JULY 9–13, 2023

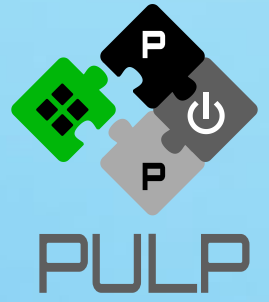
MOSCONE WEST CENTER

SAN FRANCISCO, CA, USA

Specialisation meets Flexibility: A Heterogeneous Architecture for High-Efficiency, High-Flexibility AR/VR Processing

Arpan Suravi Prasad[1], Luca Benini[1,2], Francesco Conti[2]

[1]ETH Zürich, [2]University of Bologna



ETH zürich

Outline

- Introduction & Motivation
- ArchiMEDES: Heterogeneous Cluster Template
  - NEureka DNN Accelerator
    - Scalability
    - Performance
  - HCI interconnect
    - Motivation
    - Architecture
- AR/VR case study
- State of the Art Comparison
- Conclusion

“Smart” Glasses Today

Socially Acceptable form factor → **like regular glasses**

Lightweight → **<50 gram**

All-day battery → LiPo **167mAh @ 3.7V → 25mW for 24-h operation**

Meta Ray-Ban Stories

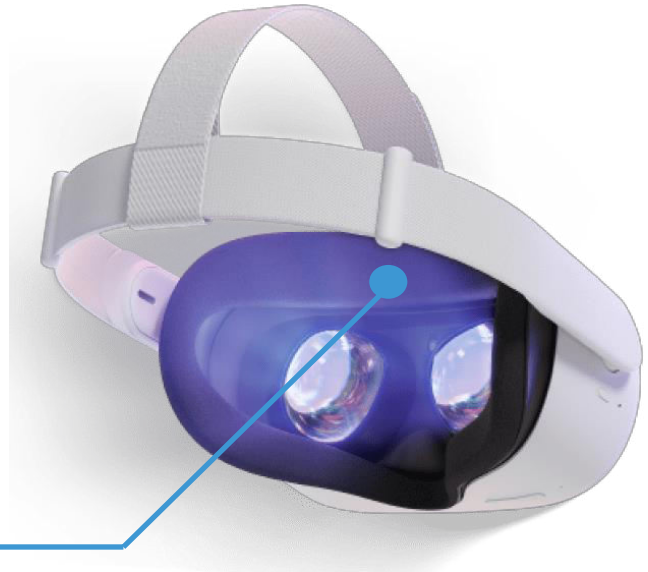
[https://www.techinsights.com/blog/ray-ban-stories-smart-glasses-cameras]

eXtended Reality Glasses, today

Dedicated form factor → **cumbersome and uncomfortable in the long run**

Heavyweight → **~500 gram**

2/3-hours battery → Li-ion **3640mAh @ 3.85V → ~5W operation**



Meta Quest 2

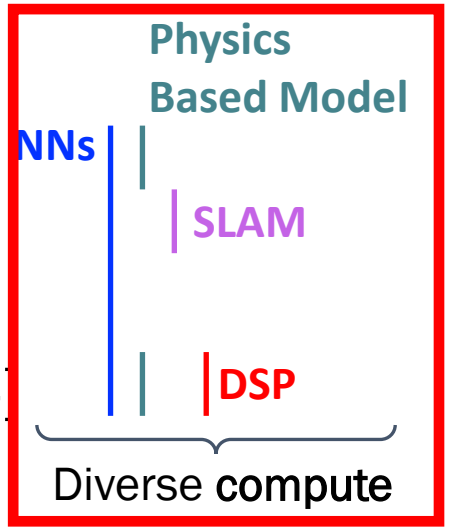
Apple Vision Pro

Microsoft HoloLens 2

[https://en.wikichip.org/wiki/qualcomm/snapdragon_800/865] [https://arstechnica.com/] [https://www.apple.com/newsroom]

How to “fold” XR functionality into smart glasses?

- **Many tasks to manage…**
  - Eye gaze tracking [1]
  - Head tracking [2]
  - Hand tracking [3]
  - Spatial beamforming [4]
  - …



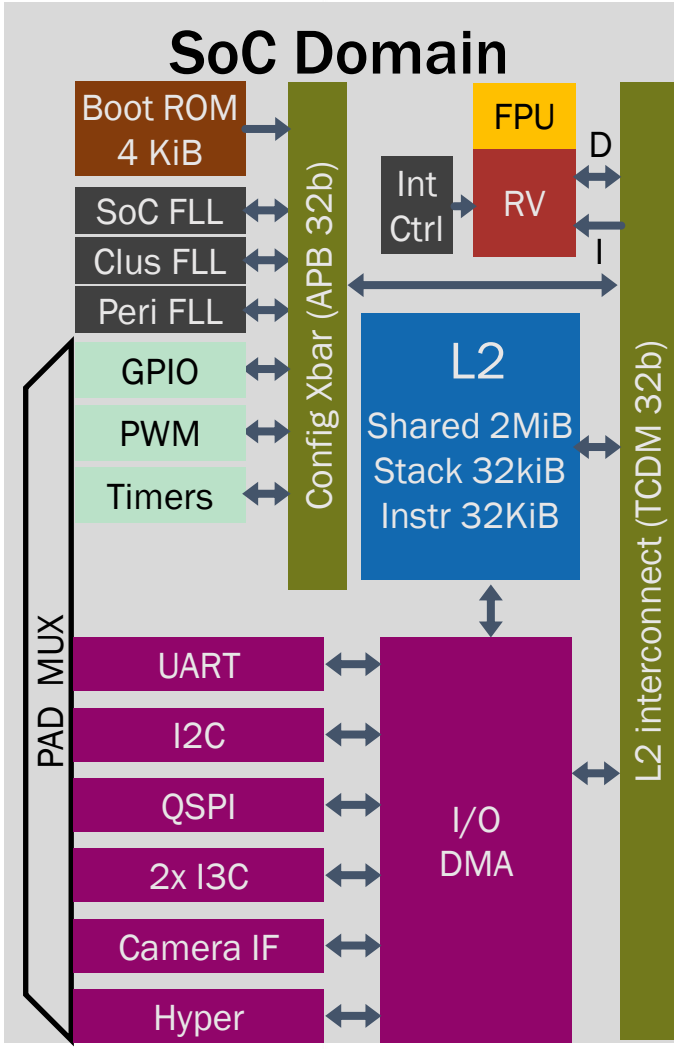
- **… hard constraints to meet on computation!**
  - a starting point: think of MobileNet-V2 like @ 500fps in 25mW
  - **~600 MOps → 300 GOPS → 12 TOPS/W**

Many technological hurdles to address as well!
E.g., see-through hi-res displays [0]

Outline

- Introduction & Motivation
- ArchiMEDES: Heterogeneous Cluster Template
  - NEureka DNN Accelerator
    - Scalability
    - Performance
  - HCI interconnect
    - Motivation
    - Architecture
- AR/VR case study
- State-of-the-Art Comparison
- Conclusion

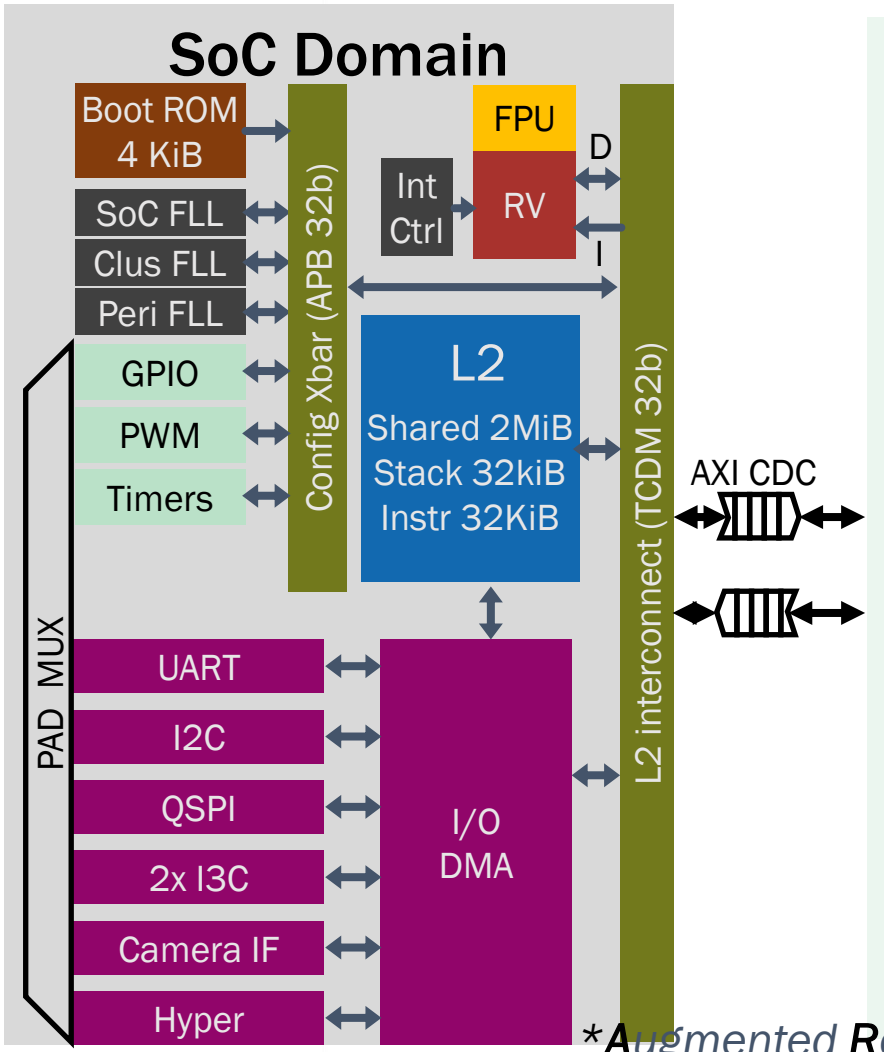
PULP(Parallel Ultra Low Power) SoC



How can we achieve necessary speedup?

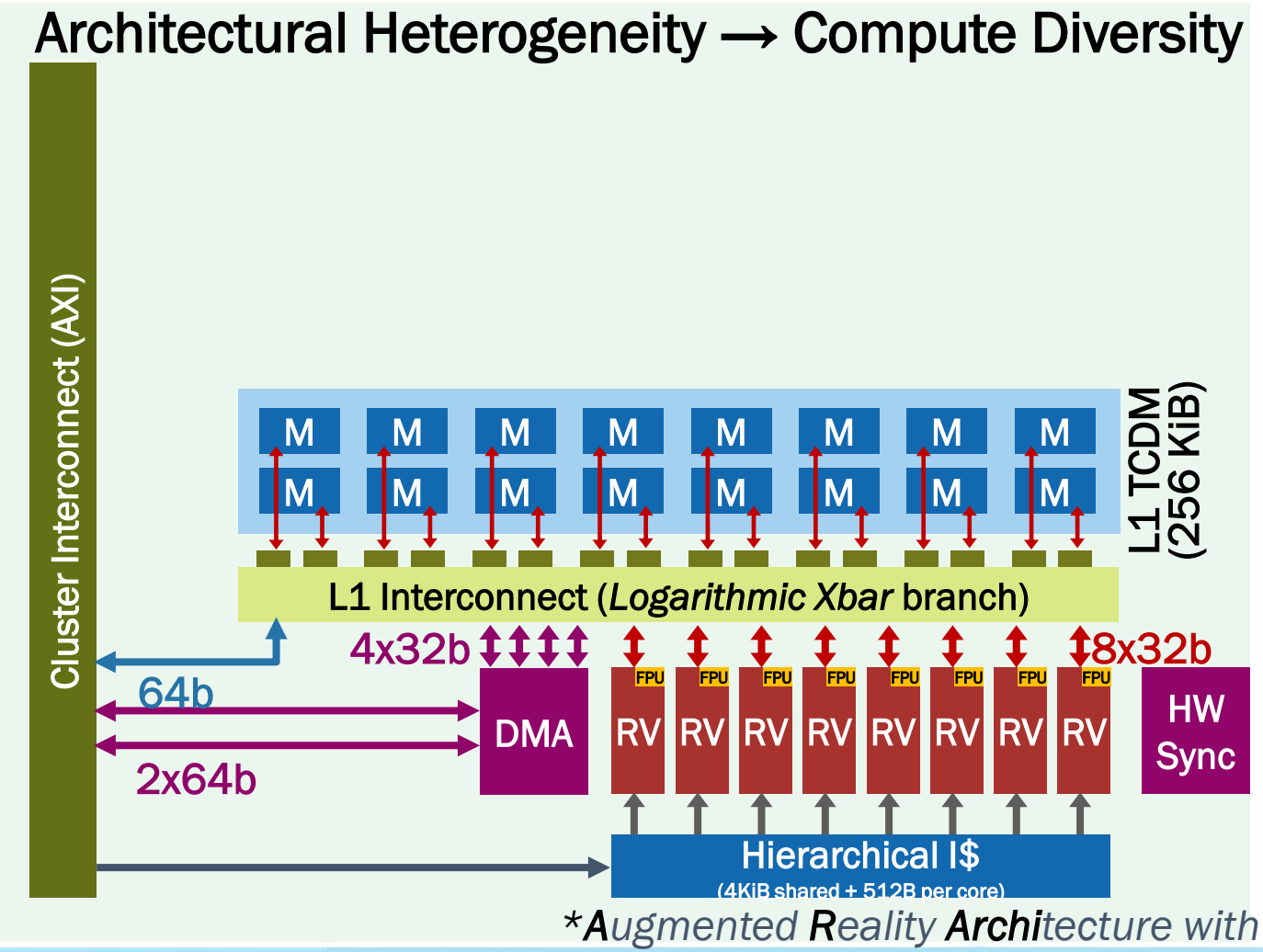
PULP approach is through SW cluster

PULP(Parallel Ultra Low Power) SoC



*Augmented Reality Architecture with Minimum Energy DNNs Embedded Specialization

ArchiMEDES* Cluster Template



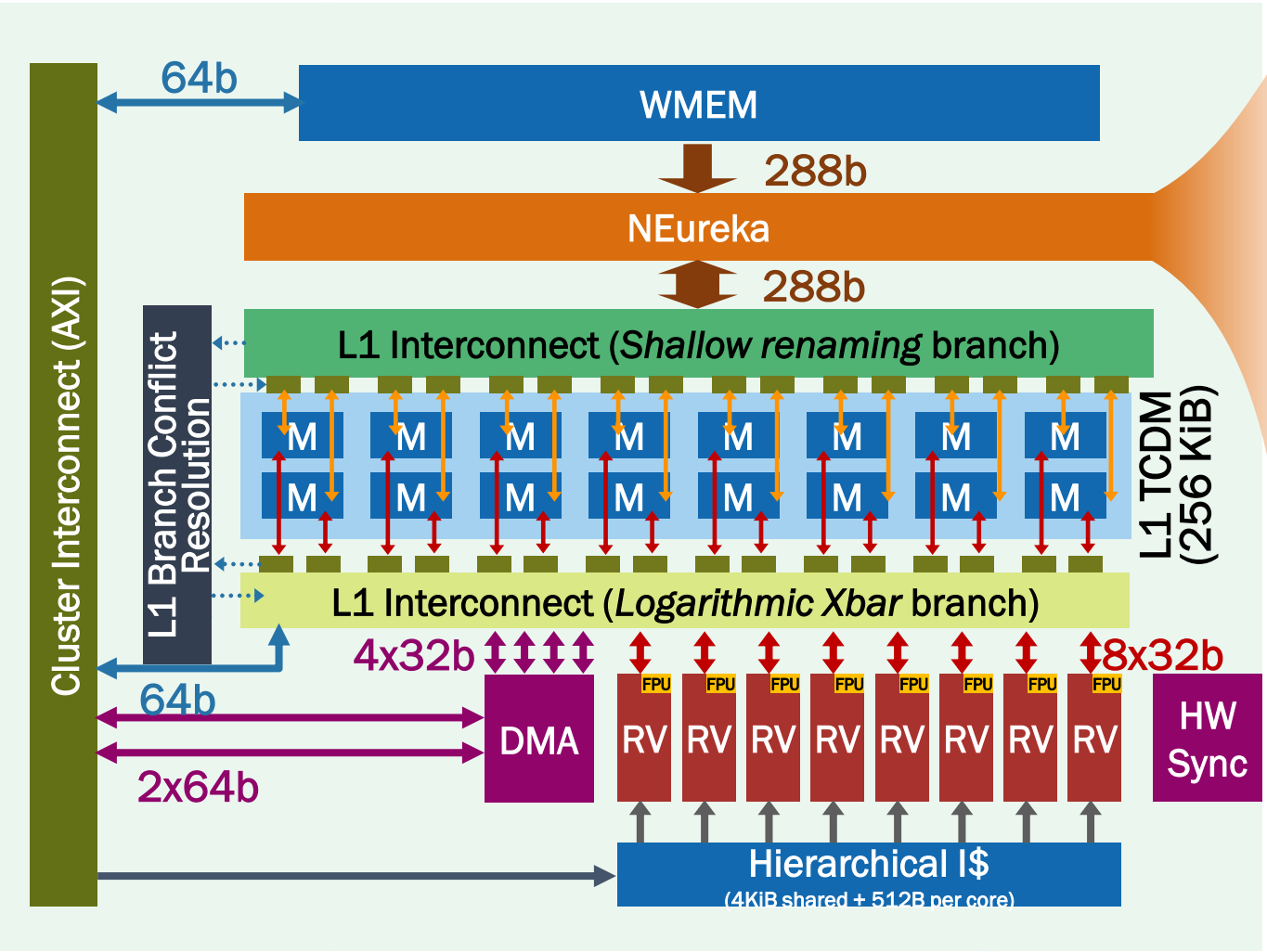
- PULP cluster with 8 **RV32IMCF** cores
- Private FPUs per core
- hierarchical instruction cache (4 KiB + 512B per core)
- **Xpulpv2** extensions [5] for integer mixed-precision DSP + DNNs
- 256 KiB of Tightly-Coupled Data Memory (TCDM) divided in 16 word-interleaved SRAM banks
- The Logarithmic Interconnect

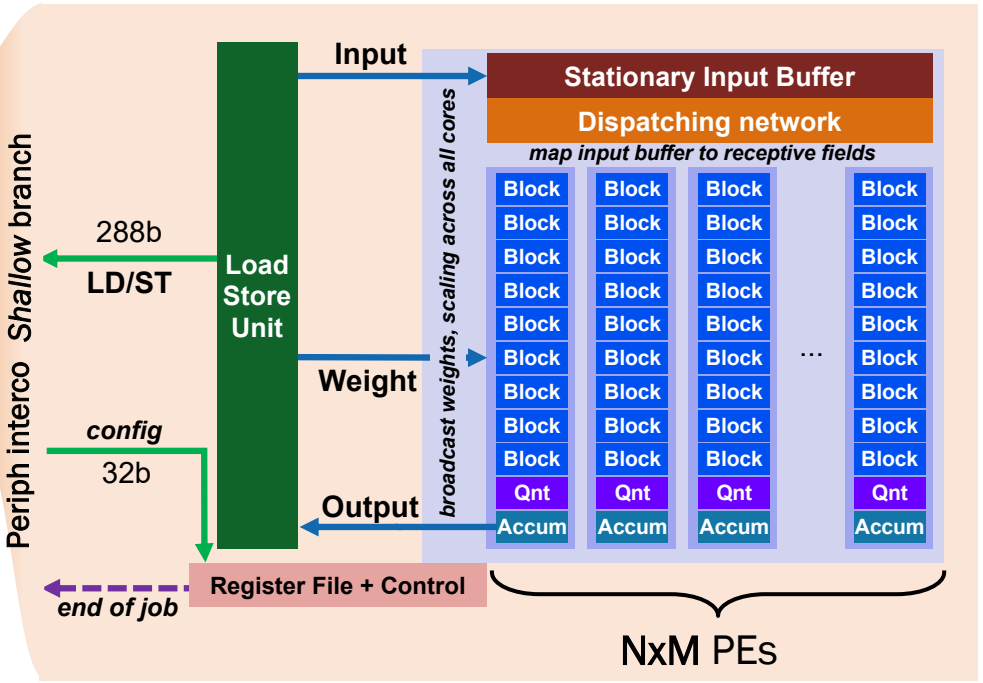
Augmented **Reality **Archi**tecture with **M**inimum **E**nergy **D**NNs **E**mbedded **S**pecialization*

Outline

- Introduction & Motivation
- ArchiMEDES: Heterogeneous Cluster Template
  - NEureka DNN Accelerator
    - Scalability
    - Performance
  - HCI interconnect
    - Motivation
    - Architecture
- AR/VR case study
- State-of-the-Art Comparison
- Conclusion

ArchiMEDES Cluster Template



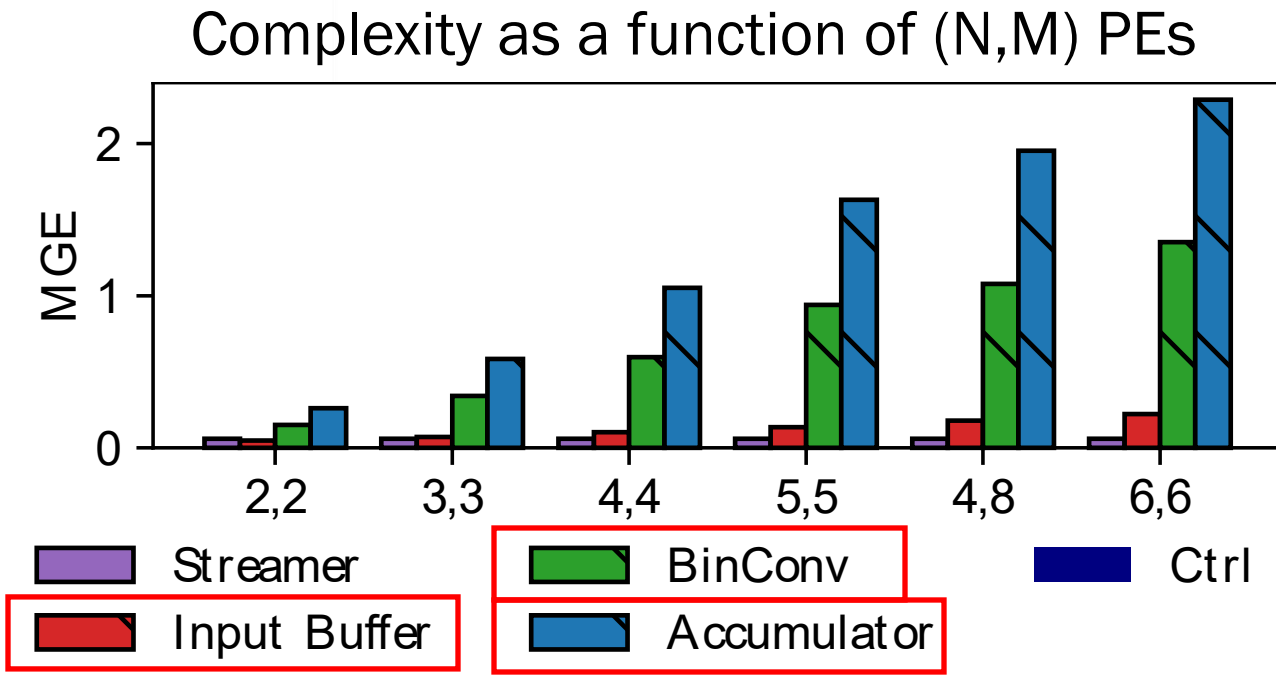


- 1 PE = receptive field of 1x1 px in output
- Output stationary, Input quasi-stationary
- Parametric number of Cores (NxM out-px)
- 8b activations, 2-8b weights

Outline

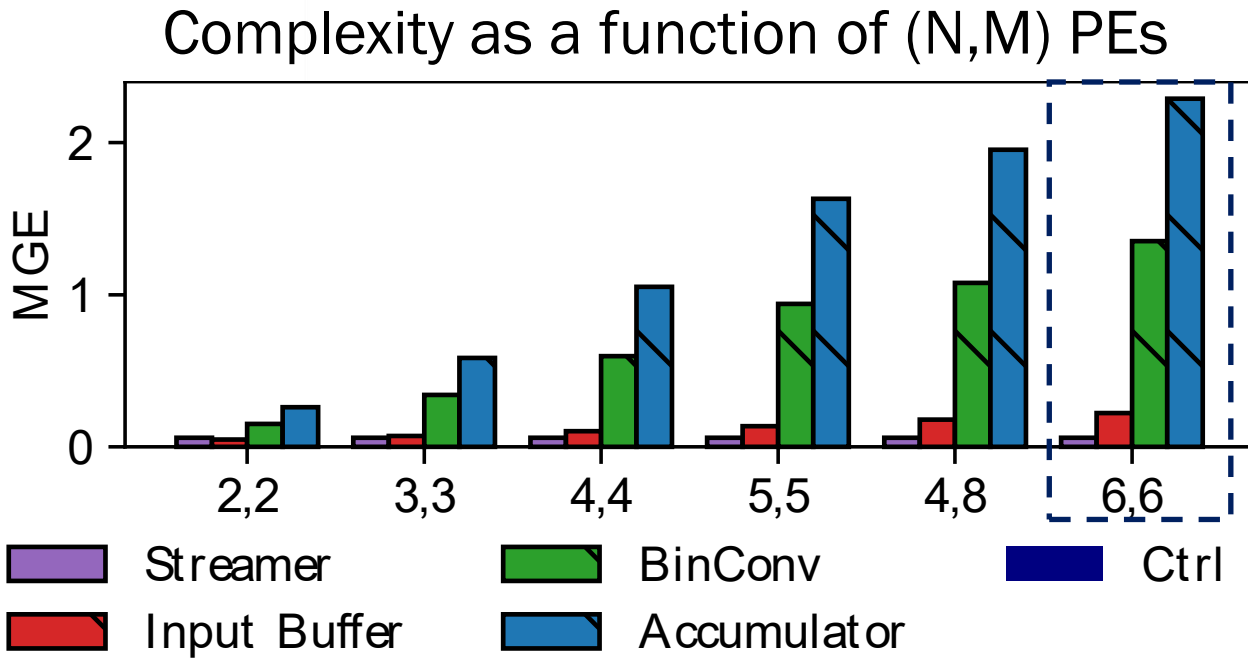
- Introduction & Motivation
- ArchiMEDES: Heterogeneous Cluster Template
  - NEureka DNN Accelerator
    - Scalability
    - Performance
  - HCI interconnect
    - Motivation
    - Architecture
- AR/VR case study
- State-of-the-Art Comparison
- Conclusion

NEureka – Performance Across Kernels



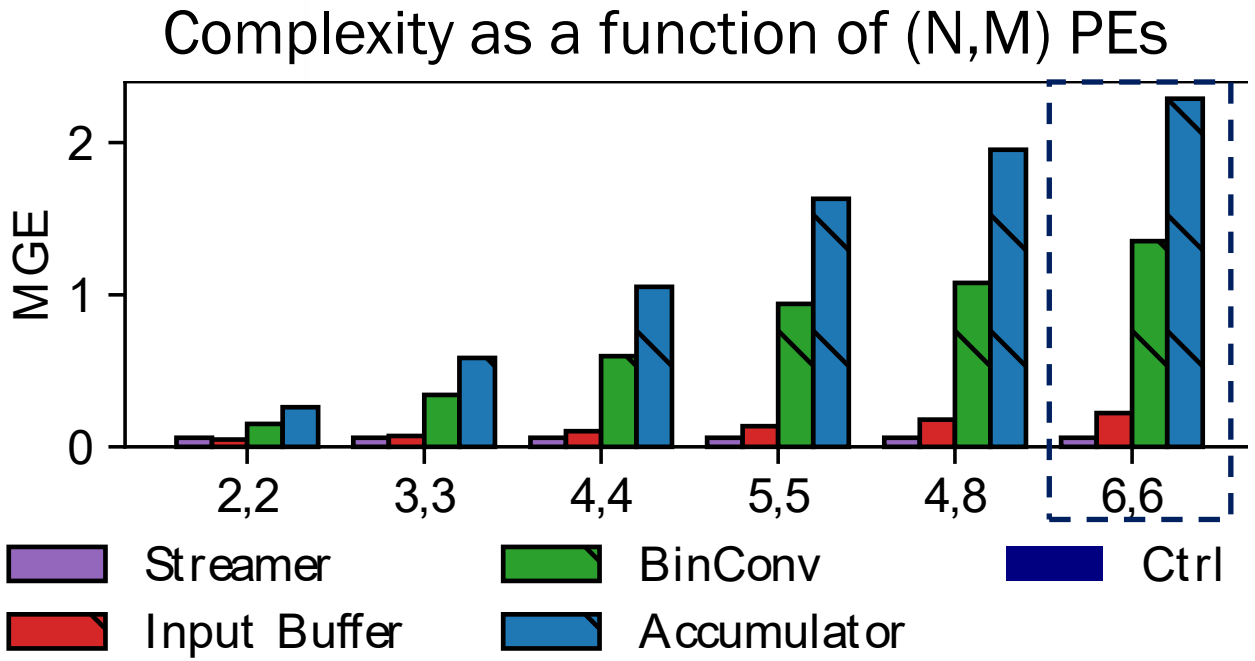
Complexity ∝ #PEs ⟶
Area increases linearly with #PEs

NEureka – Performance Across Kernels

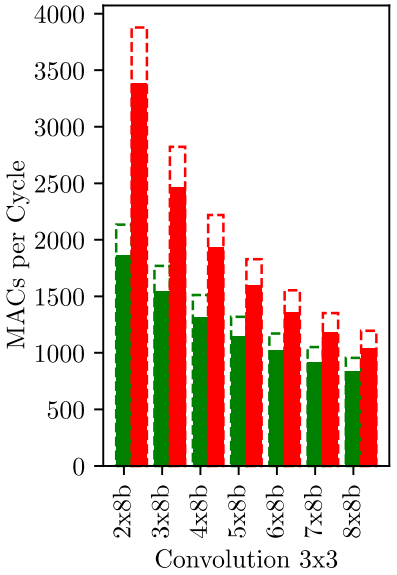


Complexity ∝ #PEs ⟶
Area increases linearly with #PEs

NEureka – Performance Across Kernels



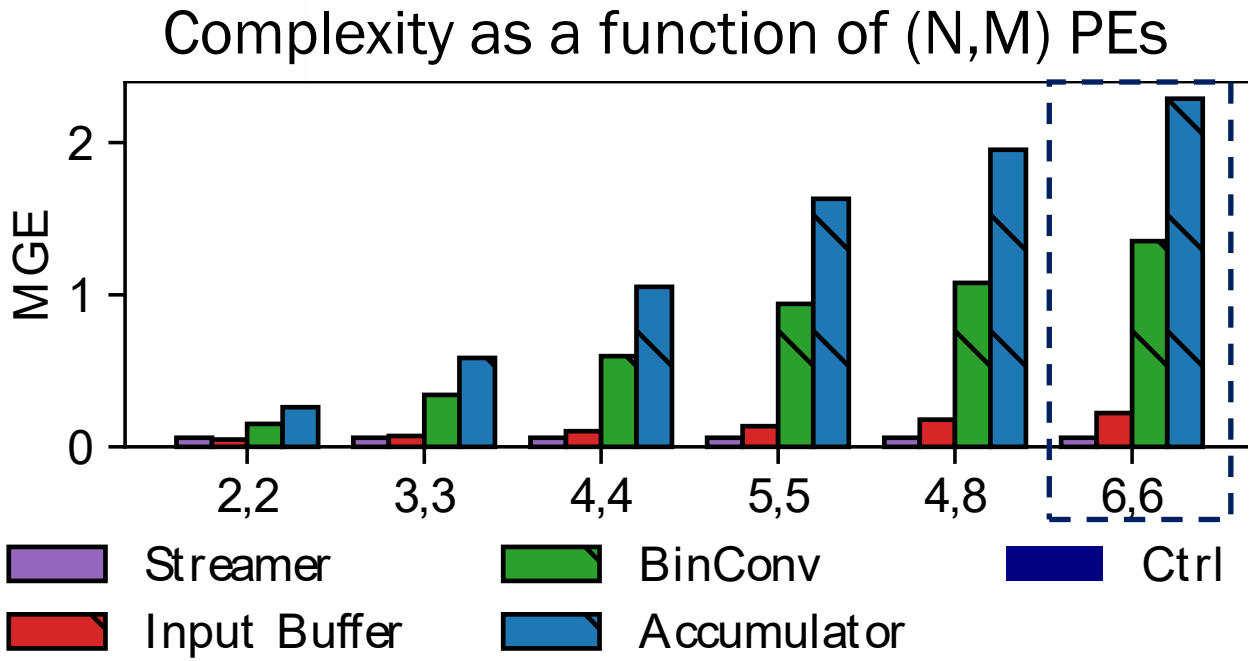
- **Bit-Serial** mode of operation
- Execution cycle ∝ weight-bit precision



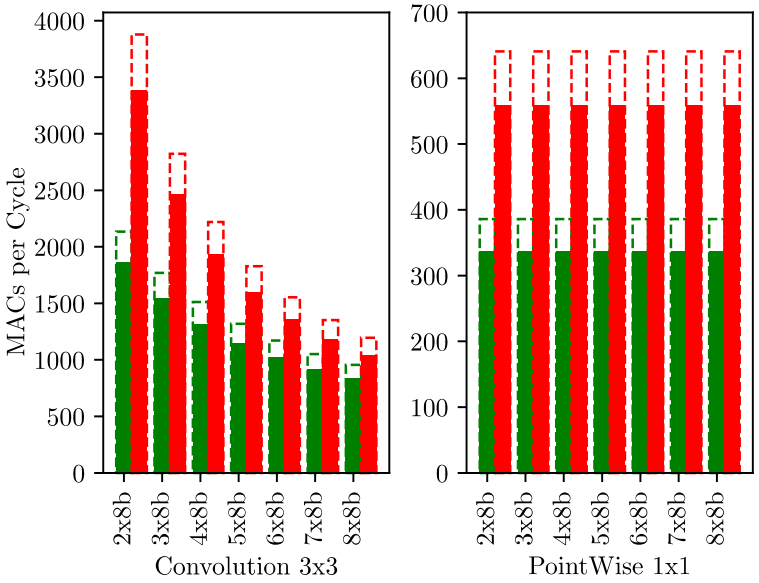
without prefetch | With WMEM | ideal

Complexity ∝ #PEs ⟶
Area increases linearly with #PEs

NEureka – Performance Across Kernels



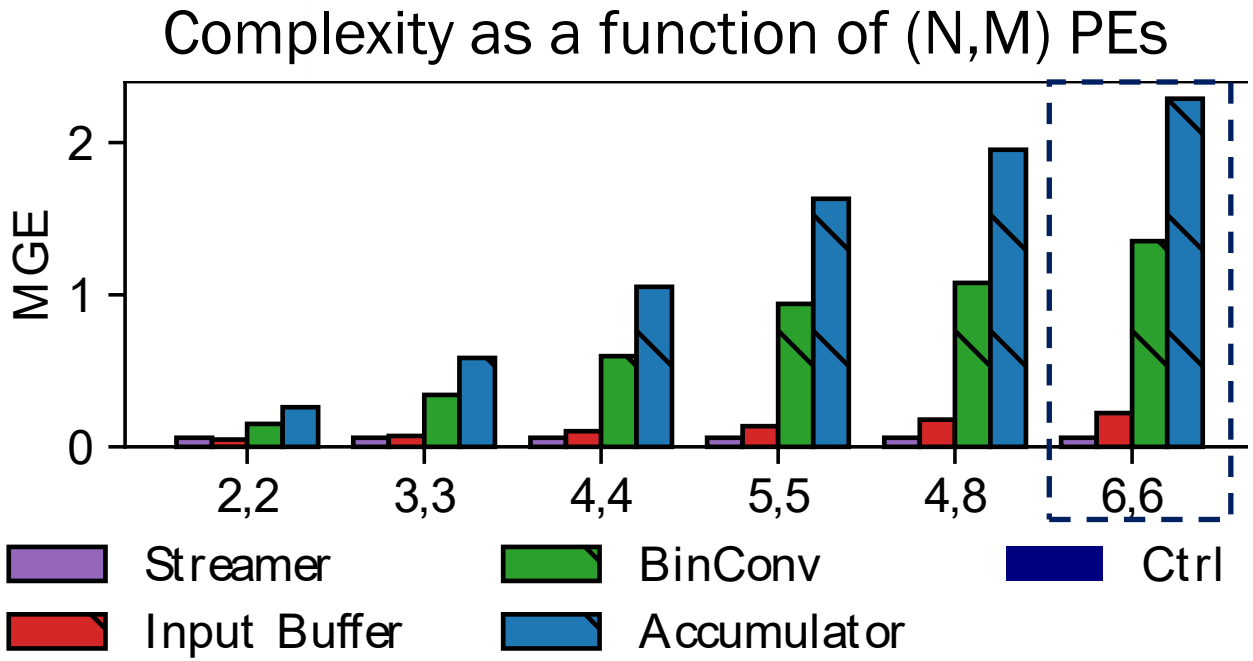
- Bit-Parallel mode of operation
- Constant throughput



without prefetch | With WMEM | ideal

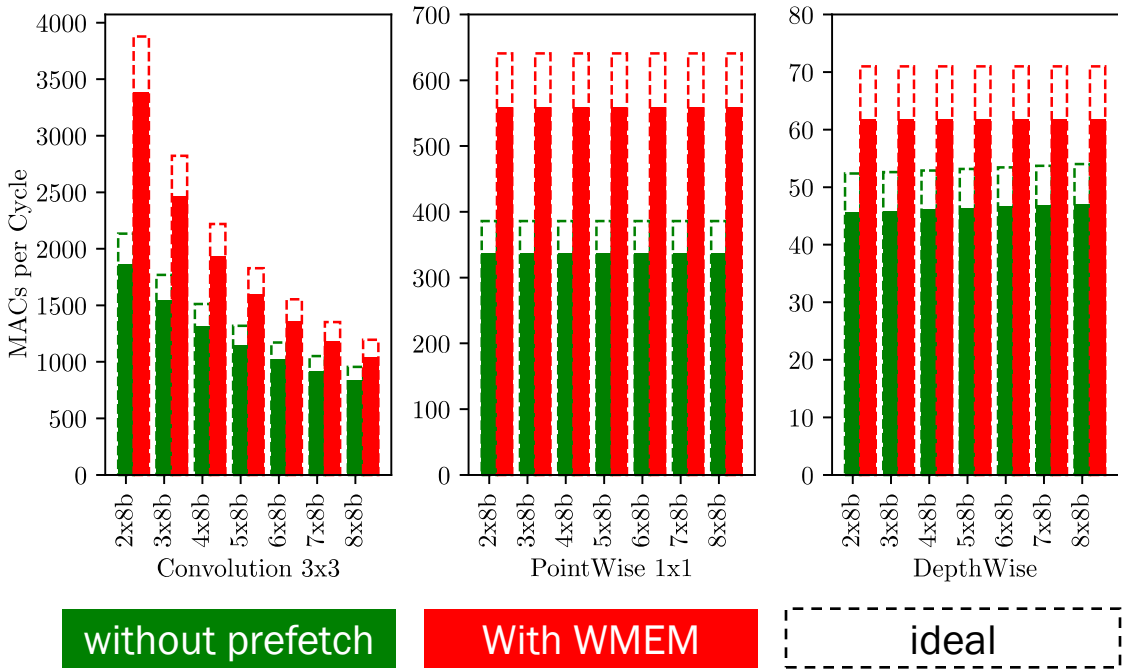
Complexity ∝ #PEs ⟶
Area increases linearly with #PEs

NEureka – Performance Across Kernels

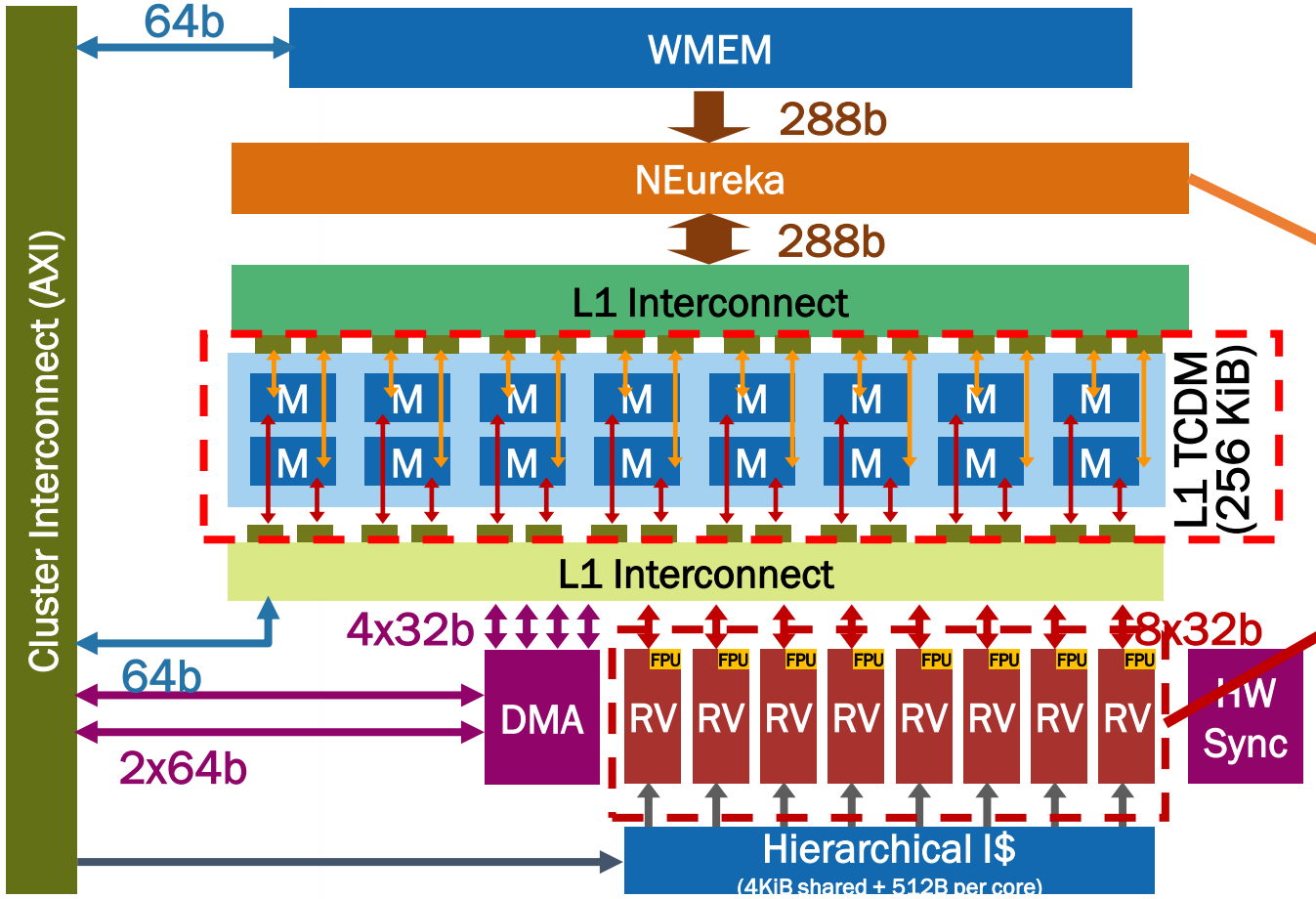


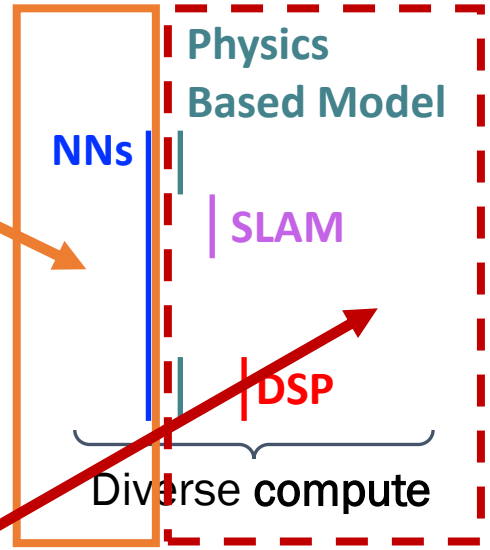
Complexity ∝ #PEs ⟶
Area increases linearly with #PEs

- Bit-Serial mode of operation
- low input reuse ⟶ small compute cycle



Summarizing Until Now



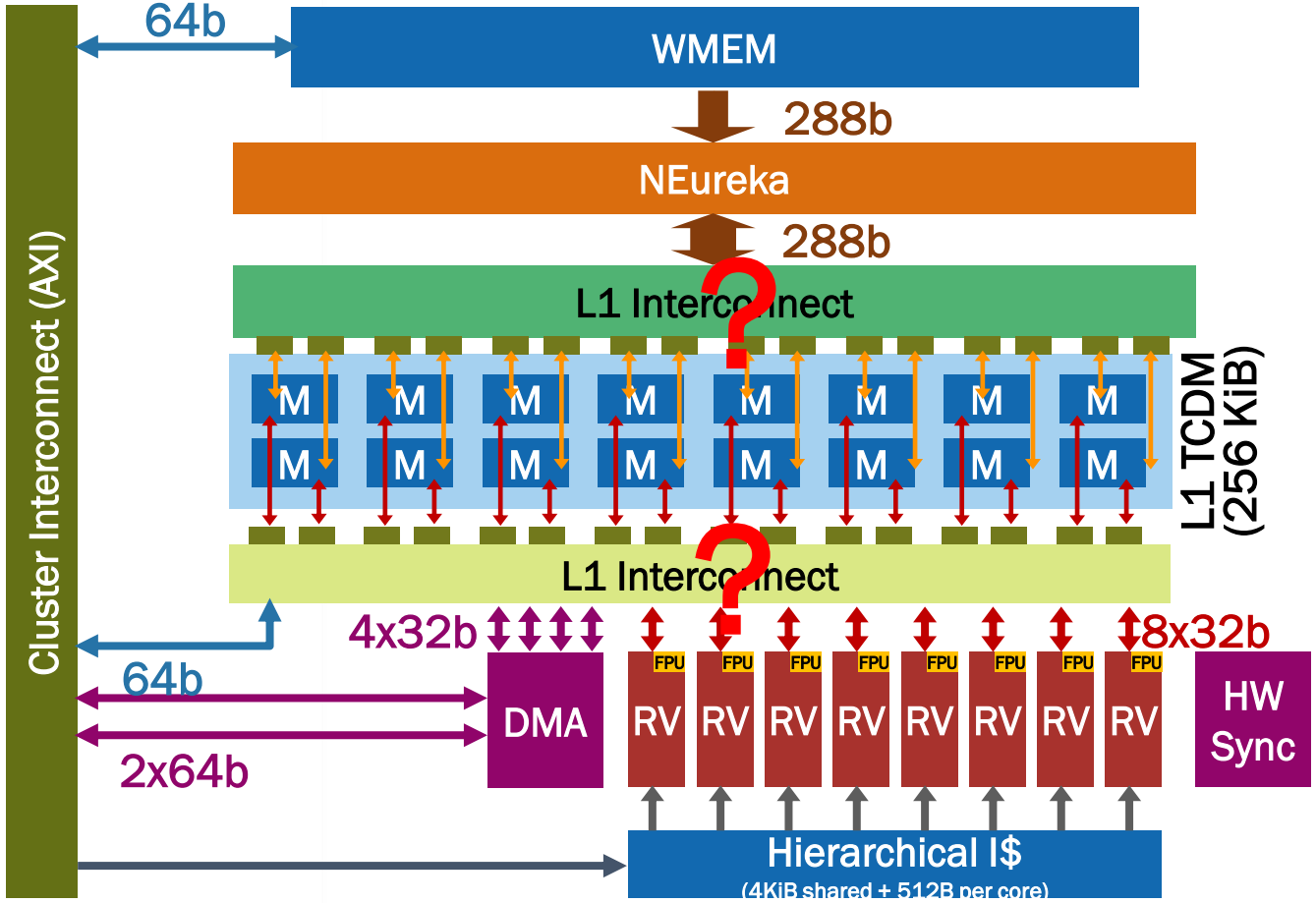


Is this enough?

Outline

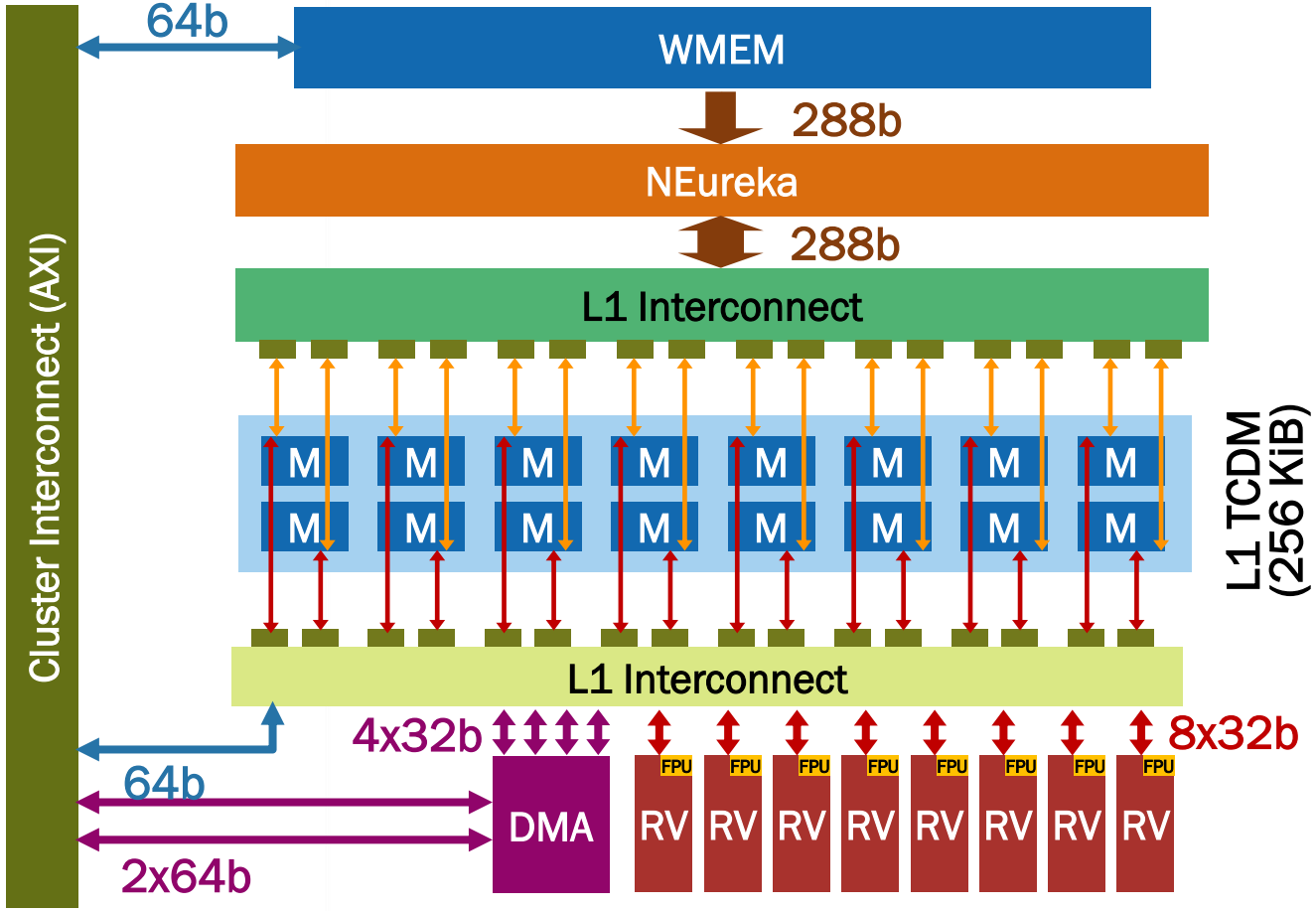
- Introduction & Motivation
- ArchiMEDES: Heterogeneous Cluster Template
  - NEureka DNN Accelerator
    - Scalability
    - Performance
  - HCI interconnect
    - Motivation
    - Architecture
- AR/VR case study
- State-of-the-Art Comparison
- Conclusion

How to efficiently access Shared TCDM?

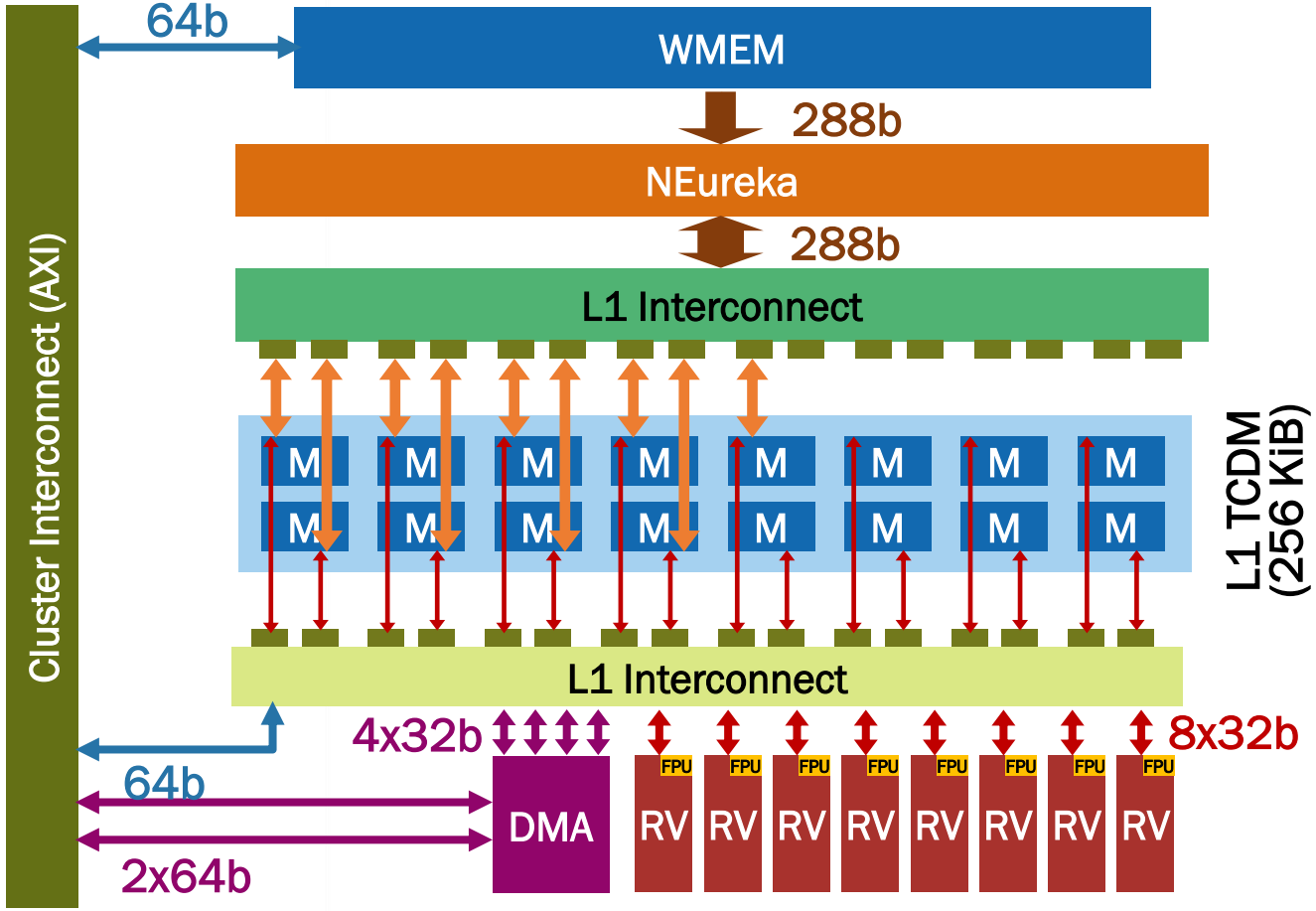


- Different bandwidths towards TCDM
  - Cores → 32-bit
  - NEureka → 288-bit
- Diverse Data Access Pattern
  - Cores → Algorithm(1 bank)
  - NEureka → Contiguous (9 bank)
- Sensitivity to stalls

How to efficiently access Shared TCDM?

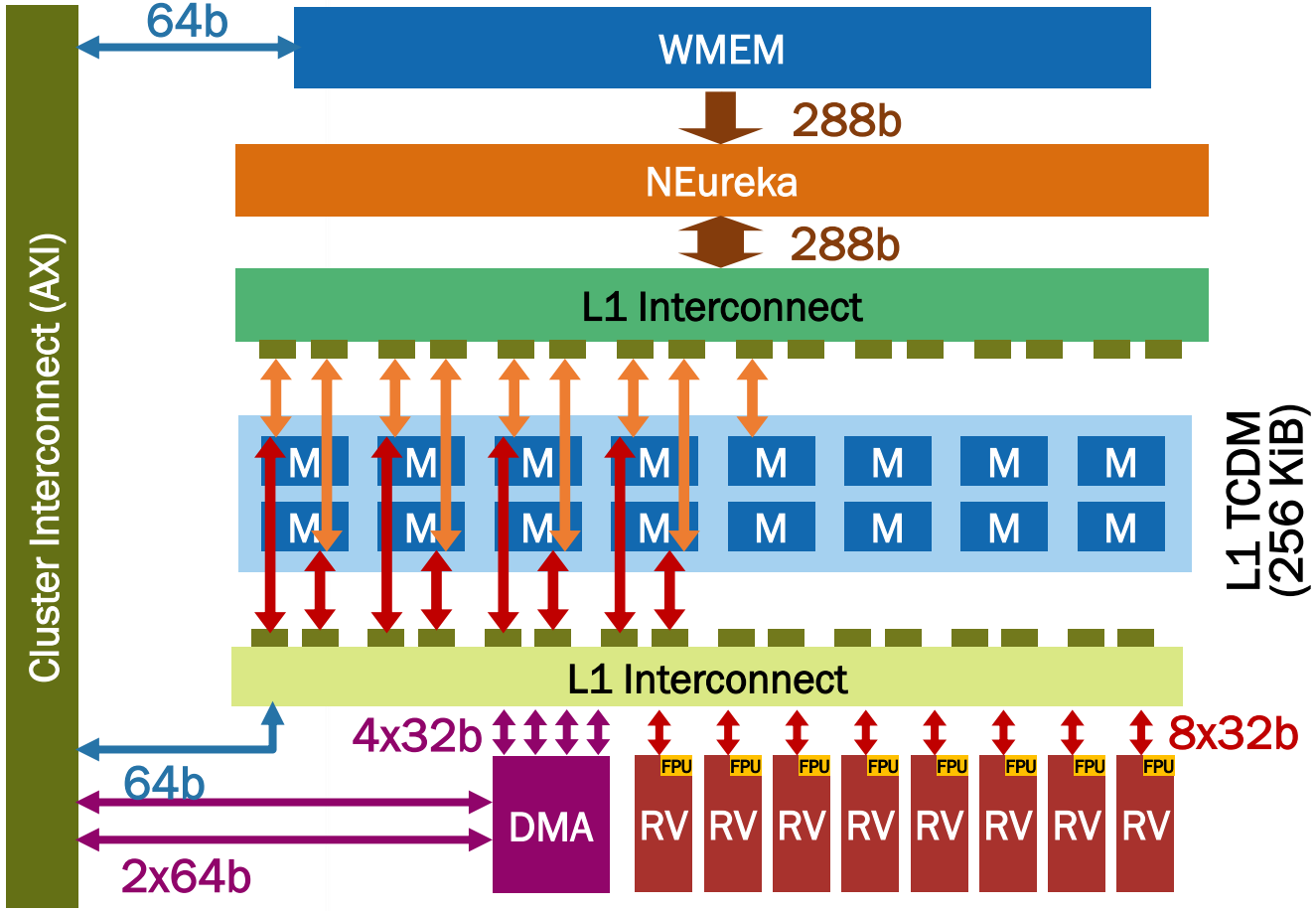


How to efficiently access Shared TCDM?



NEureka 288-bit access →9 banks

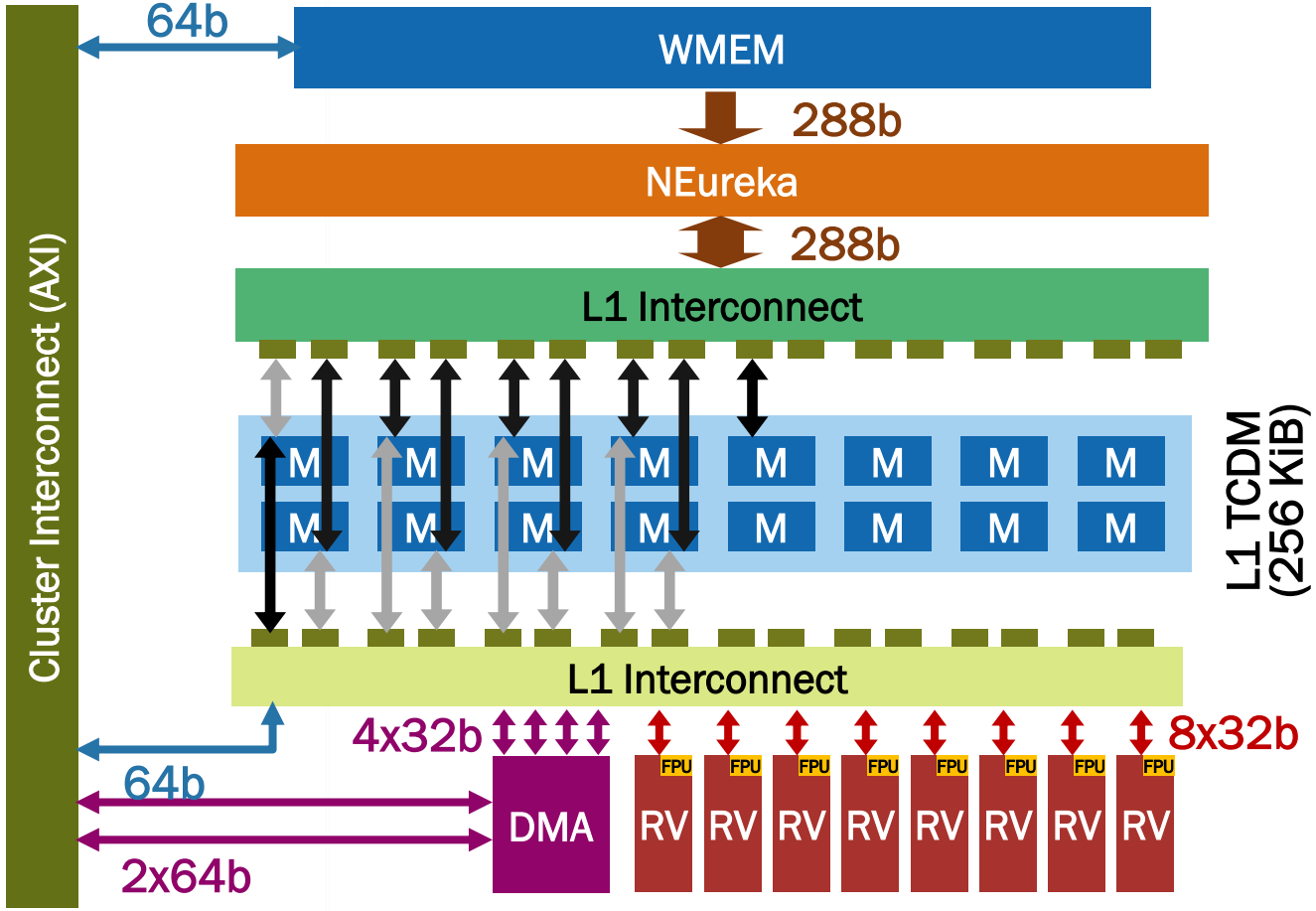
How to efficiently access Shared TCDM?



NEureka 288-bit access →9 banks

RV cores 8x32-bit access →8 banks

How to efficiently access Shared TCDM?

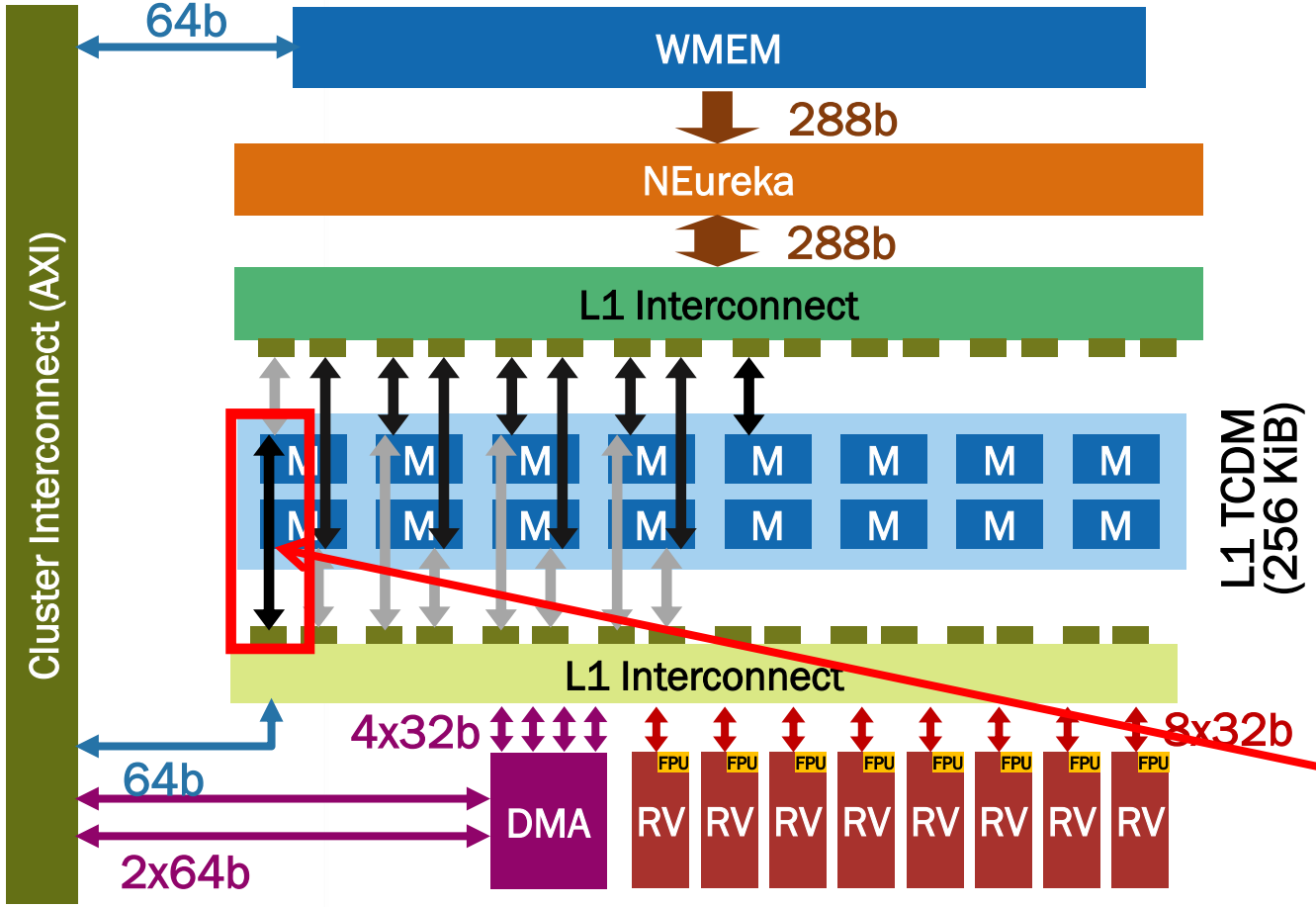


NEureka 288-bit access →9 banks

Arbitration

RV cores 8x32-bit access →8 banks

How to efficiently access Shared TCDM?



NEureka 288-bit access →9 banks

Win 8/9 banks

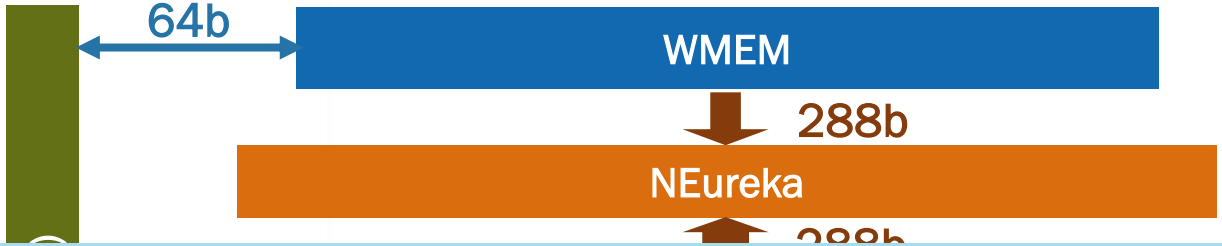
Arbitration

Win 1/8 cores

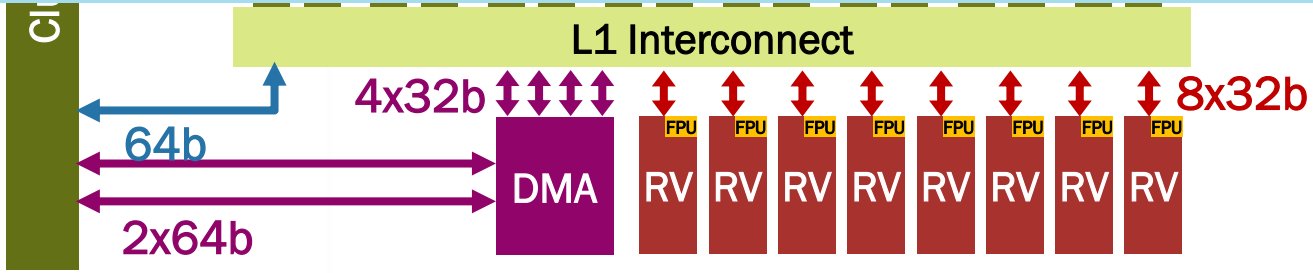
RV cores 8x32-bit access →8 banks

Only 32-bits useful data

How to efficiently access Shared TCDM?



Efficient data sharing is crucial

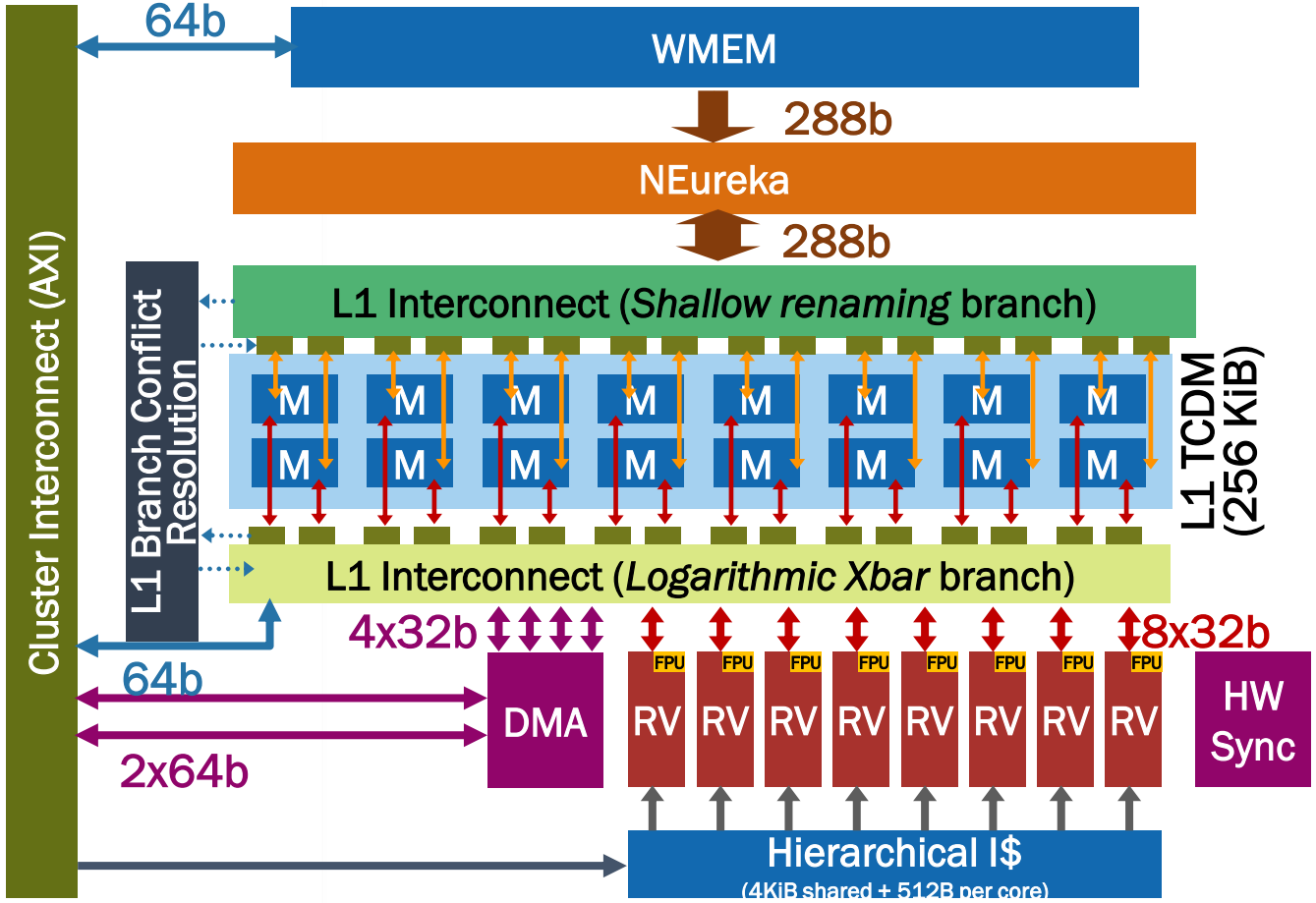


RV cores 8x32-bit access →8 banks

Outline

- Introduction & Motivation
- ArchiMEDES: Heterogeneous Cluster Template
  - NEureka DNN Accelerator
    - Scalability
    - Performance
  - HCI interconnect
    - Motivation
    - Architecture
- AR/VR case study
- State-of-the-Art Comparison
- Conclusion

How to efficiently access Shared TCDM?



Hierarchical heterogenous Interconnect

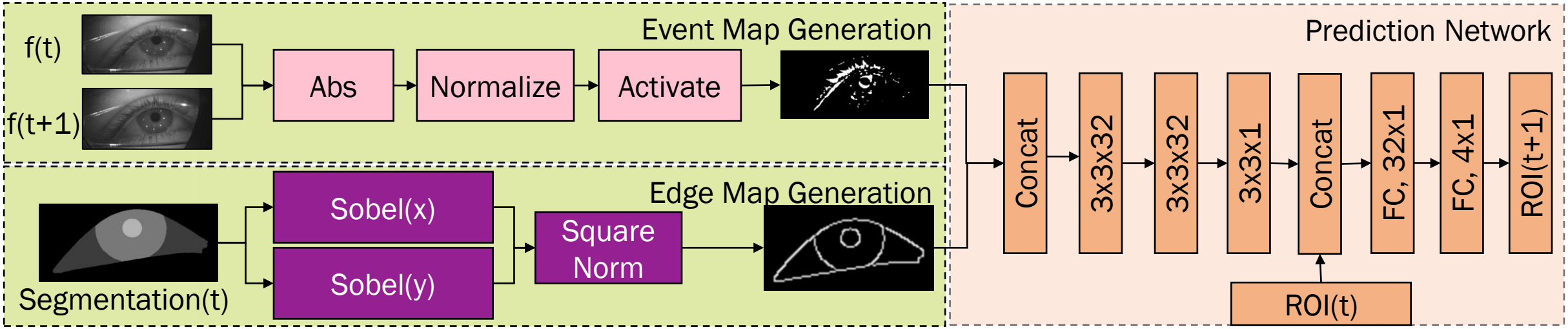
Dedicated Interconnect for NEureka

Stall Tuning knob →Conflict Resolution

Outline

- Introduction & Motivation
- ArchiMEDES: Heterogeneous Cluster Template
  - NEureka DNN Accelerator
    - Scalability
    - Performance
  - HCI interconnect
    - Motivation
    - Architecture
- AR/VR case study
- State-of-the-Art Comparison
- Conclusion

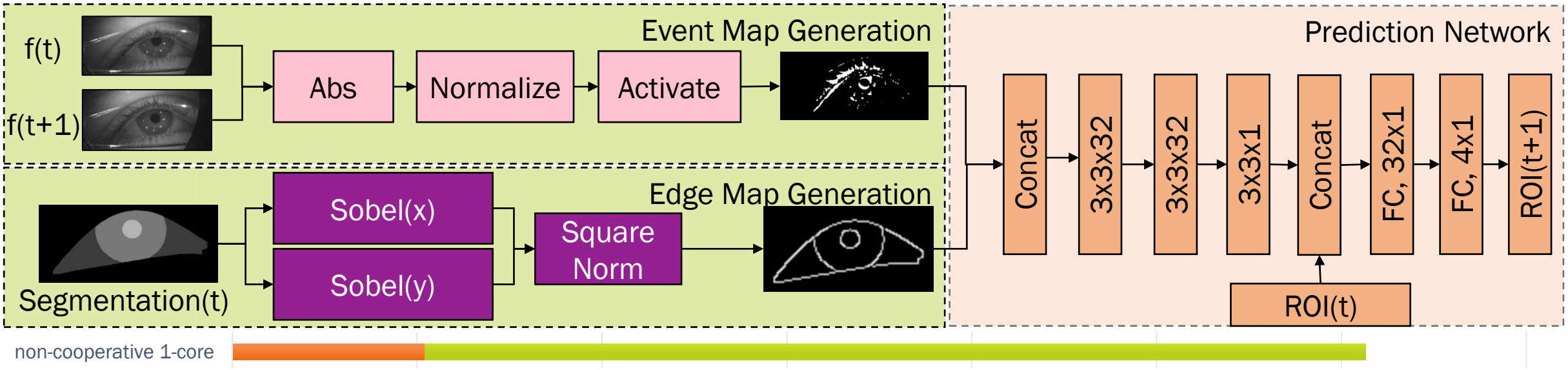
Eyetracking(ROI) execution example



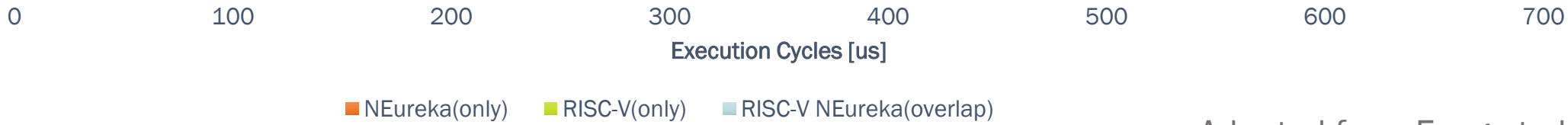
Eye tracking ⟶ gaze tracking ⟶ Only render the area of interest in high resolution

Adapted from Feng et al.[1]

Eyetracking(ROI) execution example

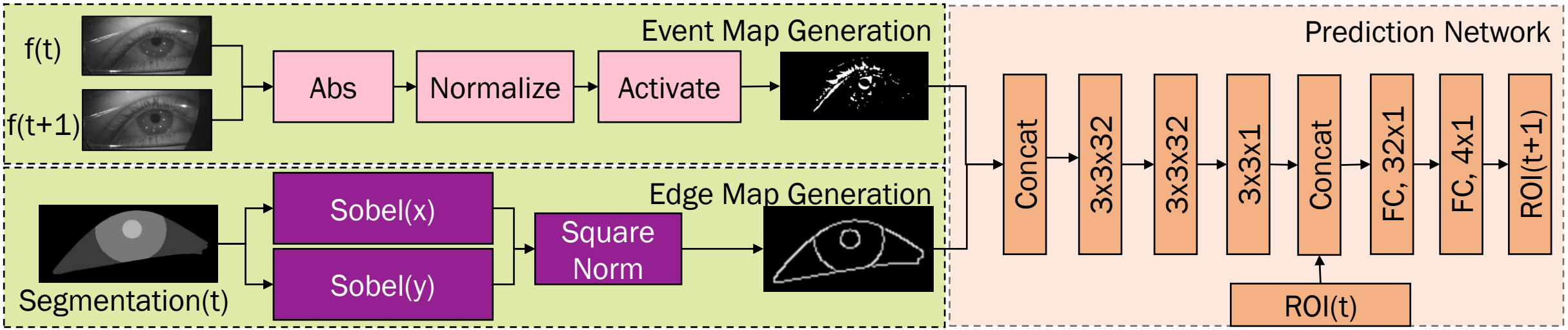


DNN Accelerator(NEureka) + Host Core (sequential execution) → non-cooperative one core

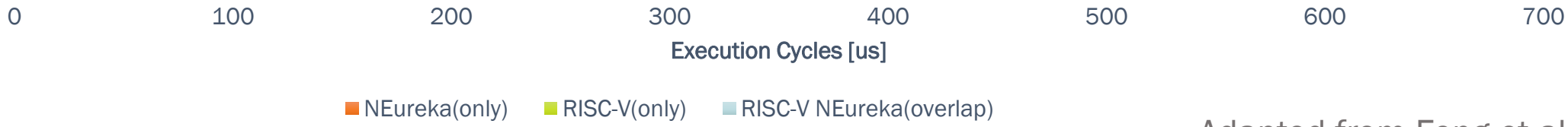


Adapted from Feng et al.[1]

Eyetracking(ROI) execution example

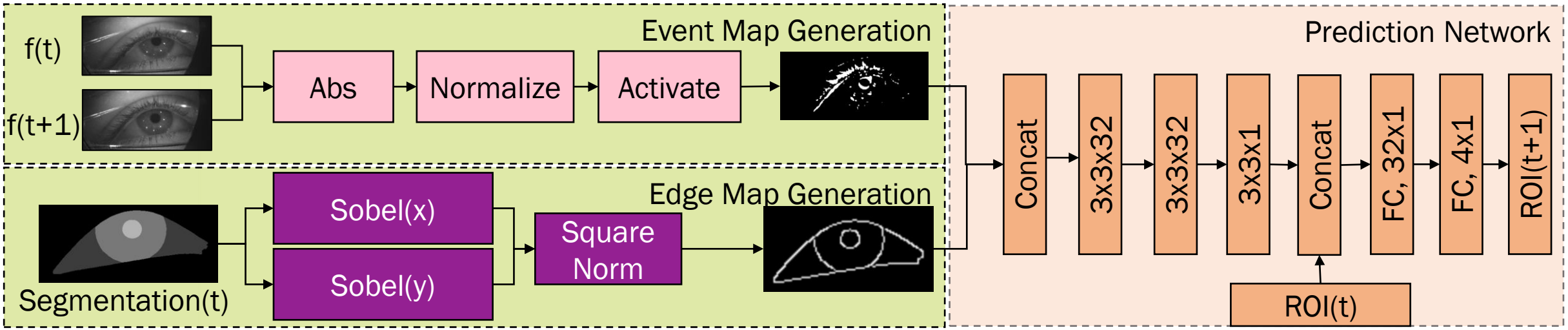


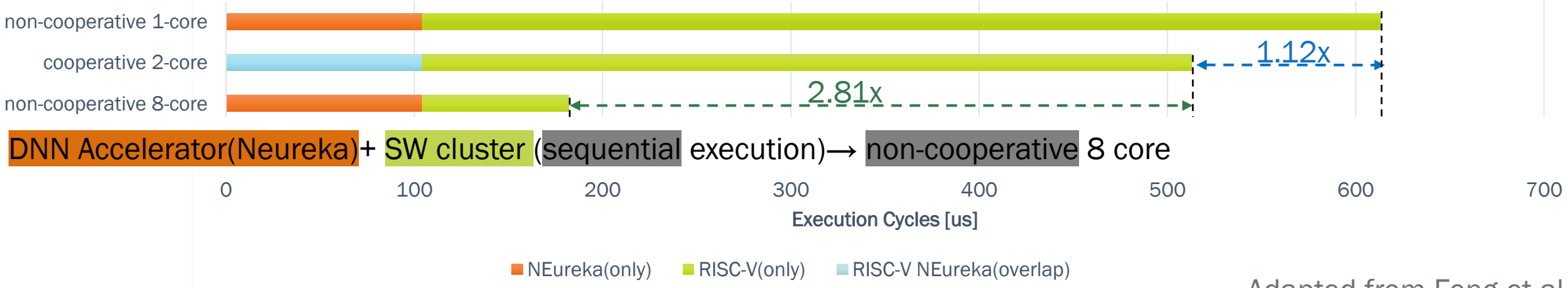
DNN Accelerator(NEureka) + Host Core (overlapped execution) → cooperative 2 core



Adapted from Feng et al.[1]

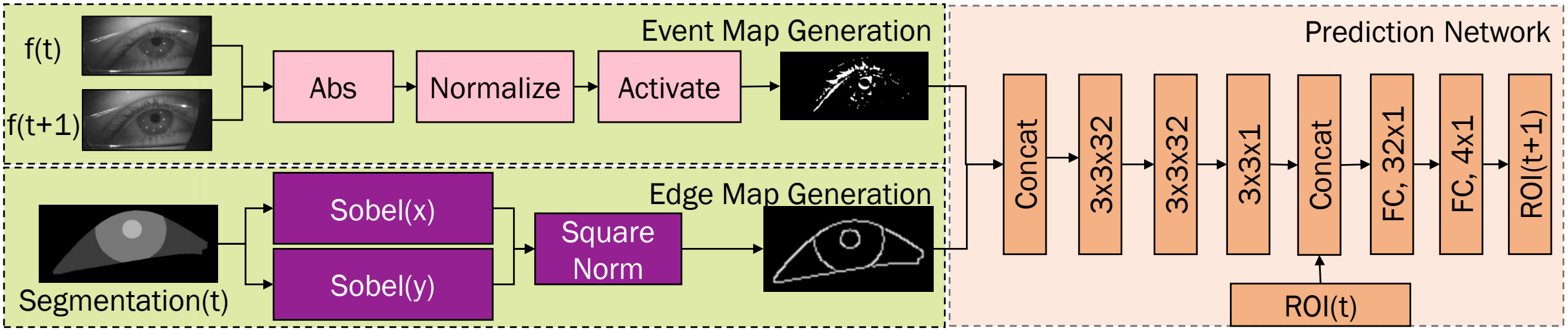
Eyetracking(ROI) execution example

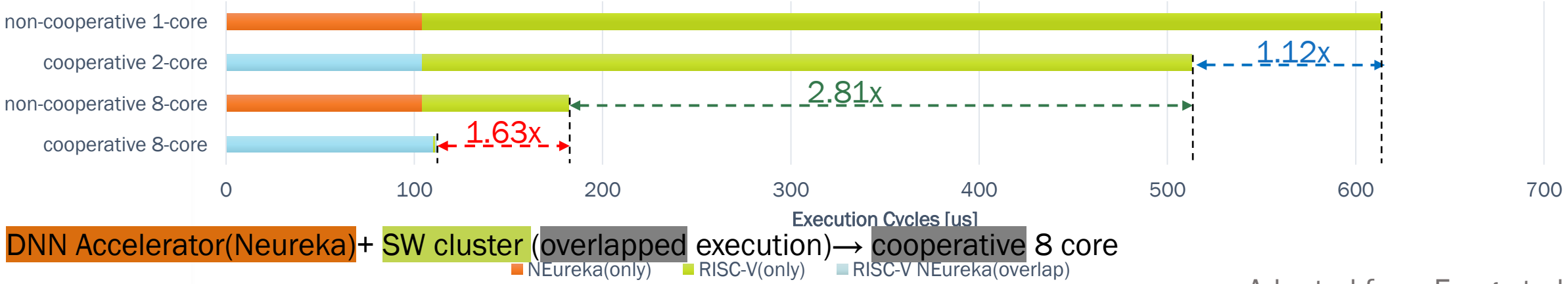




Adapted from Feng et al.[1]

Eyetracking(ROI) execution example



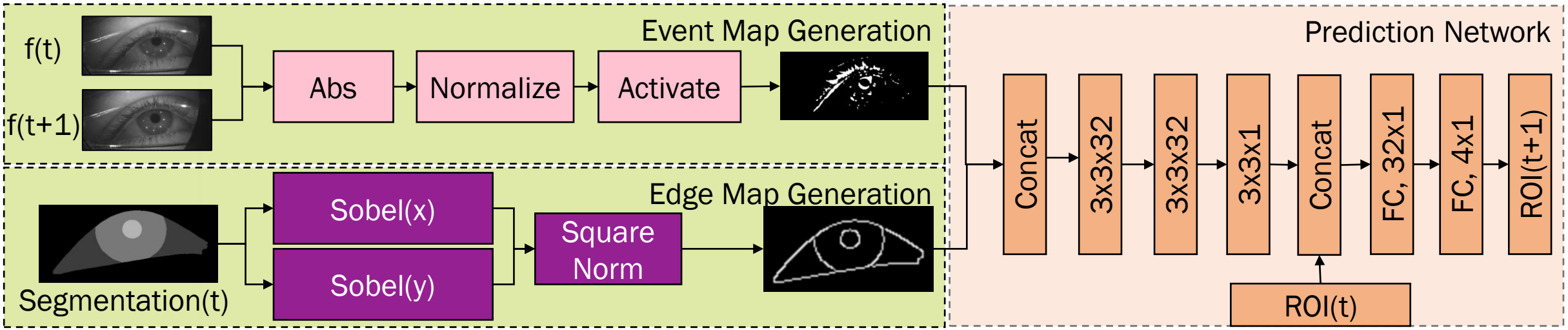


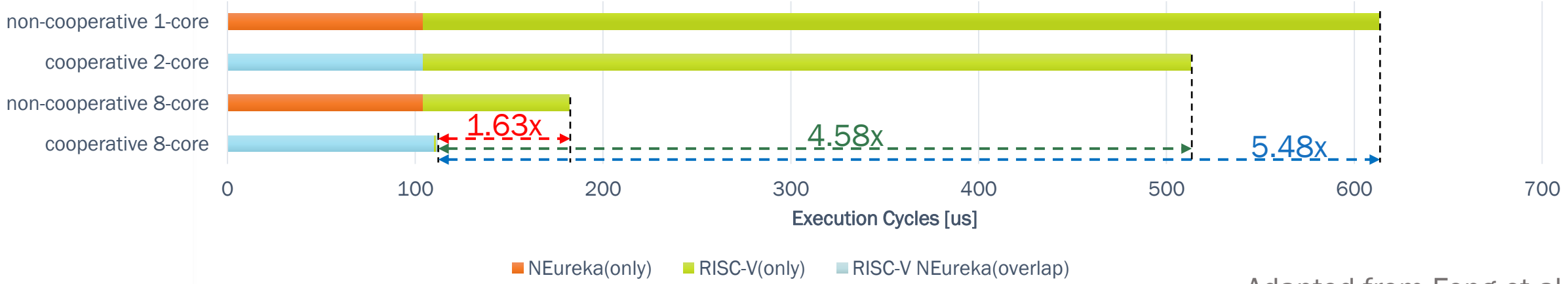
DNN Accelerator(Neureka)+ SW cluster (overlapped execution)→ cooperative 8 core

NEureka(only) RISC-V(only) RISC-V NEureka(overlap)

Adapted from Feng et al.[1]

Eyetracking(ROI) execution example





Adapted from Feng et al.[1]

Outline

- Introduction & Motivation
- ArchiMEDES: Heterogeneous Cluster Template
  - NEureka DNN Accelerator
    - Scalability
    - Performance
  - HCI interconnect
    - Motivation
    - Architecture
- AR/VR case study
- State-of-the-Art Comparison
- Conclusion

Comparision with SoA

Postlayout results, GF22FDx, SSG corner, 125MHz, 0.65V, 125C

Metrics	Sumbul et al[6] [2022]	Huang et al[7] [2022]	Yoo et al[8] [2020]	George et al[9] [2022]	ArchiMEDES
Frequency [MHz]	500	10-200	150	1000	270
Technology [nm]	7	28	65	7	22
Voltage [V]	0.75	–	0.78 – 1.2	0.67	0.65
Area [mm2]	2.56	8.28	16.0	0.20	3.38
Power [mW]	22.7	6.5 – 415.9	35 – 273	10	112
General Purpose Cores	1RV	1RV	–	1	8RV
Performance [GOPS]	146.5	410	737.5	256	1198
Efficiency [TOPS/W]	4.98	3.1	5.62	25.6	10.6

Comparision with SoA

Postlayout results, GF22FDx, SSG corner, 125MHz, 0.65V, 125C

Metrics	Sumbul et al[6] [2022]	Huang et al[7] [2022]	Im et al[8] [2020]	George et al[9] [2022]	ArchiMEDES
Frequency [MHz]	500	10-200	150		
Technology [nm]	7	28	65		
Voltage [V]	0.75	–	0.78 – 1.2	0.6	0.65
Area [mm2]	2.56	8.28	16.0	0.2(	3.38
Power [mW]	22.7	6.5 – 415.9	35 – 273	10	112
General Purpose Cores	1RV	1RV	–	1	8RV
Performance [GOPS]	146.5	410	737.5	256	1198
Efficiency [TOPS/W]	4.98	3.1	5.62	25.6	10.6

General Purpose compute Acceleration

Comparision with SoA

Postlayout results, GF22FDx, SSG corner, 125MHz, 0.65V, 125C

Metrics	Sumbul et al[6] [2022]	Huang et al[7] [2022]	Im et al[8] [2020]	George et al[9] [2022]	ArchiMEDES
Frequency [MHz]	500	10-200	150		
Technology [nm]	7				
Voltage [V]	0.75			0.6	0.65
Area [mm2]	2.56	8.28	16.0	0.2	3.38
Power [mW]	22.7	6.5 – 415.	35 – 273	10	112
General Purpose Cores	1RV	1RV	–	1	8RV
Performance [GOPS]	146.5	410	737.5	256	1198
Efficiency [TOPS/W]	4.98	3.1	5.62	25.6	10.6

Highest Throughput

General Purpose compute Acceleration

Comparision with SoA

Postlayout results, GF22FDx, SSG corner, 125MHz, 0.65V, 125C

Metrics	Sumbul et al[6] [2022]	Huang et al[7] [2022]	Im et al[8] [2020]	George et al[9] [2022]	ArchiMEDES
Frequency [MHz]	500	10-200	150		
Technology [nm]	7				
Voltage [V]				0.6	0.65
Area [mm2]		.28	16.0	0.2(	3.38
Power [mW]		– 415.5	35 – 273	10	112
General Purpose Cores	1RV	1RV	–	1	8RV
Performance [GOPS]	146.5	410	737.5	256	1198
Efficiency [TOPS/W]	4.98	3.1	5.62	25.6	10.6

Highest Efficiency except [9]
[9] only core power

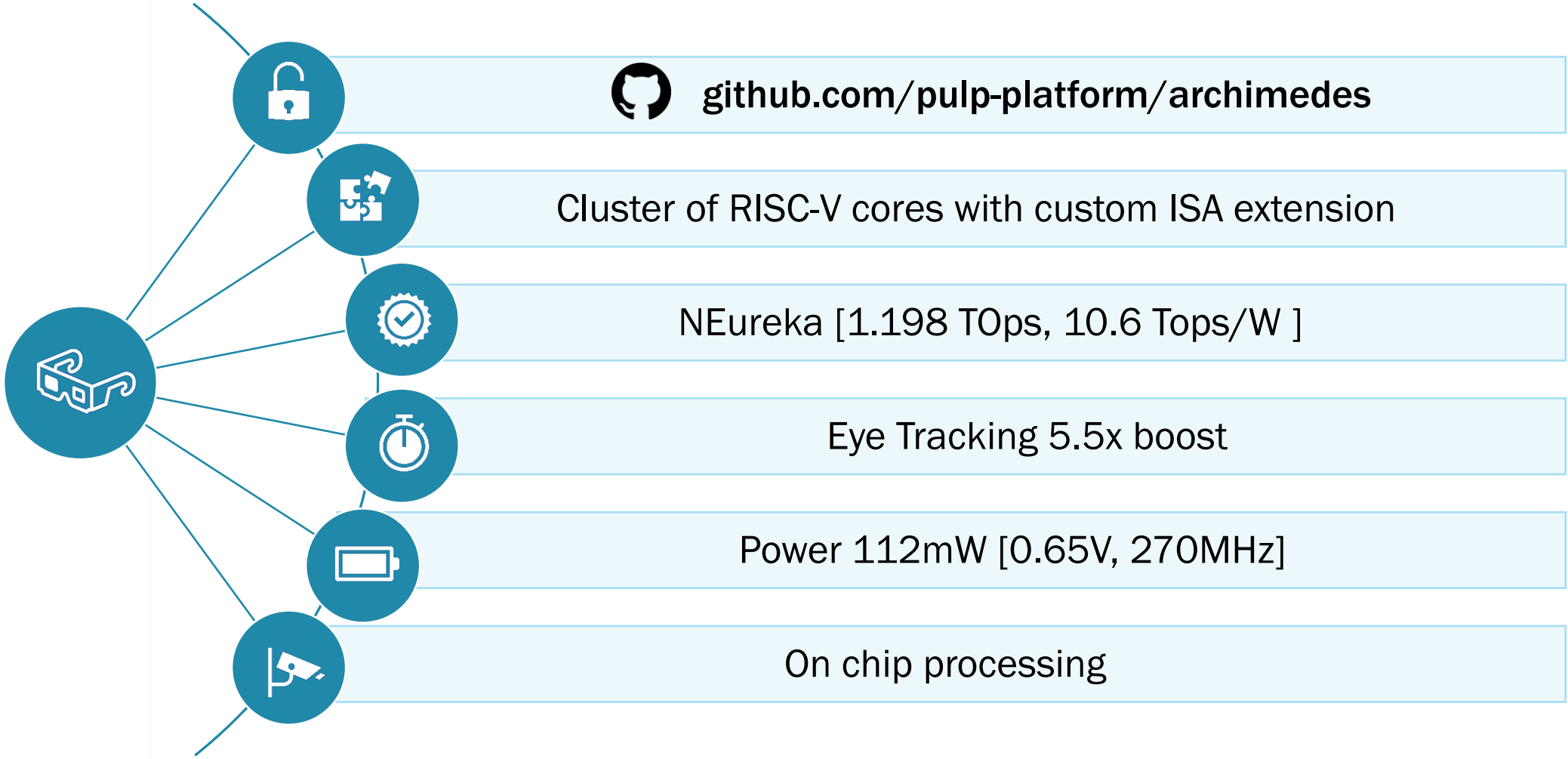
Highest Throughput

General Purpose compute Acceleration

Outline

- Introduction & Motivation
- ArchiMEDES: Heterogeneous Cluster Template
  - NEureka DNN Accelerator
    - Scalability
    - Performance
  - HCI interconnect
    - Motivation
    - Architecture
- AR/VR case study
- State-of-the-Art Comparison
- Conclusion

Conclusion



github.com/pulp-platform/archimedes

Cluster of RISC-V cores with custom ISA extension

NEureka [1.198 TOps, 10.6 Tops/W]

Eye Tracking 5.5x boost

Power 112mW [0.65V, 270MHz]

On chip processing

We would like to thank Meta for their support for this research project

References

- [0] M. Abrash, “Creating the Future: Augmented Reality, the next Human-Machine Interface,” in *2021 IEEE International Electron Devices Meeting (IEDM)*, Dec. 2021, p. 1.2.1-1.2.11. doi: 10.1109/IEDM19574.2021.9720526.
- [5] Y. Feng, N. Goulding-Hotta, A. Khan, H. Reyserhove, and Y. Zhu, “Real-Time Gaze Tracking with Event-Driven Eye Segmentation,” in *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, Mar. 2022, pp. 399–408. doi: 10.1109/VR51125.2022.00059.
- [2] S. Huang *et al.*, “A new head pose tracking method based on stereo visual SLAM,” *Journal of Visual Communication and Image Representation*, vol. 82, p. 103402, Jan. 2022, doi: 10.1016/j.jvcir.2021.103402.
- [3] F. Zhang *et al.*, “MediaPipe Hands: On-device Real-time Hand Tracking.” arXiv, Jun. 17, 2020. doi: 10.48550/arXiv.2006.10214.
- [4] A. Li, W. Liu, C. Zheng, and X. Li, “Embedding and Beamforming: All-Neural Causal Beamformer for Multichannel Speech Enhancement,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 6487–6491. doi: 10.1109/ICASSP43922.2022.9746432.
- [5] A. Pullini, D. Rossi, I. Loi, G. Tagliavini and L. Benini, "Mr.Wolf: An Energy-Precision Scalable Parallel Ultra Low Power SoC for IoT Edge Processing," in IEEE Journal of Solid-State Circuits, vol. 54, no. 7, pp. 1970-1981, July 2019, doi: 10.1109/JSSC.2019.2912307.

References

- [6] H. E. Sumbul et al., "System-Level Design and Integration of a Prototype AR/VR Hardware Featuring a Custom Low-Power DNN Accelerator Chip in 7nm Technology for Codec Avatars," 2022 IEEE Custom Integrated Circuits Conference (CICC), Newport Beach, CA, USA, 2022, pp. 01-08, doi: 10.1109/CICC53496.2022.9772810.
- [7] W. -C. Huang et al., "A 28-nm 25.1 TOPS/W Sparsity-Aware CNN-GCN Deep Learning SoC for Mobile Augmented Reality," 2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), Honolulu, HI, USA, 2022, pp. 42-43, doi: 10.1109/VLSITechnologyandCir46769.2022.9830261.
- [8] Im, Dongseok & Kang, Sanghoon & Han, Donghyeon & Choi, Sungpill & Yoo, Hoi-Jun. (2020). A 4.45 ms Low-Latency 3D Point-Cloud-Based Neural Network Processor for Hand Pose Estimation in Immersive Wearable Devices. 1-2. 10.1109/VLSICircuits18222.2020.9162895.
- [9] Biji George, Om Ji Omer, Ziaul Choudhury, Anoop V, and Sreenivas Subramoney. 2022. A Unified Programmable Edge Matrix Processor for Deep Neural Networks and Matrix Algebra. ACM Trans. Embed. Comput. Syst. 21, 5, Article 63 (September 2022), 30 pages. https://doi.org/10.1145/3524453