# An Energy-Efficient Hardware Architecture for Data Intensive Edge Devices

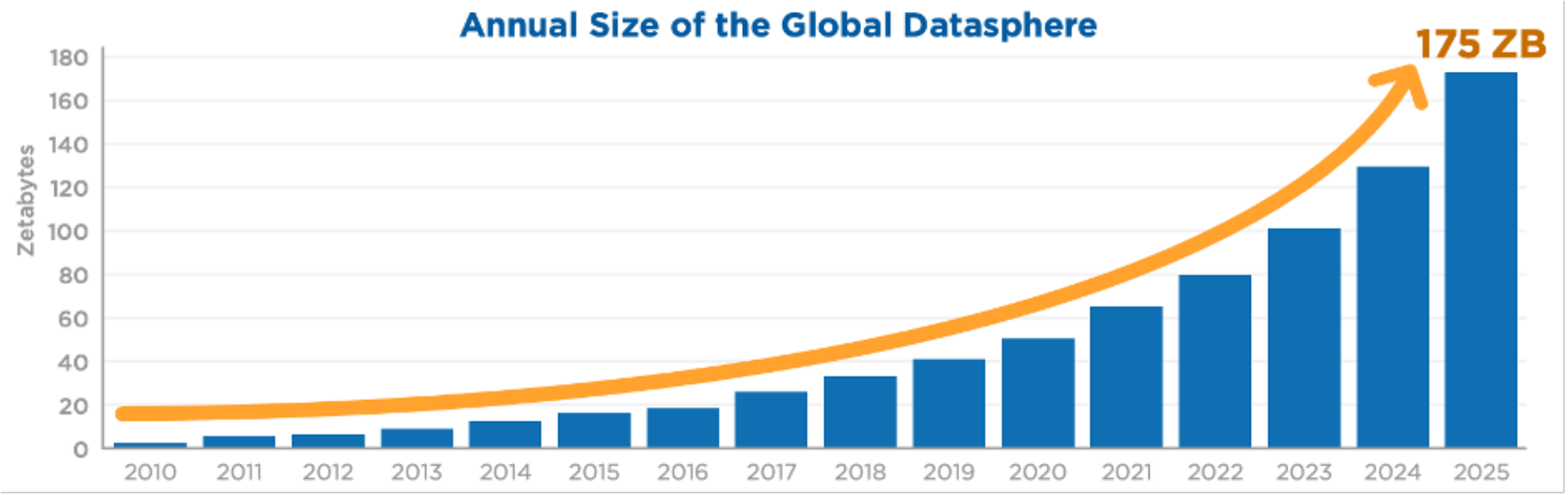
Arturo Prieto, Integrated Electronic Systems, EIT, LTH

# Team Profile

- Circuits for energy efficient Edge-AI processing
- RISC-V systems
  - Dedicated HW-accelerators
  - Custom instructions
- ASIC and FPGA integration flows (65nm, 28nm, 22FDX)
- Open-source & Commercial RISC-V IP

# An Energy-Efficient Hardware Architecture for Data Intensive Edge Devices

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



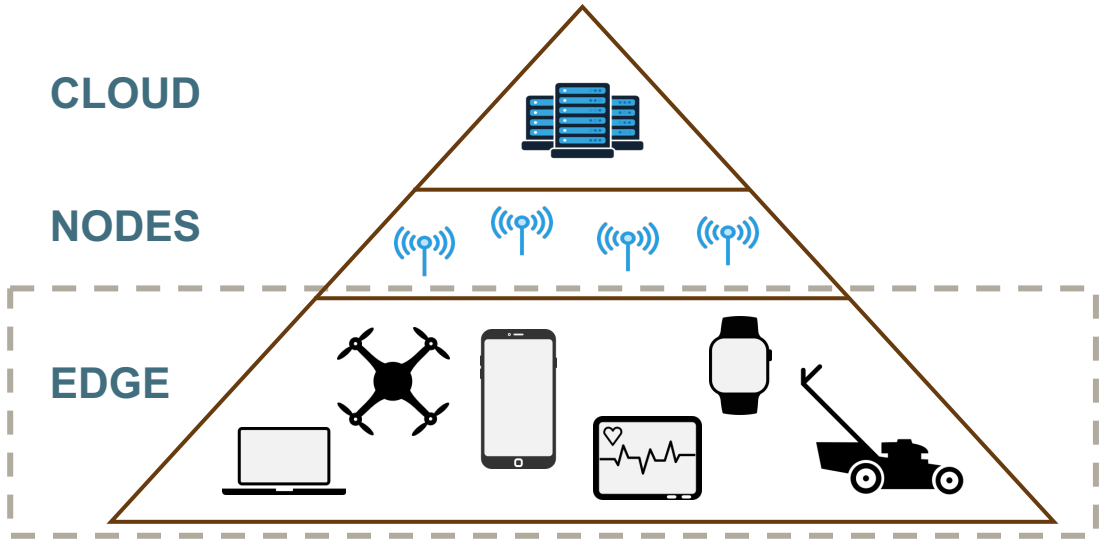

Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere



LUND UNIVERSITY

# An Energy-Efficient Hardware Architecture for Data Intensive Edge Devices

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




**Limited energy budget**
**Limited on-chip memory**

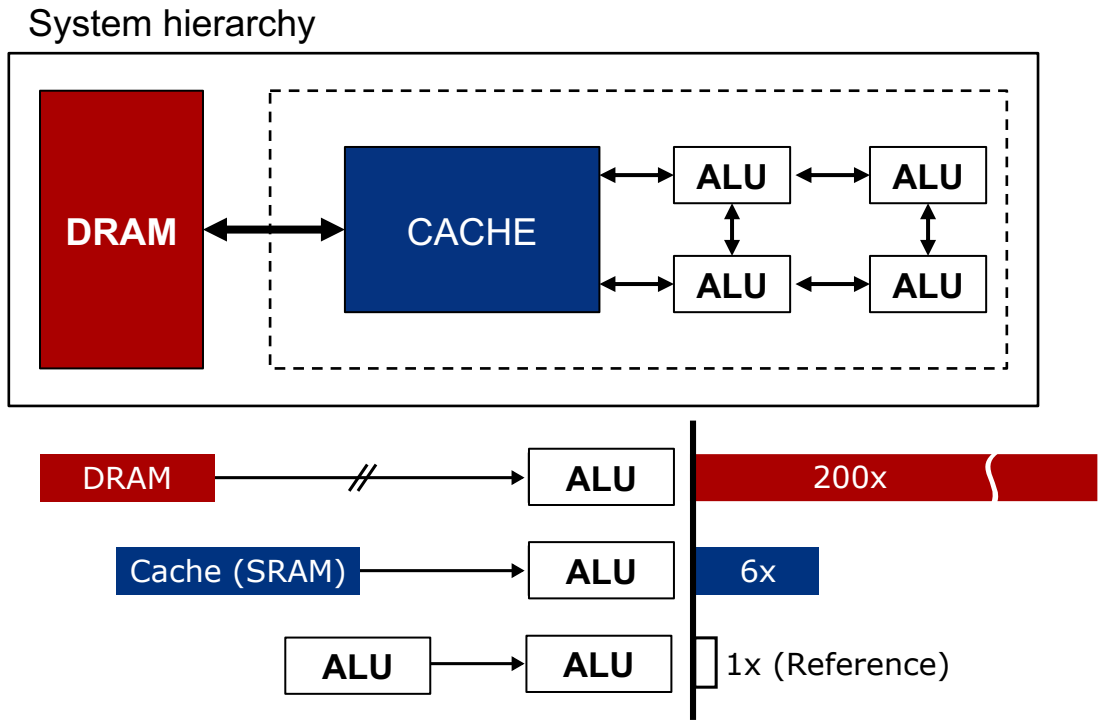
**Low power computation**
**Efficient memory utilization**
**Flexibility and scalability**



LUND UNIVERSITY

# An Energy-Efficient Hardware Architecture for Data Intensive Edge Devices

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



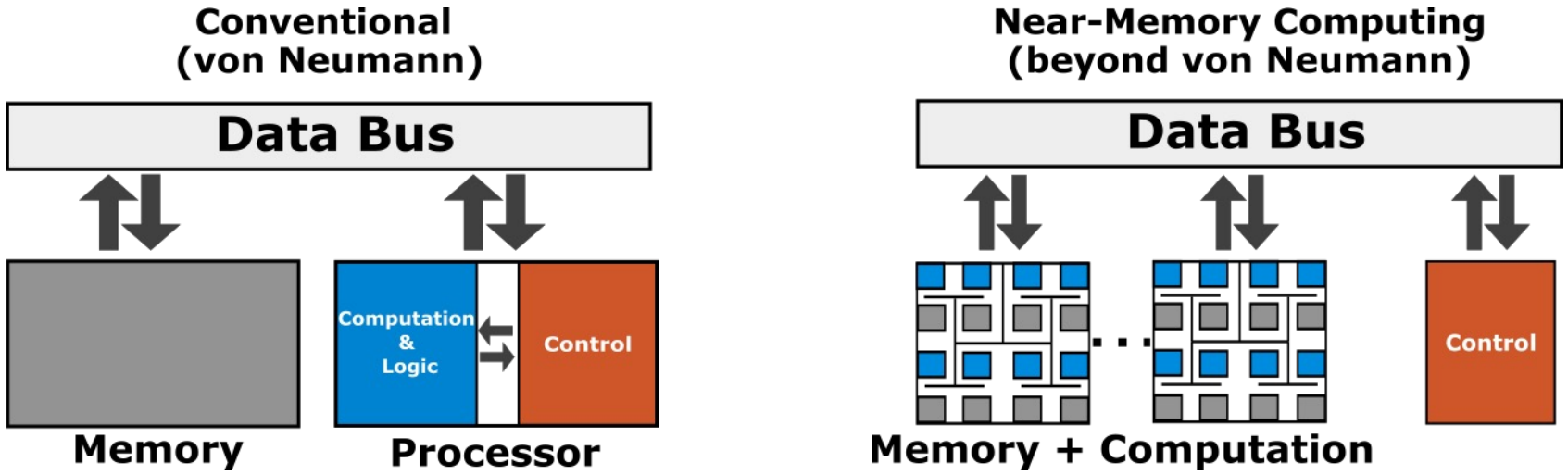

Cost of data handling in Machine Learning [1]



[1] 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

# An Energy-Efficient Hardware Architecture for Data Intensive Edge Devices

- Beyond-von Neumann architecture
  - Data-centric approach



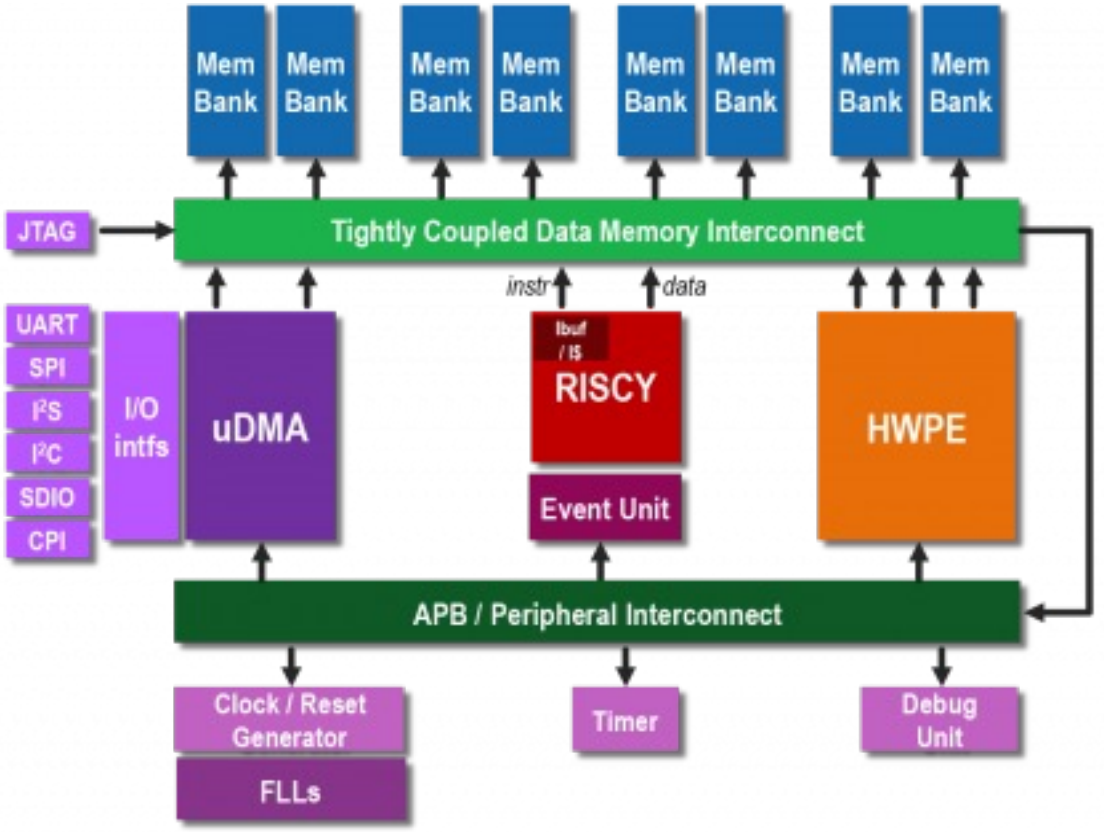

- NMC architecture realized as co-processor integrated in a microcontroller unit (MCU) cache memory using SRAM



LUND UNIVERSITY

# An Energy-Efficient Hardware Architecture for Data Intensive Edge Devices

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



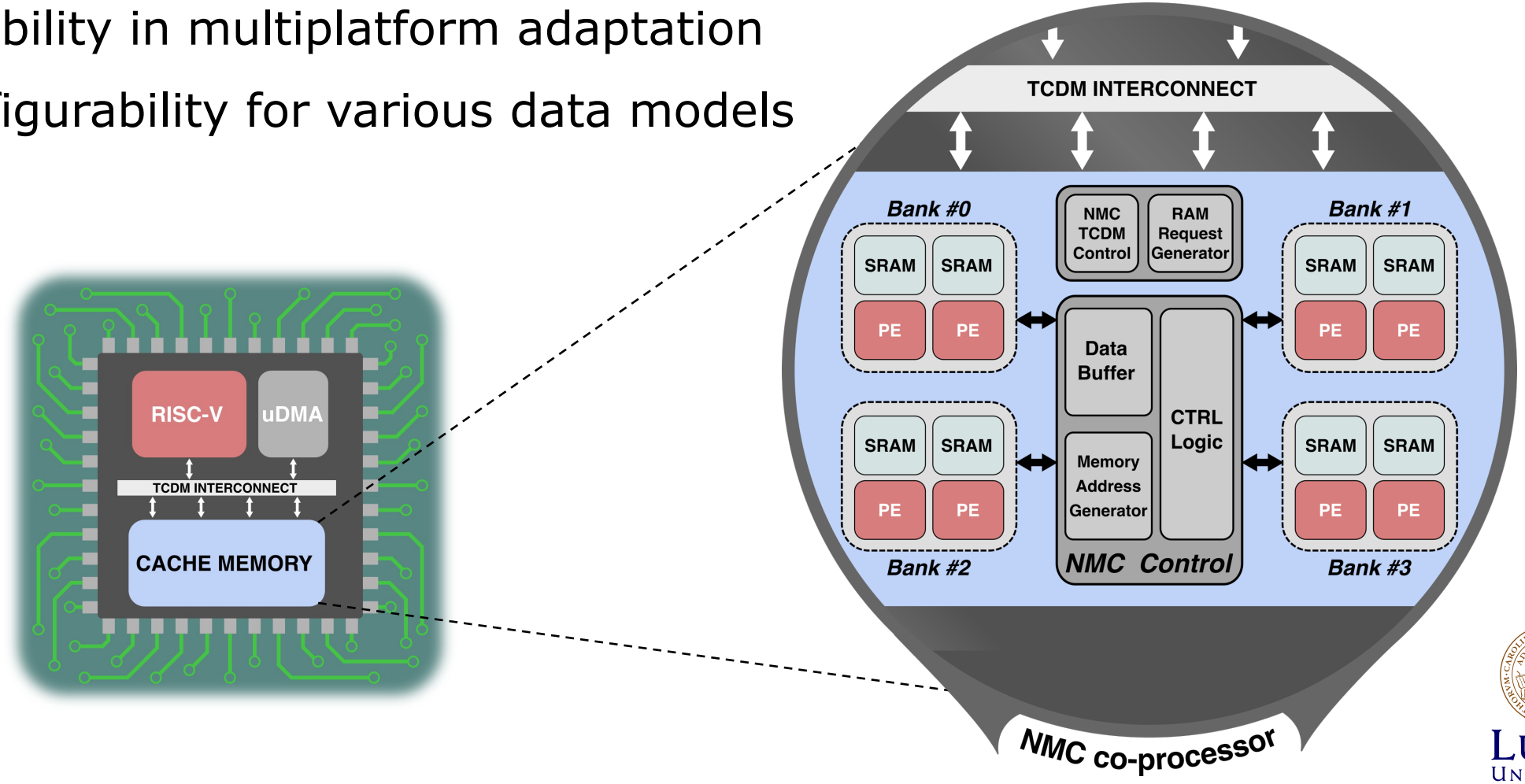

[2] P. D. Schiavone et al., "Quentin: an Ultra-Low-Power PULPissimo SoC in 22nm FDX," *IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, pp. 1-3, 2018



LUND UNIVERSITY

# An Energy-Efficient Hardware Architecture for Data Intensive Edge Devices

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

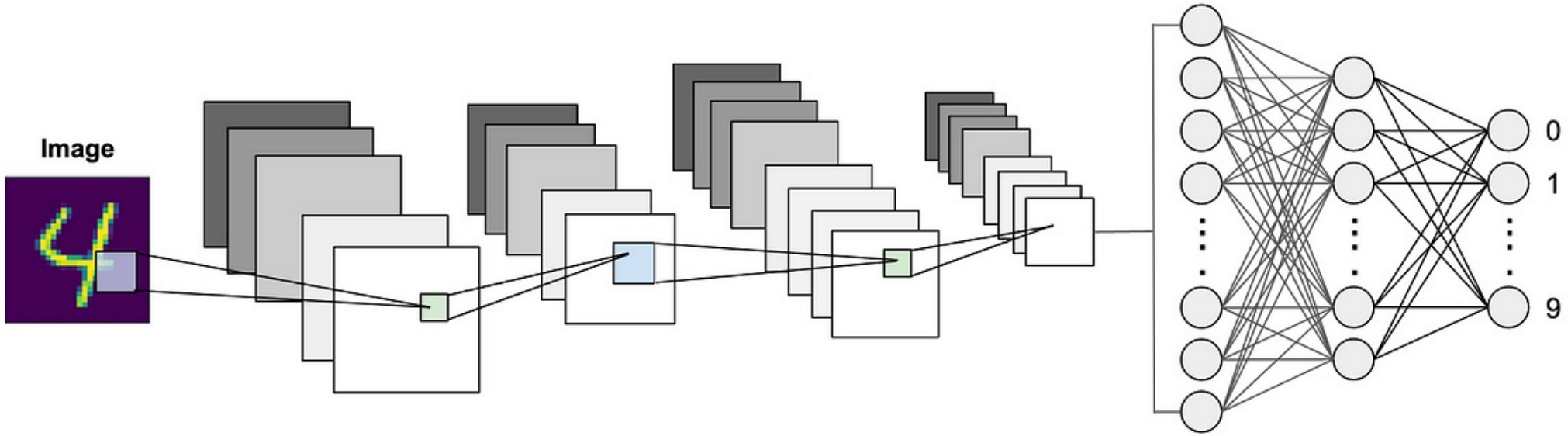





LUND UNIVERSITY

# An Energy-Efficient Hardware Architecture – Evaluation

- CNN model for image classification on MNIST dataset
  - Convolution + Pooling + Fully Connected layers




**Model accuracy > 98%**

**# Multiply-Accumulate operations ≈ 55K**

**8-bit fixed-point**



LUND UNIVERSITY

# An Energy-Efficient Hardware Architecture – Discussion

- 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 hardware-accelerator alternatives

- Synthesis in 28 nm process technology shows that the NMC unit of this work is realized with a negligible area overhead

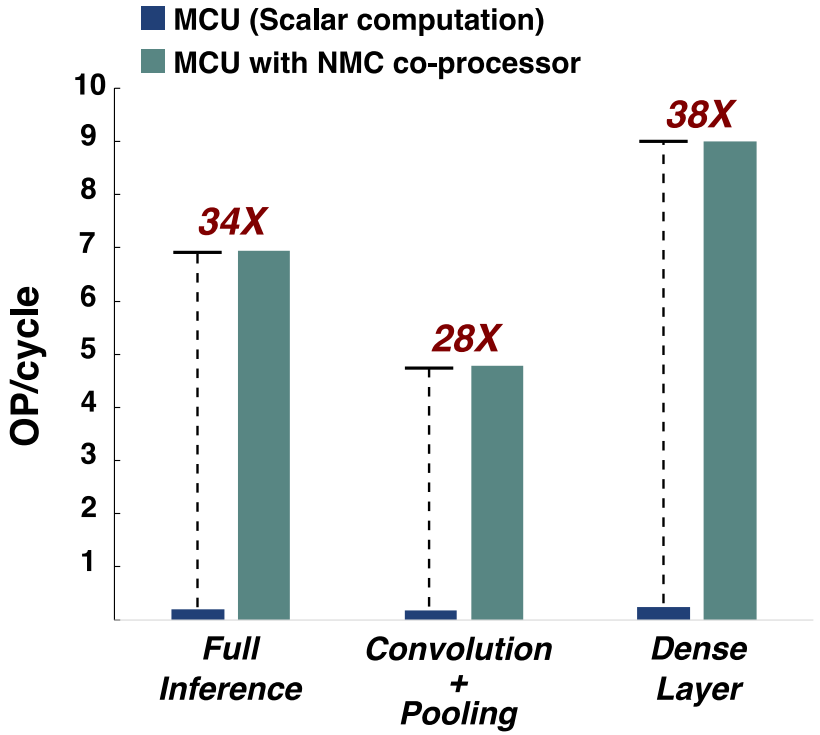
# An Energy-Efficient Hardware Architecture – Results

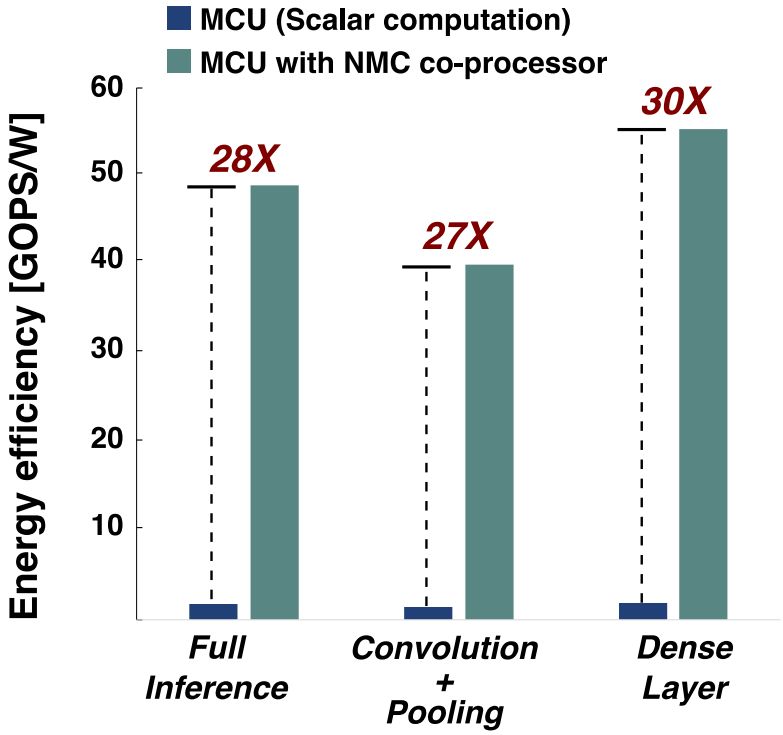
| Source | ICECS 2019 | ISLPED 2020 | ICECS 2019 | A-SSCC 2018 | VLSIC 2020 | This work |
|---|---|---|---|---|---|---|
| **Design Level** | MCU | | NN Library | Accelerator | | NMC @ Cache |
| **Clock [MHz]** | 480 | | 170 | 650 | | 200 |
| **Performance [GOPS]** | 0.03 | 0.15 | 1.07 | 20.3 | 46.15 | **1.39** |
| **Energy Efficiency [GOPS/W]** | 1.7 | 0.61 | 16.1 | 30.03 | 16.19 | **49** |

- The proposed NMC architecture improves:
  - **9-46x** performance and **28-80x** energy-efficiency to MCUs
  - **1.3x** performance and **3x** energy-efficiency to NN Library
  - **1.6-3x** energy-efficiency compared to hardware-accelerators

# An Energy-Efficient Hardware Architecture – Results

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






LUND UNIVERSITY

# An Energy-Efficient Hardware Architecture – Conclusion

- A 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 negligible area overhead
- Scalable and MCU platform-agnostic architecture


LUND
UNIVERSITY