Reconfigurable Tiles of Computing-In-Memory SRAM Architecture for Scalable Vectorization

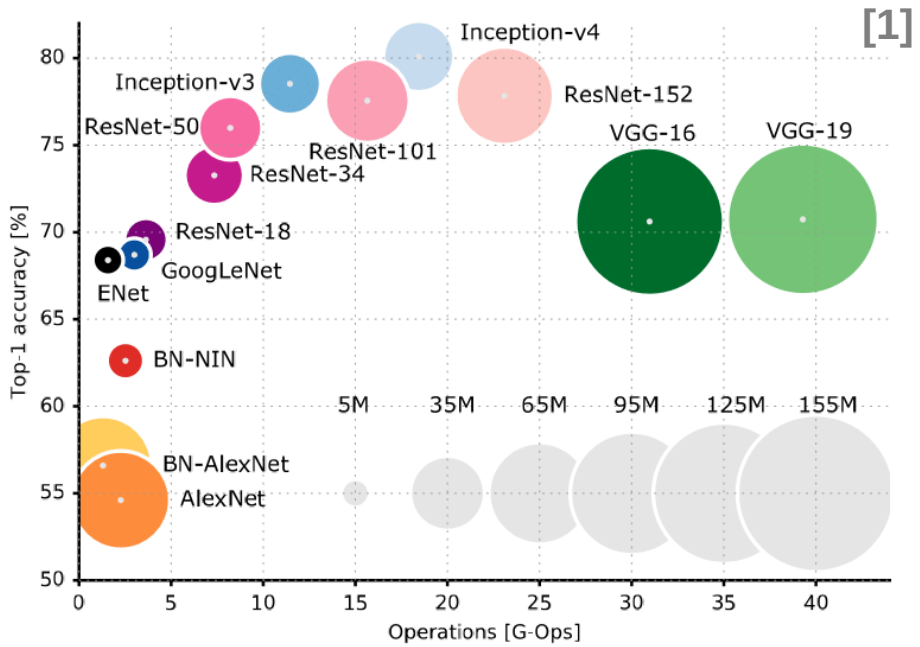
Roman Gauchi[1] (Speaker), V. Egloff[1], M. Kooli[1], J.-P. Noel[1], B. Giraud[1], P. Vivet[1], S. Mitra[2], H.-P. Charles[1]

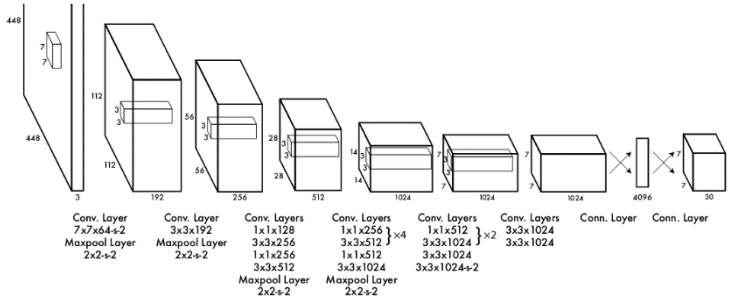
[1] Université Grenoble Alpes, CEA List, Grenoble, France
[2] Stanford University, Palo Alto, CA, USA

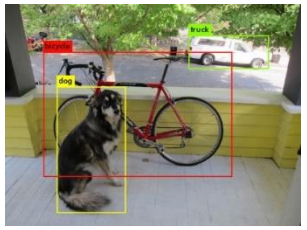
ISLPED - August 10-12, 2020

MOTIVATIONS

➔ Modern applications require **more and more data** to be processed







Darknet, YOLO v1 [2]

- **Data-centric Applications**
  - Big amount of data (**~MB** of weights for Neural Networks)
  - Mostly Logic & Arithmetic operations (**~95%**)

[1] "Gesture Recognition for Robotic Control Using Deep Learning", C. Kawatsu & al., AGS Technical Session, 2017
[2] "You Only Look Once: Unified, Real-Time Object Detection", J. Redmon, A. Farhadi & al., arXiv, 2016

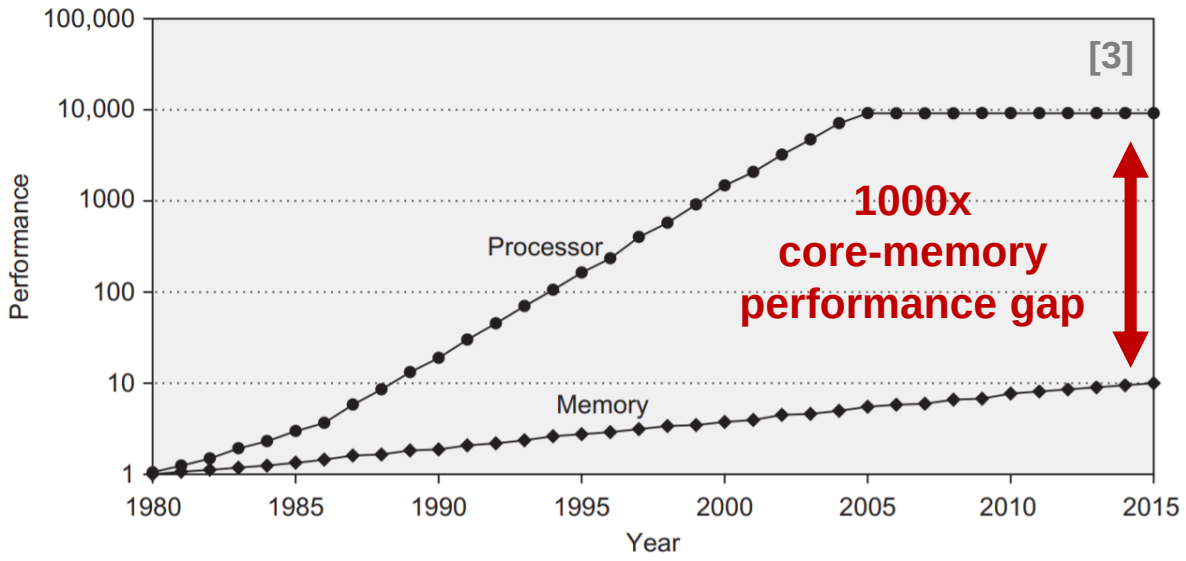
MOTIVATIONS

➔ **Computation latency** becomes limited by the **memory access time**…

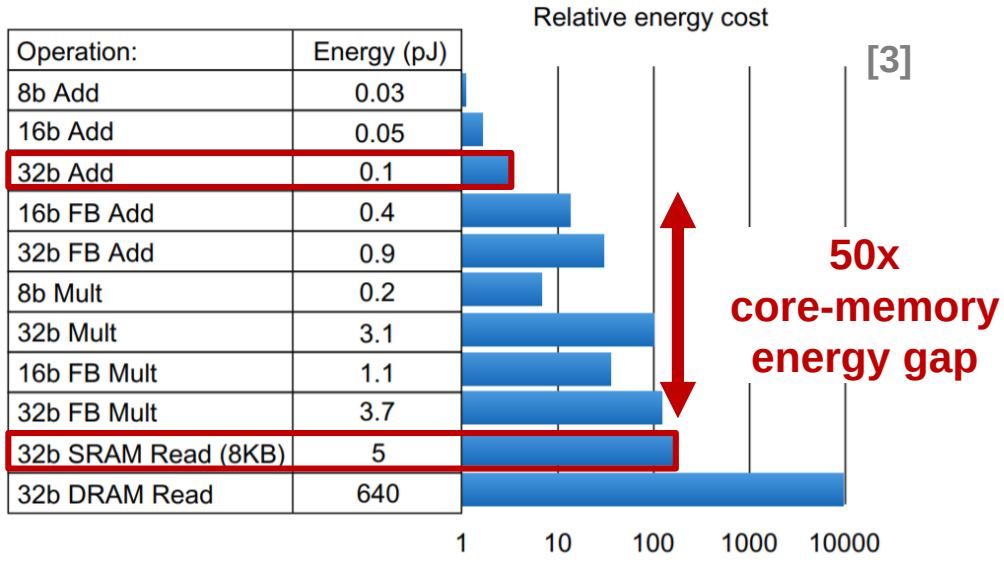
…and **memory energy** is much higher than the **computational energy**

➔ Moore's Law predicts that technology will not scale anymore (CMOS)

The Memory Wall: **Performance impacts**



The Memory Wall: **Energy impacts**



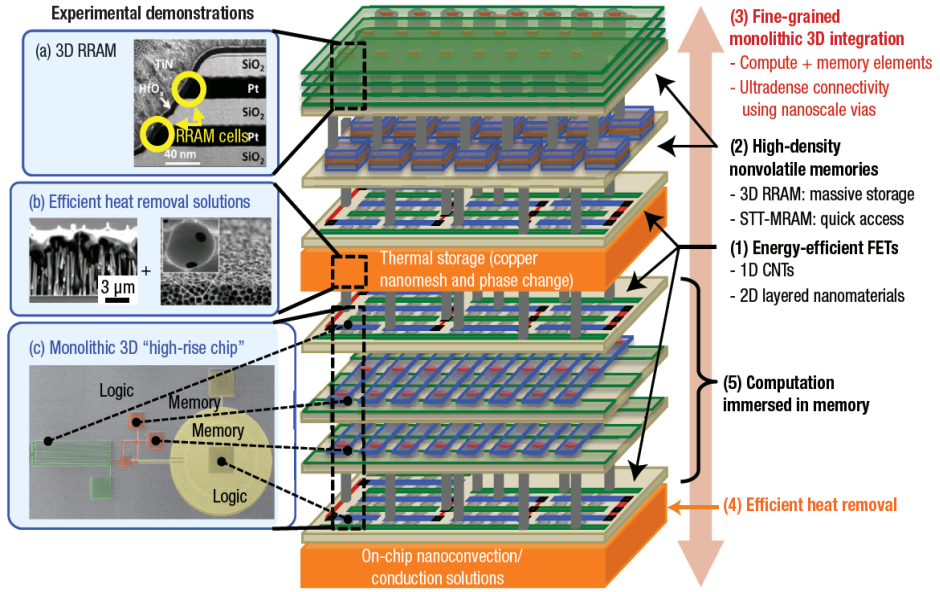
[3] "Computer Architecture: A Quantitative Approach", J. L. Hennessy and D. A. Patterson, 6th edition, 2018

MOTIVATIONS

➔ **Possible solution to break the "Memory Wall"**

- **Emerging architectures coupled with emerging technologies**

[4]



(1) Fine-grained monolithic 3D integration

(2) High-density on-chip nonvolatile memories

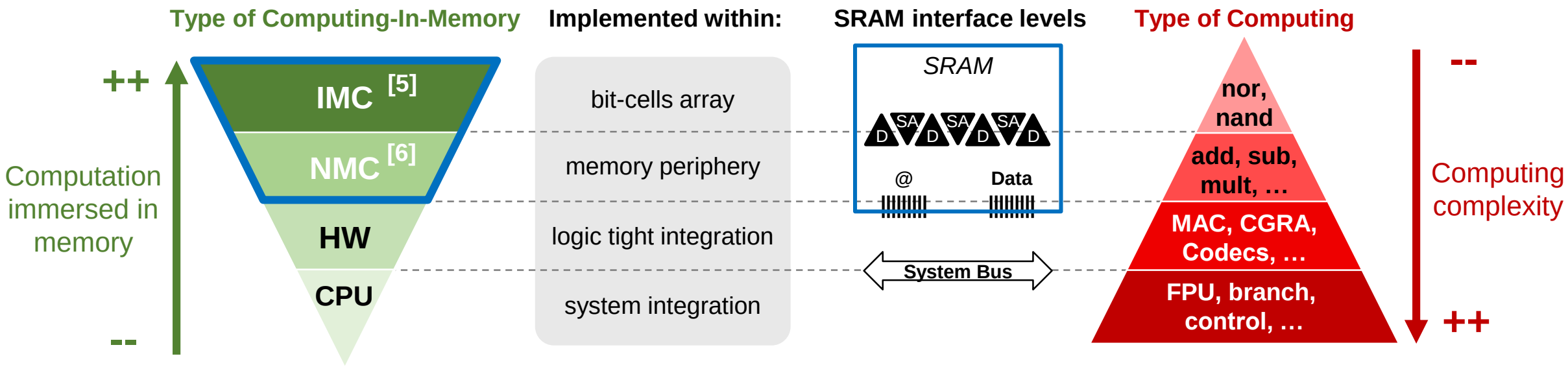
(3) Distributed computing

(4) Computation immersed in memory (CIM)

[4] "Energy-Efficient Abundant-Data Computing: The N3XT 1000x", M. M. Sabry Aly et al., Computer, 2015

STATE-OF-THE-ART
ON COMPUTING-IN-MEMORY (CIM)

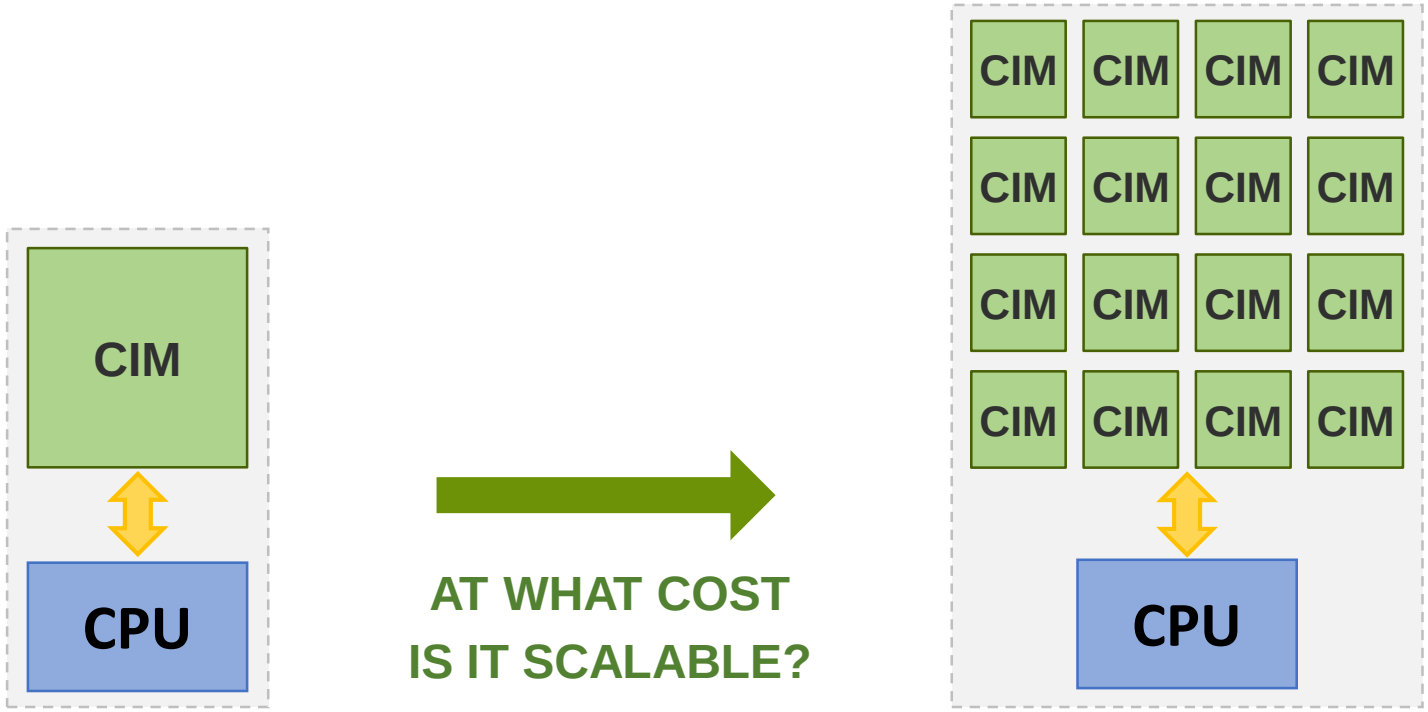
➔ **Concept**: Bring computation closer to the memory (*SRAM-based technology*)



IMC	*In-Memory Computing*
NMC	*Near-Memory Computing*
HW	*Hardware accelerator integrating SRAM*
CPU	*Central Processing Unit*

➔ **Focus on SRAM-based CIM concepts**

[5] "DRC2: Dynamically Reconfigurable Computing Circuit based on Memory Architecture", K. C. Akyel et al., ICRC, 2016
[6] "Neural cache: bit-serial in-cache acceleration of deep neural networks", R. Das et al., ISCA, 2018



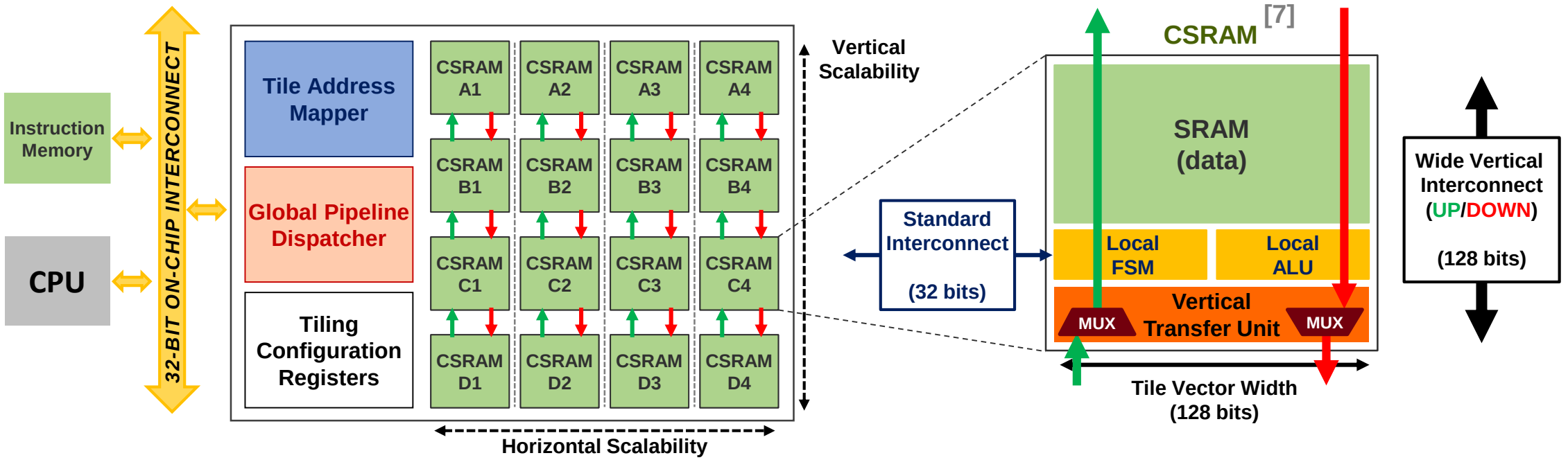
➔ **Single memory is not enough, we need more !**

OUTLINE

1. Motivations & State-of-the-Art
2. A Reconfigurable Tiles of CIM SRAM Architecture
3. A Hardware/Software Simulation Platform
4. Architecture Benchmarking Results
5. Conclusions

A RECONFIGURABLE TILES OF C-SRAM ARCHITECTURE

ARCHITECTURE LEVEL



2D Scalable Vector Cluster

- **Tile Address Mapper**
  - Horizontal & Vertical configurability
  - Scalable vector extension
- **Global Pipeline Dispatcher**
  - Data Hazards & address conflicts
- **Tiling Configuration Registers**
  - Define grid shape and vector size

CIM Tile

- **CSRAM (Compute-SRAM)**
  - SRAM data memory
  - Arithmetic & Logical instructions
  - Pipelined instructions (5 stages)
- **Vertical Transfer Unit**
  - Inter-tile vertical communication
  - Interleaved data transfers

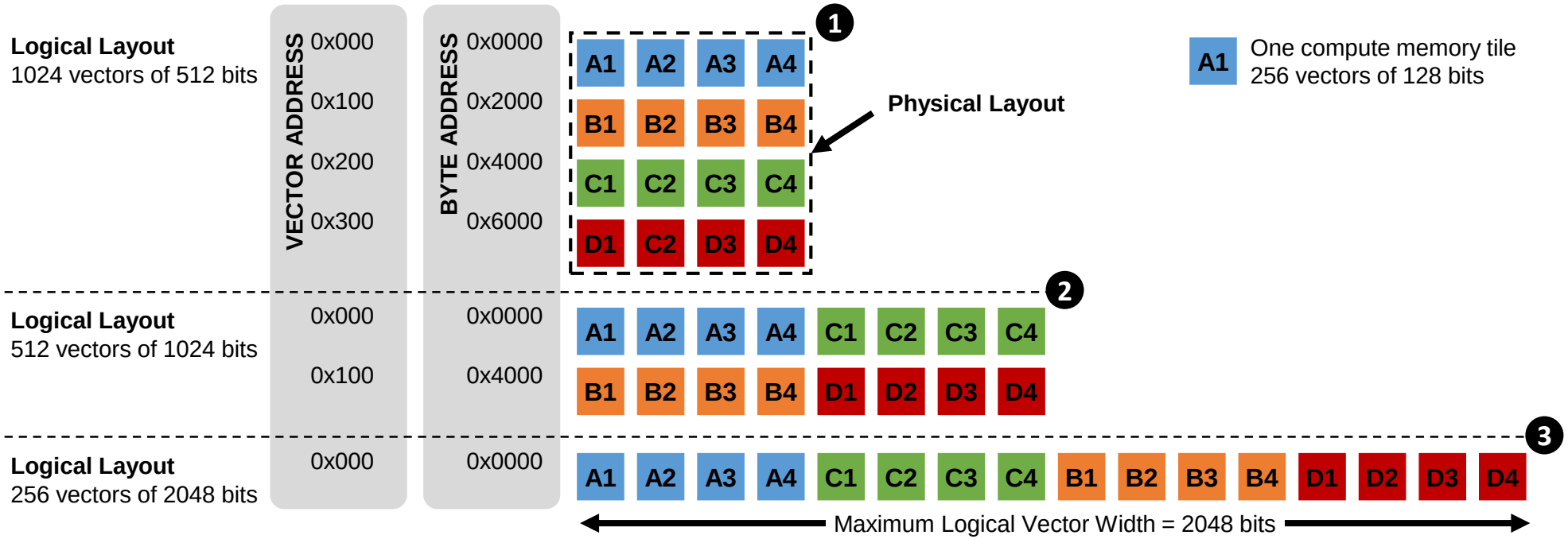
Interconnects

- **Standard interconnect**
  - 32-bit Read access
  - 32-bit Write access
- **Wide Vertical Interconnect**
  - 128-bit Upward data transfers
  - 128-bit Downward data transfers

[7] "Computational SRAM Design Automation using Pushed-Rule Bitcells […]", J.-P. Noel et al., DATE, 2020

A RECONFIGURABLE TILES OF C-SRAM ARCHITECTURE

SOFTWARE LEVEL



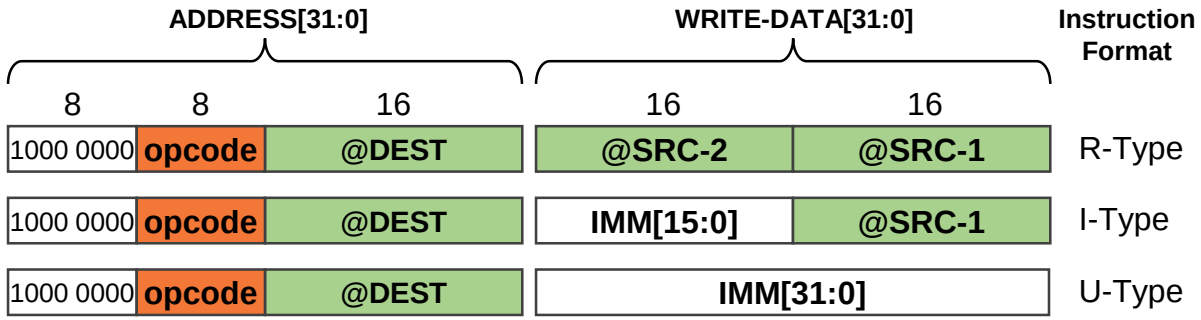
- **Parallelism at word level,** using wide vector computation (configurable)
- **Horizontal configurability :** to extend vector size
- **Vertical configurability :** to extend number of vectors
- **Both computation and memory extension performed through multiple tiles**

A RECONFIGURABLE TILES OF C-SRAM ARCHITECTURE

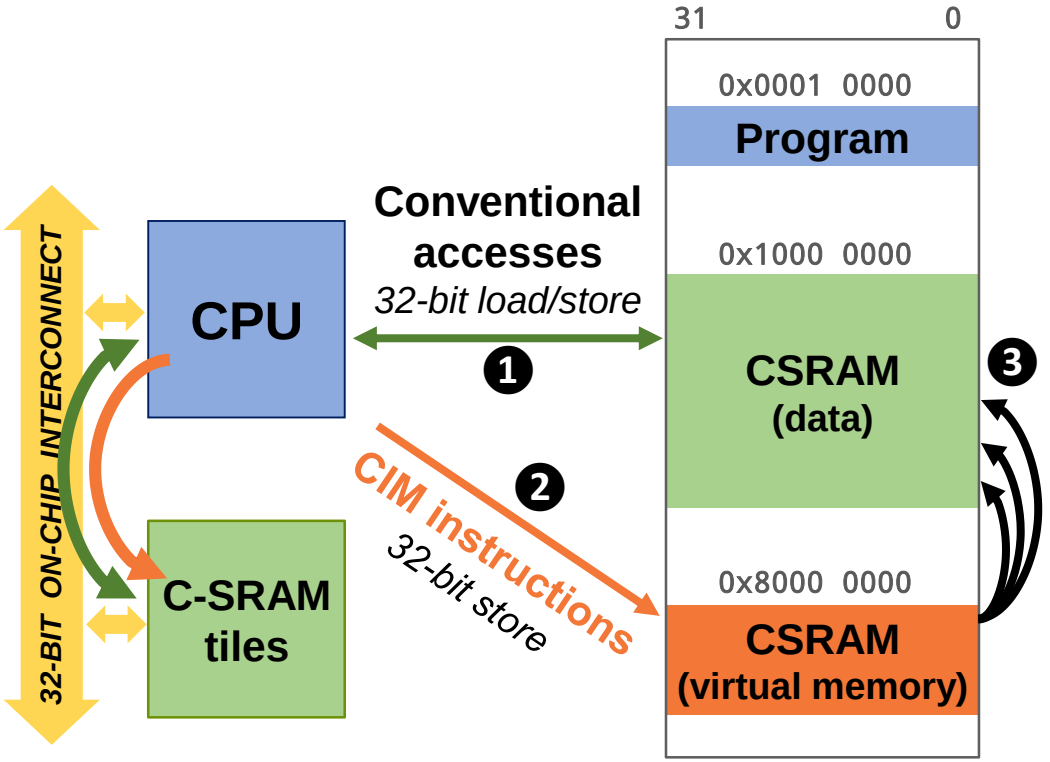
INSTRUCTION SET ARCHITECTURE

Instruction Formats [8]

(*integrated on 32-bit system bus*)



System Memory Mapping



Programming model

- The proposed CIM architecture refers as a memory
  - From the **architecture level** and the **software level**
- ISA for CIM can be integrated on a 32-bit system bus
- ❶ 32-bit standard read/write in the CSRAM data
- ❷ CIM instructions ⇔ standard store instructions
- ❸ Dispatch the workload on CSRAM data

➔ **How to evaluate the architecture?**

[8] "Smart instruction codes for in-memory computing architectures […]", M. Kooli et al., DATE, 2018

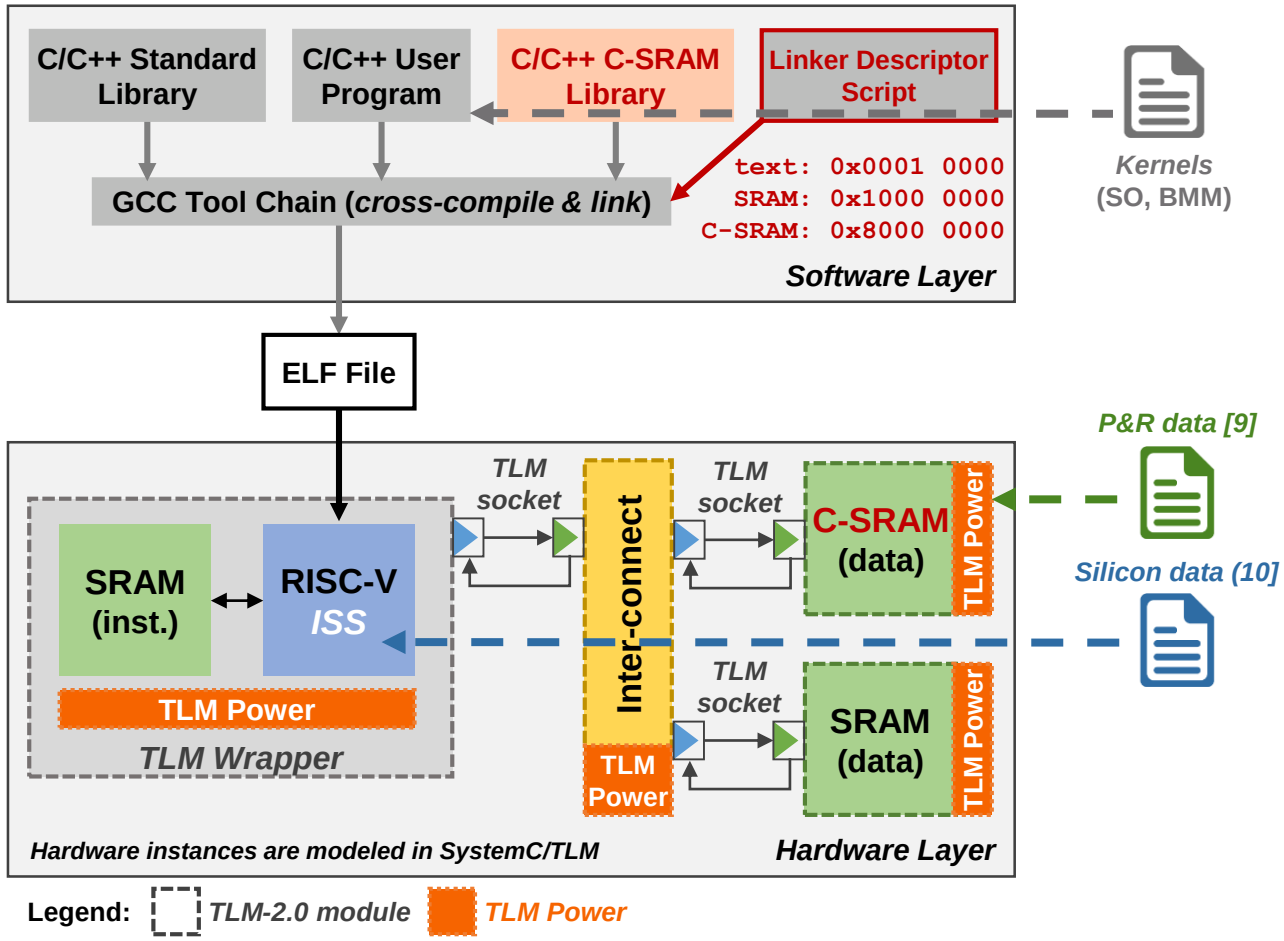
OUTLINE

1. Motivations & State-of-the-Art
2. A Reconfigurable Tiles of CIM SRAM Architecture
3. A Hardware/Software Simulation Platform
4. Architecture Benchmarking Results
5. Conclusions

A HARDWARE/SOFTWARE SIMULATION PLATFORM
FOR CIM EVALUATION

- **Software Layer**
  - C-SRAM ISA Library
    ➔ CIM instructions
  - Linker Script
    ➔ data & virtual memory sections
  - User Program (kernels)
    ➔ SIMD engines compatibility

- **Hardware Layer**
  - **Timing Modeling**
    - Instruction Set Simulator (ISS)
    - SystemC/TLM sockets
  - **Energy models**
    - Place&Route extraction
    - Silicon data extraction



[9] "Computational SRAM Design Automation using Pushed-Rule Bitcells for Energy-Efficient Vector Processing", J.-P. Noel & al., DATE, 2020
[10] "A 50.5 ns wake-up-latency 11.2 pj/inst asynchronous wake-up controller in fdsoi 28 nm", J.-F. Christmann et al., JLPEA, 2019

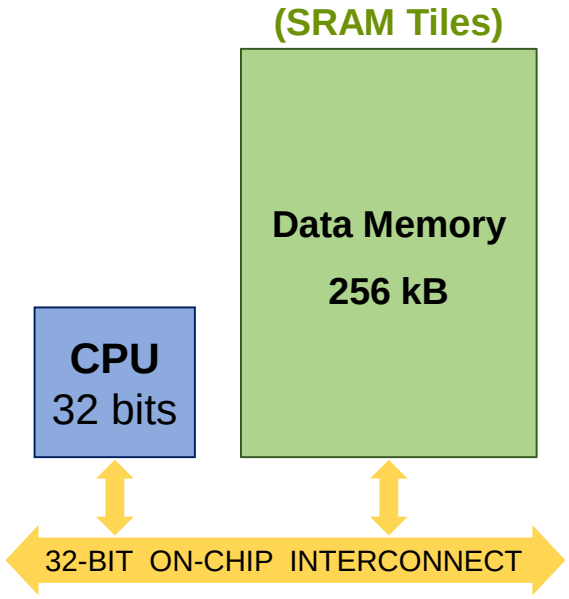
OUTLINE

1. Motivations & State-of-the-Art
2. A Reconfigurable Tiles of CIM SRAM Architecture
3. A Hardware/Software Simulation Platform
4. Architecture Benchmarking Results
5. Conclusions

ARCHITECTURE BENCHMARKING

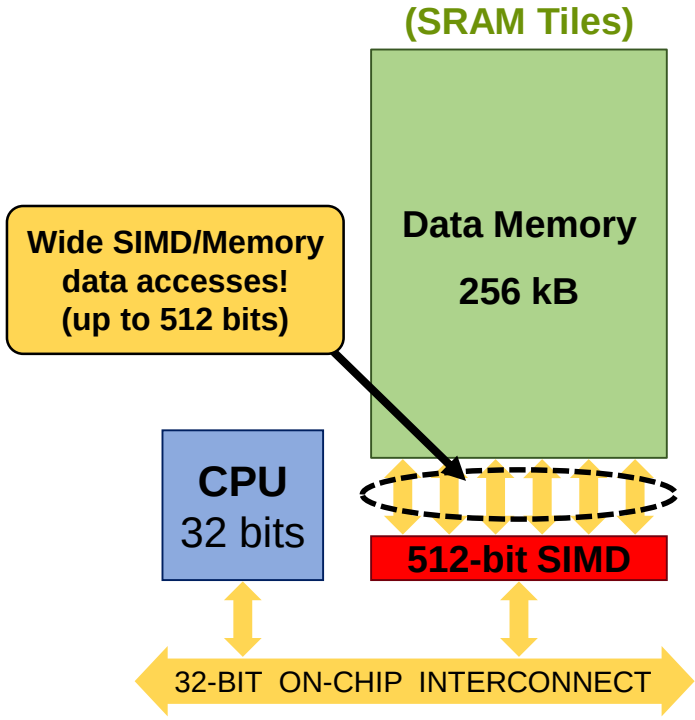
ARCHITECTURE SET-UP

Scalar CPU Architecture



Evaluated using the proposed HW/SW simulation platform

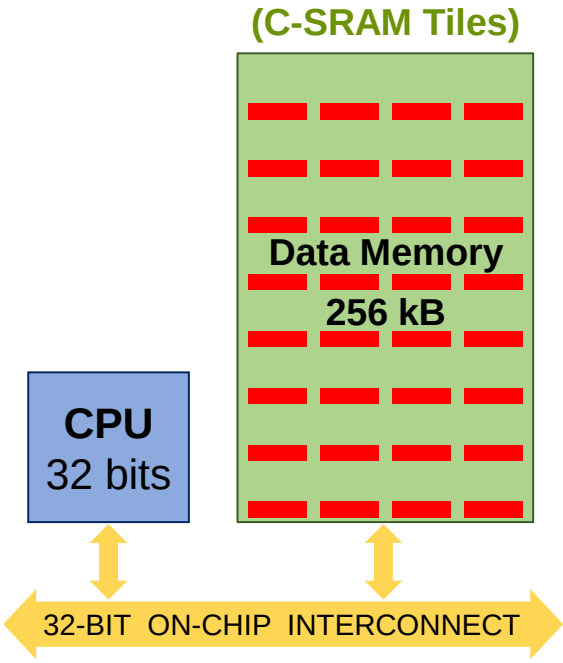
512-bit SIMD Architecture



Evaluated using Pin [11], a dynamic binary instrumentation framework developed by Intel

CIM Architecture

(up to 8192 bits)



Evaluated using the proposed HW/SW simulation platform

[11] "PinADX: An interface for customizable debugging with dynamic instrumentation", G. Lueck et al., CGO, 2012

ARCHITECTURE BENCHMARKING

KERNEL OVERVIEW

Compute bound kernels

1. Shift-OR (**so**)
   - **Application:** Bio-informatics
   - **Memory Access:** **2 %**
   - **CIM instructions:** OR, AND, SRL, SLL
   - **Description:** *Match a pattern in a DNA sequence*

2. Hamming Weight (**hw**)
   - **Application:** Information theory
   - **Memory Access:** **9 %**
   - **CIM instructions:** XOR, AND, ADD8, SUB8, SRL
   - **Description:** *Count the number of differences after perform a XOR between two strings*

Memory bound kernels

3. Atax (**atax**) [12]
   - **Application:** Image computing
   - **Memory Access:** **28 %**
   - **CIM instructions:** ADD8, MUL8
   - **Description:** *matrix transpose and vector multiplication*

4. General Matrix to Matrix Multiplication (**gemm**) [12]
   - **Application:** Neural Network
   - **Memory Access:** **37 %**
   - **CIM instructions:** ADD8, MUL8
   - **Description:** *matrix multiplication as C = α.A.B + β.C*

[12] "PolyBench/C, the Polyhedral Benchmark suite", L.-N. Pouchet., 2015

ARCHITECTURE BENCHMARKING

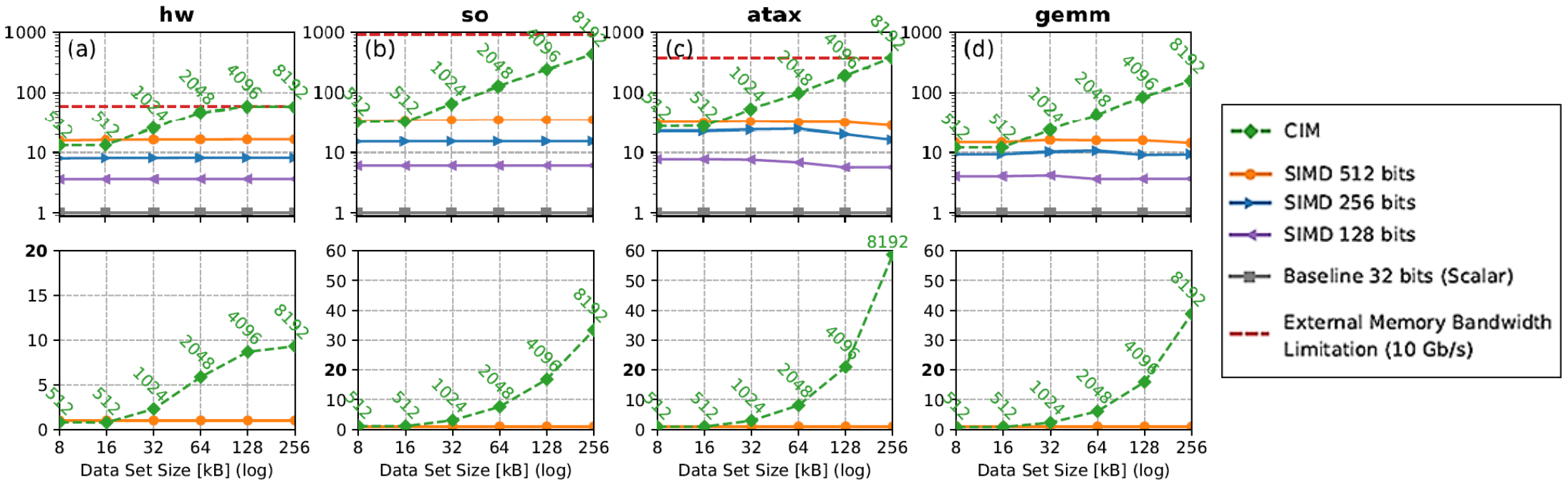
RESULTS

Speed Up (Ratio)

Baseline: Scalar CPU

EDP (Ratio)

Baseline: 512-bit SIMD



	CIM vs. 32-bit Scalar		CIM vs. 512-bit SIMD	
	Speed Up	*EDP*	*Speed Up*	*EDP*
Memory bounds kernels	**380x**	**28000x**	**14x**	**60x**
Compute bounds kernels	**440x**	**22000x**	**13x**	**34x**

- ➔ Large **Speed up** & **EDP** gains since reduced data movement in CIM architecture
- ➔ Maximum performance gains, thanks to:
  - **Large vectorization** in the application kernel
  - Select the **optimal vectorization size**

OUTLINE

1. Motivations & State-of-the-Art
2. A Reconfigurable Tiles of CIM SRAM Architecture
3. A Hardware/Software Simulation Platform
4. Architecture Benchmarking Results
5. Conclusions

CONCLUSIONS

- **We propose**
  - A reconfigurable Tiles of **Computing-In-Memory SRAM-based Architecture**
  - An hardware & software **simulation platform** calibrated on P&R extractions
- **Conclusions**
  - To break the "**Memory Wall**": emerging architectures & technologies
  - CIM architectures compared to a 512-bit SIMD architecture achieves
    - EDP gain up to **60x** and **34x** for **memory bound** and **compute bound** kernels respectively
- **Future works**
  - Circuit ➔ Physical implementation of the proposed architecture
  - Architecture ➔ Explore 3D CIM architecture

THANK YOU

FOR YOUR ATTENTION

Commissariat à l'énergie atomique et aux énergies alternatives
Institut List | CEA SACLAY NANO-INNOV | BAT. 861 – PC142
91191 Gif-sur-Yvette Cedex - FRANCE
www-list.cea.fr