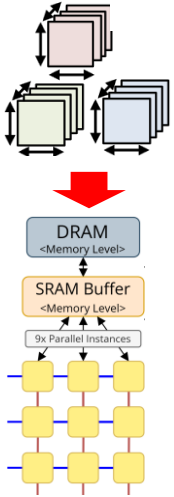
Performance improvement

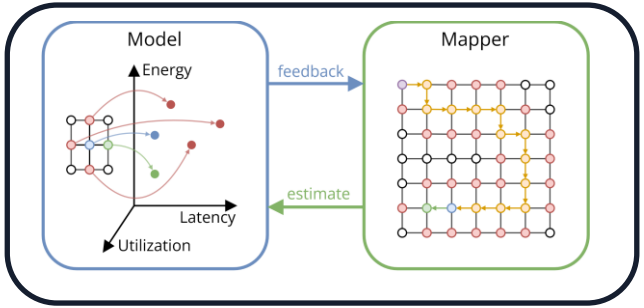
Automatic Exploration of Mapping AI Workloads to Digital SRAM-based IMC

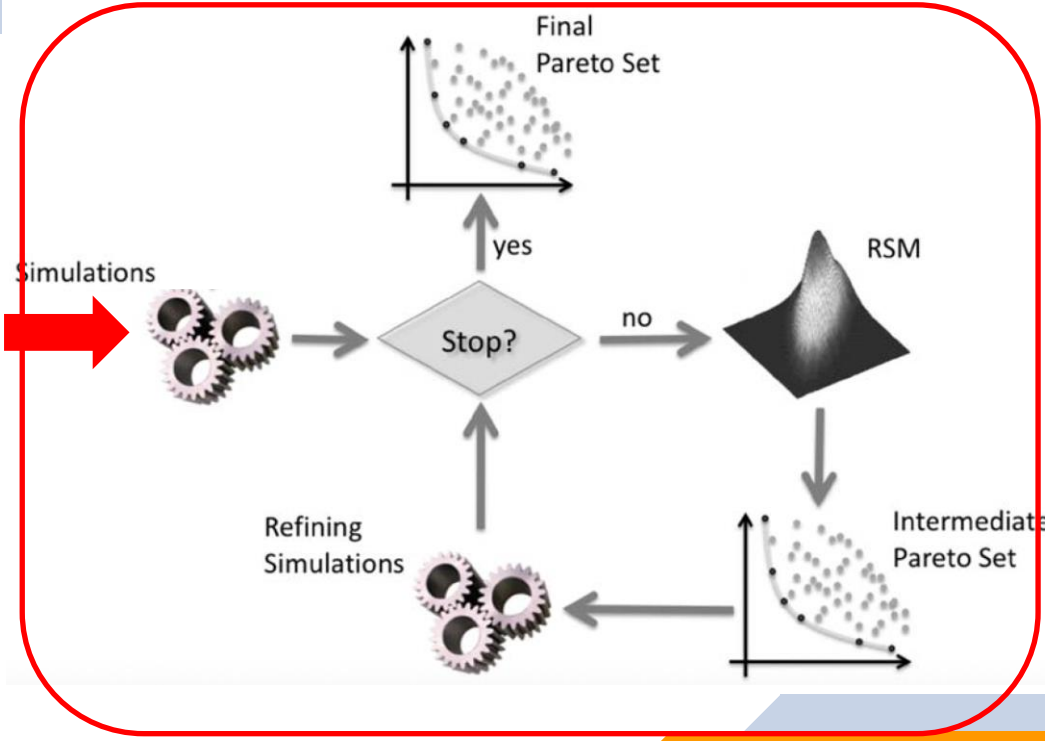


- IMC can deliver massive amounts of OPS/cycle by reducing data movement and exploiting parallelism.
- Per-layer mapping exploration space is huge!
- Need of automatic exploration of AI mapping strategies to better exploit the computing resources of IMC-based NPUs.

Automatic Exploration of Mapping AI Workloads to Digital SRAM-based IMC

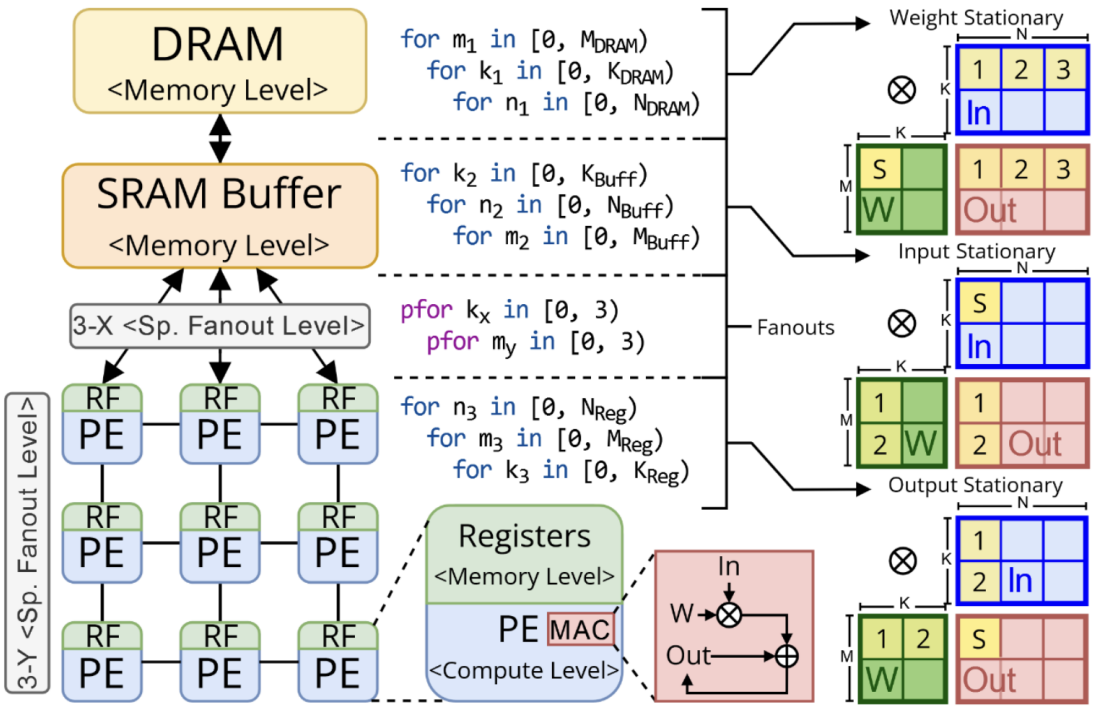






- T. Andrulis, J. S. Emer, V. Sze, "CiMLoop: A Flexible, Accurate, and Fast Compute-In-Memory Modeling Tool", 2024, arxiv.org/abs/2405.07259 https://github.com/mit-emze/cimloop
- G. Palermo, C. Silvano, V. Zaccaria: «ReSPIR: A Response Surface-Based Pareto Iterative Refinement for Application-Specific Design Space Exploration». IEEE Trans. CAD, (2009)

Mapping GEMMs on Spatial Architectures

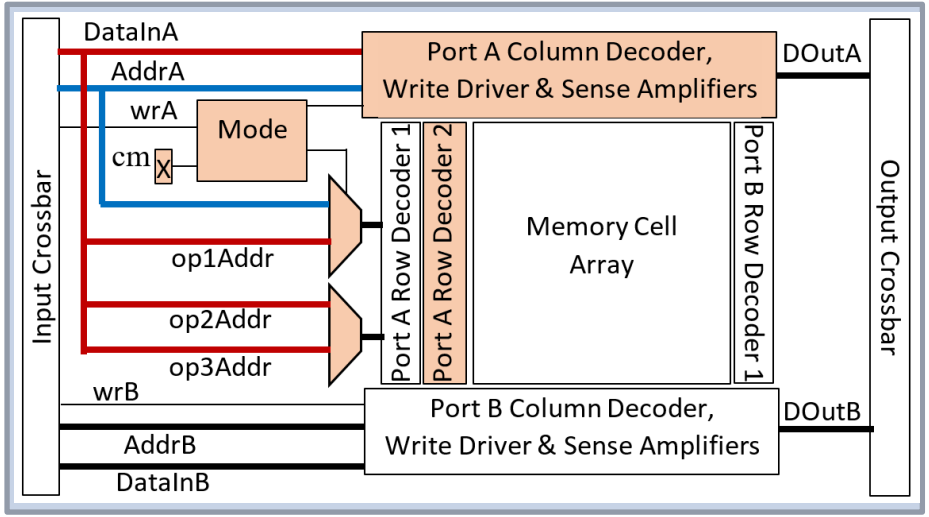


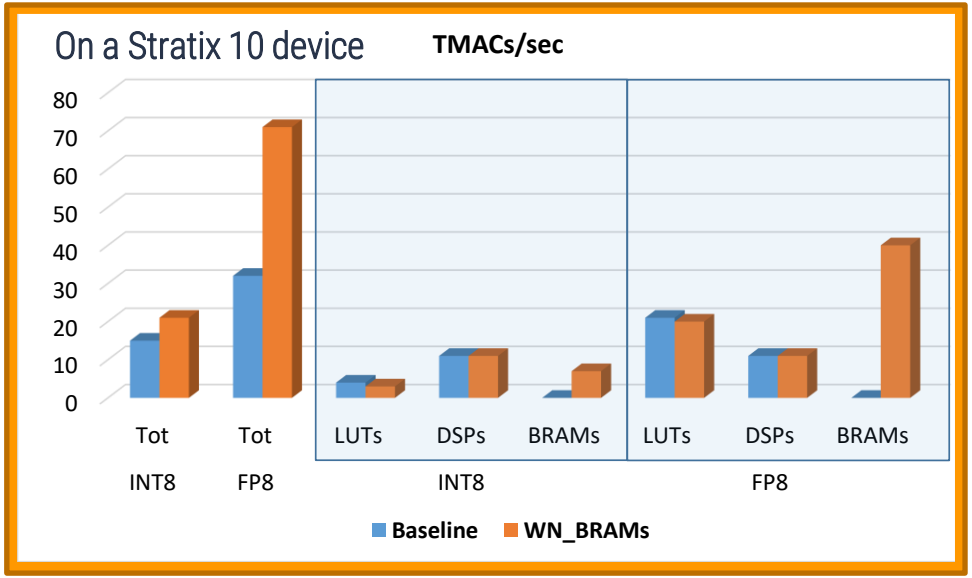
Marco Ronzani, Cristina Silvano «FactorFlow: Mapping GEMMs on Spatial Architectures through Adaptive Programming and Greedy Optimization Author(s)", Accepted to ASP-DAC 2025

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/

Proposals to exploit IMC on FPGAs

Digital IMC architectures on FPGAs (1)

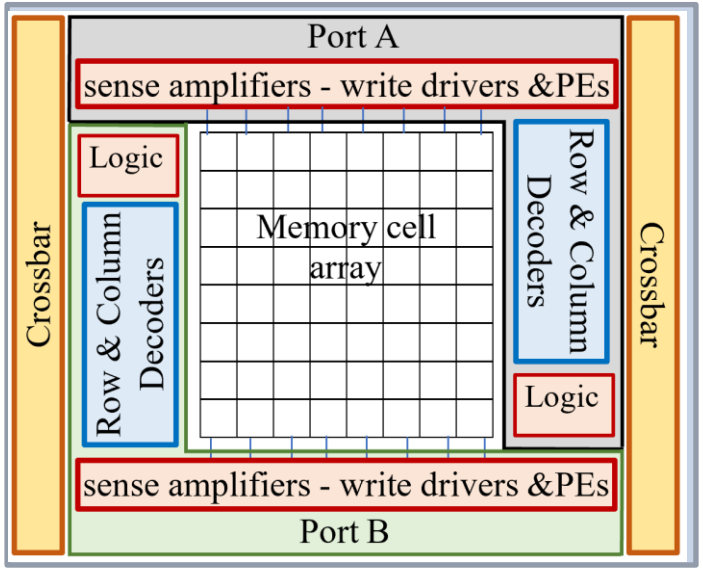




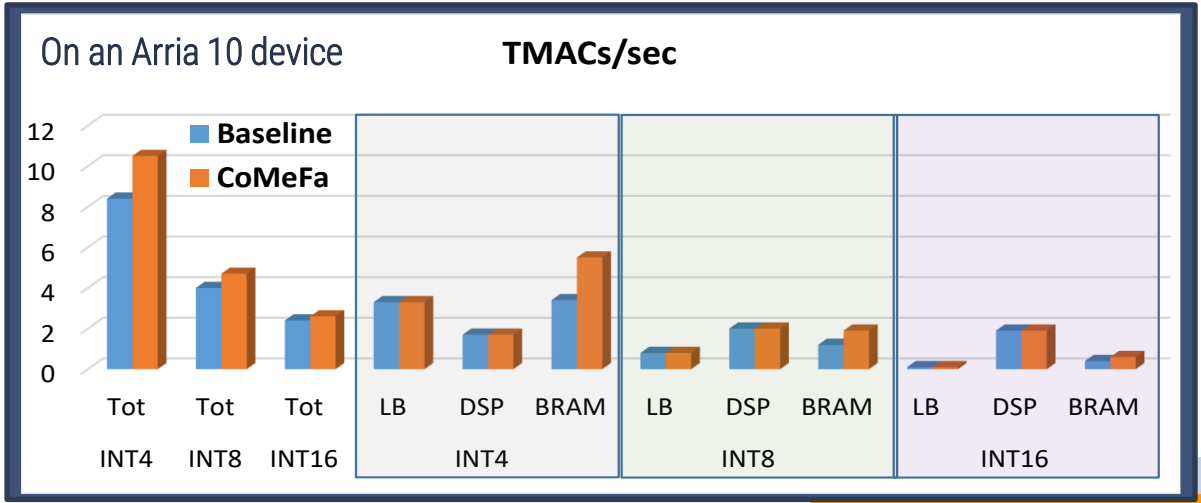
Spice simulations on a 28nm process technology at 0.9V have shown that the cycle time in the computational mode is 1.6x higher than the memory mode. For a 64Kb BRAM the area overhead is less than 8%

[X. Wang et al.] “Compute-Capable Block RAMs for Efficient Deep Learning Acceleration on FPGAs”, IEEE Int. Symp. On Field-Programmable Custom Computing Machine (FCCM), 2021

Digital IMC architectures on FPGAs (2)



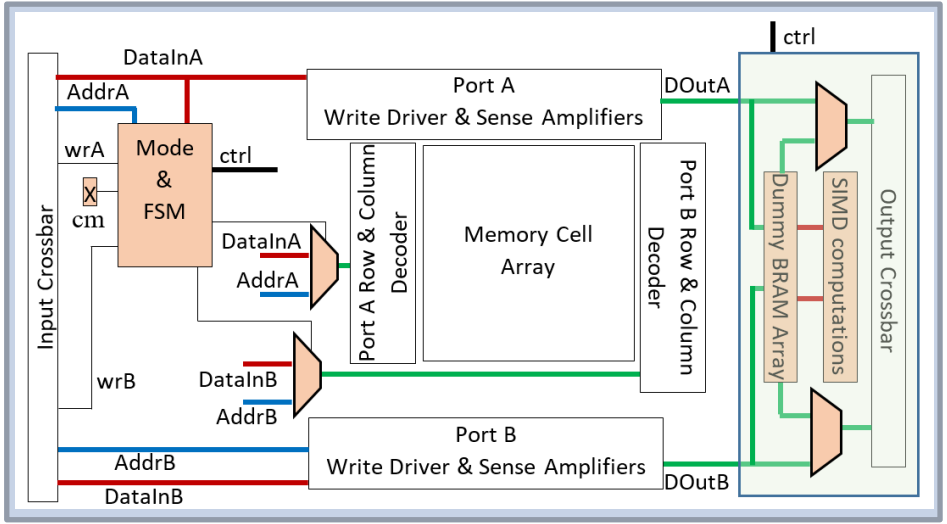
- The Compute-in-memory block integrates PEs in the sense amplifiers;
- One bit-serial PE for each bitline;
- Additional logic (comparators, multiplexers, etc.) are introduced;



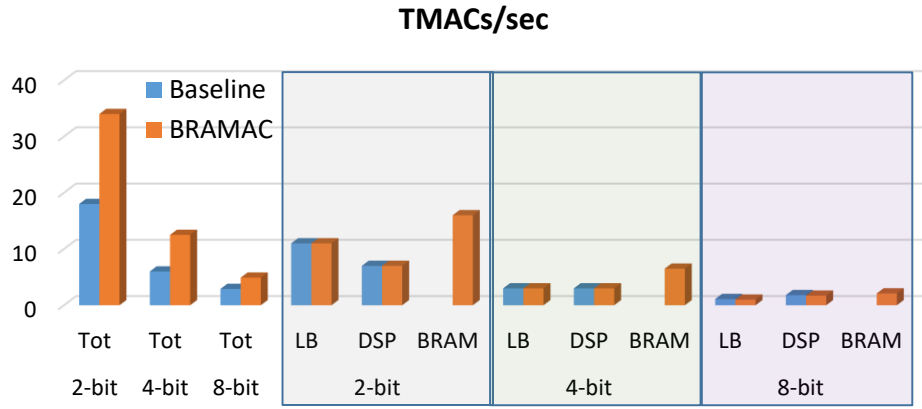
Verilog designs evaluated on a 20kb BRAM for several applications

[A. Arora et al.] “CoMeFa: Deploynig Compute-in-Memory on FPGAs for Deep Learning Acceleration”, ACM Trans. On Rec. Tech. and Sys., Vol. 16, n°3, 2023

Digital IMC architectures on FPGAs (3)



Two operation modes:
- Memory mode: a conventional BRAM
- Computation mode: data are copied to the dummy BRAM and processed in SIMD fashion (bit-serial multiplication & bit-parallel addition)

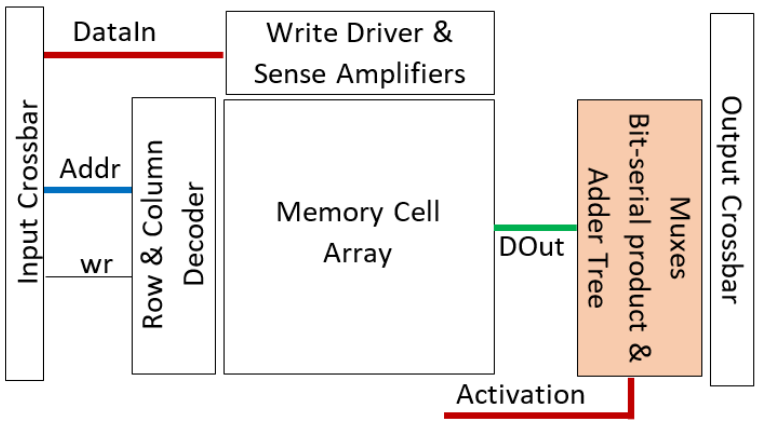


The BRAMAC architecture introduces an area overhead of 17%

When applied to accelerate AlexNet and ResNet-34 DNNs the achieved speed up is higher than 1.5

[Y. Chen et al.] “BRAMAC: Compute-in-BRAM Architectures for Multiply-Accumulate on FPGA”, IEEE Int. Symp. On Field-Programmable Custom Computing Machine (FCCM), 2023

Digital IMC architectures on FPGAs (4)



- The Compute-in-memory block uses 10T memory cells;
- In the compute mode, DataIn transfers Activations;
- Activations are multiplied by the weights stored in the memory;
- The products furnished by bit-serial multipliers are then accumulated by an adder tree;
- In the memory mode, the weighs are stored
- 8-bit activations and weights

For experiments on DL models:
- 45nm technology process @100MHz
- 64 x 512 memory cell array
- The counterparts are DSP-based architectures (Intel Agilex)

Some sample results

	Power [mW]	Delay [sec]	Area Utilization (%)	BRAMs	CIM_BRAMs
Conv_Layer (DSP)	17.68	26	58.6	5	0
Conv_Layer (IMC)	7.64	15	23.5	1	1
Lenet (DSP)	516.8	3444	57.9	162	0
Lenet (IMC)	227.2	1122	12.7	9	22
ResNet18 (DSP)	147.7	382	28.8	67	0
ResNet18 (IMC)	114.8	241	15.4	48	3

[Y. Liet al.] “An All-digital Compute-in-memory FPGA Architecture for Deep Learning Acceleration”, ACM Trans. On Rec. Tech. and Sys., Vol. 17, n°1, 2024

Open challenges, trends and conclusions

1. Need of specific design methodologies and EDA tools

Automatic compilation and mapping frameworks are needed to exploit DIMC

Expertise in DL algorithms, hardware architectures, compilers.

2. Make DIMC able to operate at different data precision

Introducing the SIMD paradigm to achieve accuracy/power/latency/area tradeoffs

3. Increase reconfigurability

For ASICs introducing multiple operating modes through emerging device technolgies, e.g. RFETs

For both ASICs and FPGAs adopting *runtime resources management* to optimize power and speed at each layer and to support tensor transformations

Finanziato dall'Unione europea
NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Thank you for your attention

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