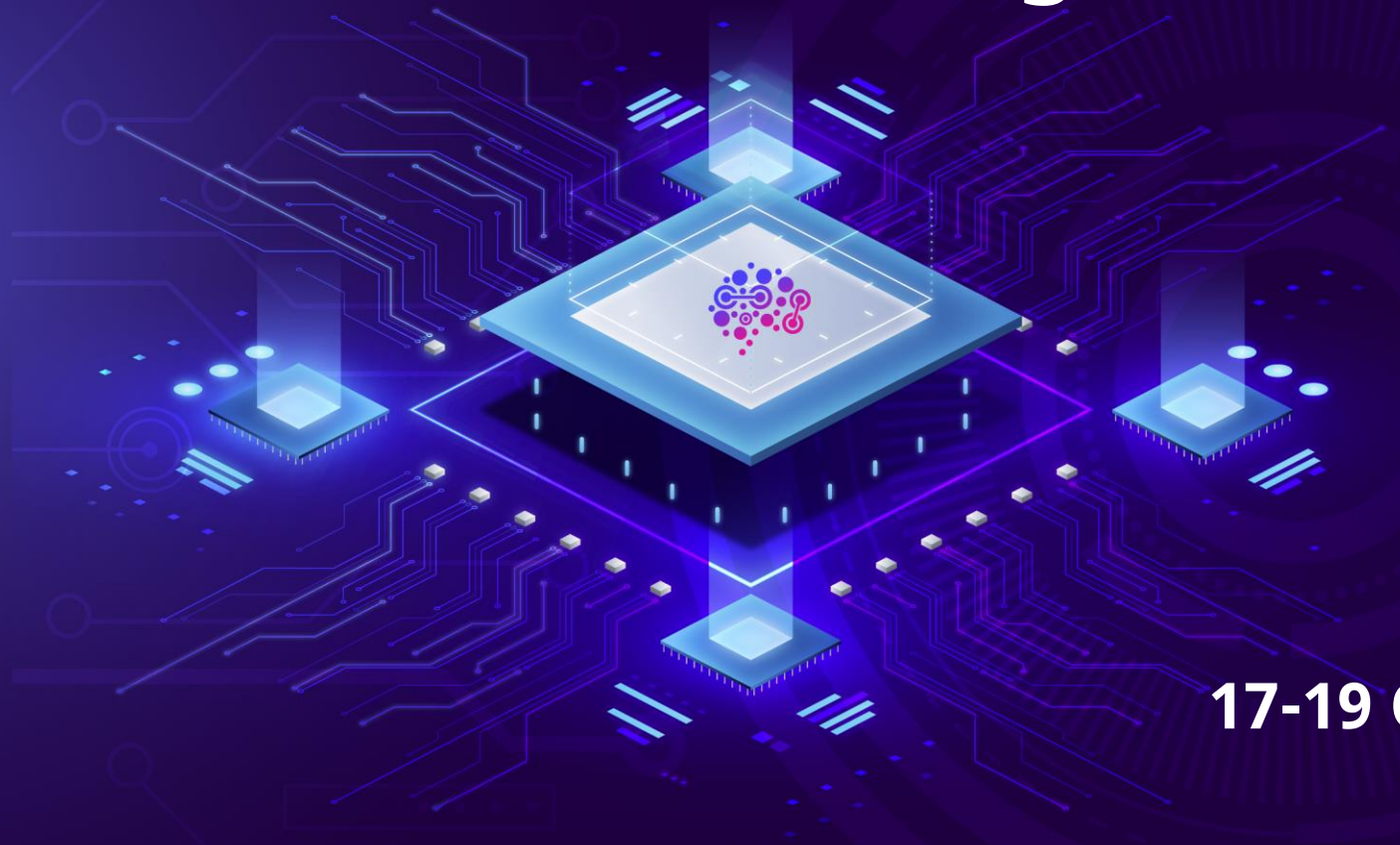
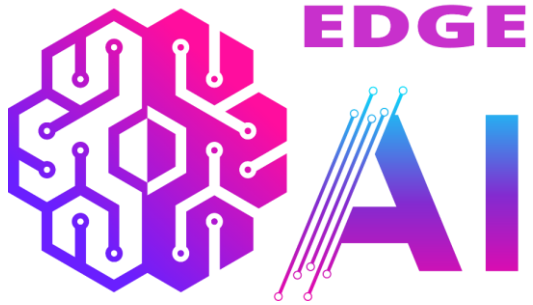
European Conference on EDGE AI Technologies and Applications - EEAI



Advancing emerging edge AI technologies and driving next generation intelligent applications.

17-19 October 2023, Athens, Greece

European Conference on EDGE AI Technologies and Applications - EEAI



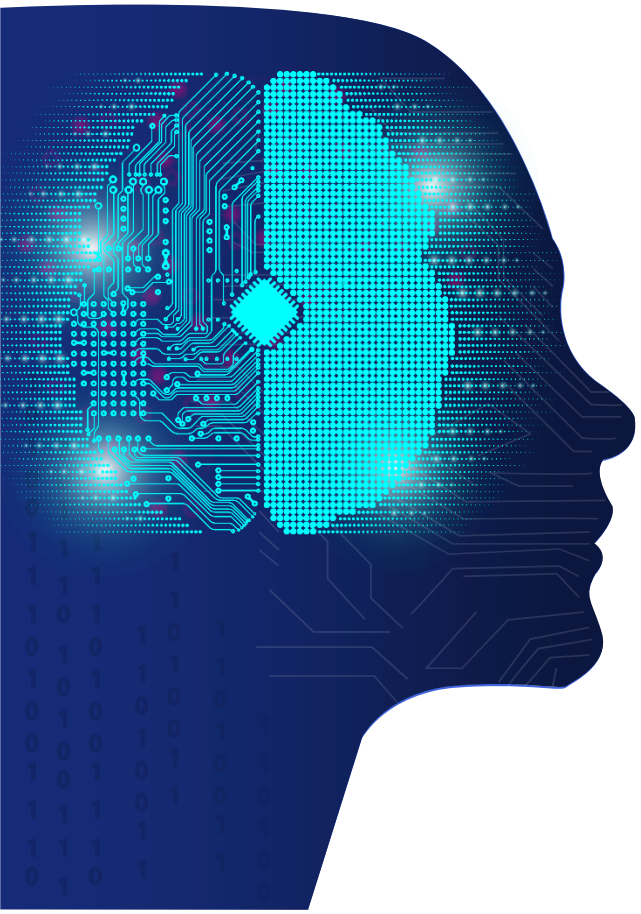
The Beauty of RISC-V Customization for Edge AI HW Acceleration

Arturo Prieto, Lund University, Sweden

17-19 October 2023 Athens, Greece

Copyright © 2023

Presentation Outline

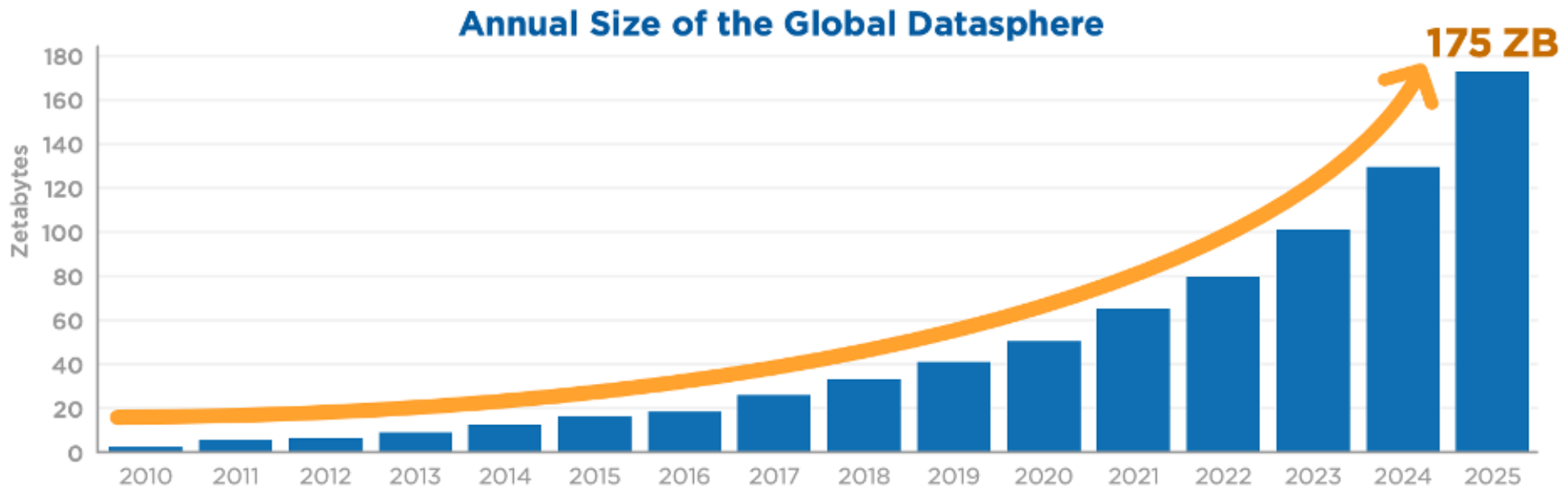


- Introduction
- State-of-the-art
- Research Goal
- Methodology
- Design and Implementation
- Experimental Results
- Future Research
- Conclusion

Copyright © 2023

Introduction

- Today's system design is challenged by the amount of data
- Applications require intense usage of data

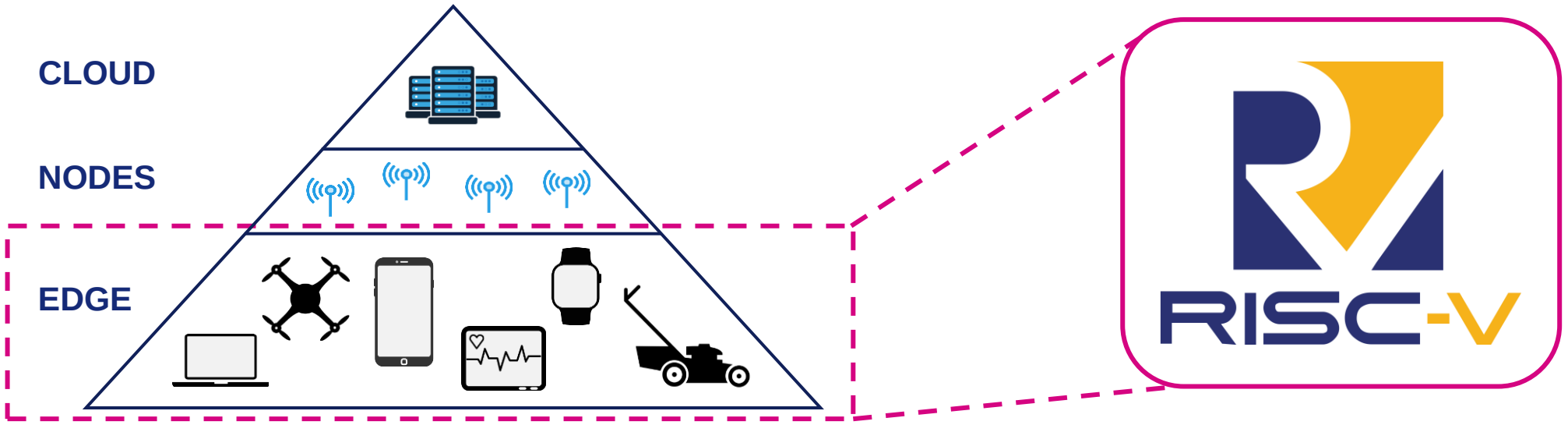


Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere

Copyright © 2023

State-of-the-art

- Edge computing devices perform local data processing
- Smartness by Neural Networks (NN) on edge devices



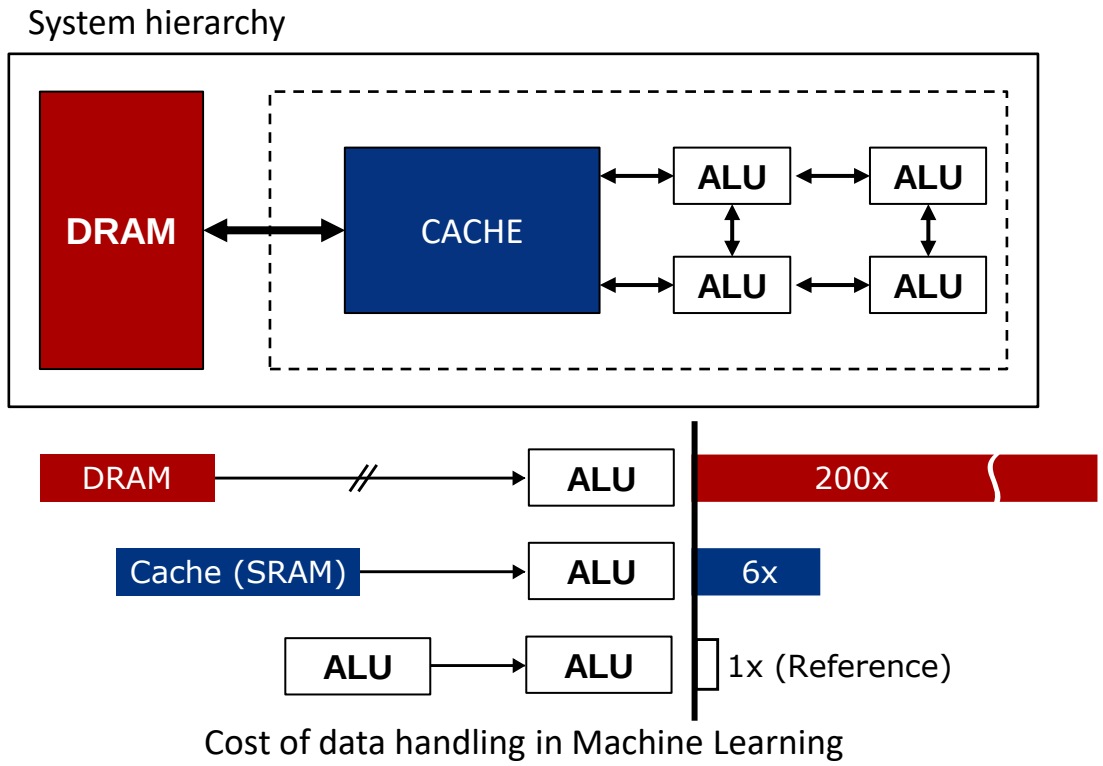
Limited energy budget
Limited on-chip memory

Low power computation
Efficient memory utilization
Flexibility and scalability

Copyright © 2023

State-of-the-art

- Cost of data movement can dominate the energy and throughput
- Limited bandwidth in system memory hierarchy

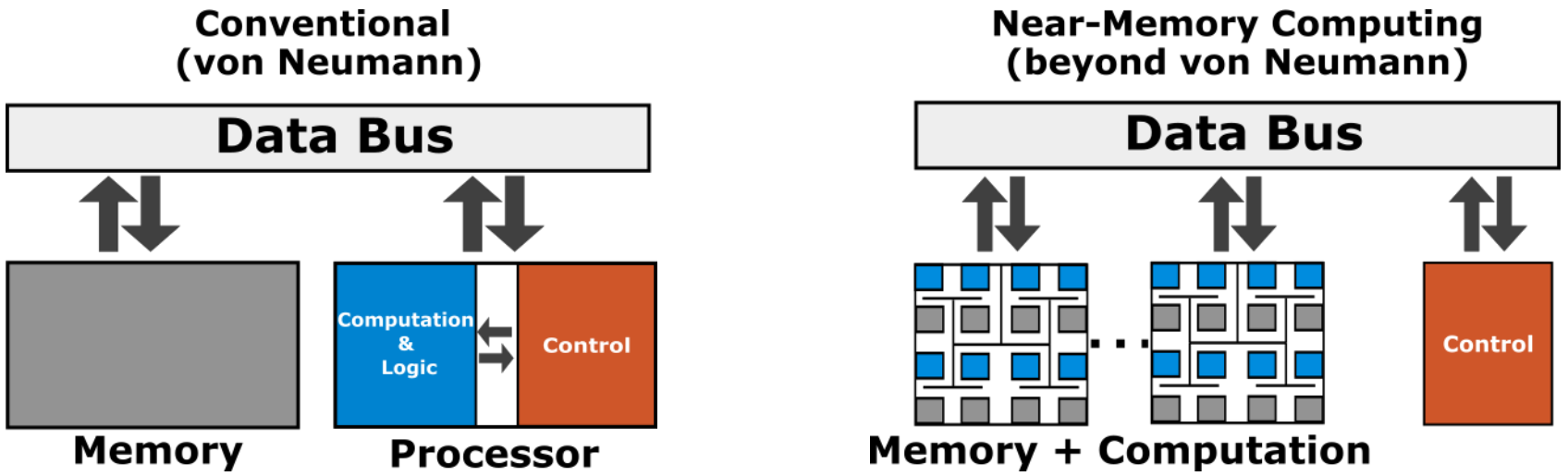


Y. Chen, J. Emer and V. Sze, "Using Dataflow to Optimize Energy Efficiency of Deep Neural Network Accelerators," in *IEEE Micro*, vol. 37, no. 3, pp. 12-21, 2017

Copyright © 2023

Research Goal

- Beyond-von Neumann architectures with data-centric approach

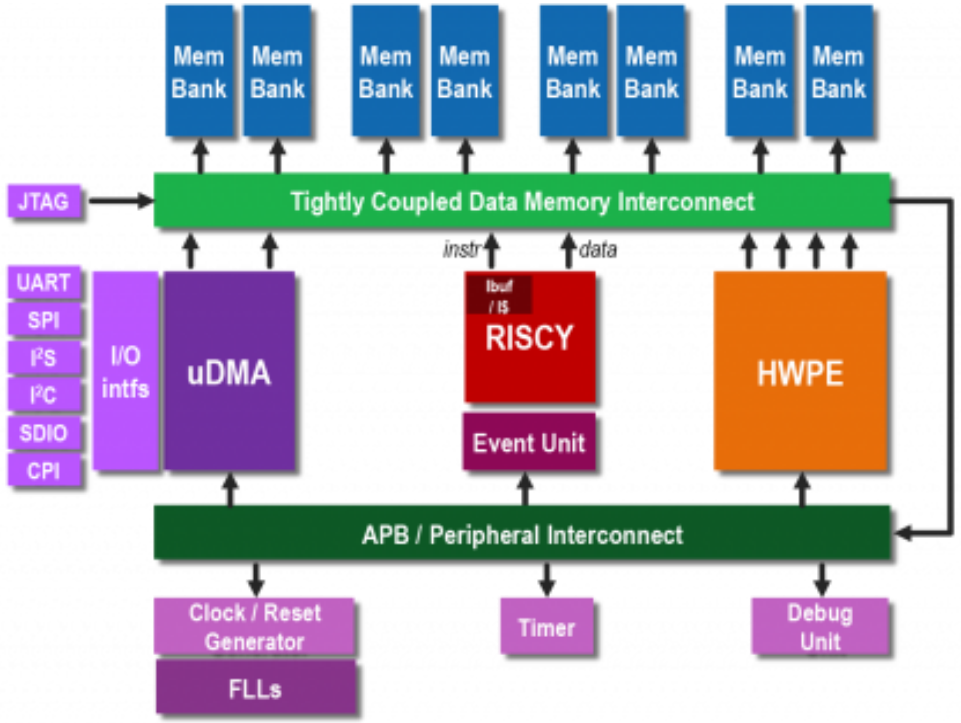


- NMC architecture realized as co-processor integrated in an MCU cache memory using SRAM

Copyright © 2023

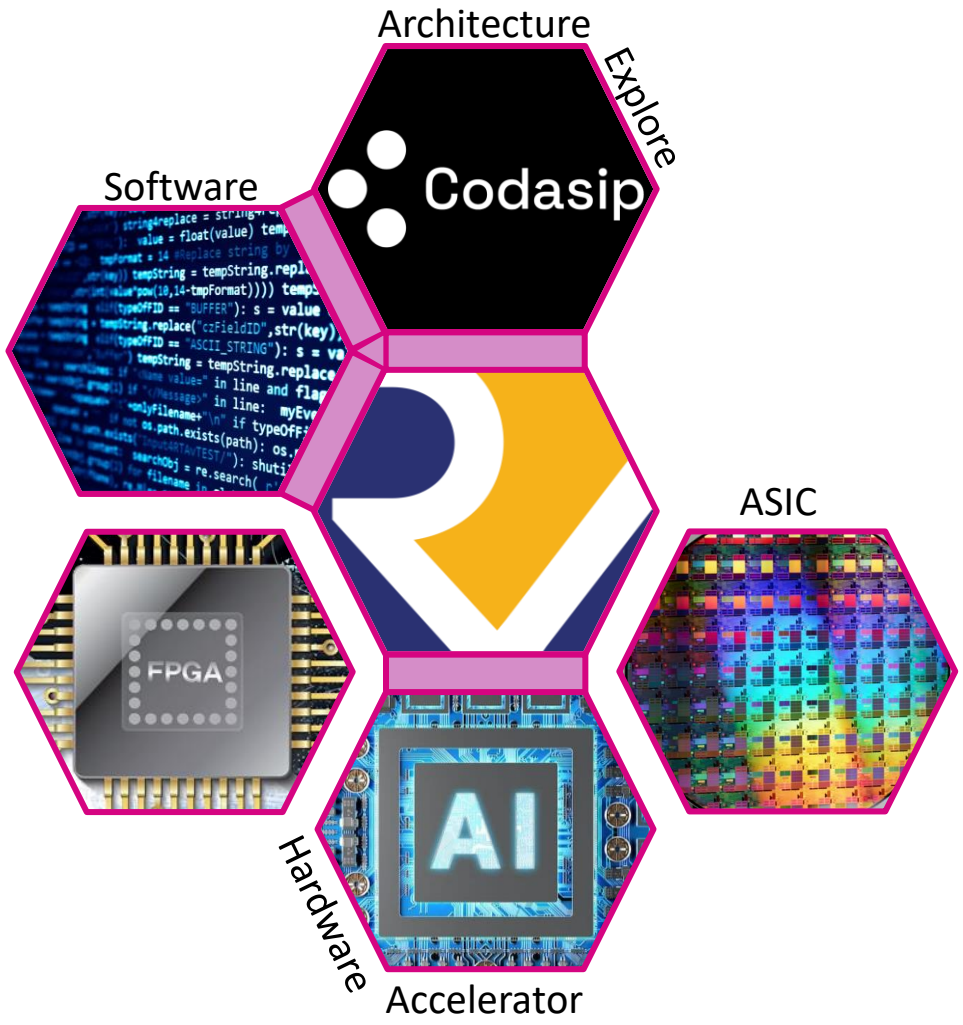
Methodology

Open-source low-power MCU with RISC-V core (base case)



P. D. Schiavone et al., "Quentin: and Ultra-Low-Power PULPissimo SoC in 22nm FDX," *IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, pp. 1-3, 2018

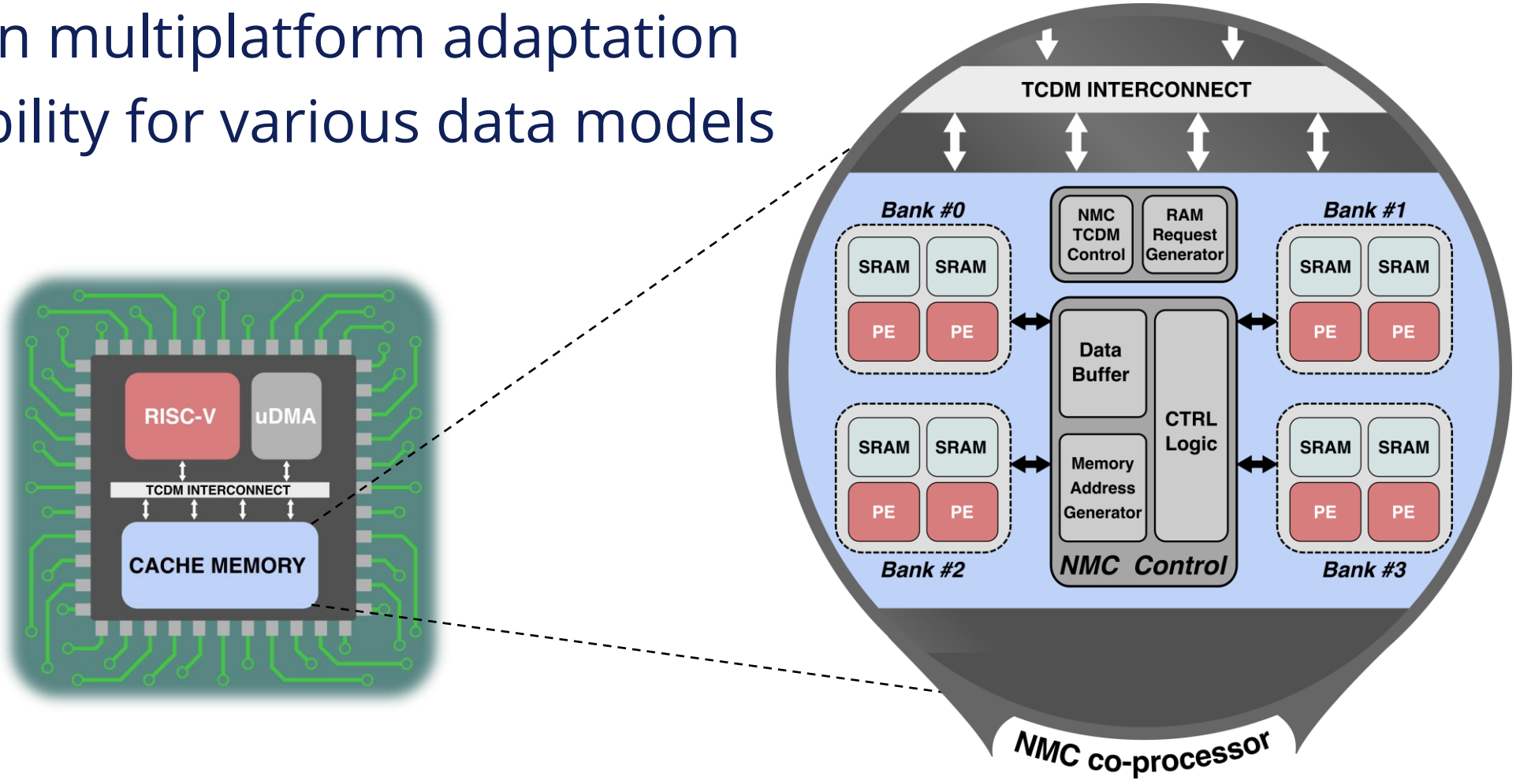
Use of a processor design platform



Copyright © 2023

Design and Implementation

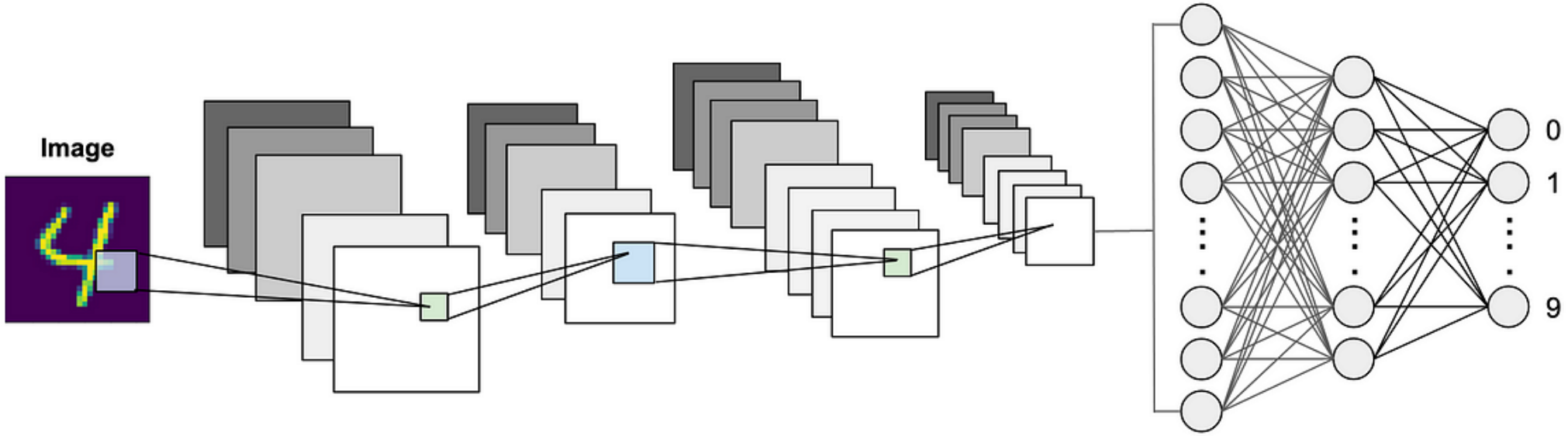
- Co-processor for data intensive computations
- Flexibility in multiplatform adaptation
- Configurability for various data models



Copyright © 2023

Design and Implementation

- Hardware (HW) Accelerator for CNN models on image classification
  - Convolution + Pooling + Fully Connected layers



  - 8-bit fixed-point data
  - Model distribution across HW accelerator

Copyright © 2023

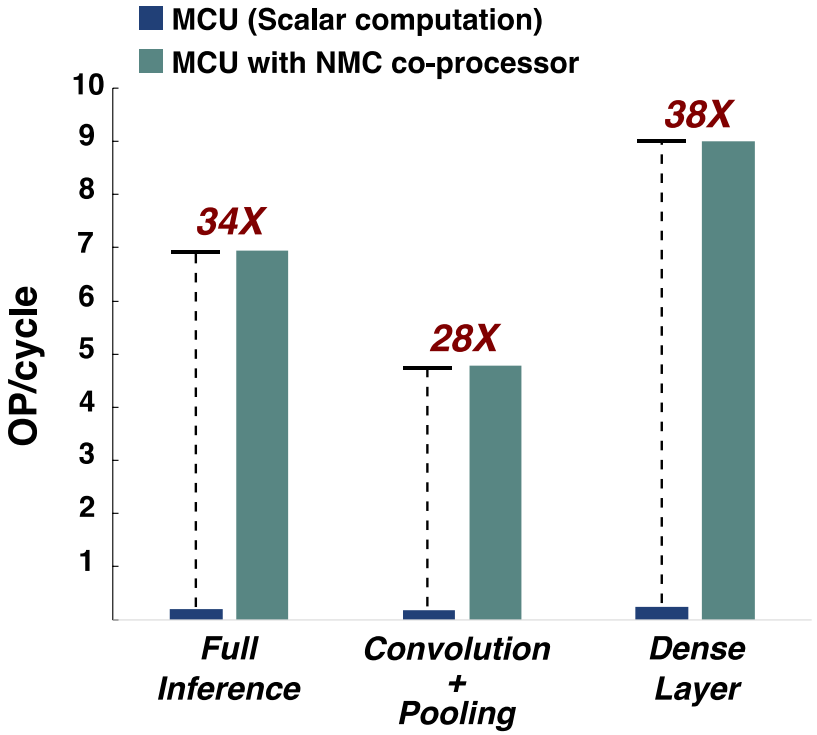
Design and Implementation

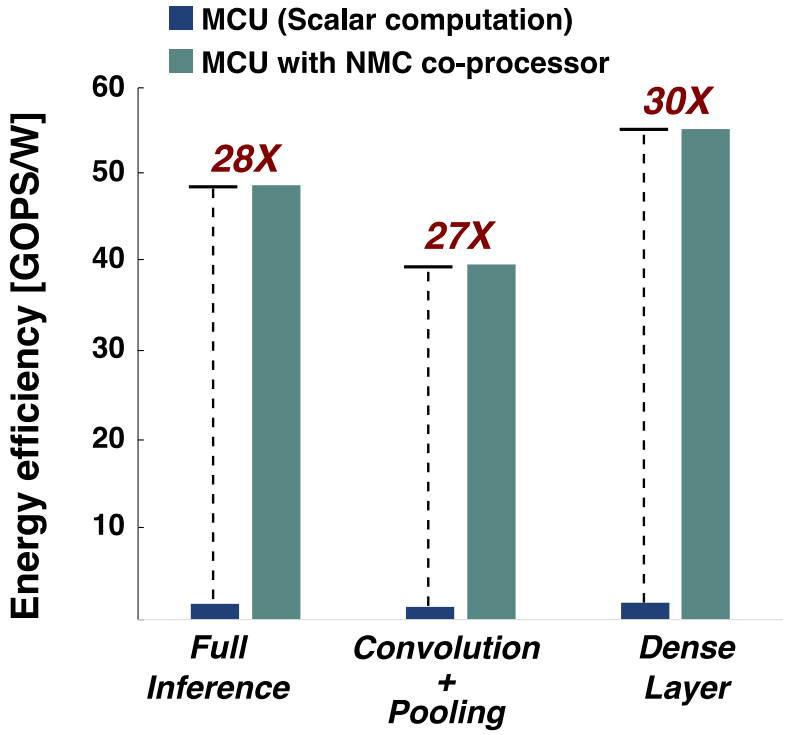
- Integration on MCU platform with RISC-V core
  - 512 kB Cache Memory
    - 4 re-purposed memory banks for CNN inference computations
- The NMC design accommodates PE units in the memory-subsystem achieving a lower area cost than other HW accelerator alternatives
- Synthesis in 28 nm process technology shows that the NMC unit of this work is realized with limited area overhead

Copyright © 2023

Experimental Results

- Performance and energy efficiency improvement compared to baseline scenario (MNIST full inference on 1 core MCU)





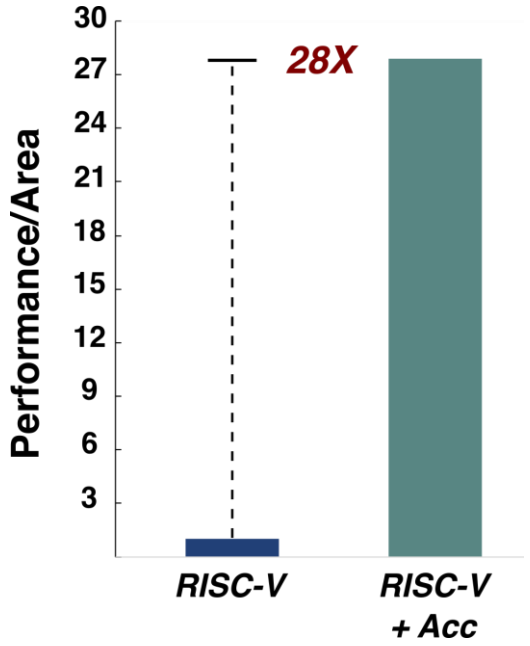
Performance gain: reduced execution time

Energy efficiency: reduced energy-hungry memory accesses

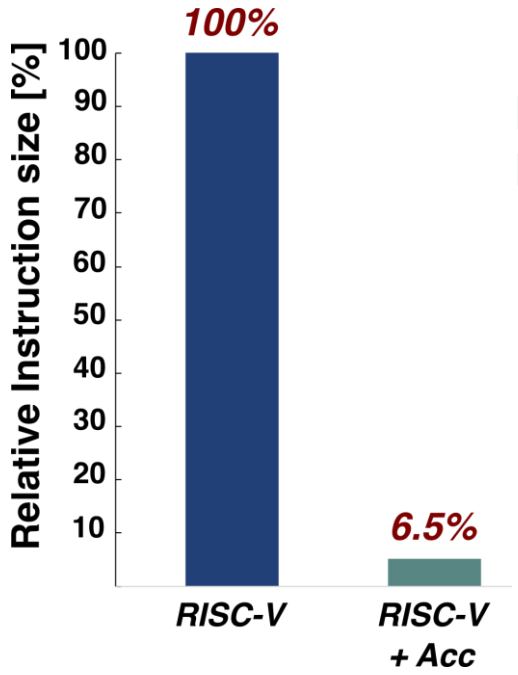
Copyright © 2023

Experimental Results

- Performance/Area & Instruction size improvement compared to baseline scenario (MNIST full inference on 1 core MCU)



Performance gain with limited area overhead



Reduced workload of RISC-V core

Copyright © 2023

Future Research

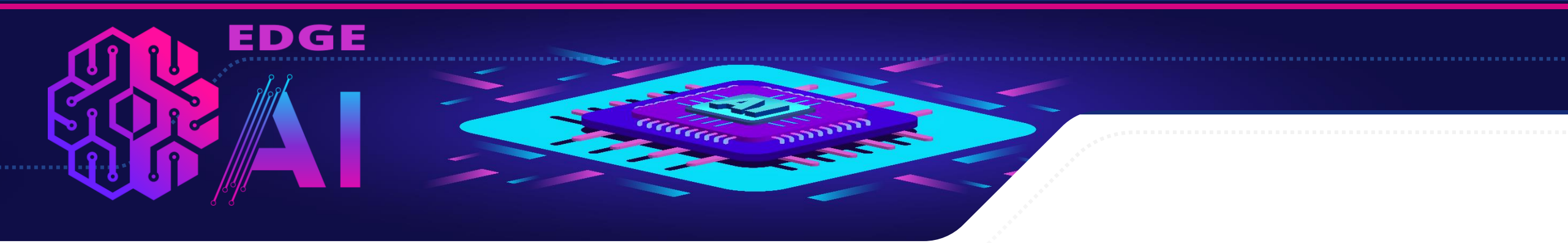
- Design of other HW accelerators for other applications
- Optimize the complete design stack (from model to HW acceleration)
- Compute-In-memory system integration
- Silicon prototype
- Software development

Copyright © 2023

Conclusion

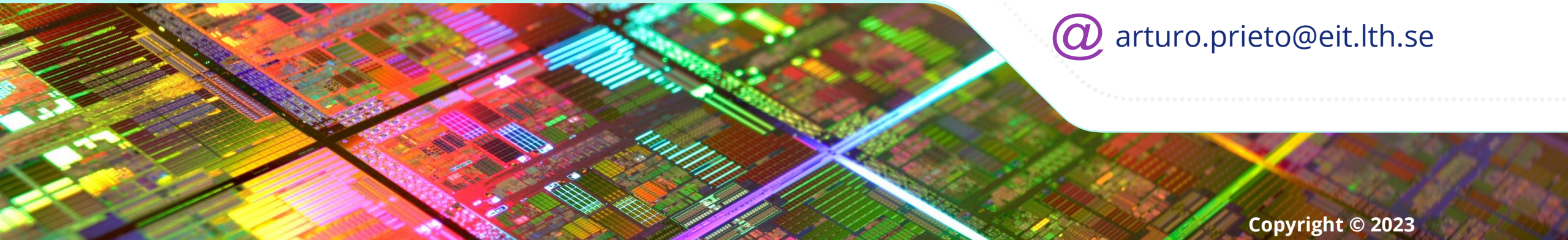
- Beyond-von Neumann NMC architecture for CNN
- Integration as a co-processor in the cache level
- Increased memory bandwidth by re-purposing conventional SRAM
- Improve energy-efficiency with reduced area overhead
- Scalable and platform-agnostic architecture

Copyright © 2023



Thank You

For your attention



@ arturo.prieto@eit.lth.se

Copyright © 2023