ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

XXXV Cycle

Exploration of Beyond von Neumann Computing to solve the Memory-Wall

Andrea Coluccio

Supervisor: Prof. Mariagrazia Graziano

Research context and motivation

- The exponential development of transistor technology has been the main driving force behind modern electronics. However, this process has slowed over time, introducing **performance bottlenecks** in **data-intensive applications**. The leading cause is the classical von Neumann architecture, which entails constant data exchanges between the processing unit and data memory, wasting time and power.
- The main bottleneck is the **Memory-Wall:** CPUs are becoming more efficient and faster, but the memories cannot follow the same trend.

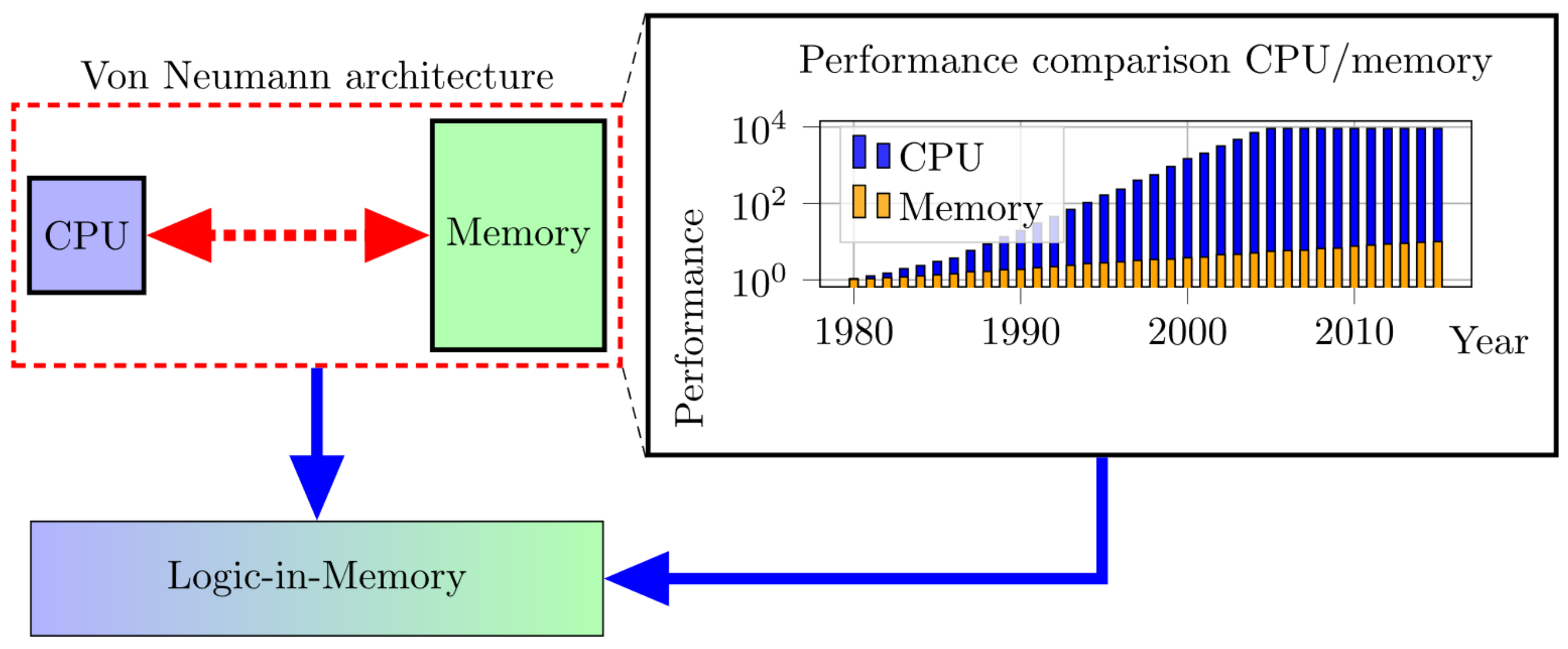


Fig.1: von Neumann Bottleneck. Performance comparison CPU-Memory [4]

- **Logic-in-Memory** is rapidly spreading, bringing computing elements as near as possible to memory while inserting customized processing elements to elaborate more data.
- **Energy** and **time** are saved through **parallel execution** and usage of processing components with **local memory elements**.

Addressed research questions/problems

- Modern and emerging computing approaches, especially Logic-in-Memory, usually require a complete **redefinition** of the design paradigm, resulting in a **very long and complex process**.
- For this reason, in recent years, engineers have realized specific software or **CADs** to assist designers in emerging computing paradigms or technologies.
- This thesis work focuses on the **architectural model** shown in Fig. 2 and answers the following question.

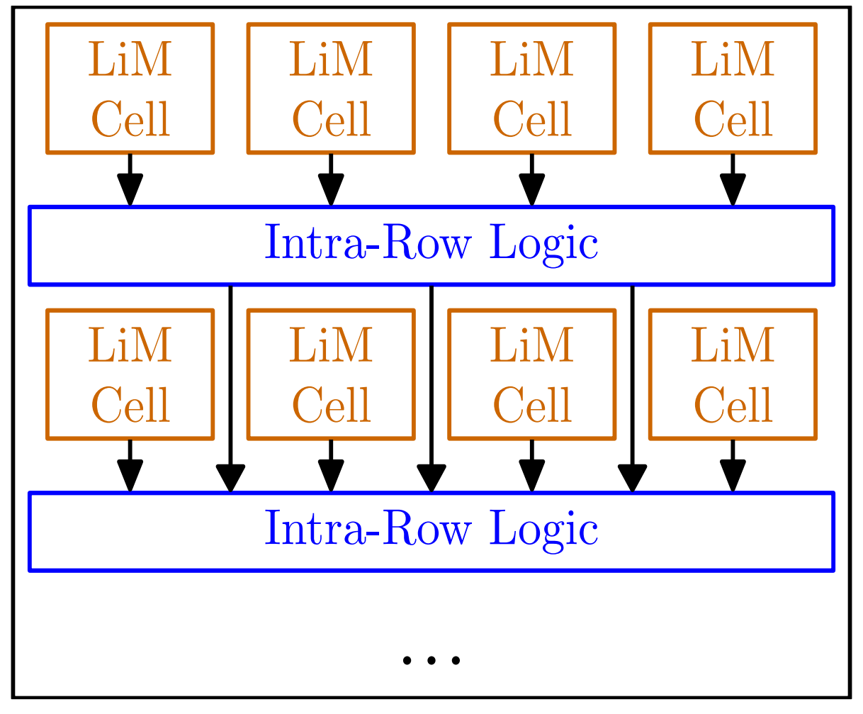
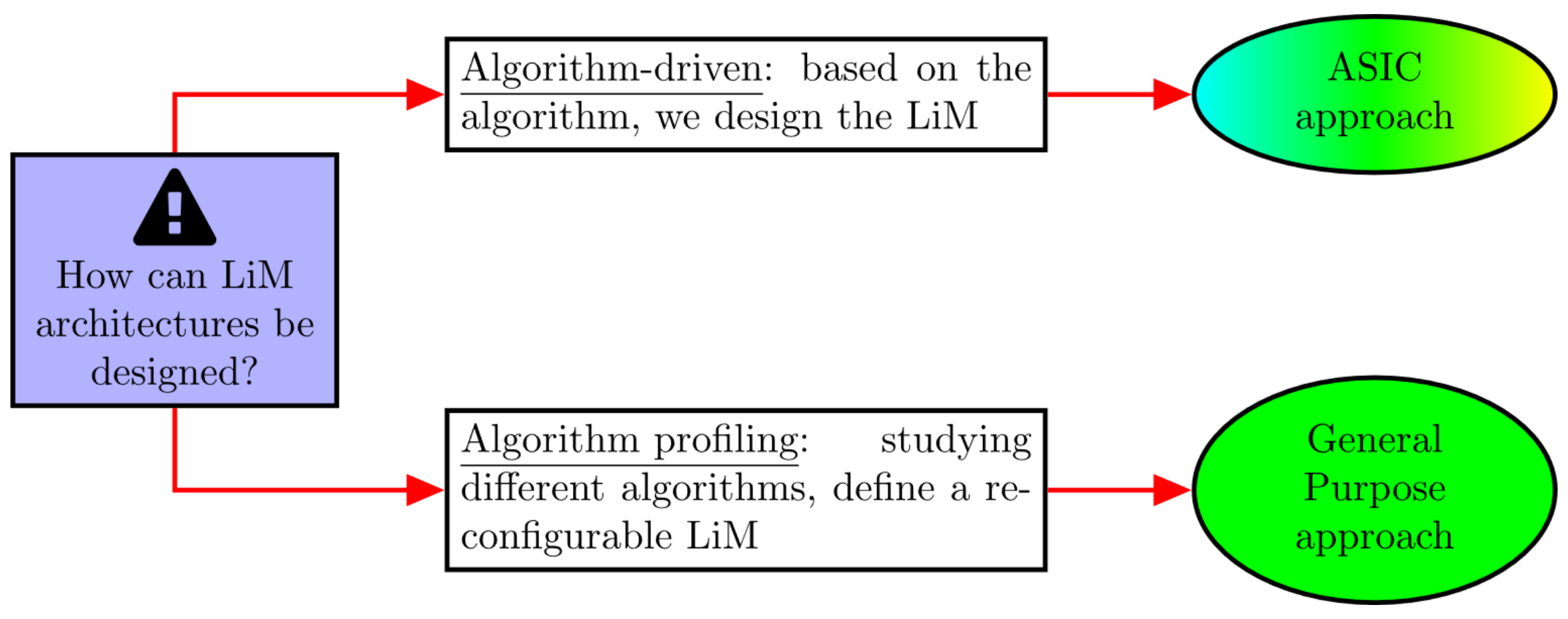


Fig.2: LiM architectural model [4-7]


How can LiM architectures be designed?
Algorithm-driven: based on the algorithm, we design the LiM
ASIC approach
Algorithm profiling: studying different algorithms, define a re-configurable LiM
General Purpose approach

- However, **ASIC** (Fig. 3) and **General Purpose** (Fig. 4) solutions require lots of **manual work**, starting from the design of the single LiM Cell, the Intra-Row Logic blocks, the top-entity architecture, and the control unit.

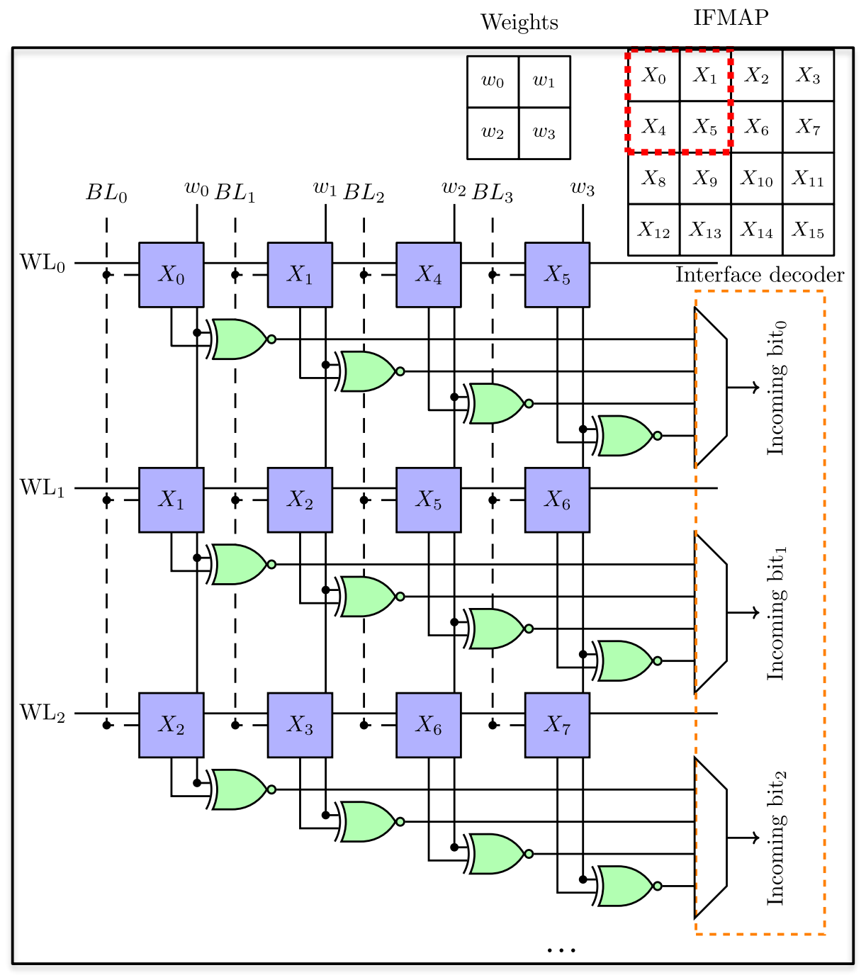


Fig. 3: ASIC LiM implementation of a XNOR-Net [2]

- In this work, a tool called **D**esign **Ex**plorer for **I**n-**M**emory **A**rchitectures (**DExIMA**) is proposed, which is able to assist the designer in the realization of the Logic-in-Memory structures.

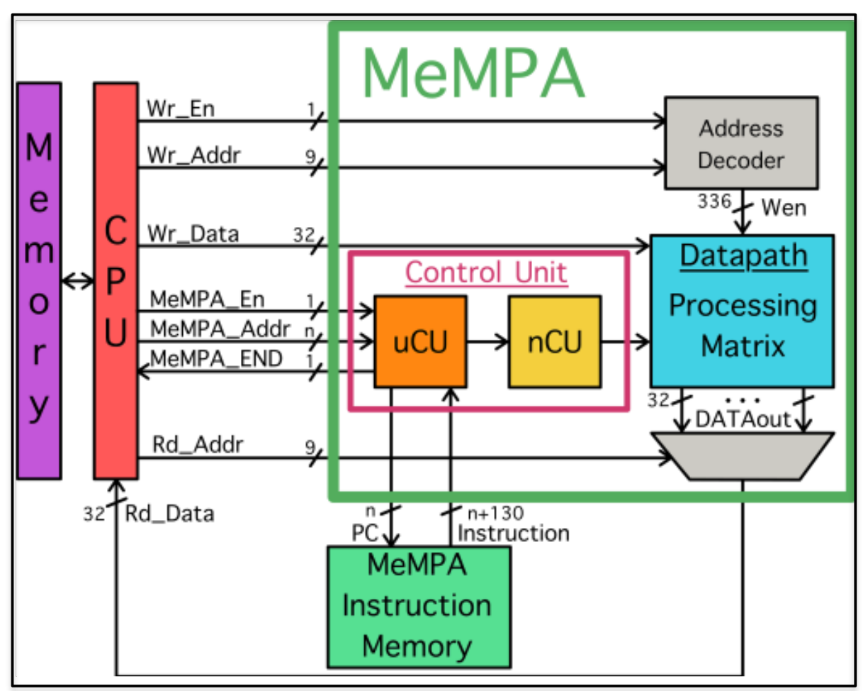


Fig. 4: General Purpose LiM implementation [6]

- **DExIMA** design flow starts from the **LiM architecture definition** and implements **automatic verification** and **performance estimations.**

Submitted and published works

- [1] S. D. Antonietta, A. Coluccio et al., "WINNER: a high speed high energy efficient Neural Network implementation for image classification," 2019 26th IEEE International Conference on Electronics, Circuits and Systems (ICECS), 2019, pp. 29-32
- [2] Coluccio, et. al, G. Logic-in-Memory Computation: Is It Worth It? A Binary Neural Network Case Study. J. Low Power Electron. Appl. 2020, 10, 7
- [3] A. Marchesin, A. Coluccio et al., "Octantis: An Exploration Tool for Beyond von Neumann architectures," 2021 16th International Conference on Design & Technology of Integrated Systems in Nanoscale Era (DTIS), 2021, pp. 1-5
- [4] A. Coluccio et al., "Hybrid-SIMD: A Modular and Reconfigurable Approach to Beyond von Neumann Computing," in IEEE Transactions on Computers, vol. 71, no. 9, pp. 2287-2299, 1 Sept. 2022
- [5] A. Coluccio et al., ""RISC-Vlim, a RISC-V Framework for Logic-in-Memory Architectures", accepted in MDPI Electronics
- [6] A. Guastamacchia, A. Coluccio et al., "MeMPA: a Memory Mapped M-SIMD Co-processor to cope with the Memory-Wall Issue", submitted to ACM Transactions on Computer Systems
- [7] A.Coluccio et al., "DExIMA: Design Explorer for In-Memory Architectures", submitted to IEEE Transactions on Computer-Aided Design (TCAD)

Novel contributions

- **General Purpose** LiM architectures are defined by employing the **Algorithm Profiling** approach.
  - ➢ Benchmarks are executed with a standard **CPU-Memory** architecture and profiled in terms of executed instructions, implementing the most recurrent in LiM[6].

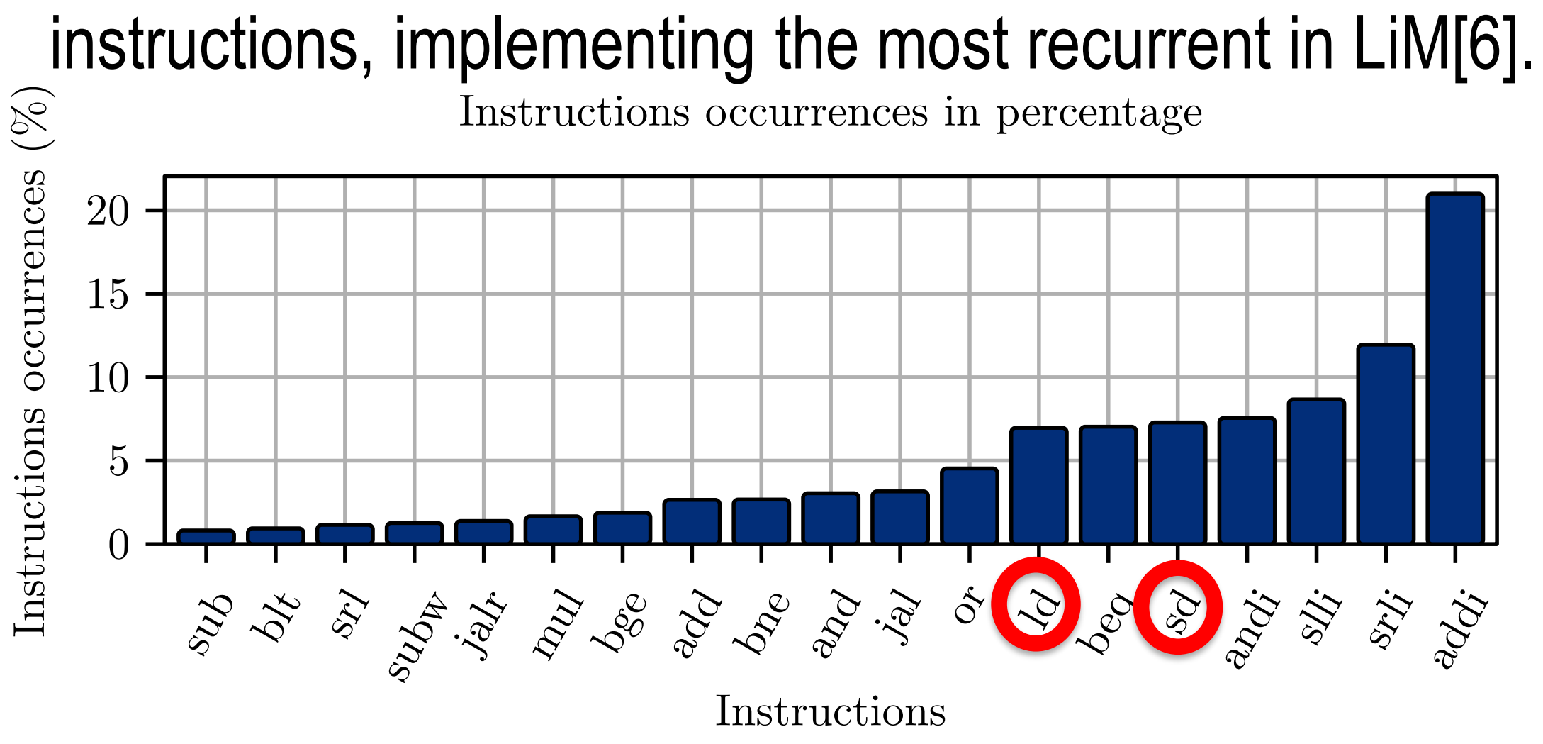


Fig.5: Instruction profiling of SPLASH-2 benchmarks.* [6]

* Simulation with a RISC-V based In-Order processor with 2 levels of caches using Gem5

- LiM Design Flow is automatized with **DExIMA**, and the LiM results are **compared** automatically with a von Neumann Architecture.

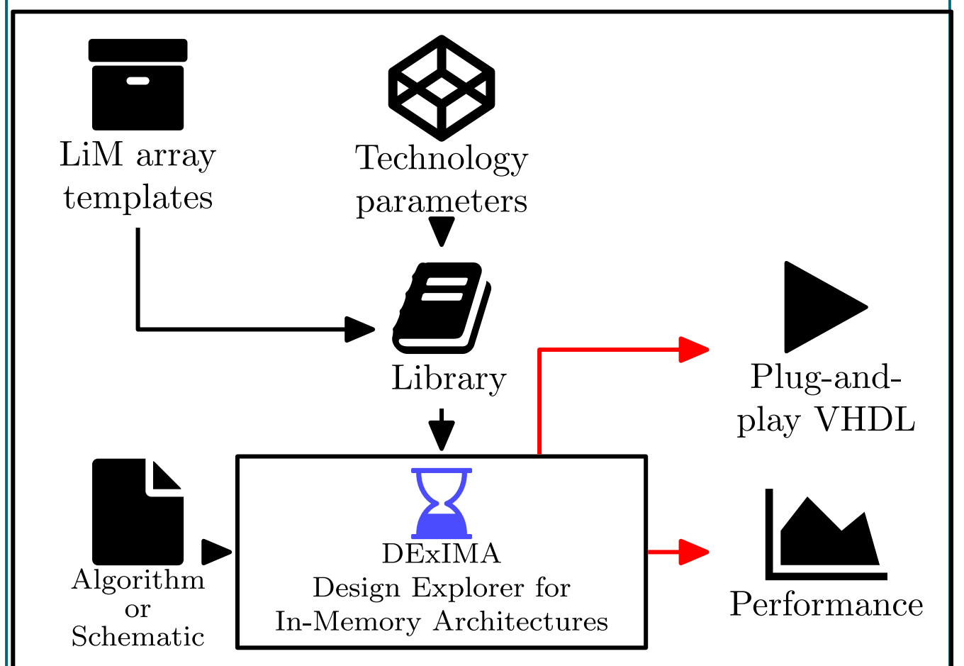


Fig.6: DExIMA High-Level scheme

Adopted methodologies & Results

DExIMA tool implements the **LiM design flow** shown in Fig. 7. The results in Fig.8 refer to Matrix-Vector Multiplication (**MVM**) algorithm. Fig. 8 (a) shows the performance results of the LiM array. Figs. 8(b-c) show the instruction count for the CPU-Memory and CPU-Memory-LiM solutions, respectively. Finally, Fig. 8 (d) illustrates the performance comparisons.

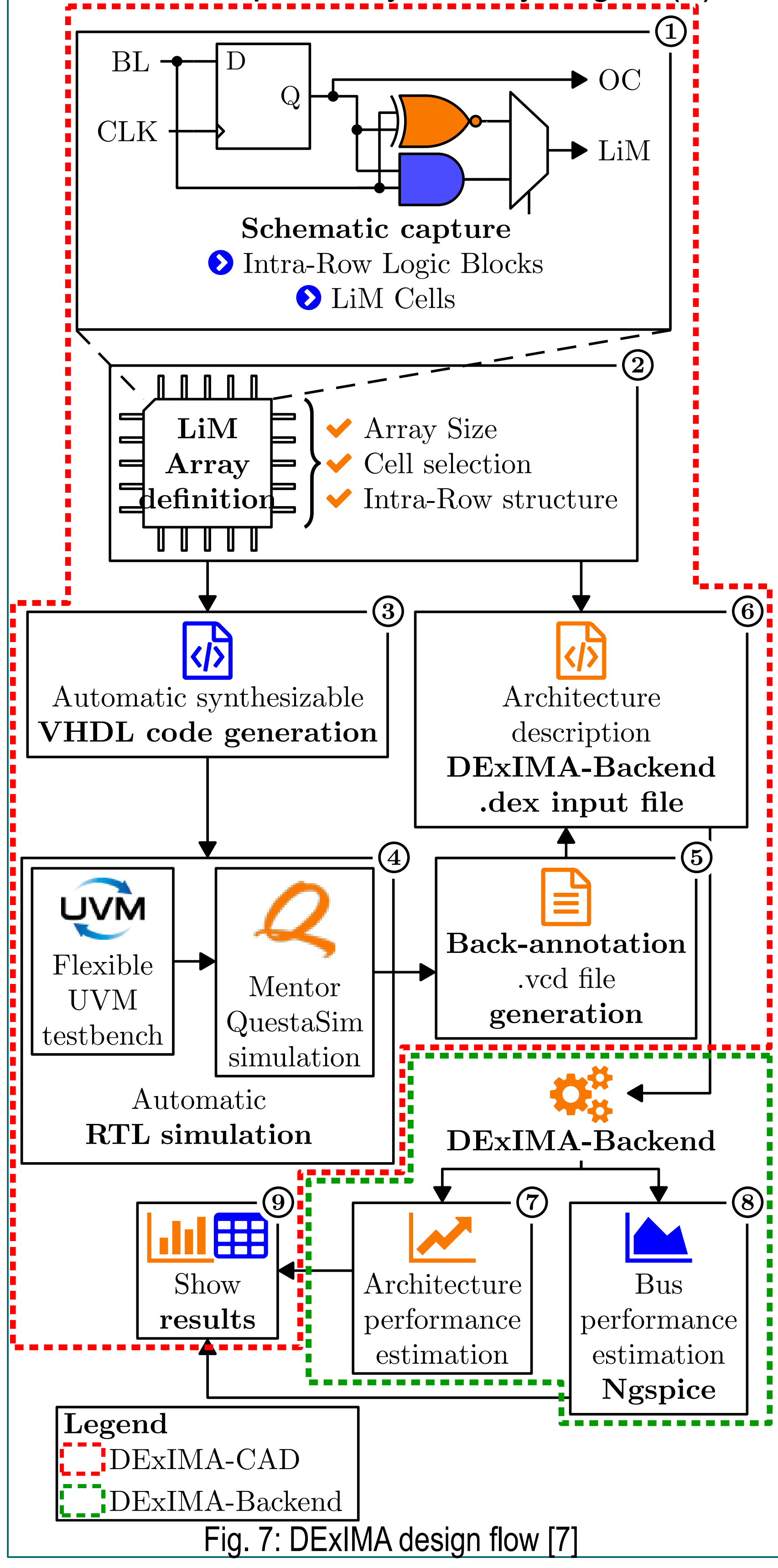


Fig. 7: DExIMA design flow [7]

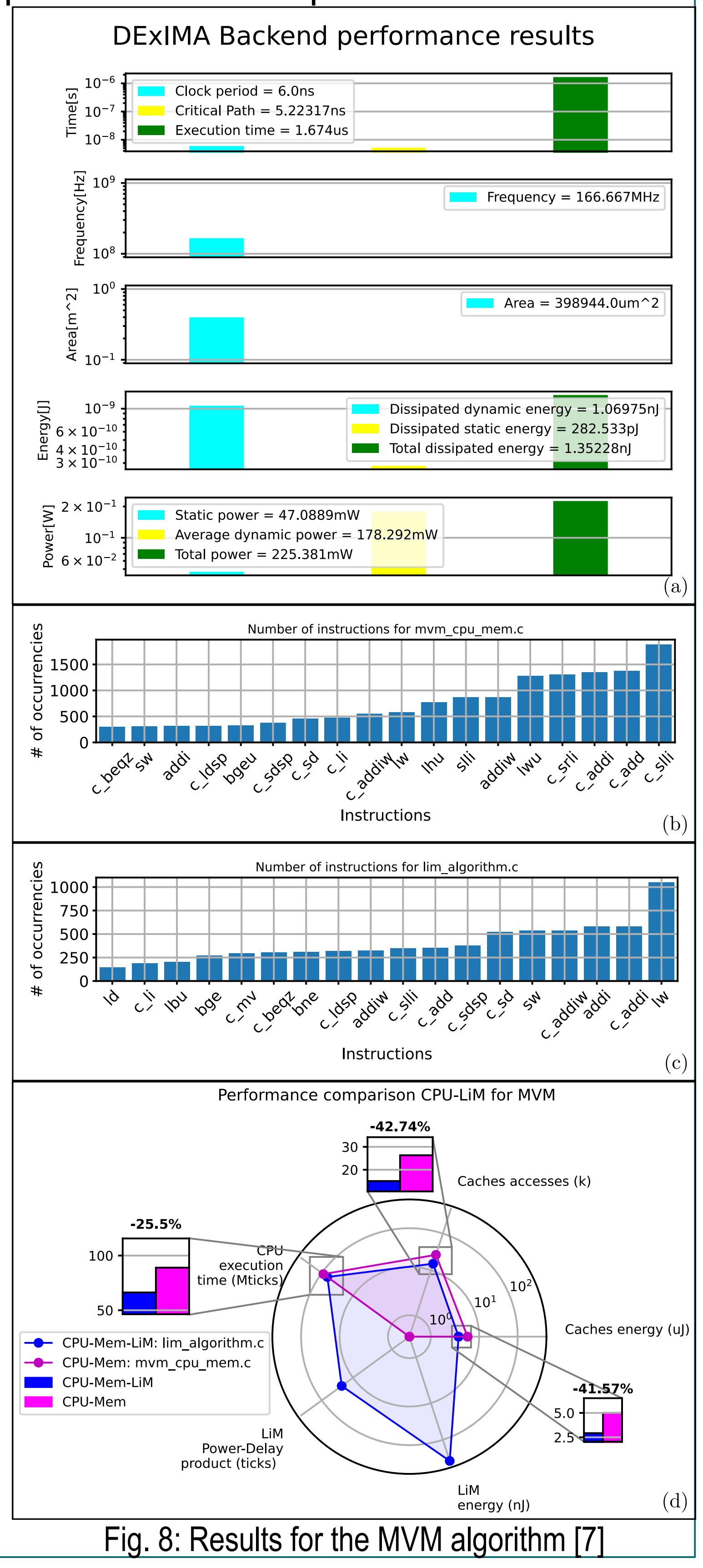


Fig. 8: Results for the MVM algorithm [7]

Future work

- Implementation of **beyond-CMOS** emerging technologies on DExIMA
- Implementation of different **LiM computing paradigms**
- **Algorithmic exploration** to improve DExIMA capabilities

List of attended classes

- 02LWHRV – Communication (21/11/2019, 1)
- 01QTEIU - Data mining concepts and algorithms (20/1/2020, 4)
- 01UJARV - Data science for networks (23/7/2020, 4)
- 01TZFOQ - Engineering Empathy (16/7/2020, 6)
- 02LCPRV - Experimental modeling: costruzione di modelli da dati sperimentali (9/2/2021, 7)
- 01NOYOQ - Microelectronic systems (9/7/2020, 6)
- 01UIYRV - Physics-based modeling of semiconductor devices (1/4/2020, 3)
- 01SFURV - Programmazione scientifica avanzata in matlab (25/5/2020, 4)
- 08IXTRV - Project management (3/1/2020, 1)
- 01RISRV - Public speaking (18/11/2019, 1)
- 01TAHIU - Quantum computing (1/7/2020, 4)
- 01QSCIU - Reconfigurable computing (20/7/2020, 4)
- 01SYBRV - Research integrity (25/11/2019, 1)
- 01SWQRV - Responsible research and innovation, the impact on social challenges (4/1/2020, 1)
- 01SWPRV - Time management (19/11/2019, 1)
- 01QORRV - Writing Scientific Papers in English (20/2/2020, 3)

POLITECNICO DI TORINO

PhD program in
Electrical, Electronics and Communications Engineering