Siracusa – Towards On-Sensor Computing for Extended Reality Applications

Moritz Scherer[1], Manuel Eggimann[1], Alfio Di Mauro[1], Arpan Suravi Prasad[1], Francesco Conti[2], Davide Rossi[2], Jorge Gomez[3], Syed Shakib Sarwar[3], Zhao Wang[3], Barbara De Salvo[3] & Luca Benini[1,2]

[1]ETH Zürich, Switzerland
[2]Universita di Bologna, Italy
[3]Meta Reality Labs Research, USA
scheremo@iis.ee.ethz.ch

Smart glasses

Ray-Ban Stories

- Socially acceptable form factor
  - Like regular glasses
- Lightweight
  - < 50 grams

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

eXtended Reality Glasses

- Dedicated form factor
  - Cumbersome
  - Uncomfortable for multi-hour wear
- Heavyweight
  - ~500 grams
- 2-3 hours Battery
  - ~5 W power envelope

Meta Quest 2

Microsoft HoloLens 2

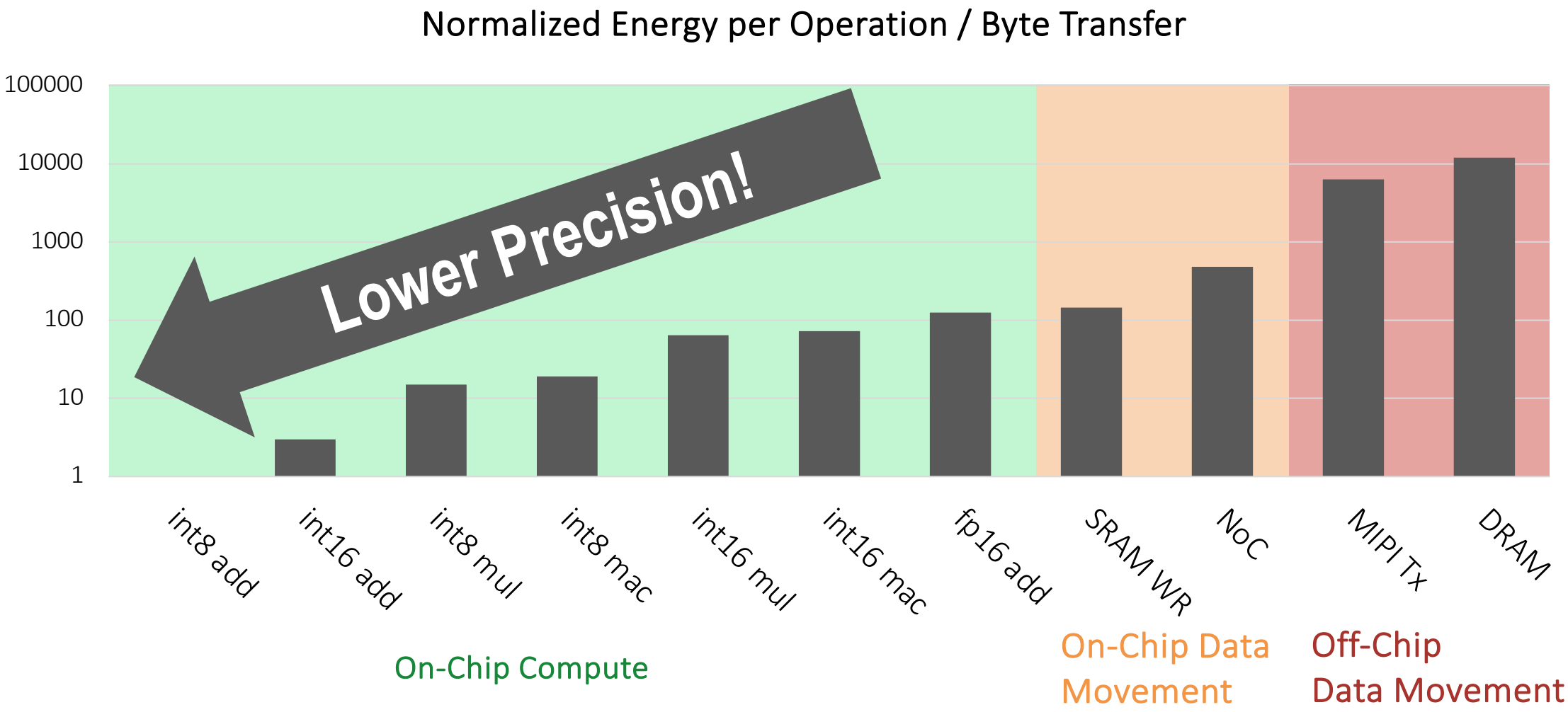
[https://en.wikichip.org/wiki/qualcomm/snapdragon_800/865] [https://arstechnica.com/]

Bringing XR to smart glasses

- Many tasks to manage...
  - Eye gaze tracking [1]
  - Head tracking [2]
  - Hand tracking [3]
  - ...
- ... hard constraints to meet on computation!
  - Real-time, multi-task operation
  - Tight power budget, 10s of mWs
  - High compute intensity NNs, 100s GOp/s

How do we address energy efficiency?

Save Energy where it counts!



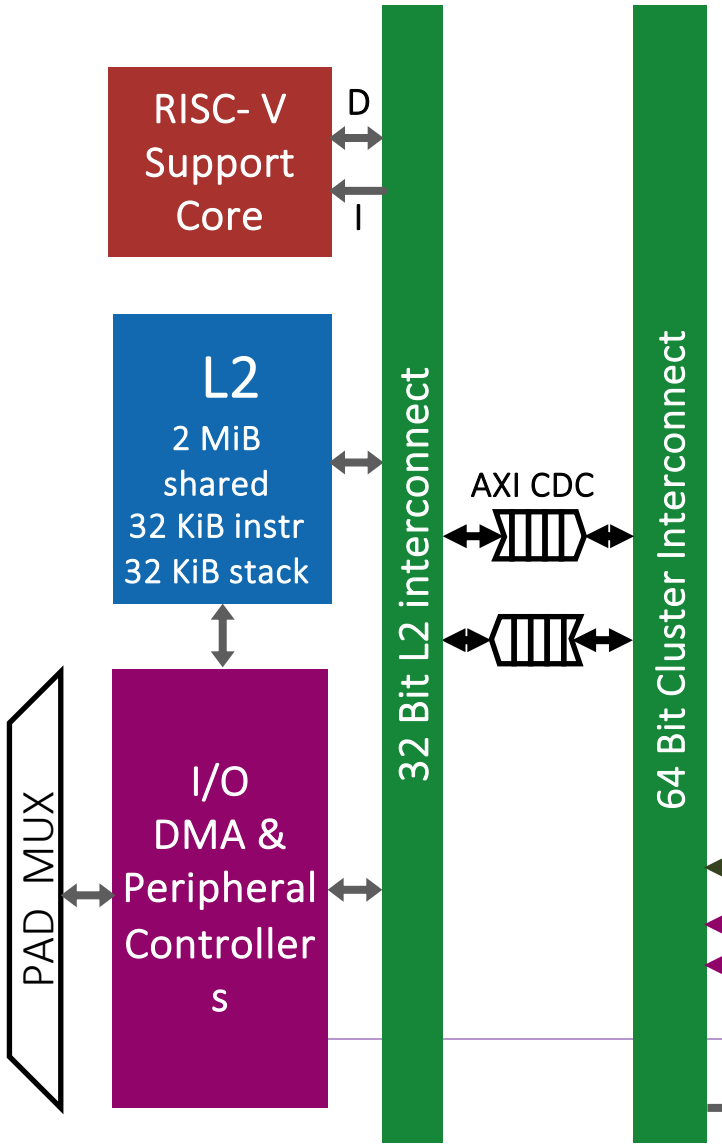
[M. Abrash: Creating the Future: Augmented Reality, the next Human-Machine Interface, IEEE IEDM 2021]

Towards Ultra-Low Power XR

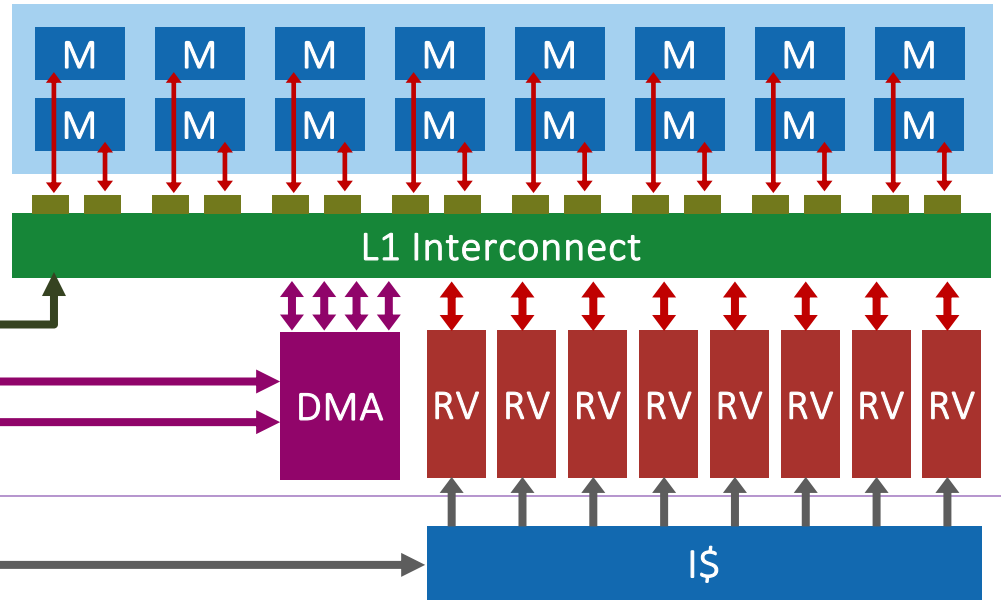
Siracusa - A PULP SoC to maximize energy efficiency for XR applications

- Integration of N-EUREKA accelerator in a PULP system
  - Weight-precision scalable CNN accelerator
- Implementation of a dedicated weight memory subsystem
  - Eliminates off-chip memory transfers for real-world workloads
  - Maximizes weight-intensive layer performance & energy efficiency
- Real-time performance in 10s of mW on a real-world hand-detection workload

The Essential PULP System



- Fully fledged, multi-core SoC
  - 8 RISC-V Cluster Cores
  - Custom DSP ISA extensions
- Software-controlled L2 & L1 memory w/ dedicated DMA
  - Enables neural network activation tiling!
- Support core & rich set of peripherals

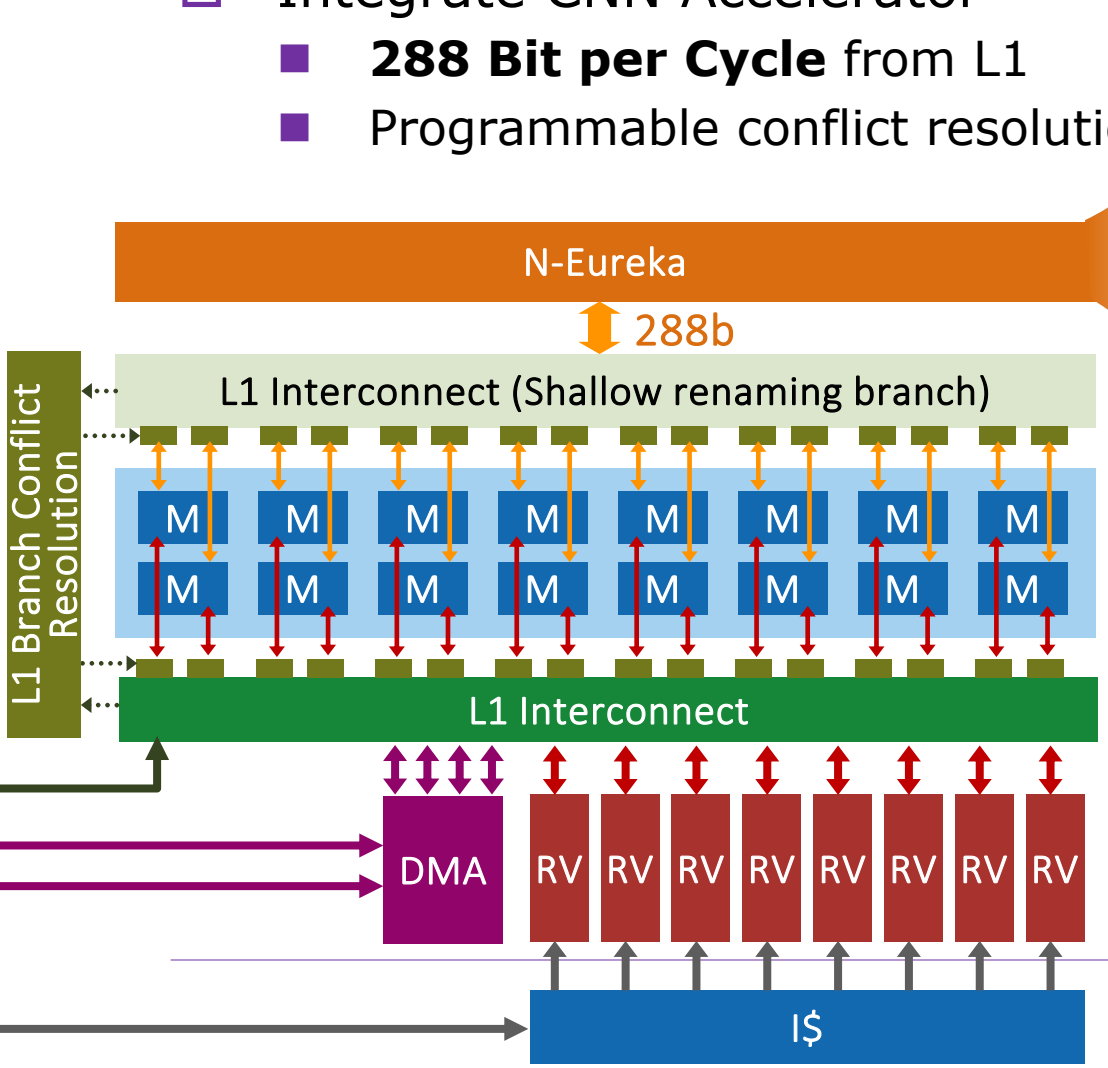


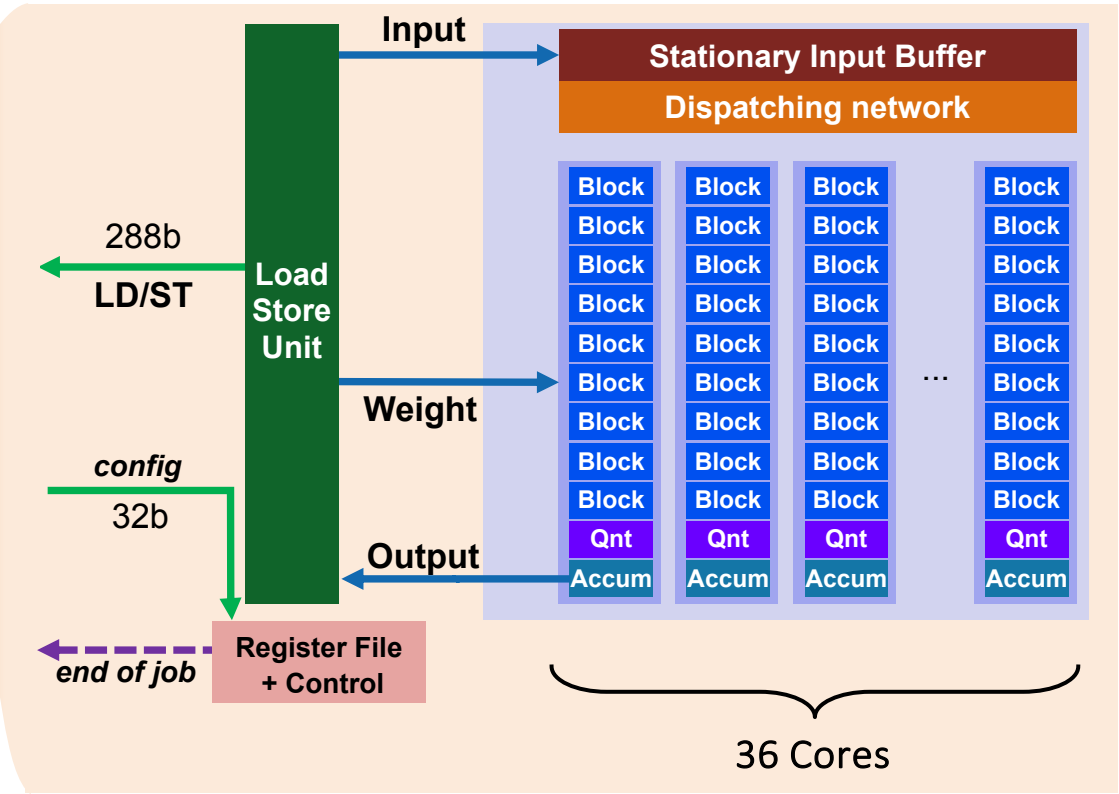
256 KiB TCDM
16x 32 Bit Ports

8 Cores
DSP & SIMD Instructions

The Siracusa Cluster

- Integrate CNN Accelerator
  - **288 Bit per Cycle** from L1
  - Programmable conflict resolution

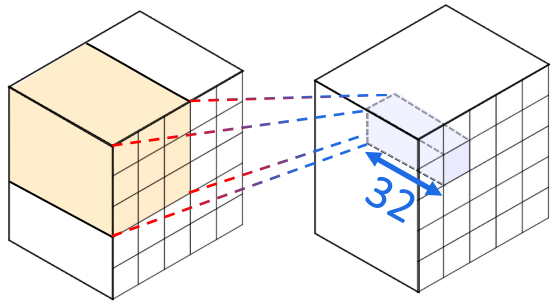


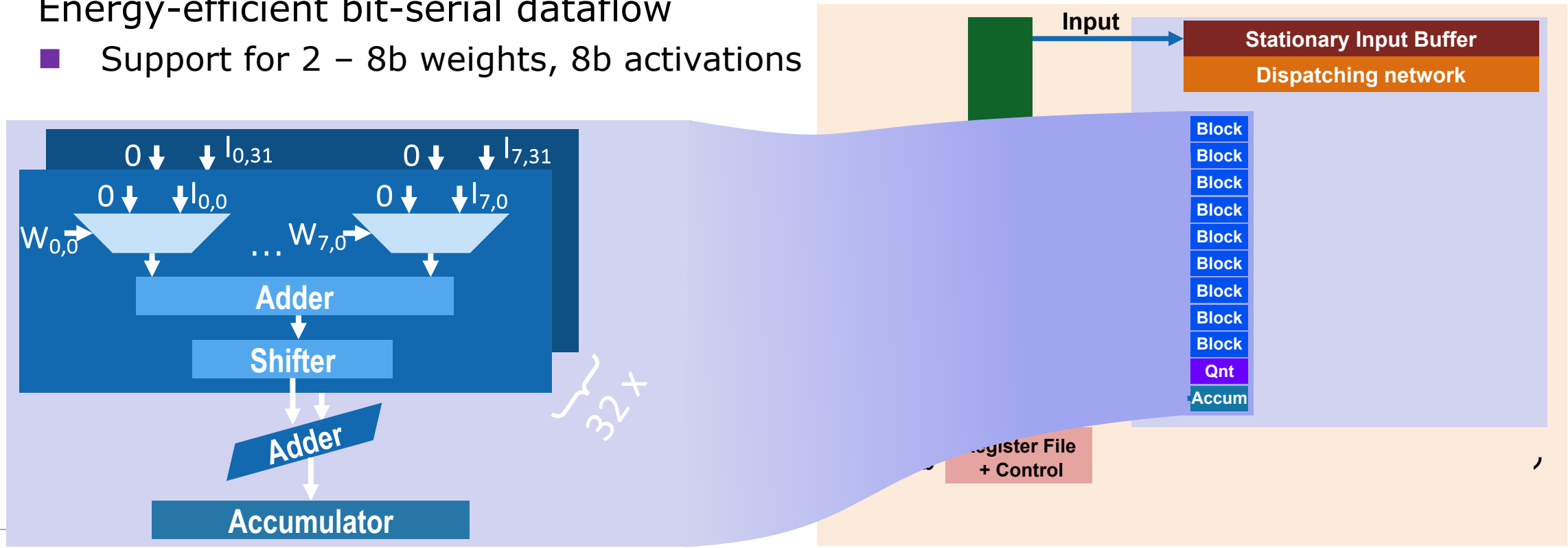


- Dense & Depthwise Convolutions
- Bit-serial Operation, scaling performance

N-EUREKA Architecture

- Computation tiled over output
  - Each **core** computes 32 channels of 1 px
  - More **cores** -> larger output tile
- Energy-efficient bit-serial dataflow
  - Support for 2 – 8b weights, 8b activations





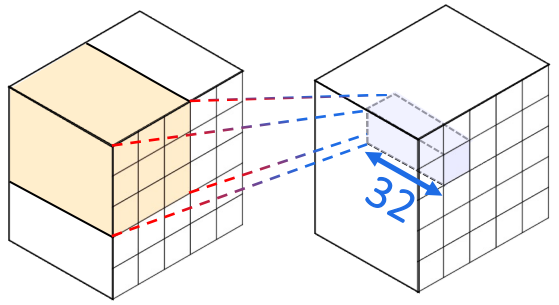
[Prasad et al.:Specialization meets Flexibility: a Heterogeneous Architecture for High-Efficiency, High-flexibility AR/VR Processing , DAC 2023]

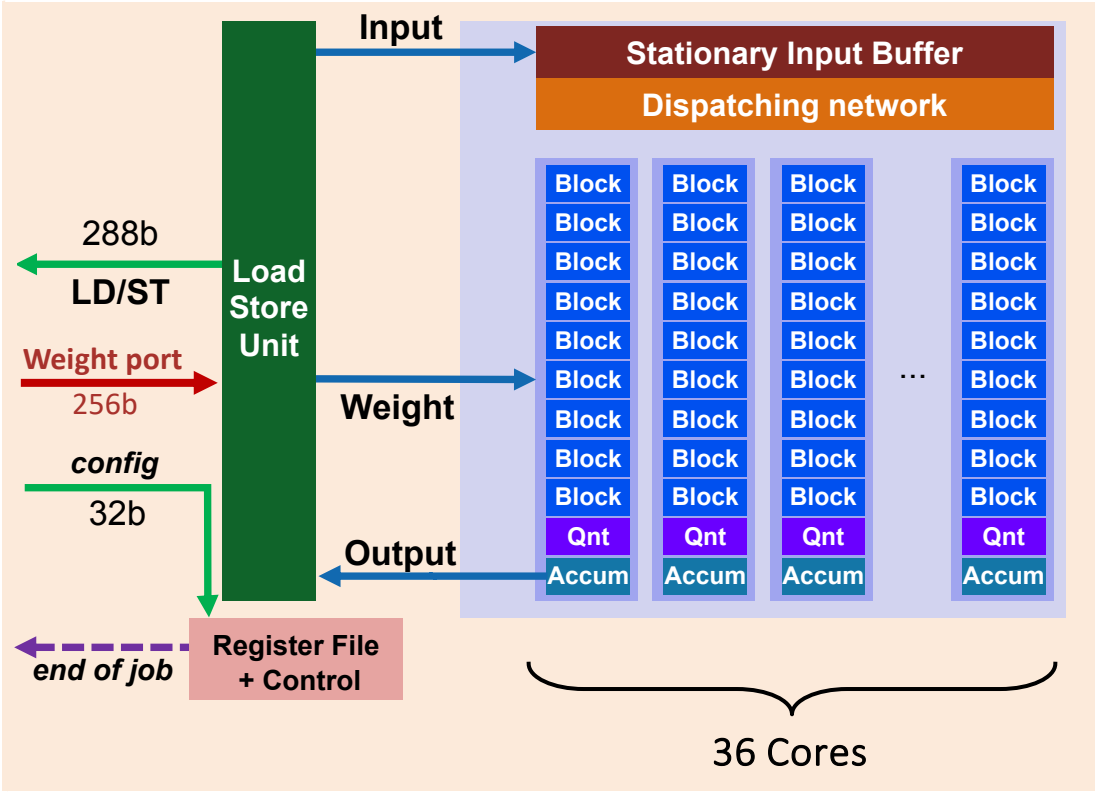
N-EUREKA Architecture

- Computation tiled over output
  - Each **core** computes 32 channels of 1 px
  - More **cores** -> larger output tile
- Energy-efficient bit-serial dataflow
  - Support for 2 – 8b weights, 8b activations
- 1x1 convolutions are crucial
  - > 90% of computations in MobileNet v2
  - High bandwidth needed!

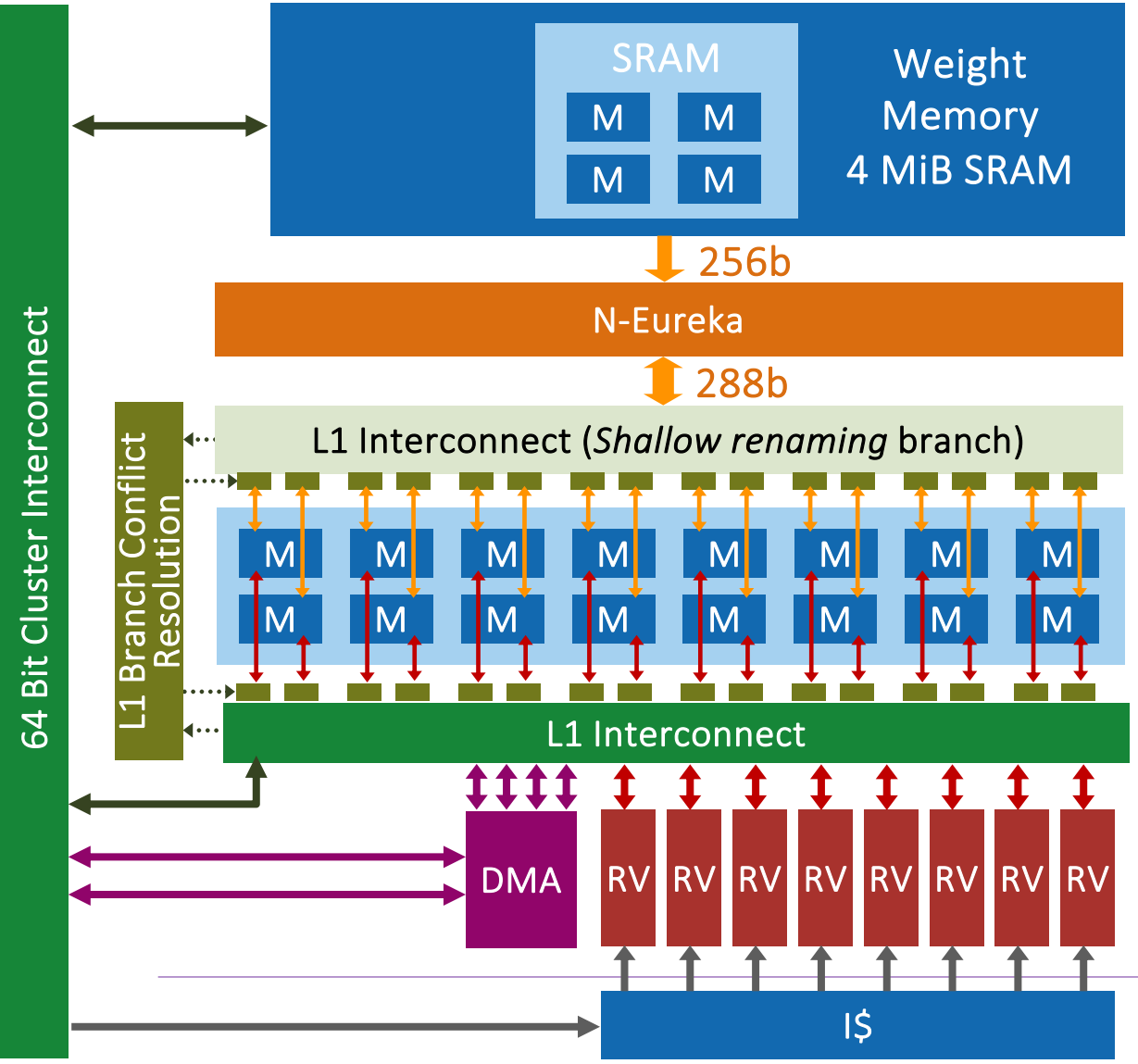
How to support 1x1-convolutions efficiently?

- Add a dedicated weight port!





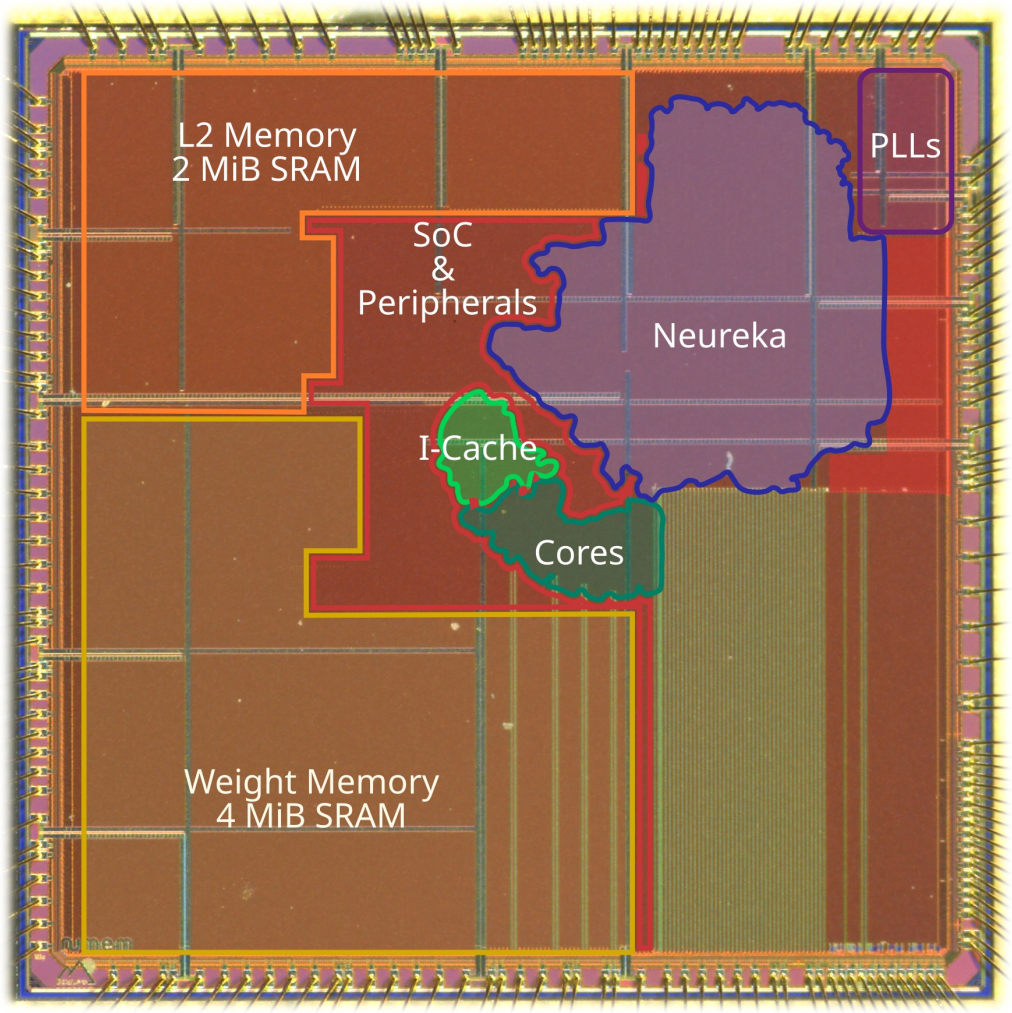
Weight Memory Subsystem



- Add wide weight access port
  - **256 Bits per cycle**
- Exploit large weight memory
  - Avoids off-chip weight accesses for NNs
  - Efficiently support bit-serial access
- Use L1 port for activations only

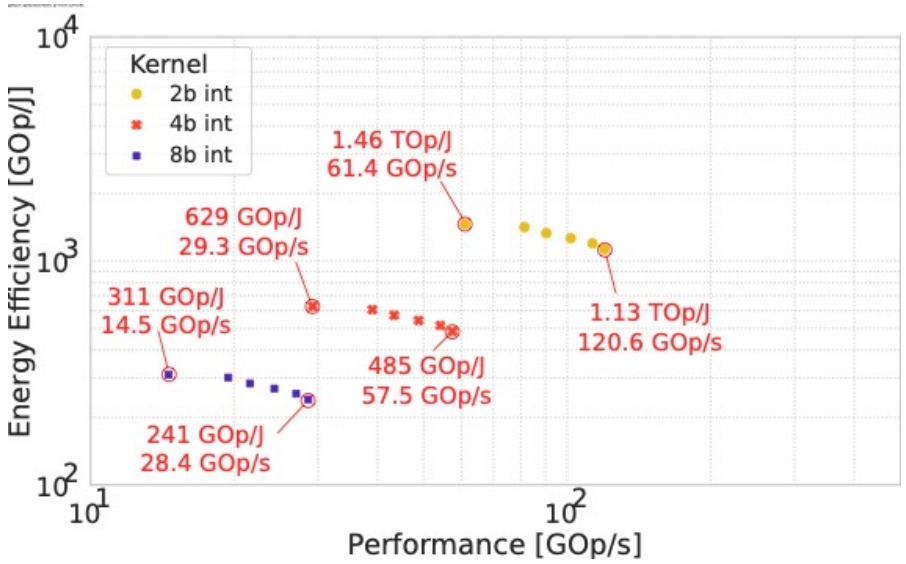
Siracusa Implementation

- TSMC 16 nm implementation
  - 4 mm x 4 mm
- A plethora of on-chip memory
  - 4 MiB SRAM weight memory
  - 2 MiB L2
  - 256 KiB L1 TCDM
- Three independent clock domains
  - SoC, Cluster & Peripherals
- N-EUREKA & 8 RISC-V Cluster Cores



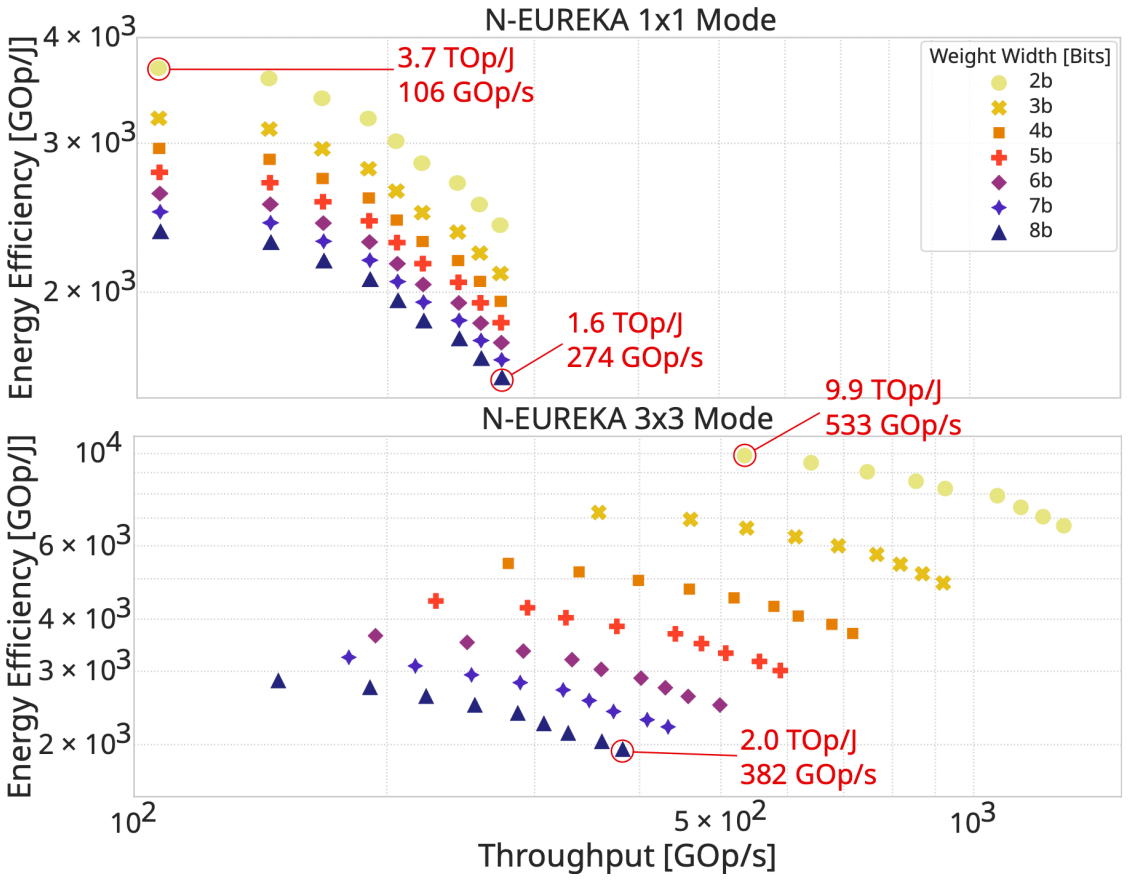
Cluster Performance

- Measured peak performance & efficiency on matrix multiplication
  - Using L1 TCDM and 8-core parallel cluster
  - Characterized for 0.65 - 0.8V
  - Characterized for 2b, 4b and 8b operands
- Efficient general purpose & DSP acceleration
  - **1.13 TOp/J @ 120.6 GOp/s** – 2b matmul
  - 241 GOp/J @ 28.4 GOp/s – 8b matmul



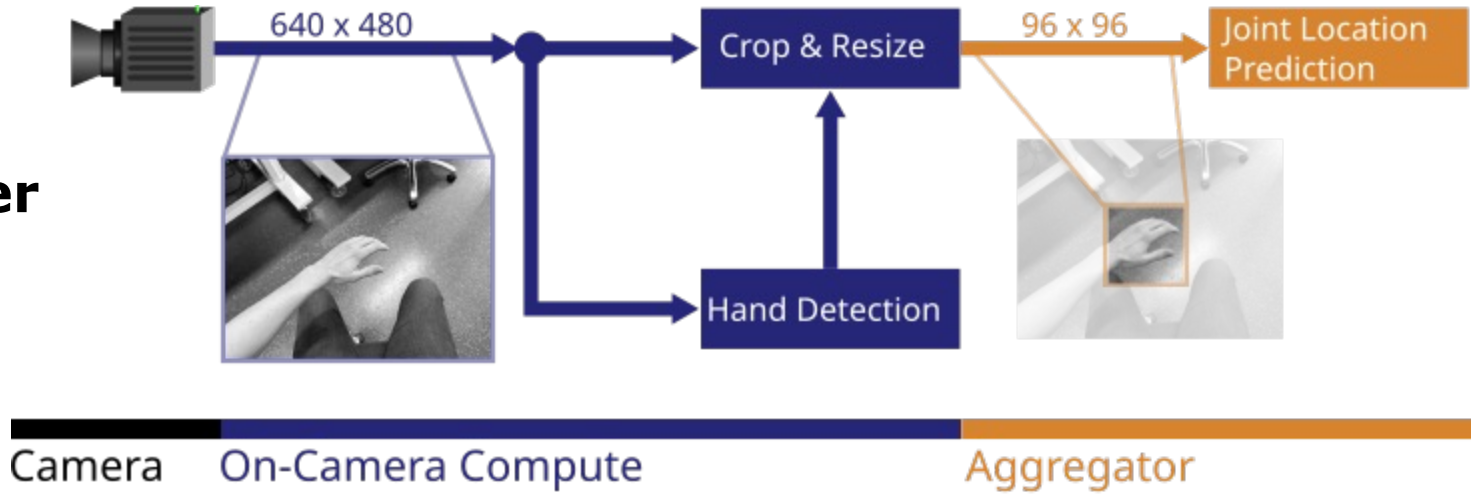
N-EUREKA Performance

- Measured peak performance & efficiency on convolutional layers
  - 1x1, 3x3 kernel dense convolutions
  - 256 channels, 6x6 feature map
  - Using SRAM weight memory
  - Characterized for 0.65 - 0.8V
- Wide range of efficiency/throughput
  - 1x1:
    - **3.7 TOp/J @ 106 GOp/s** – 2b weight
    - 1.6 TOp/J @ 274 GOp/s – 8b weight
  - 3x3 (dense):
    - **9.9 TOp/J @ 533 GOp/s** – 2b weight
    - 2.0 TOp/J @ 382 GOp/s – 8b weight



Hand Detection Results

- Hand detection for Region-of-Interest cropping
  - MobileNetv2-based network for detection
- Implemented & measured NN on N-EUREKA
  - 8 Bit activations, 8 Bit weights
  - 120 MOp per Inference
  - 1.3 Mparameters
- End-to-end performance
  - 168.5 8b-MAC/Cycle
  - **107.3 µJ/Inference**
- **33x less external transfer**

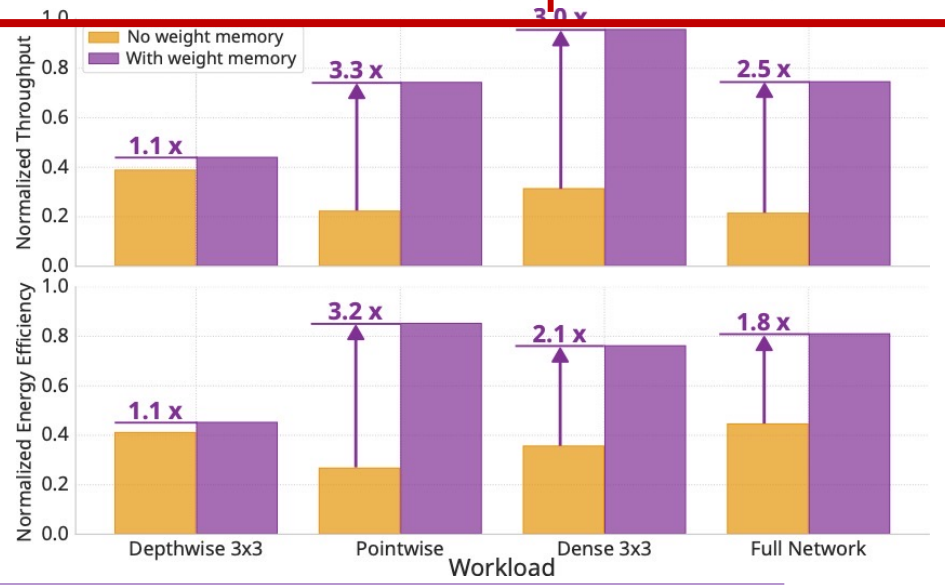


[Han et al.:MEgATrack: monochrome egocentric articulated hand-tracking for virtual reality, ACM Transactions on Graphics vol. 39]

The Impact of Weight Memory

- Measured N-EUREKA performance
  - With using weight memory
  - Without weight memory -> Weights from L1
- Weight memory increases performance
  - PW: 3.3x throughput, 3.2x energy efficiency
  - Dense: 3.0x throughput, 2.1x energy efficiency
  - **Full Network: 2.5x throughput, 1.8x EE**
- … and closes the gap to the compute bound
  - >75% for Dense 1x1 & full network

Theoretical Compute Bound



Conclusion

- Siracusa optimizes energy efficiency for NNs by
  - eliminating off-chip memory accesses thanks to large on-chip memory
  - minimizing on-chip memory movement thanks to L1 weight memory
  - exploiting mixed-precision, bit-serial accelerator dataflow
  - 9.9 TOp/J @ 533 GOp/s – **1.2x more efficient than digital state-of-the-art**[4]
- L1 weight memory maximizes end-to-end performance
  - **2.5x increase in throughput**
  - **1.8x increase in energy efficiency**
- Siracusa offers efficient general-purpose acceleration
  - 8 Core Cluster with state-of-the-art throughput & efficiency

Questions?

Distributed, on-sensor computing

- Collect raw data
- Process directly **on-sensor**
- **Aggregate** on larger computing platforms

Acceleration

- L1 **HW acceleration** for DNNs
- On-chip L1 **weight memory** for DNNs
- L0 **acceleration** for diverse processing

`asic.ethz.ch`

`pulp-platform.org`

References

- [1] 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] Conti *et al.*, "A 12.4 TOPS/W @ 136GOPS AI-IoT System-on-Chip with 16 RISC-V, 2-to-8b Precision-Scalable DNN Acceleration and 30\%-Boost Adaptive Body Biasing"