Analog memristive CAMs for area- and energy-efficient reconfigurable computing

João Paulo Cardoso de Lima, Rafael Fão de Moura and Luigi Carro

Federal University of Rio Grande do Sul

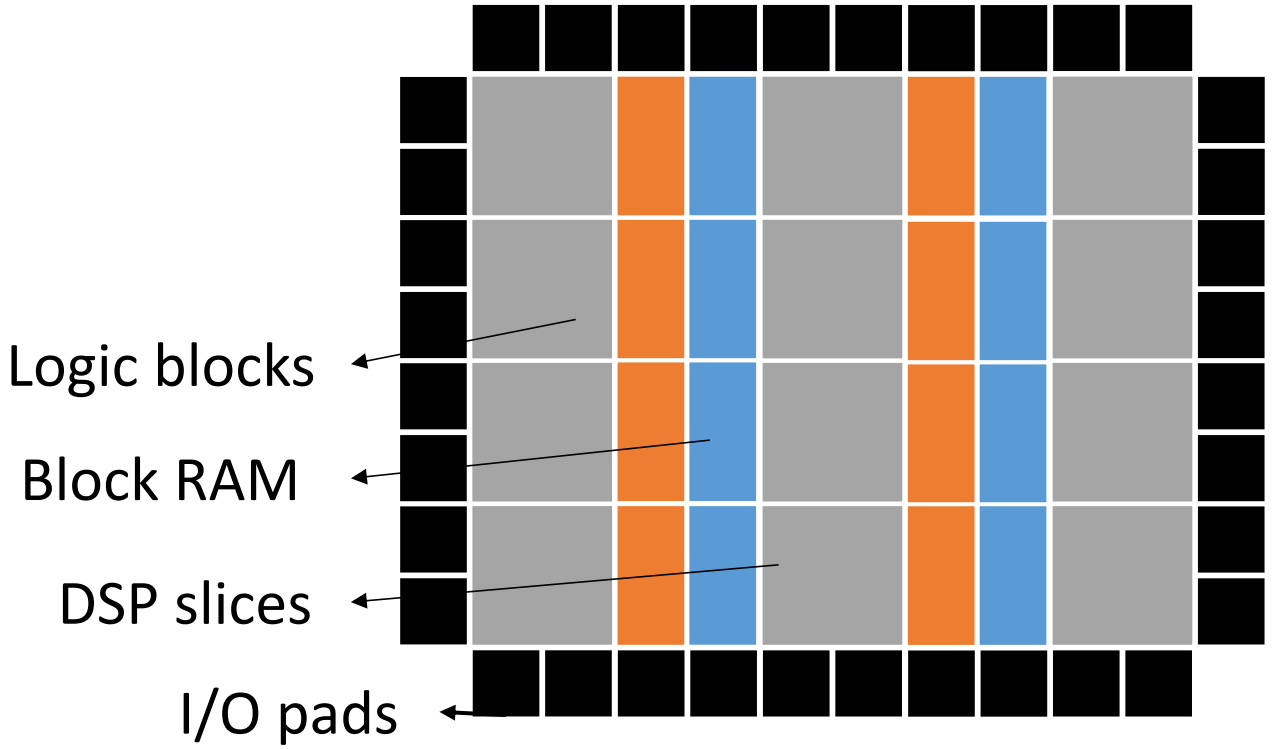
corresponding e-mail: jpclima@inf.ufrgs.br

Perks of FPGAs

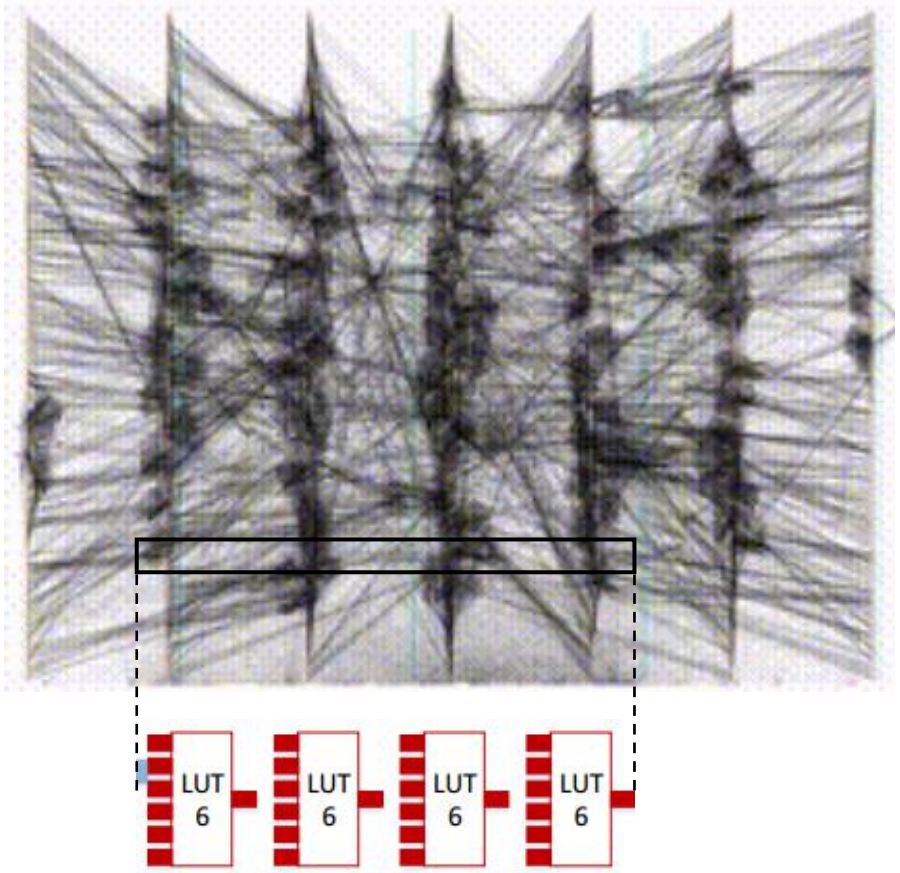
- FPGAs suffer from ...
  - Poor scalability across different process generations
  - Programmable interconnection is a bottleneck

- How to mitigate it?

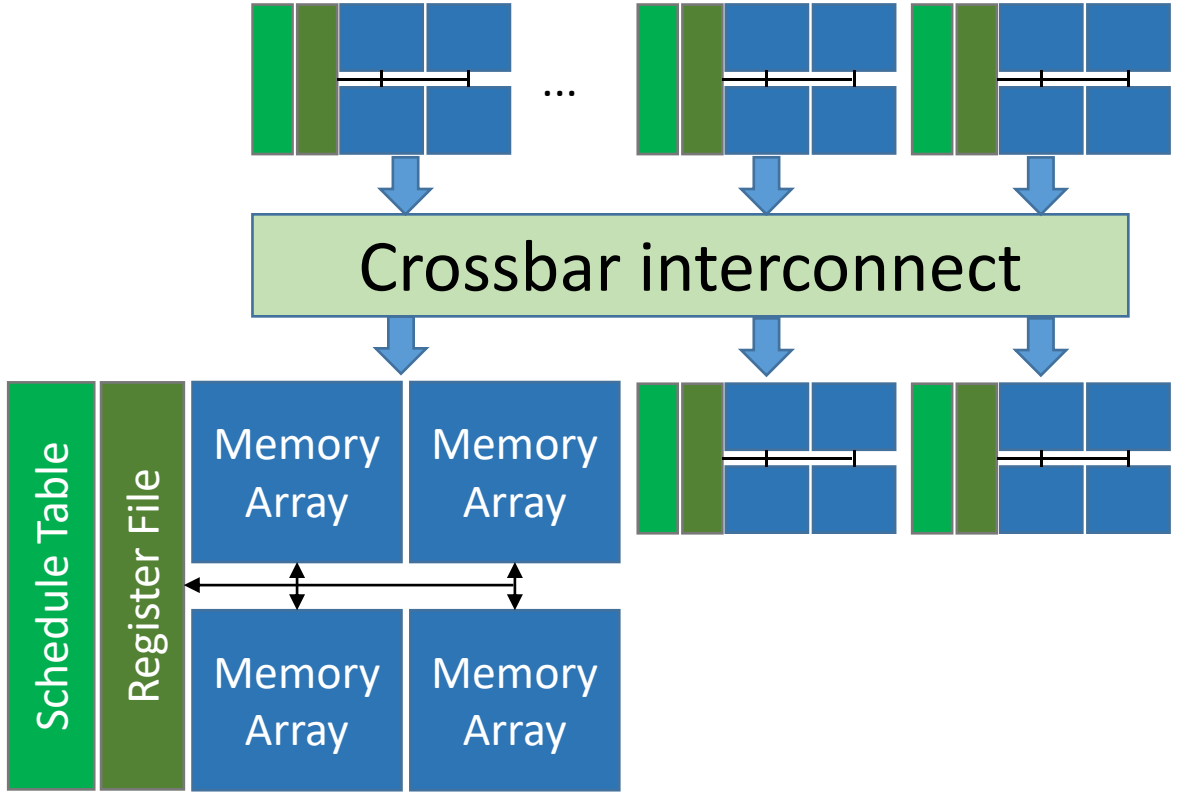
spatial + temporal computing paradigms

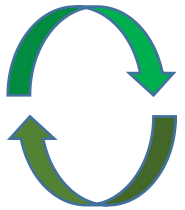


Programmable logic in spatial-temporal paradigm



Boolean functions are evaluated
in **spatial-only** manner





Boolean functions are evaluated
over multiple cycles
in **spatial-temporal** manner

Challenges of spatial-temporal architectures

1. Required memory space for logic blocks
   - Implement logic functions in LUTs, PLAs or CAMs
2. Area and leakage power of SRAM-based architectures are significant

Our contribution:

Multi-level RRAM for a CAM-based reconfigurable architecture

Compressing CAM rows based on the range of discrete resistance levels

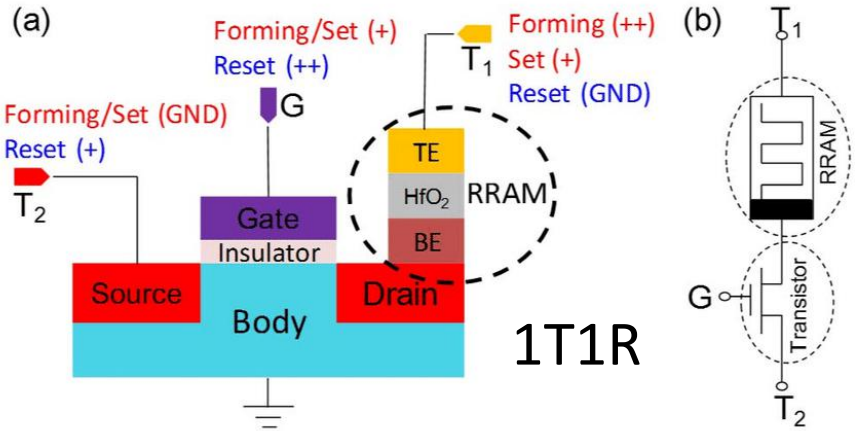
Area density and energy improvements in reconfigurable domain

Outline

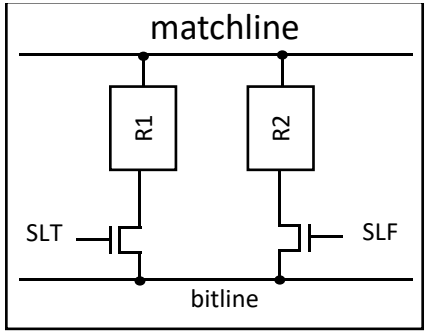
- RRAM technology
- Analog CAM operation
- CAM in programmable logic
- Memory-based Computational Block
- Application mapping
- Results for ISCAS-85 and MCNC circuits
- Main takeaways

RRAM technology

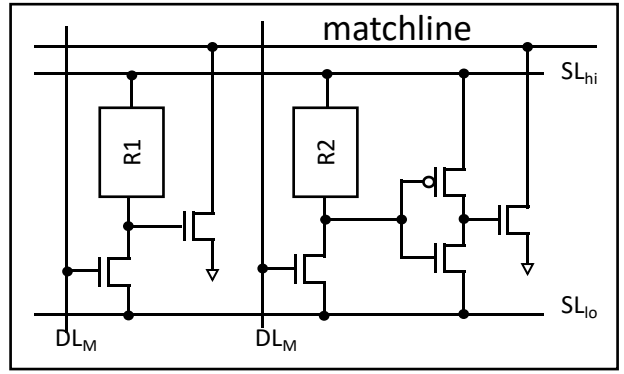
- Digital operation (HRS and LRS)
  - RAM: 1T1R
  - TCAM: 2T2R, 4T2R
- Analog operation (multi-level resistances)
  - Crossbar array: 1T1R
  - TCAM: 3T1R
  - Analog CAM: 6T2R
- Applications of RRAM analog operation
  - Dot-product engines: realizing synaptic weights in neural networks
  - CAM: partial and range matches



1T1R



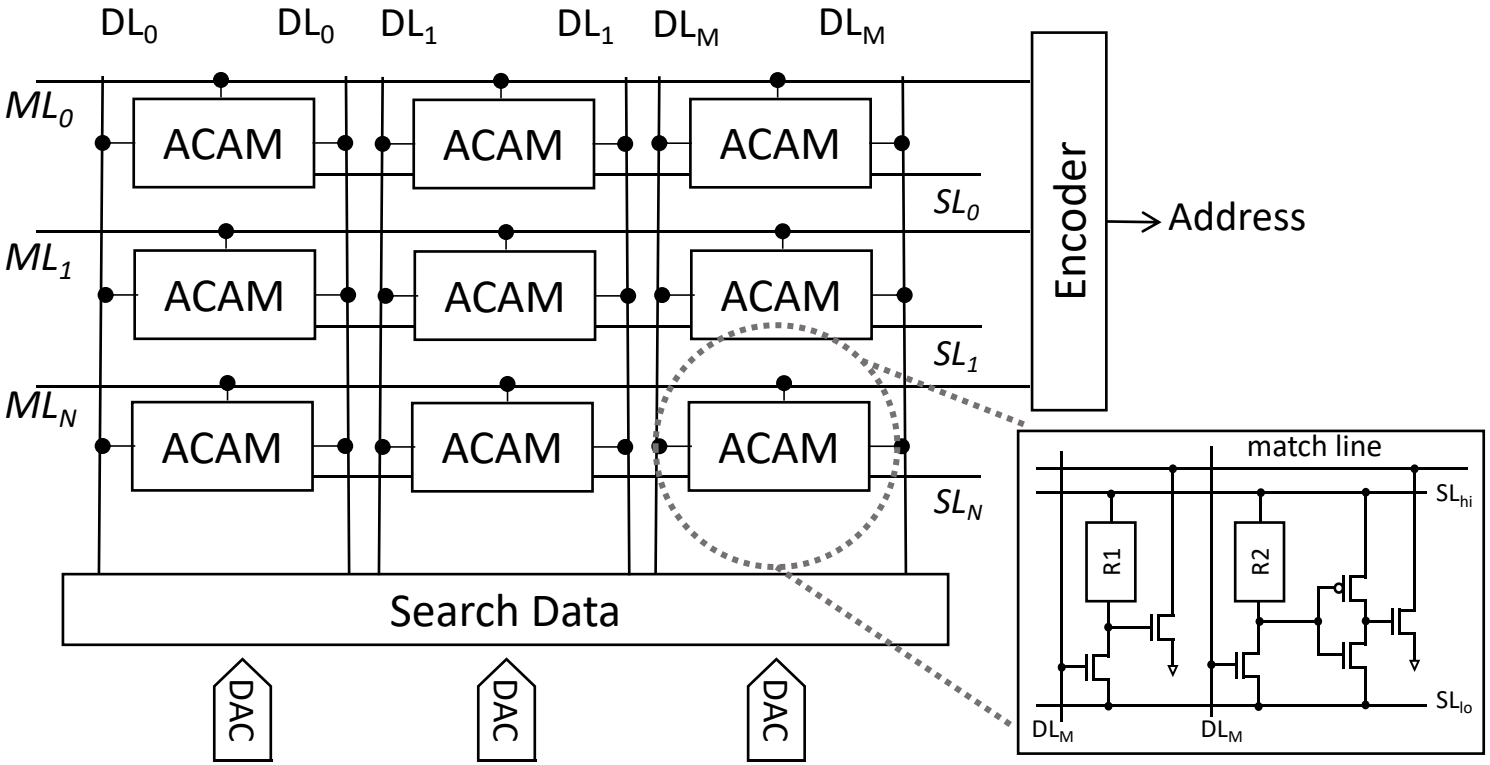
2T2R



6T2R

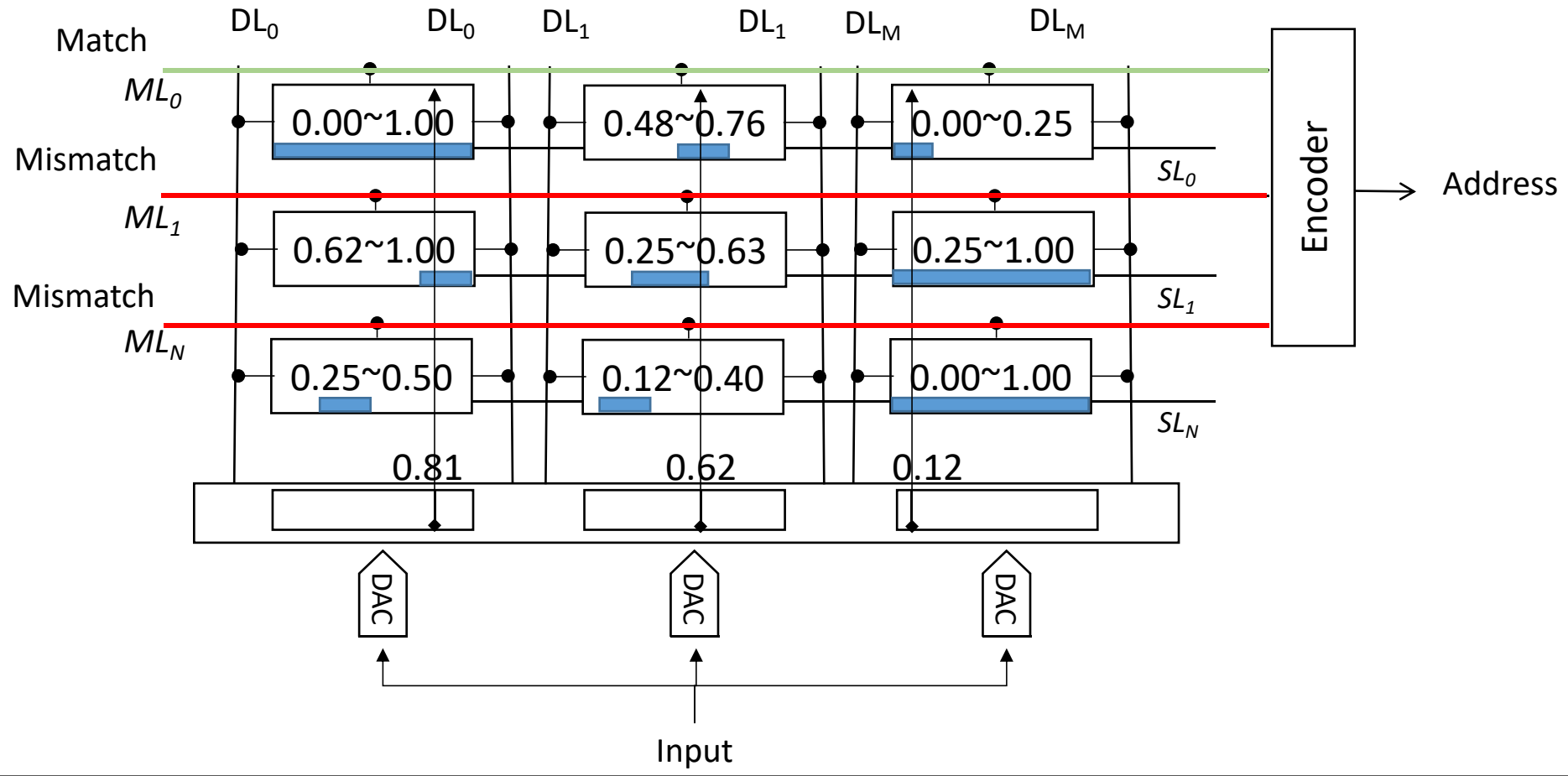
Analog CAM operation

- First analog memristive CAM was developed by HP (Li et al., 2020)
- A 6T2R cell can store a **single multi-level** resistance or a **range** of resistance
- Each ACAM cell **matches** an analog input



Analog CAM operation

- **4-bit MLC**: each cell can store an interval of resistance encoded as [m, n], where m and n range from 0.00 to 1.00



CAM in programmable logic

Let us consider function σ:

$$\sigma(a,b,c) = \neg a*b + a*\neg b*c + a*b$$

The output space is stored

a b c	*Z*
000	0
001	0
010	1
011	1
100	0
101	1
110	1
111	1

The input space is stored

a b c
010
011
101
110
111

Binary CAM

a b c
01X
101
11X

Ternary CAM

a b c
[0][3~2]
[0][7~5]

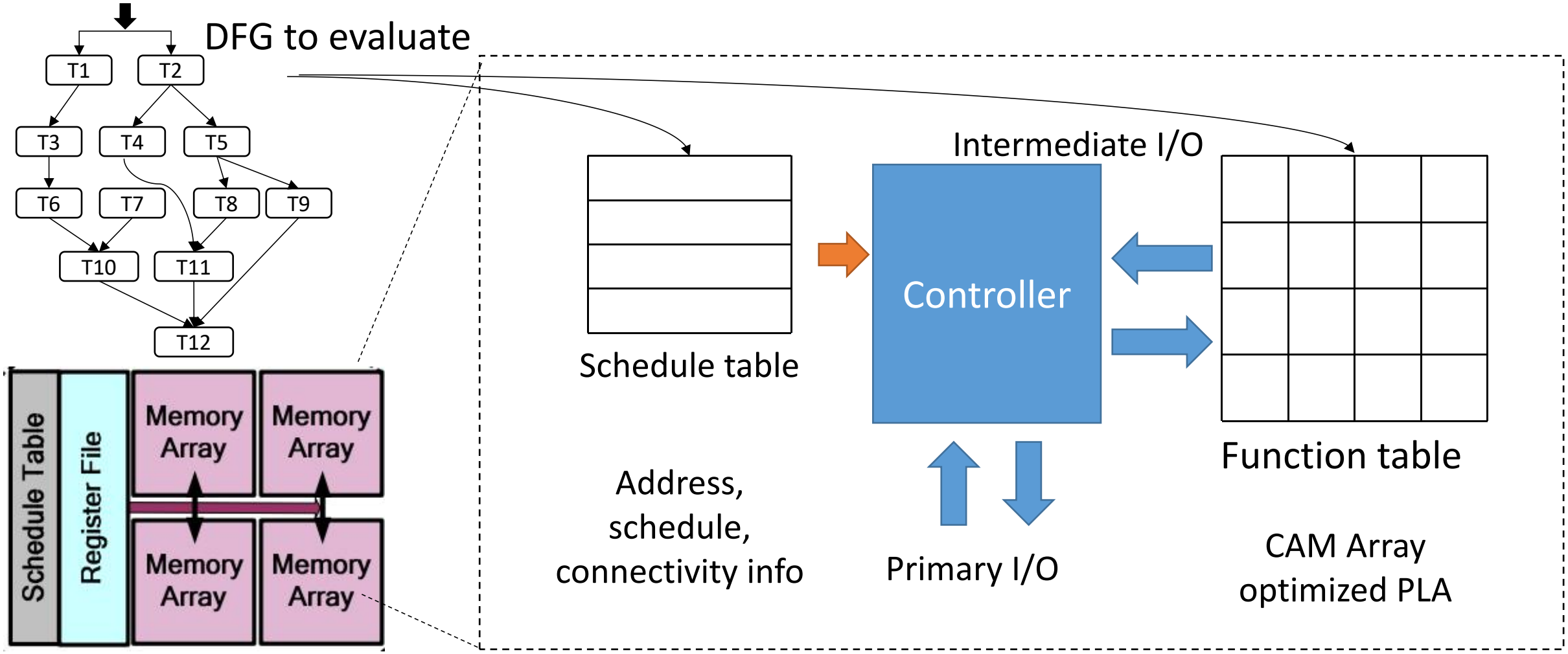
3-bit ACAM

CAM match -> output 1
CAM mismatch -> output 0

Optimized **PLA representation** minimizes **area** and **power** drawbacks of **spatial-temporal** architectures

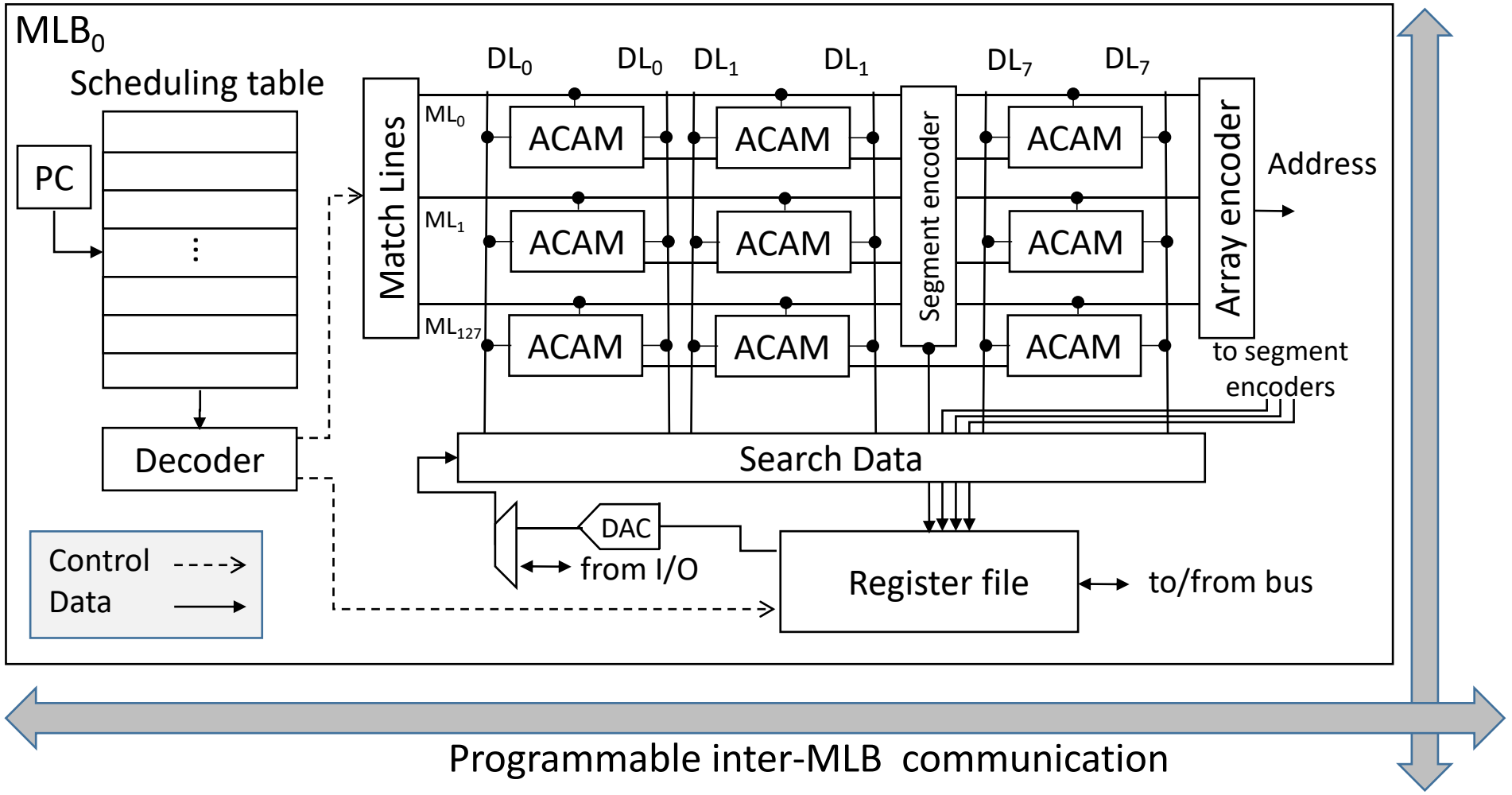
Memory-based Computational Block operation

- Optimized PLA-based logic evaluation implemented in CAM inspired by Paul & Bhunia (2008)

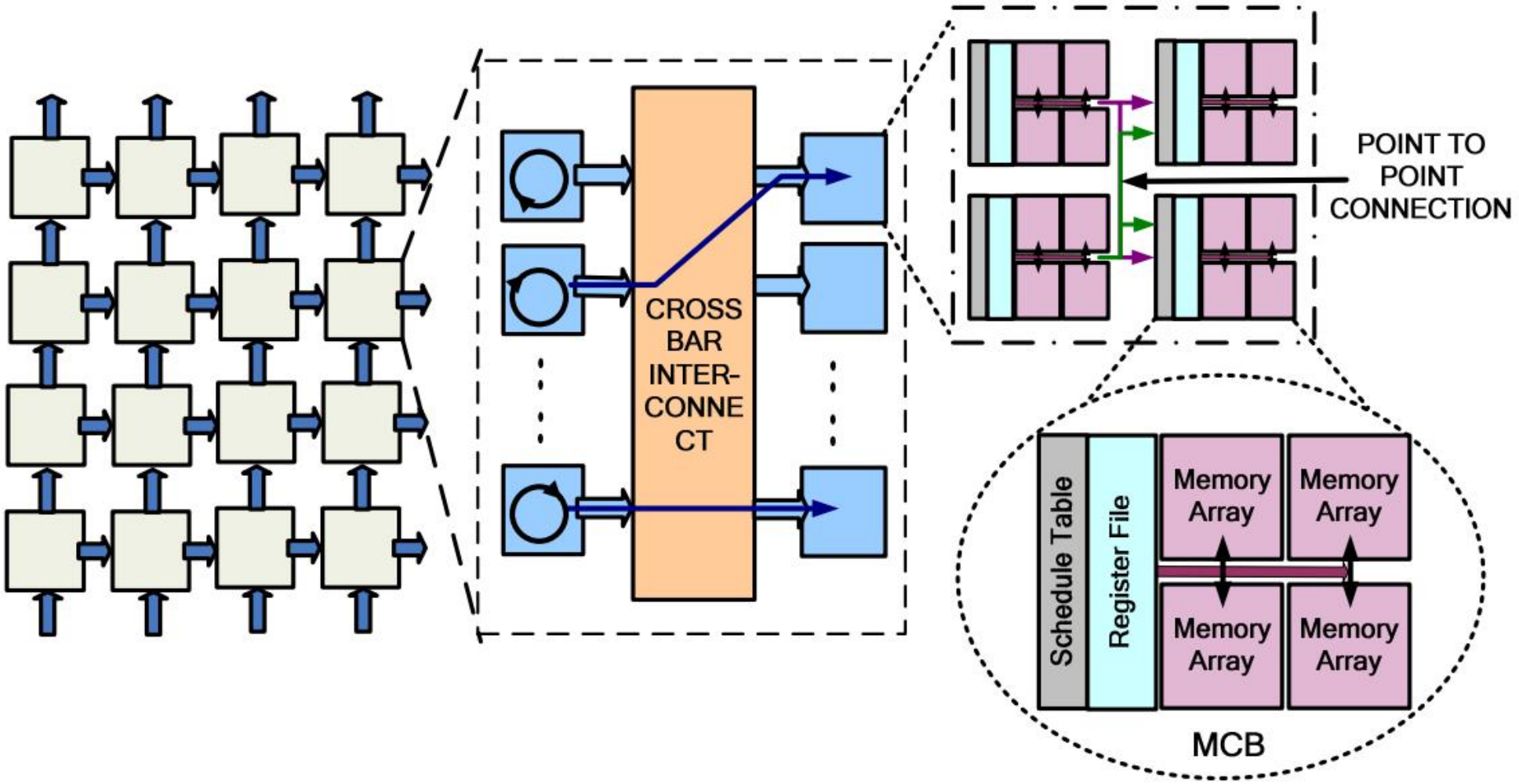


Memory-based Computational Block operation

- Proposed MCB based on analog CAMs
  - DAC overhead is not overwhelming (Li et al., 2020)



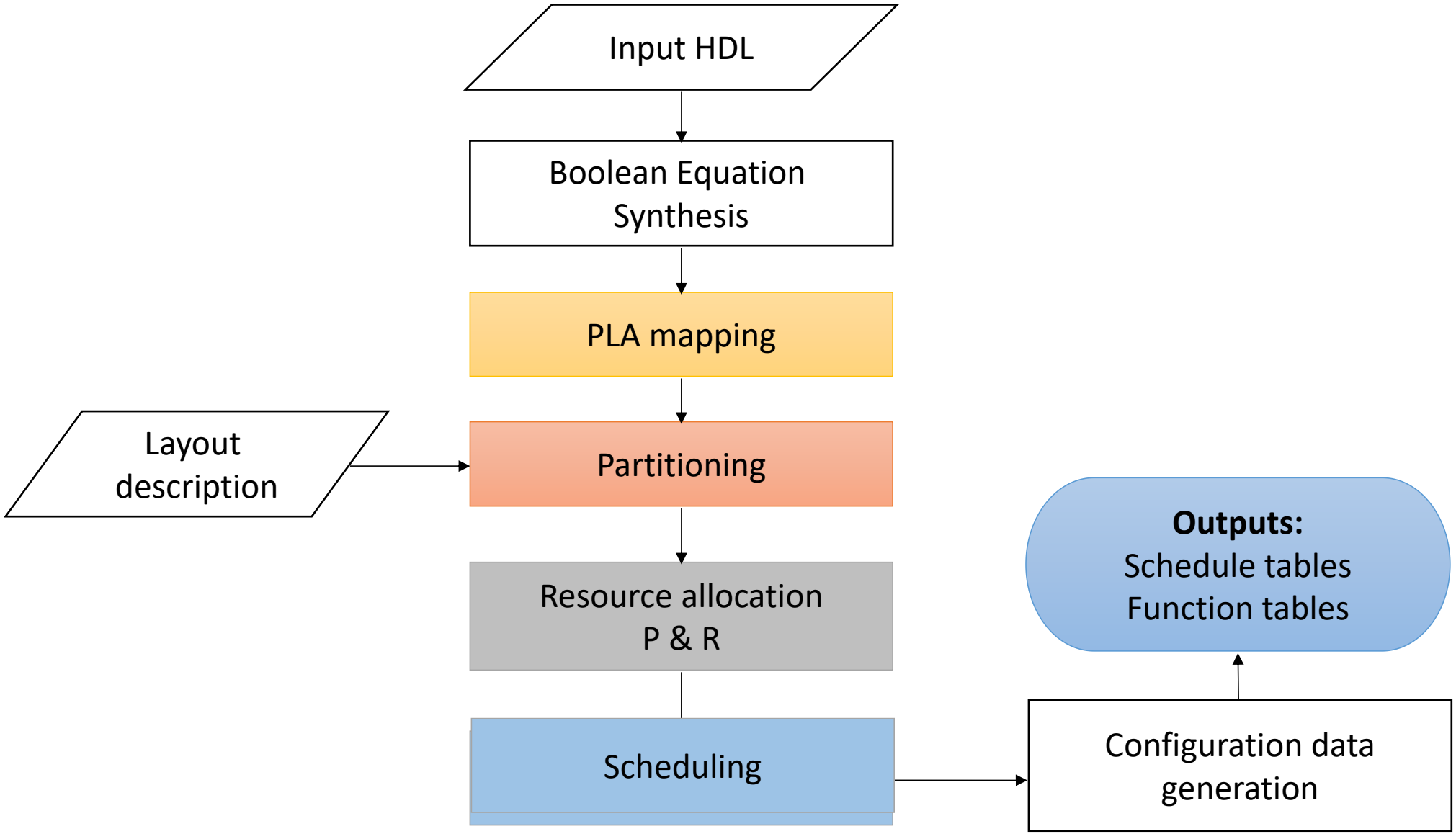
Memory-based Computational Block operation



Source: Paul & Bhunia (2008)

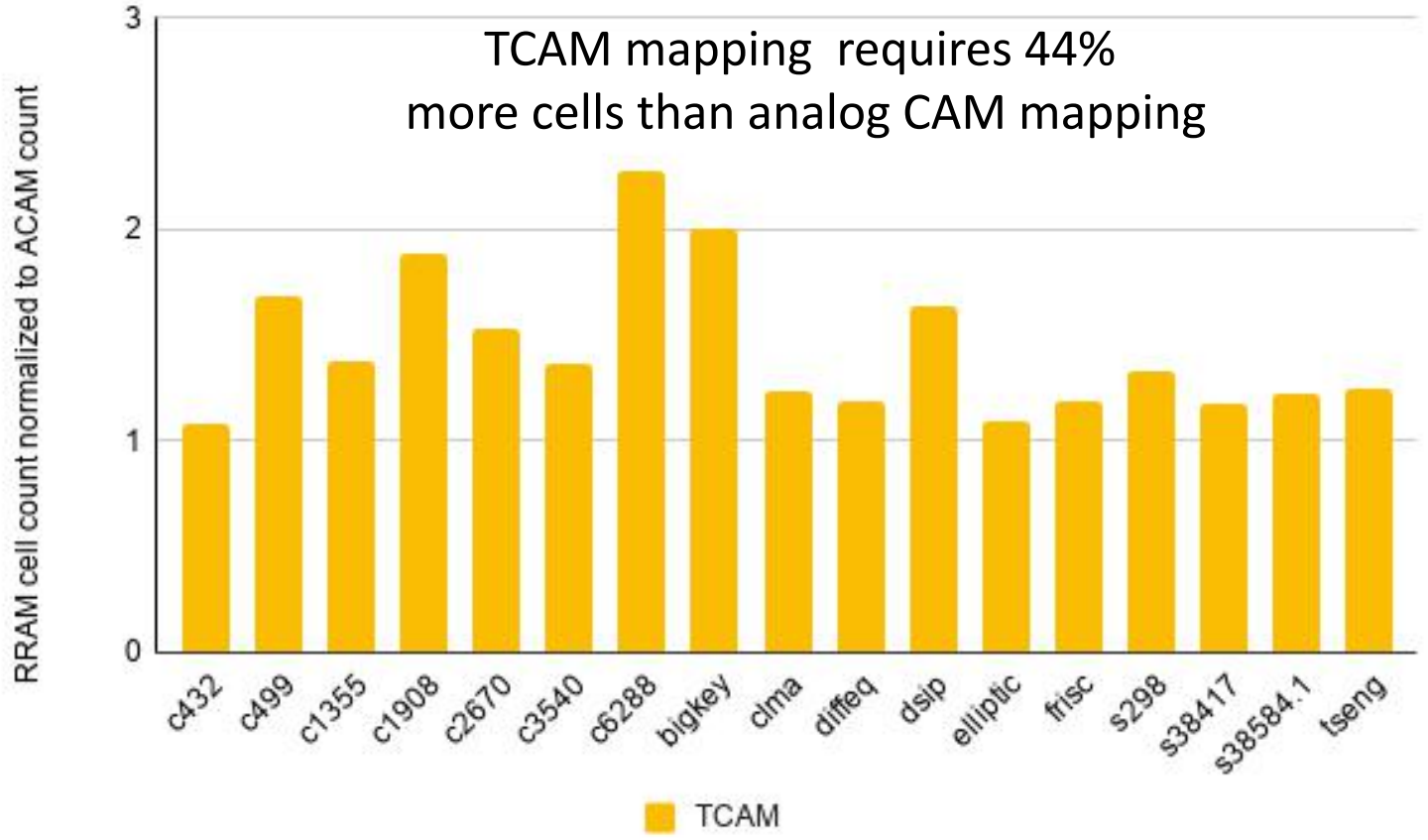
Less impacted by the interconnects because most of the computation is performed within a single partition

Application mapping



Results for ISCAS-85 and MCNC circuits

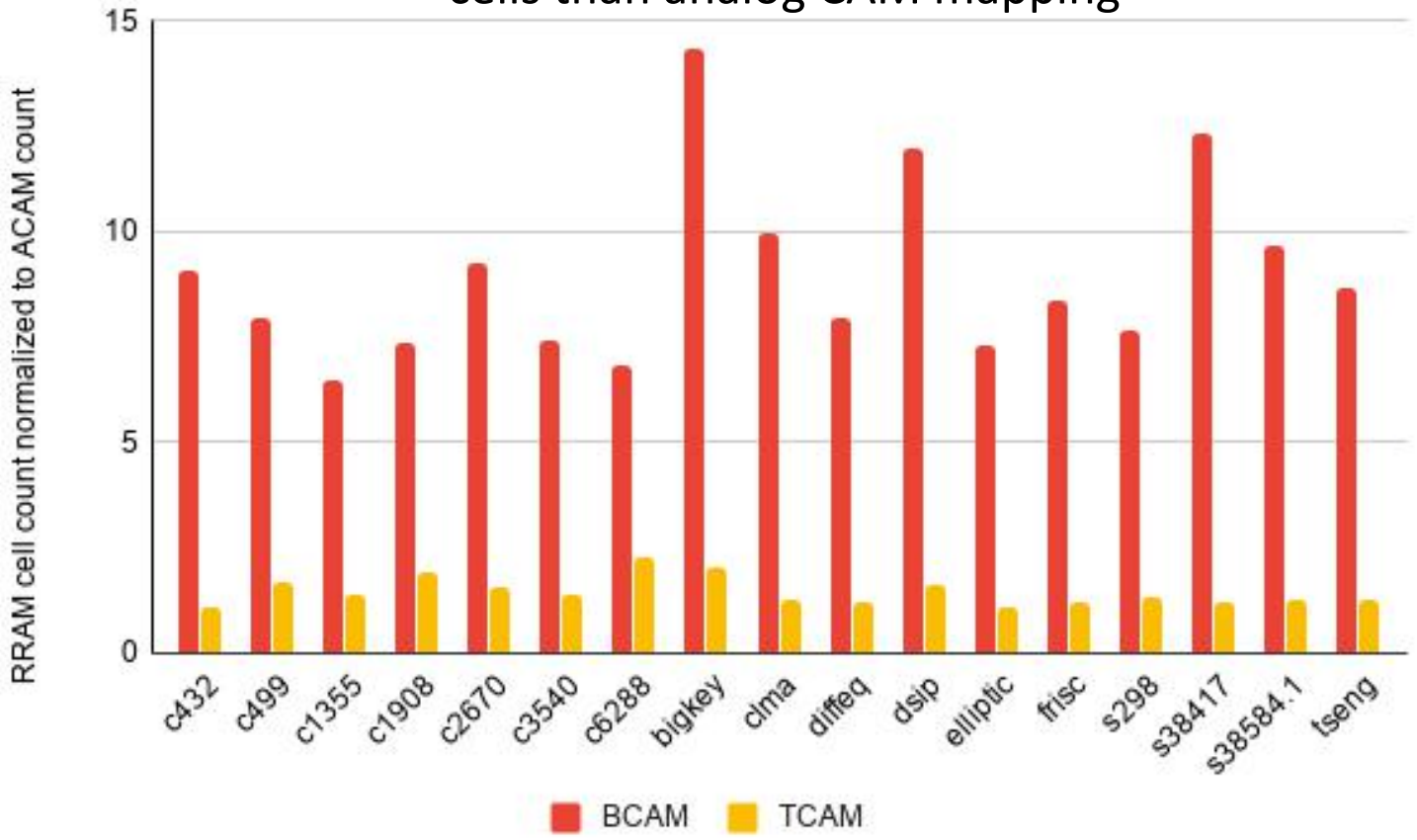
- Number of RRAM cells required to map circuits in TCAMs **over** ACAM count



Results for ISCAS-85 and MCNC circuits

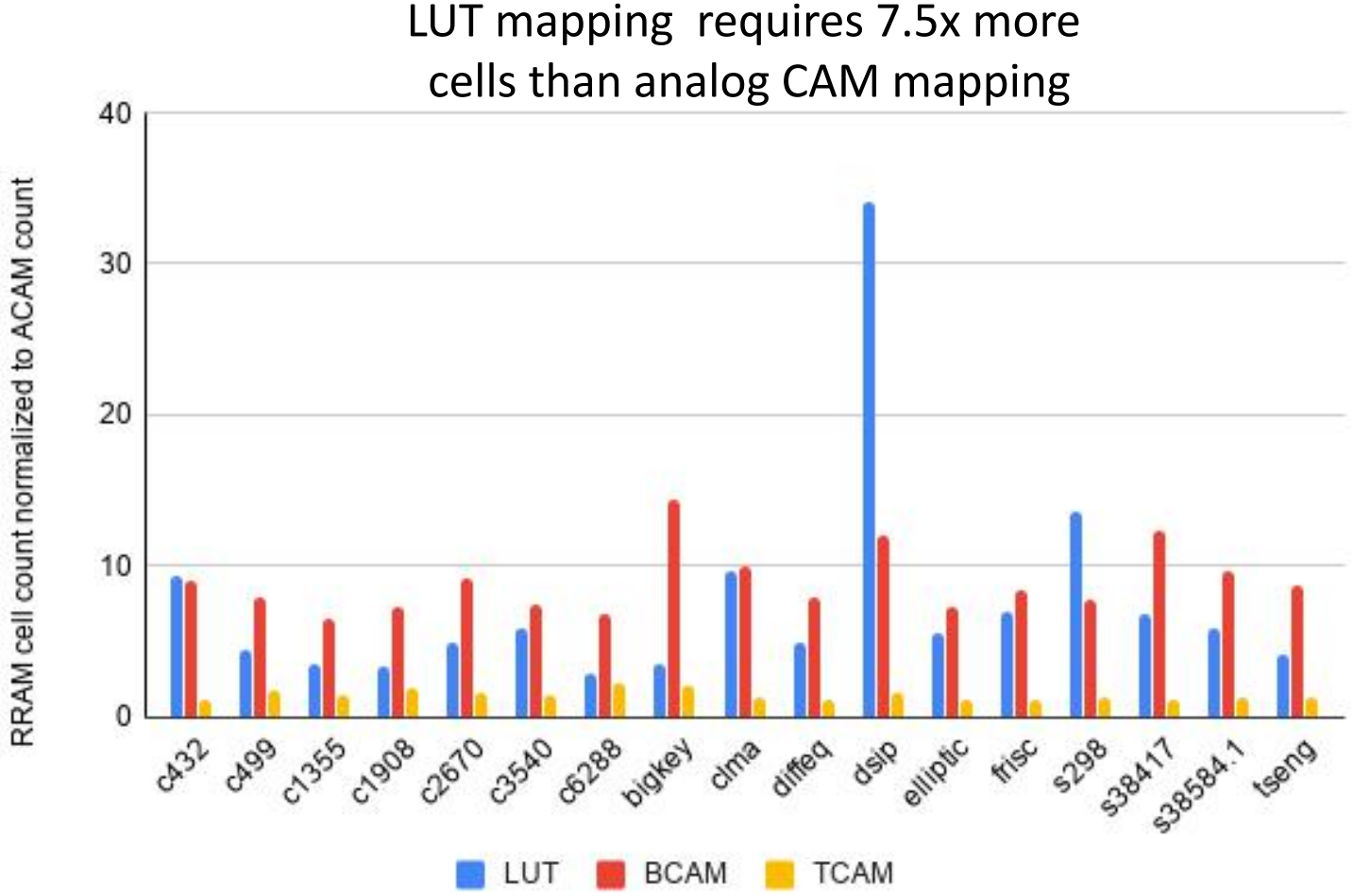
- Number of RRAM cells required to map circuits in BCAMs and TCAMs **over** ACAM count

BCAM mapping requires 9x more cells than analog CAM mapping



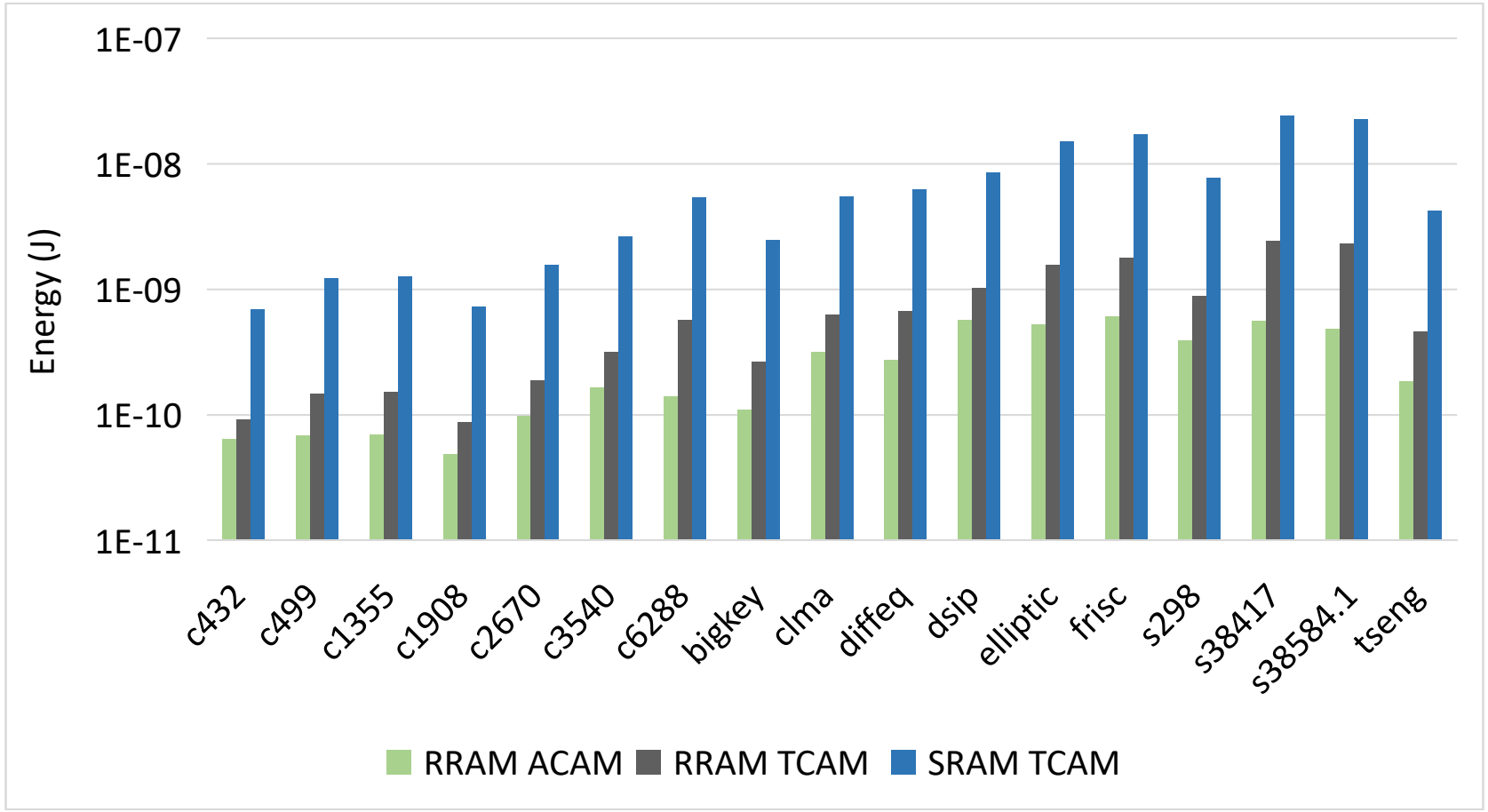
Results for ISCAS-85 and MCNC circuits

- Number of RRAM cells required to map circuits in LUTs, BCAMs and TCAMs over ACAM count



Results for ISCAS-85 and MCNC circuits

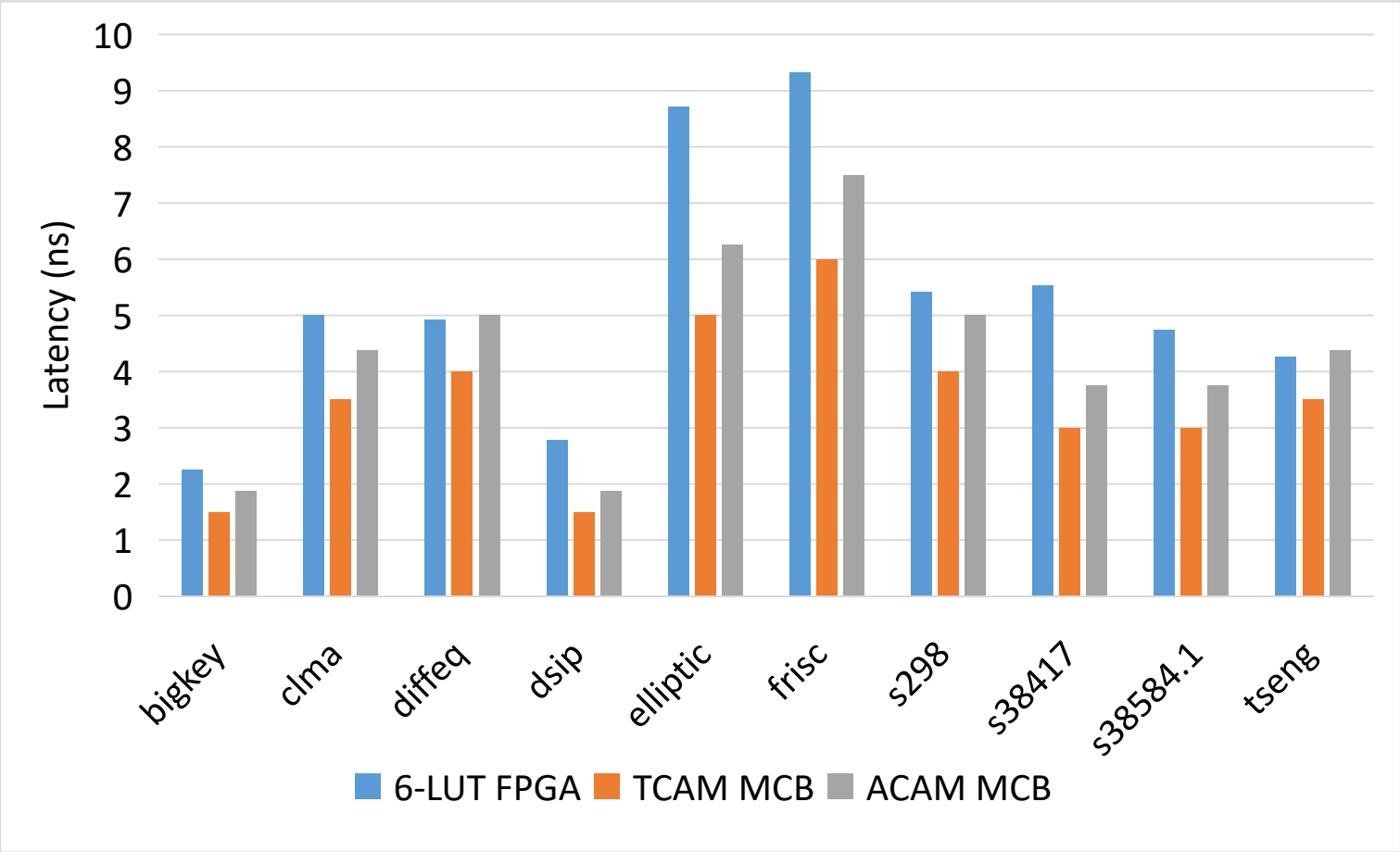
- Dynamic energy for three configurations of MCBs
  - ACAM-based MCB consumes 1.9x and 16.5x less energy than its RRAM-TCAM and SRAM-TCAM counterparts



Results for MCNC circuits

- TCAM and ACAM have similar search latency, but the DAC in the critical path adds 25% in MCB's cycle time
  - Evaluation time = cycle time x # cycles for evaluating partitions on the critical path

25% slower than the TCAM counterpart



Still 21% faster than traditional LUT FPGA

Main takeaways

- Analog RRAM CAMs improve memory density and energy of a reconfigurable fabric

- Higher density by using **multi-level cells** and **range-based compression**
  - Reducing up to 2.2× (1.4× on average) the number of required RRAM cells
- **Search energy per bit** of ACAM is **smaller** than equivalent TCAM
  - Saving 1.9×, on average, more than other non-volatile counterparts

- Future directions:
  - Use the range search capability for processing analog inputs
  - Evaluate approximate circuits and applications using range-based compression

Thank you

João Paulo Cardoso Lima
Rafael Fão de Moura
Luigi Carro

References

- Wang, Zhuo-Rui, et al. "Functionally complete Boolean logic in 1T1R resistive random access memory." IEEE Electron Device Letters 38.2 (2016): 179-182.
- Li, Can, et al. "Analog content-addressable memories with memristors." Nature communications 11.1 (2020): 1-8.
- Paul, Somnath, and Swarup, Bhunia. "Reconfigurable computing using content addressable memory for improved performance and resource usage." Proceedings of the 45th annual Design Automation Conference. 2008.
- Qian, Wenchao, et al. "ENFIRE: A spatio-temporal fine-grained reconfigurable hardware." IEEE Transactions on Very Large Scale Integration (VLSI) Systems 25.1 (2016): 177-188.
- Paul, Somnath, et al. "Energy-efficient reconfigurable computing using a circuit-architecture-software co-design approach." IEEE Journal on Emerging and Selected Topics in Circuits and Systems 1.3 (2011): 369-380.