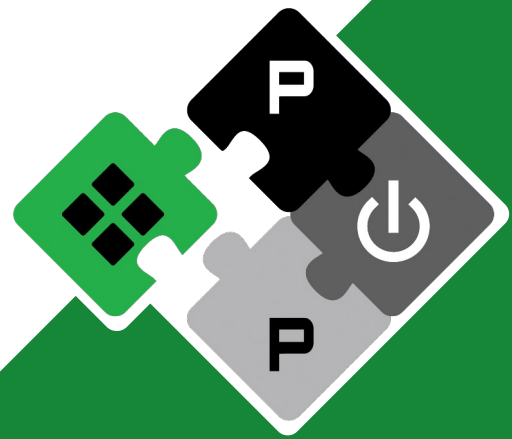
ETH zürich

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Meta

Next stop XR: towards on-sensor PULP computing for micropower eXtended Reality

Francesco Conti
f.conti@unibo.it



@pulp_platform

pulp-platform.org

youtube.com/pulp_platform

PULP Platform
Open Source Hardware, the way it should be!

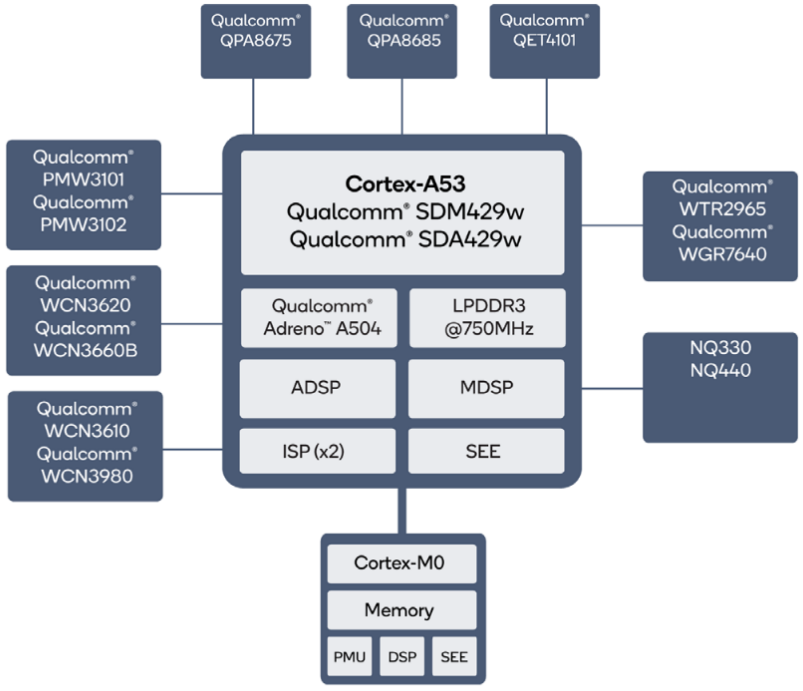
“Smart” Glasses, today

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

Lightweight → **<50 gram**

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

Ray-Ban Stories



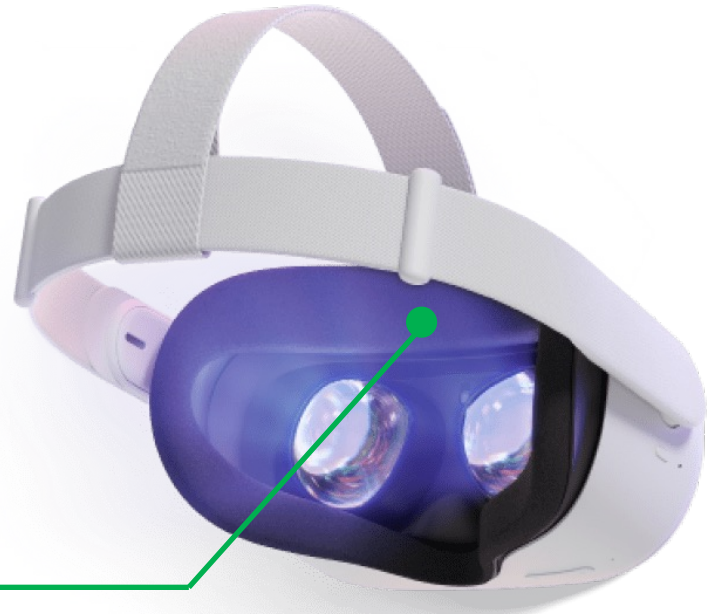
[https://www.techinsights.com/blog/ray-ban-stories-smart-glasses-cameras]

eXtended Reality Glasses, today

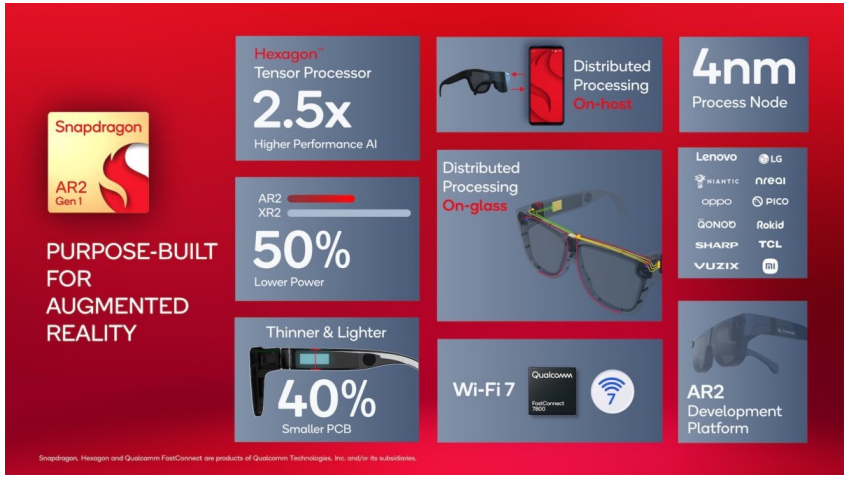
Dedicate form factor → **cumbersome and uncomfortable in the long run**

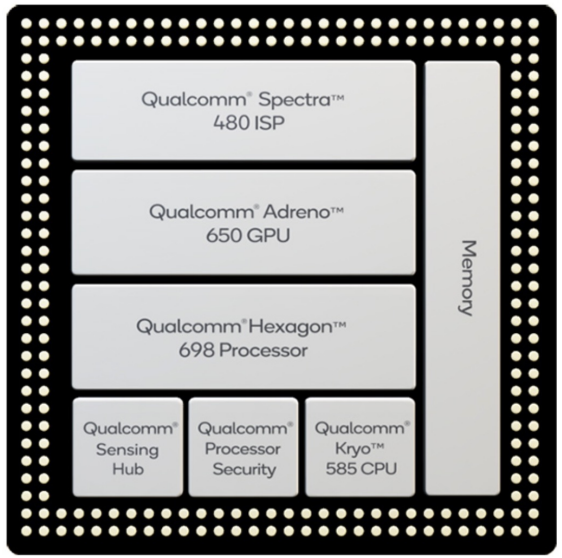
Heavyweight → **~500 gram**

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



Meta Quest 2



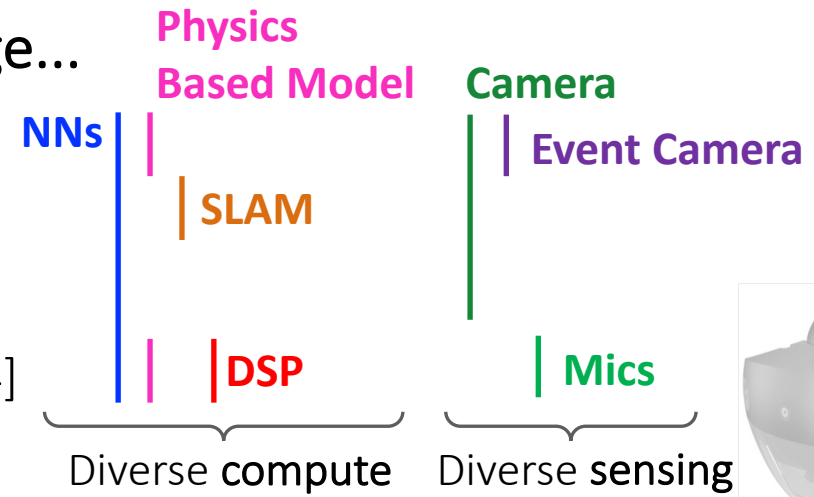


Microsoft HoloLens 2

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

How to “fold” XR functionality into smart glasses?

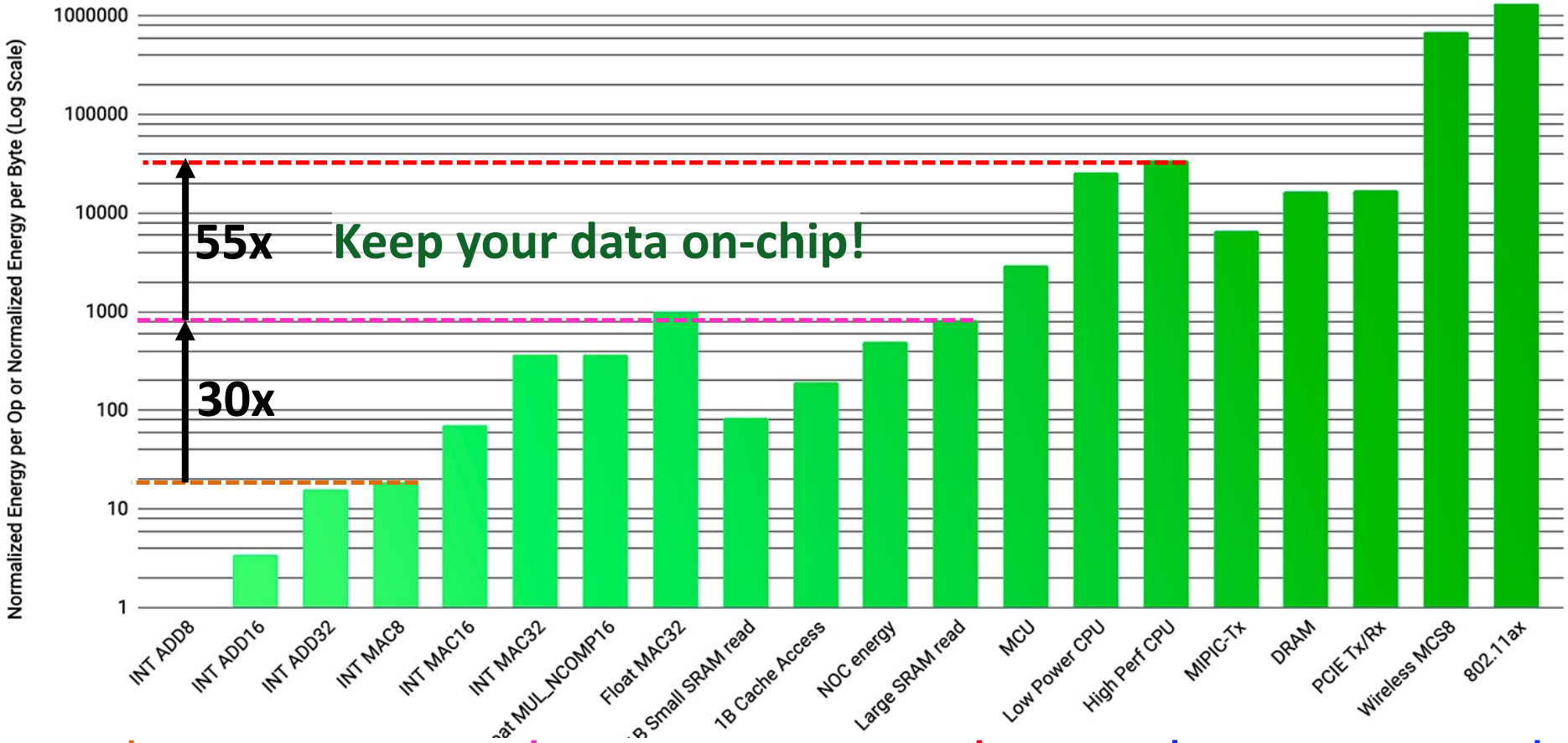
- Many tasks to manage...
  - Eye gaze tracking [1]
  - Head tracking [2]
  - Hand tracking [3]
  - Spatial beamforming [4]
  - ...



- ... 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]

Save energy where it counts!



Exploit custom units

Custom Pixel/ML Compute

Wide Operand Compute & On Chip Data Movements

General Purpose Compute

Off Chip Data Transfers

[E. Beigné, ISSCC Forum 4: Advancing Technologies for Extended Reality (XR) to Make the "Metaverse" Possible]

The PULP value

Composability

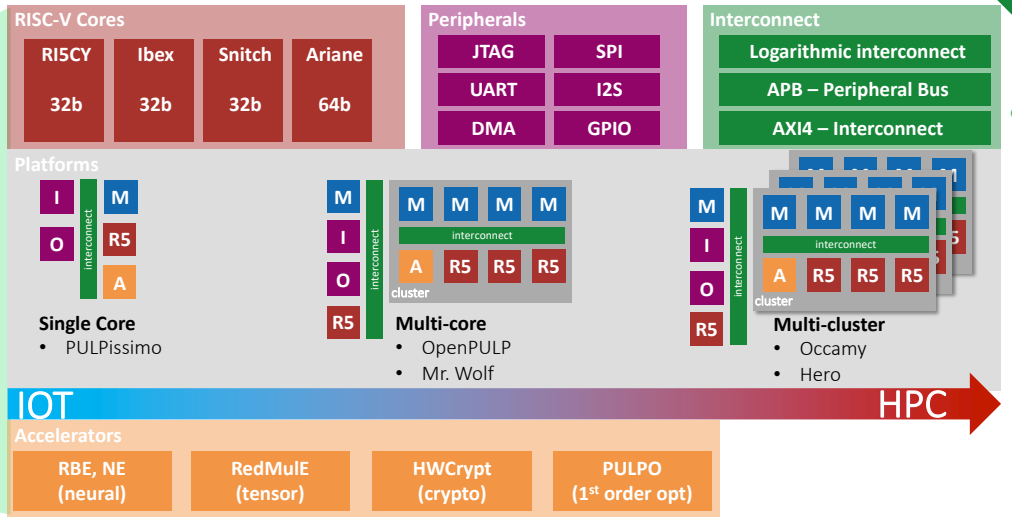
- Vast library of silicon-proven IPs
- Ranging from microcontroller to HPC

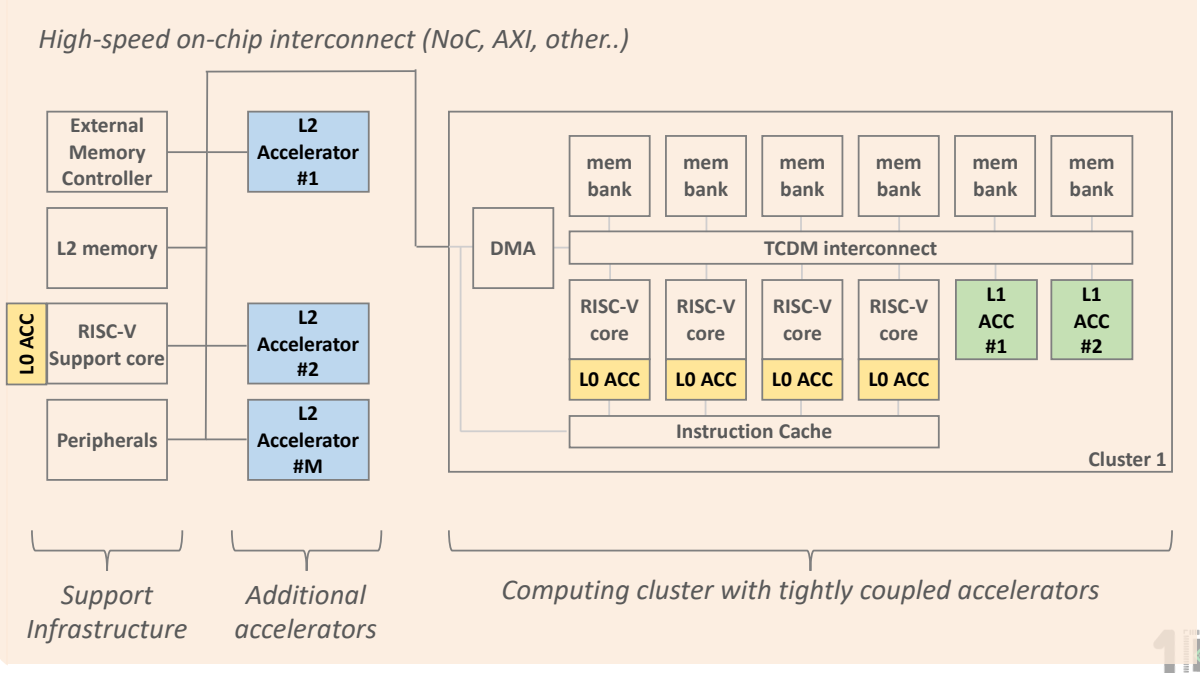
Heterogeneity

- L0 acceleration: RISC-V extensions, SSR, ...
- L1 acceleration: HWPEs (neural engines, TPEs, optimization engines...)
- L2 acceleration: AXI autonomous units (& multi-cluster)

Efficiency

- Otherwise it would be the *P___ Platform*!

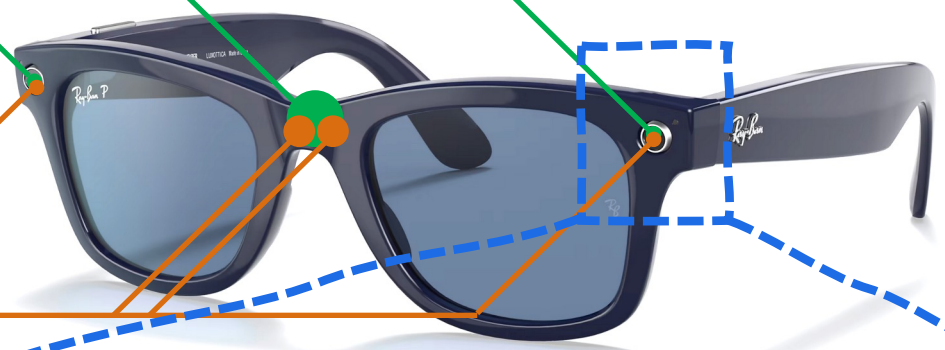




A vision for PULP-based XR glasses

Distributed, on-sensor computing

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



Acceleration

- On-chip **NVM** for DNN weights
- L1 **HW acceleration** for DNNs
- L0 **acceleration** for diverse processing

Meta Reality Labs

Siracusa: a highly heterogeneous SoC moving towards on-sensor computing

Siracusa first steps

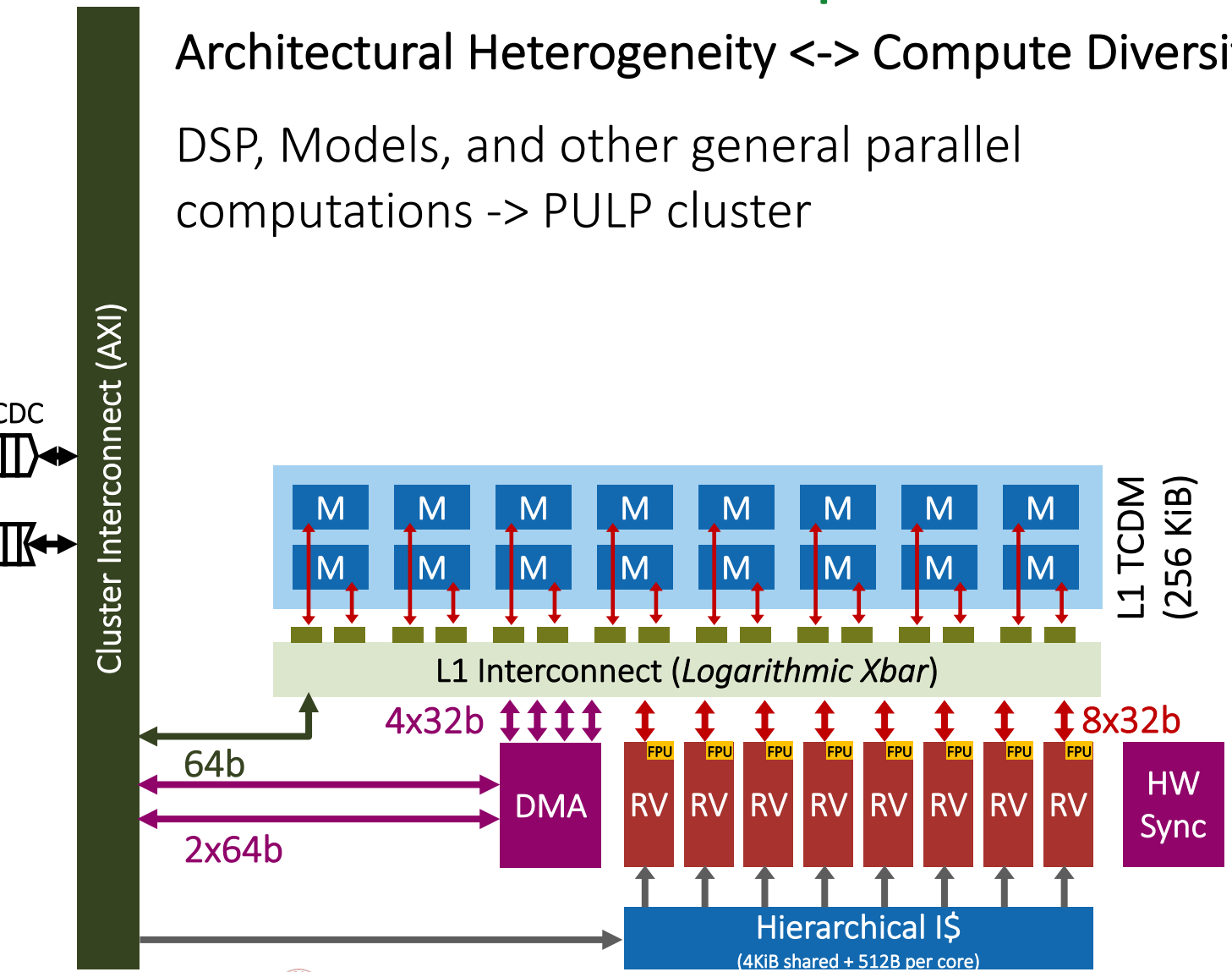
- A heterogeneous cluster template ARchiMEDES
- A novel accelerator NEureka
- A silicon prototype SoC Siracusa (TSMC16)

Disclaimer: I am not payed by the Siracusa tourist office.
But seriously, go visit!

ARchiMEDES*cluster template

Architectural Heterogeneity <-> Compute Diversity

DSP, Models, and other general parallel computations -> PULP cluster



A "classic" PULP cluster with 8 **RV32IMCFXpulpnn** cores

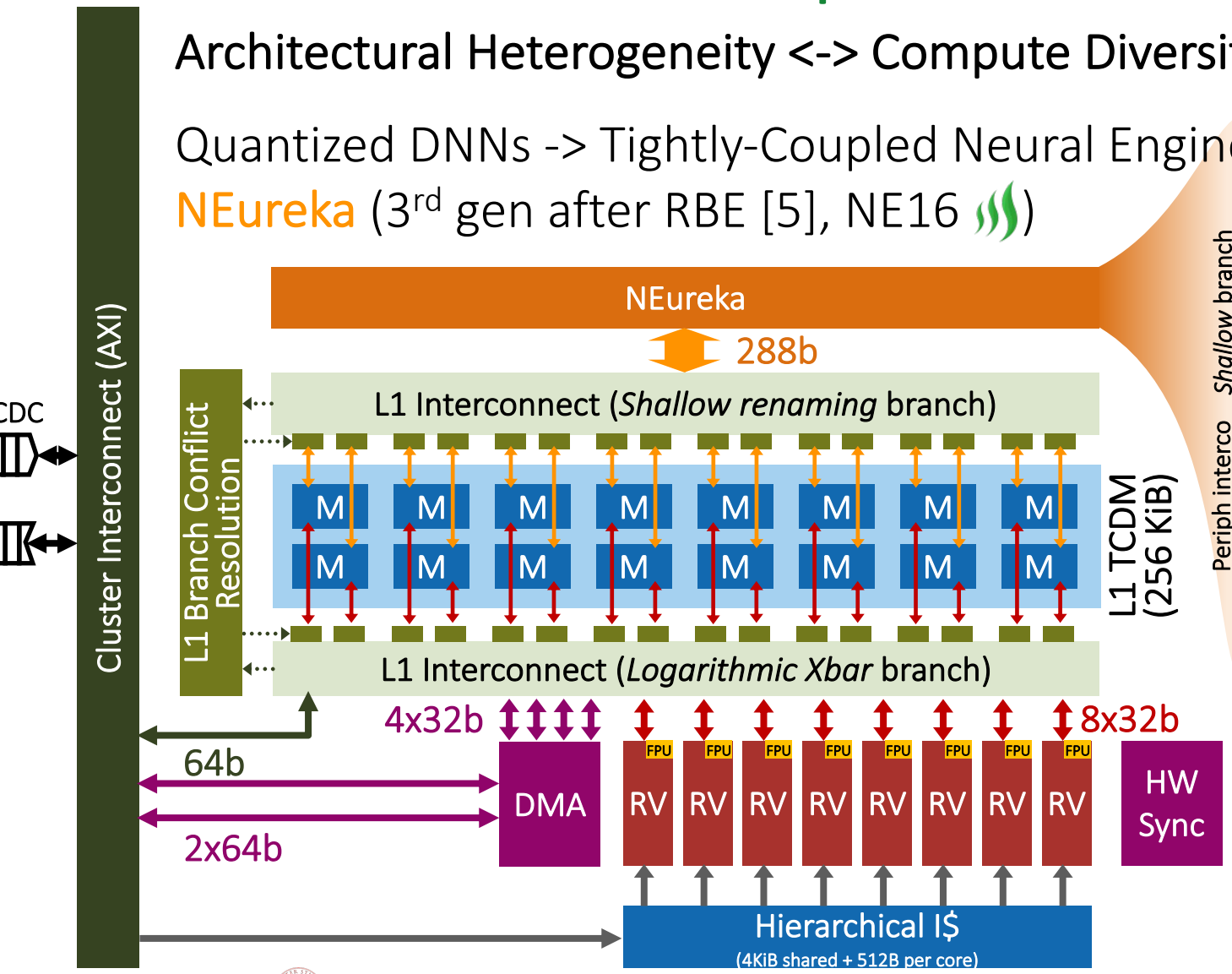
- private multi-precision FPUs
- hierarchical instruction cache (4 KiB + 512B per core)
- **Xpulpnn** extensions [5] for integer mixed-precision DSP + DNNs
- 256 KiB of Tightly-Coupled Data Memory (TCDM) divided in 16 word-interleaved SRAM banks
- The Logarithmic Interconnect we have known and loved since < 2013 ☺

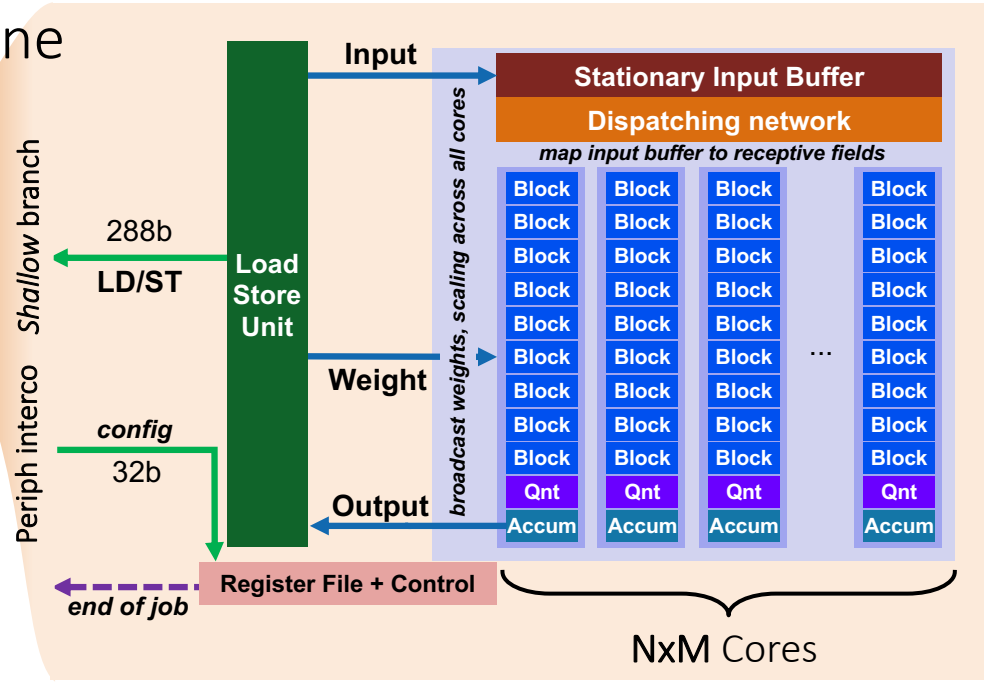
ARchiMEDES cluster template

Architectural Heterogeneity <-> Compute Diversity

Quantized DNNs -> Tightly-Coupled Neural Engine

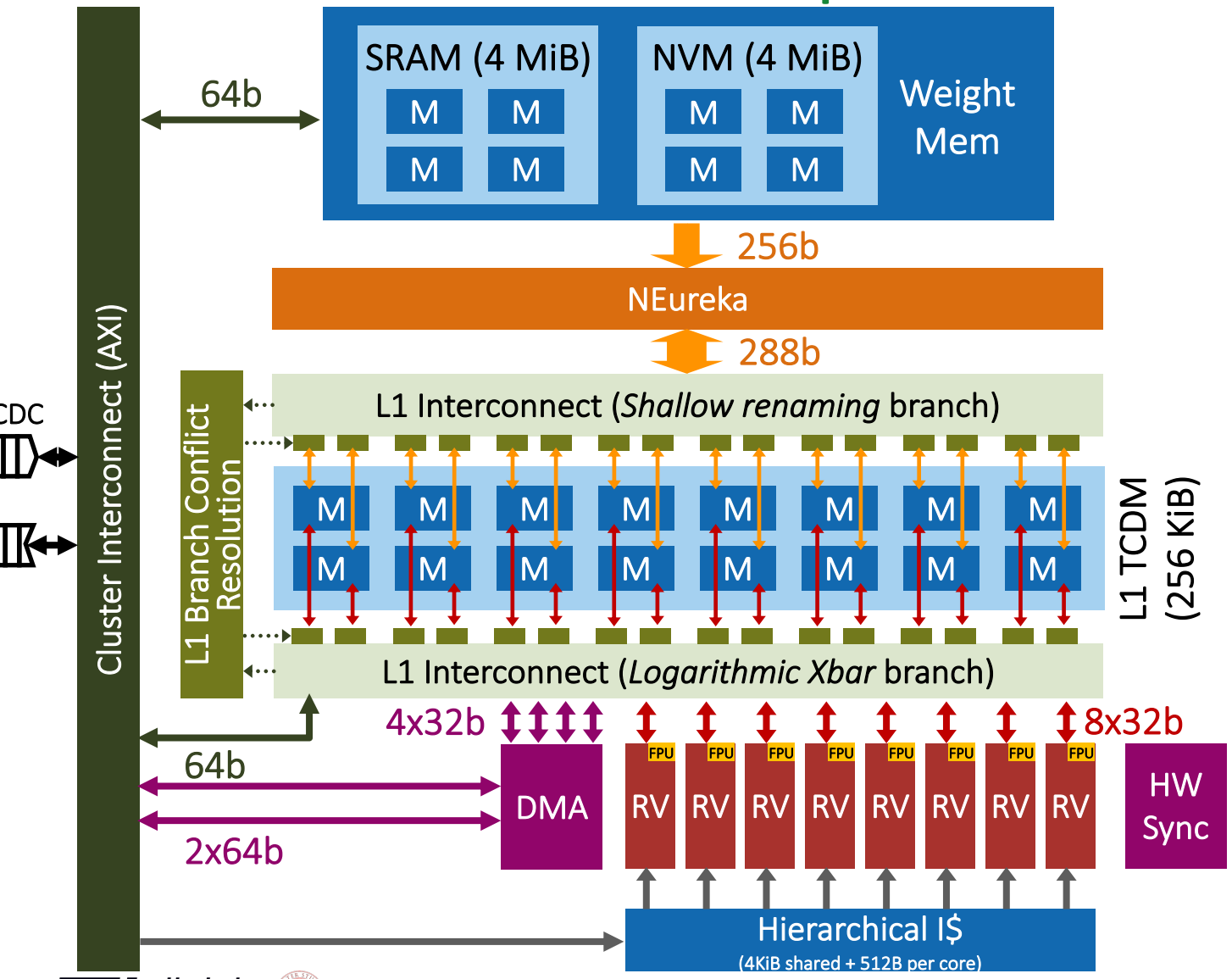
NEureka (3rd gen after RBE [5], NE16)





- 1 Core = receptive field of 1x1 px in output across 32 out-chans
- Output stationary, Input quasi-stationary
- Parametric number of Cores (NxM out-px)
- 8b activations, 2-8b weights

ARchiMEDES cluster template



Boost memory energy efficiency

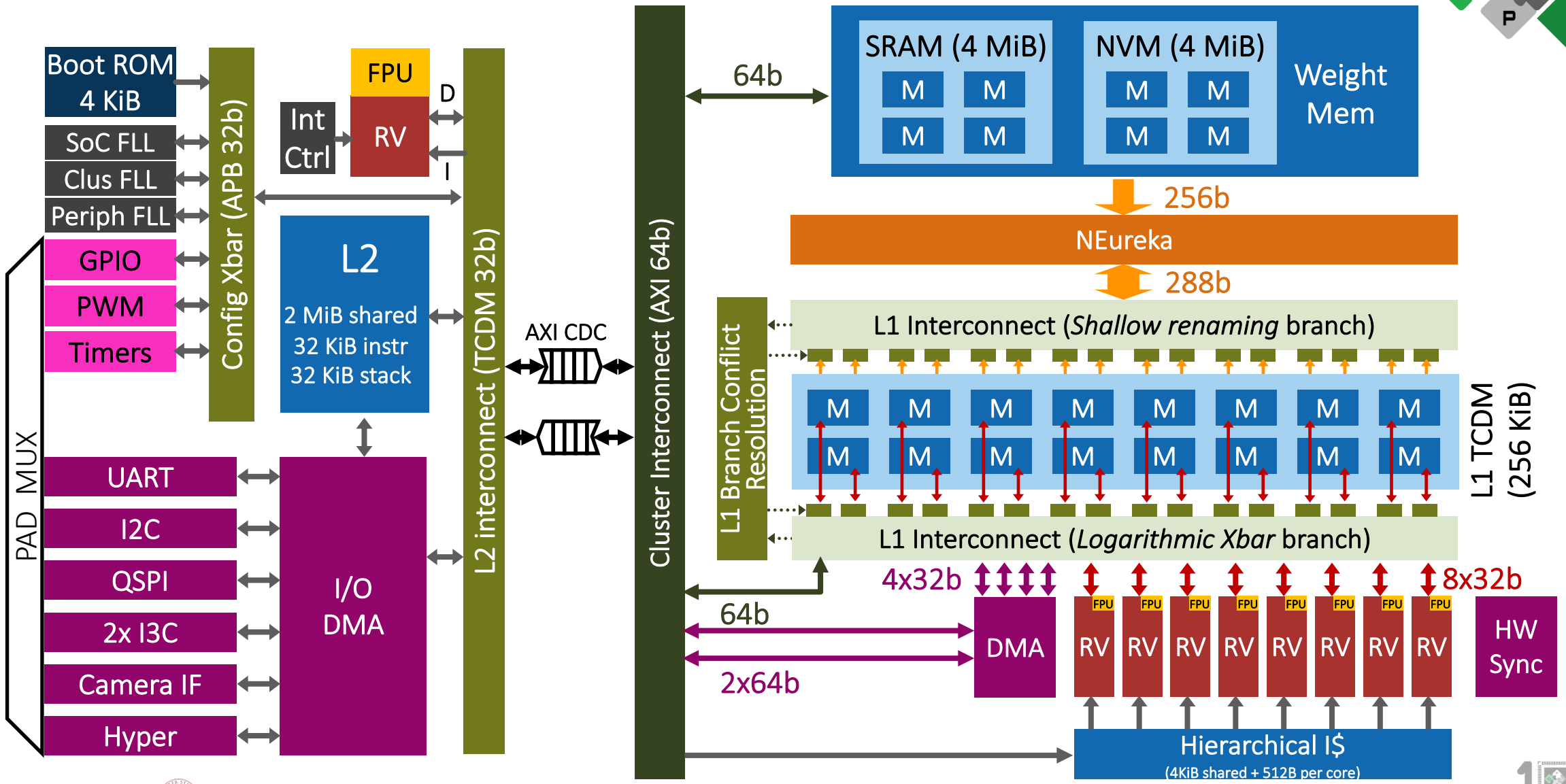
A large power-optimized on-chip memory for network weights -> cluster-level **weight stationarity**

4x 1MiB SRAM banks (64b-wide)

4x 1MiB NVM banks (64b-wide)

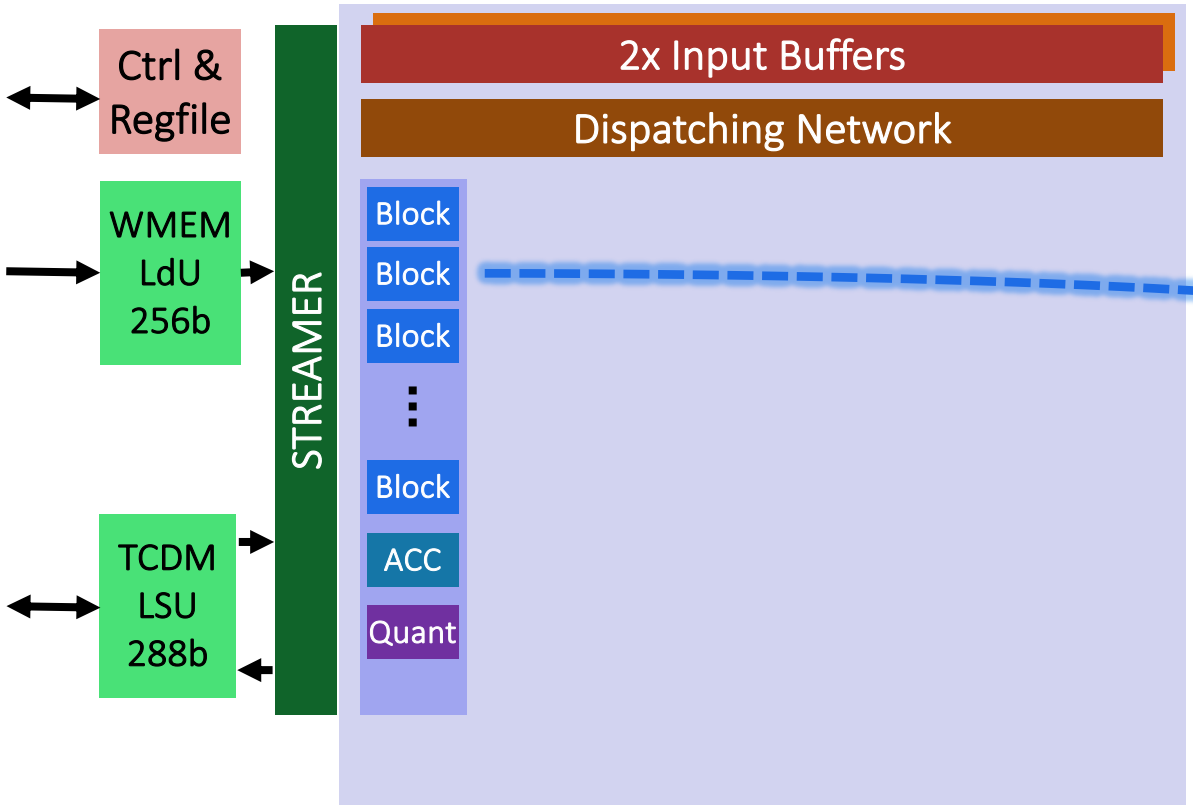
Paging support for transparent network reconfiguration with negligible increase in overall circuit area.

Siracusa SoC

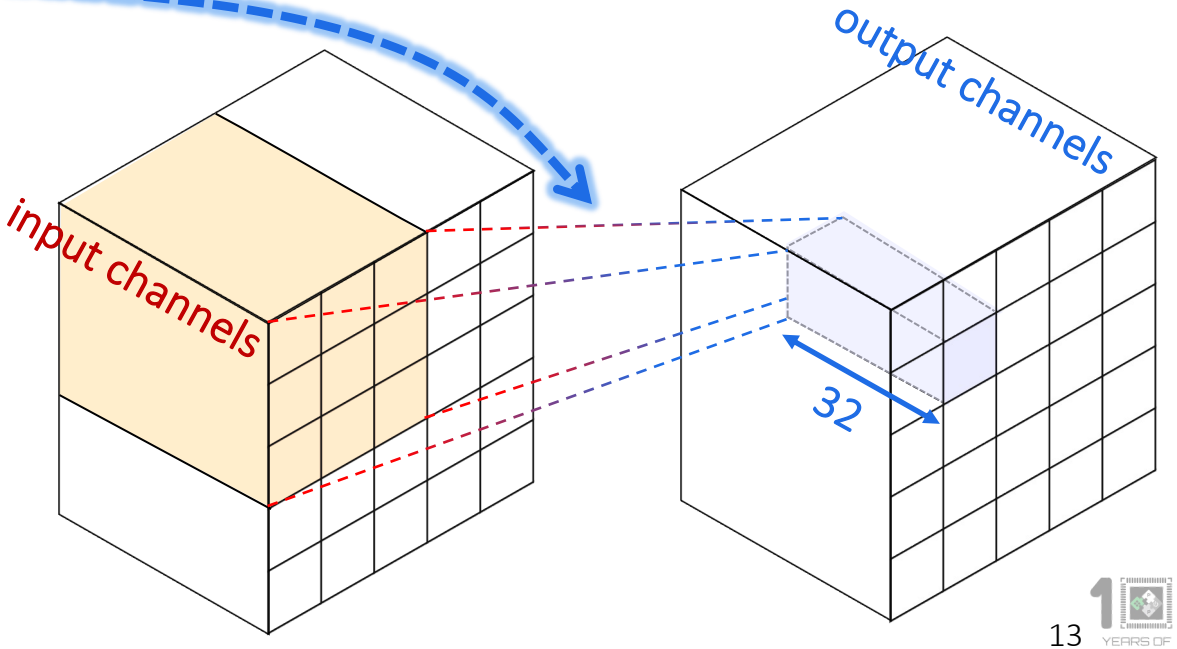


NEureka – DNN Accelerator Engine

$$\mathbf{y}(k_{out}) = \textbf{quant}\left(\sum_{i=0..Wbit}\sum_{k_{in}} 2^i\left(\mathbf{W_{bin}}(k_{out}, k_{in}) \otimes \mathbf{x}(k_{in})\right)\right)$$

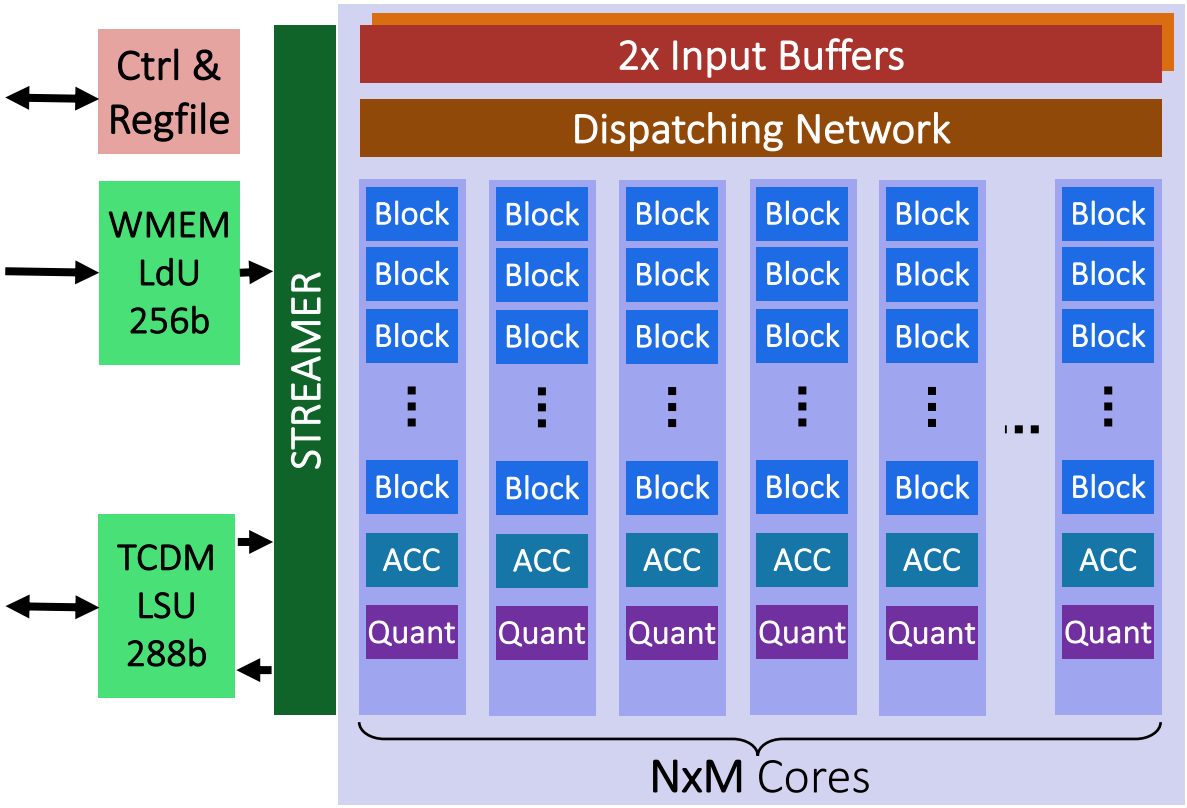


- Partially bit-serial dataflow for **CONV3x3**, **PW1x1**, **DWCONV3x3**
  - 3x3, 1x1 and 3x3 depthwise mode
  - Activations 8b, Weights 2-8b
- Core – receptive field of 1 output px across 32 output chans → more cores, larger output "tile"



NEureka – DNN Accelerator Engine

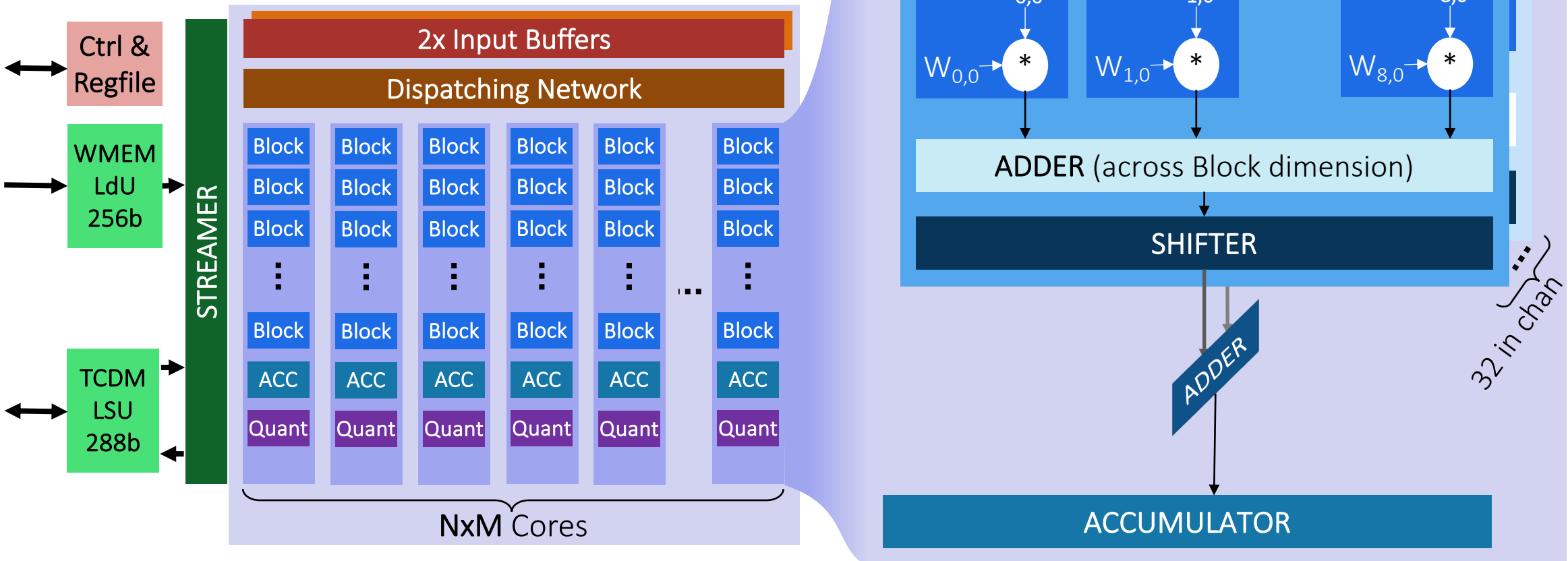
$$\mathbf{y}(k_{out}) = \textbf{quant}\left(\sum_{i=0..Wbit}\sum_{k_{in}} 2^i\left(\mathbf{W_{bin}}(k_{out}, k_{in}) \otimes \mathbf{x}(k_{in})\right)\right)$$



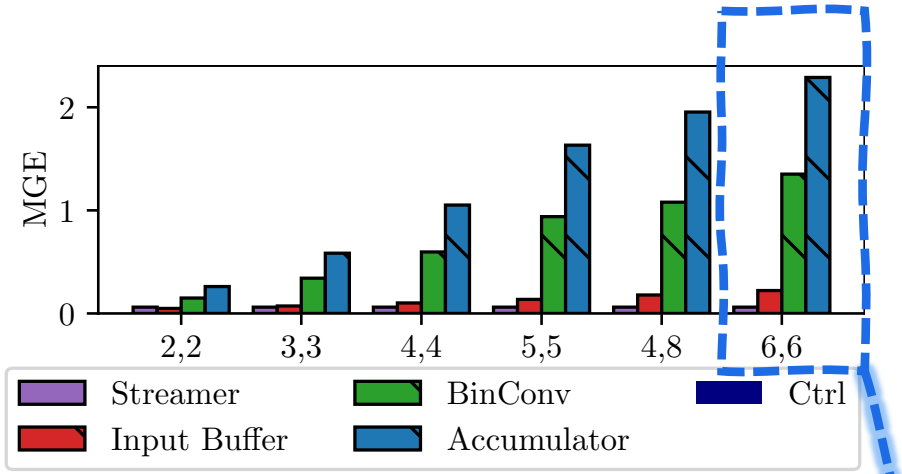
- Partially bit-serial dataflow for CONV3x3, PW1x1, DWCONV3x3
  - 3x3, 1x1 and 3x3 depthwise mode
  - Activations 8b, Weights 2-8b
- Core – receptive field of 1 output px across 32 output chans → more cores, larger output "tile"
- Stationarity
  - Output -> fully stationary in Accumulators
  - Input -> quasi-stationary in Input Buffers
  - Weights -> non-stationary (but stationary @ cluster level, thanks to WMEM!)
- Dispatching network – maps input across Cores
- Accumulator – 32x32-bit registers to store partial sums
- Quant – Normalization, Quantization, ReLU

NEureka – DNN Accelerator Engine

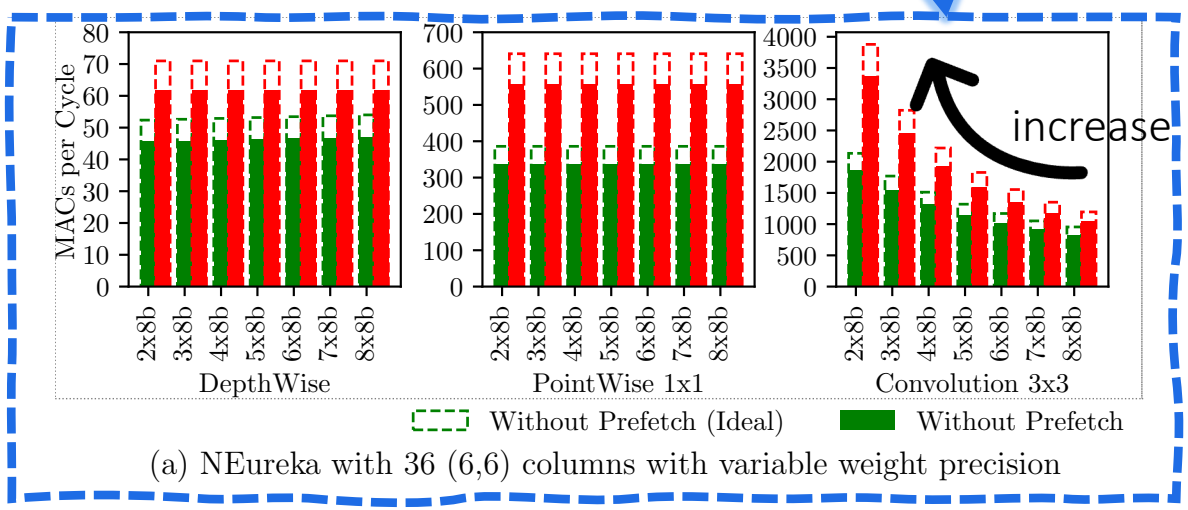
$$\mathbf{y}(k_{out}) = \textbf{quant}\left(\sum_{i=0..Wbit}\sum_{k_{in}} 2^i\left(\mathbf{W_{bin}}(k_{out},k_{in}) \otimes \mathbf{x}(k_{in})\right)\right)$$



NEureka scalability and performance

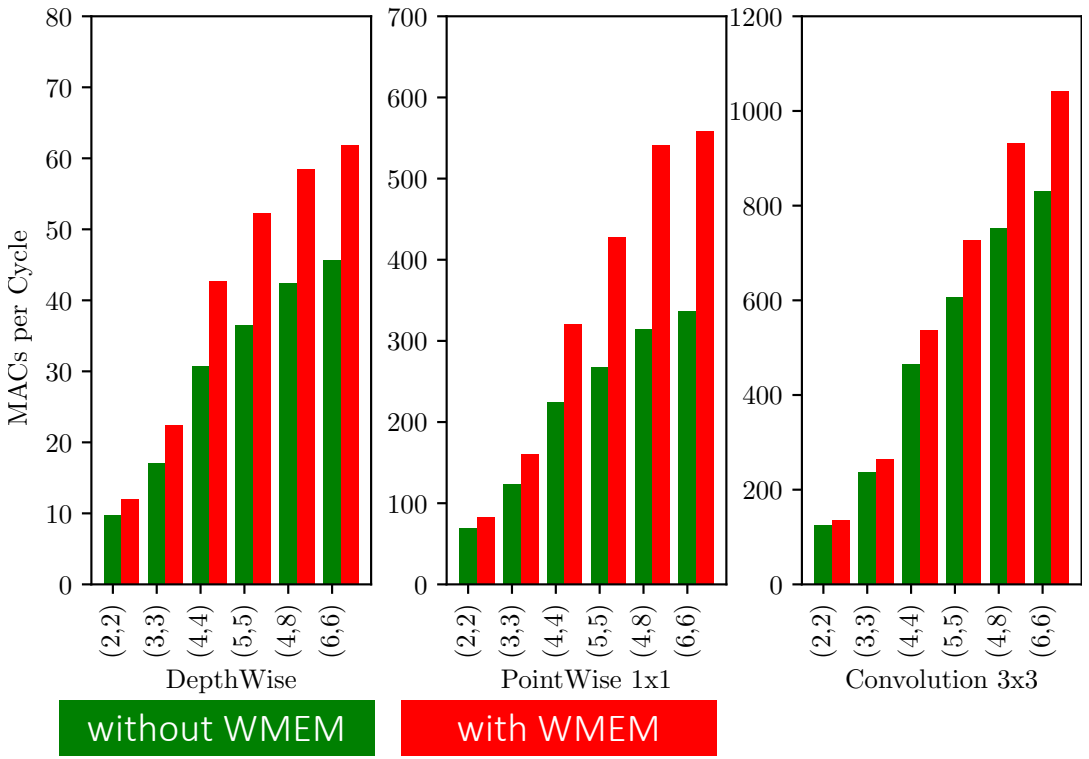


NEureka Complexity (MGE, 22nm)



(a) NEureka with 36 (6,6) columns with variable weight precision

Weight-precision scaling



Complexity scaling

To appear at **DAC'23** (*A. Prasad et al.*)

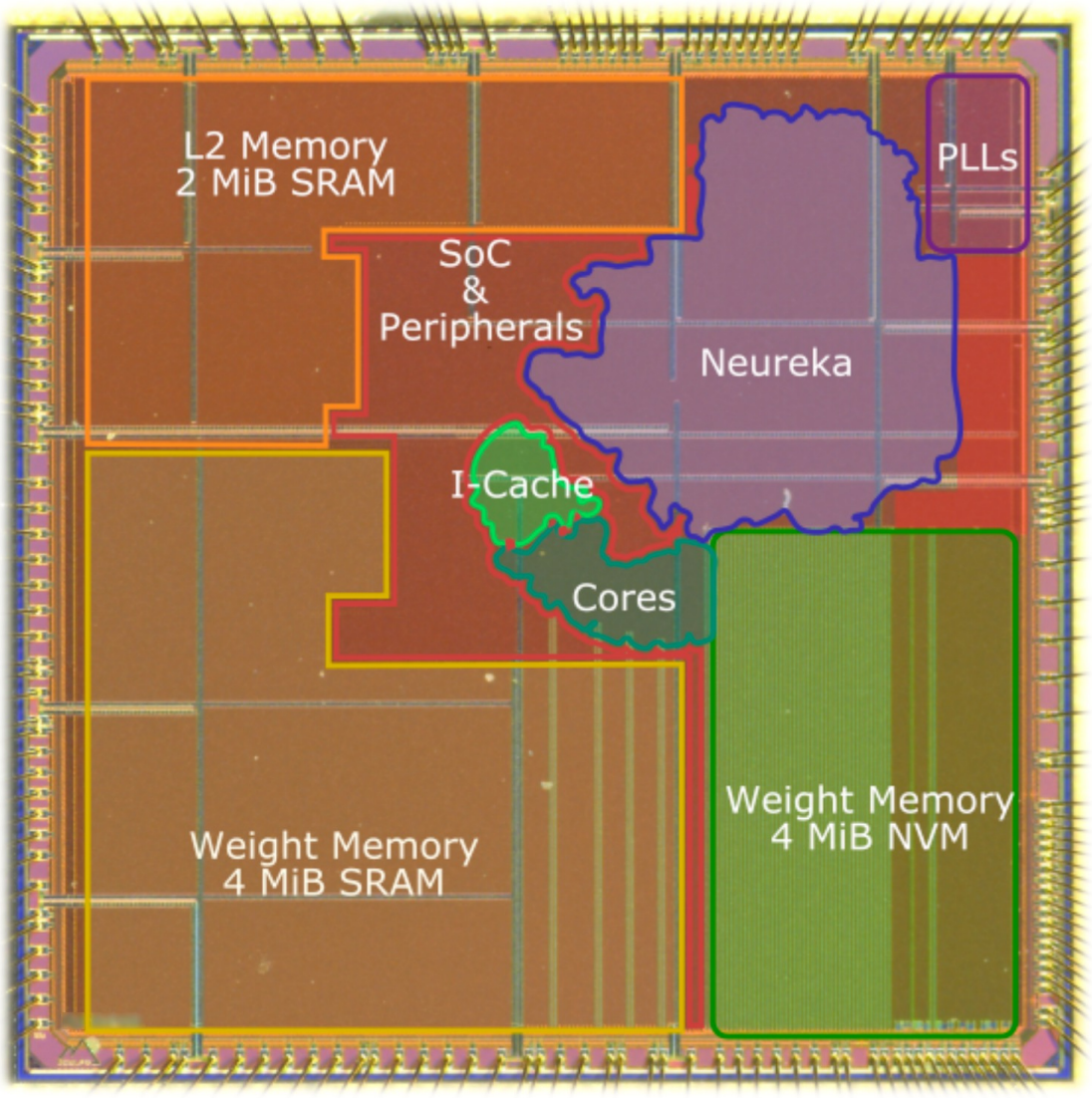
Siracusa SoC – prototype in TSMC 16nm

- **4mm x 4mm**
- A cornucopia of **memory**
  - 4 MiB of WMEM-NVM
  - 4 MiB of WMEM-SRAM
  - 2 MiB of L2 SRAM
  - 256 KiB of L1 TCDM

- Largest **NEureka** configuration
  - 6x6 = 36 Cores

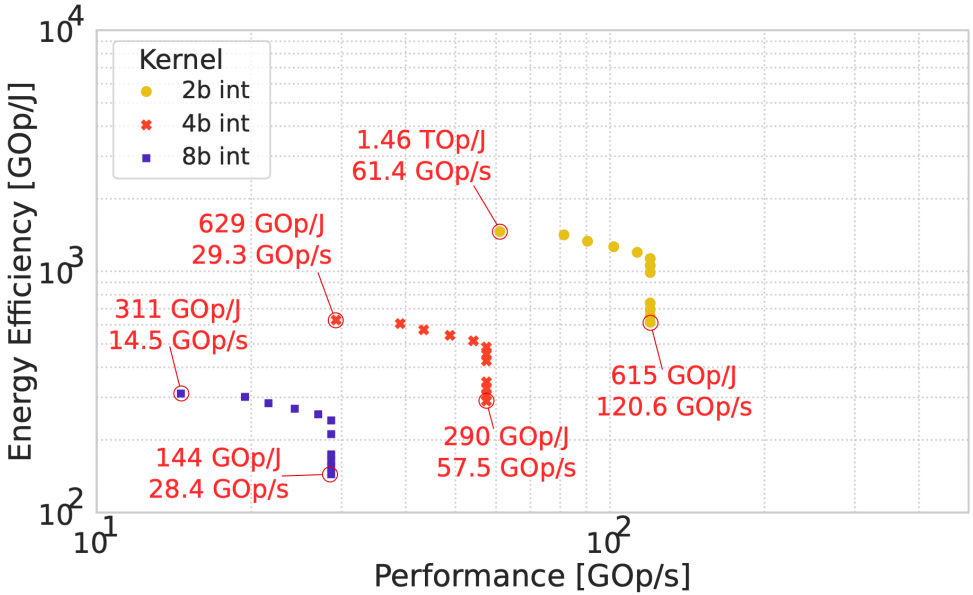
To appear at **ESSCIRC'23** (*M. Scherer et al.*)

Siracusa: A Low-Power On-Sensor RISC-V SoC for Extended Reality Visual Processing in 16nm CMOS

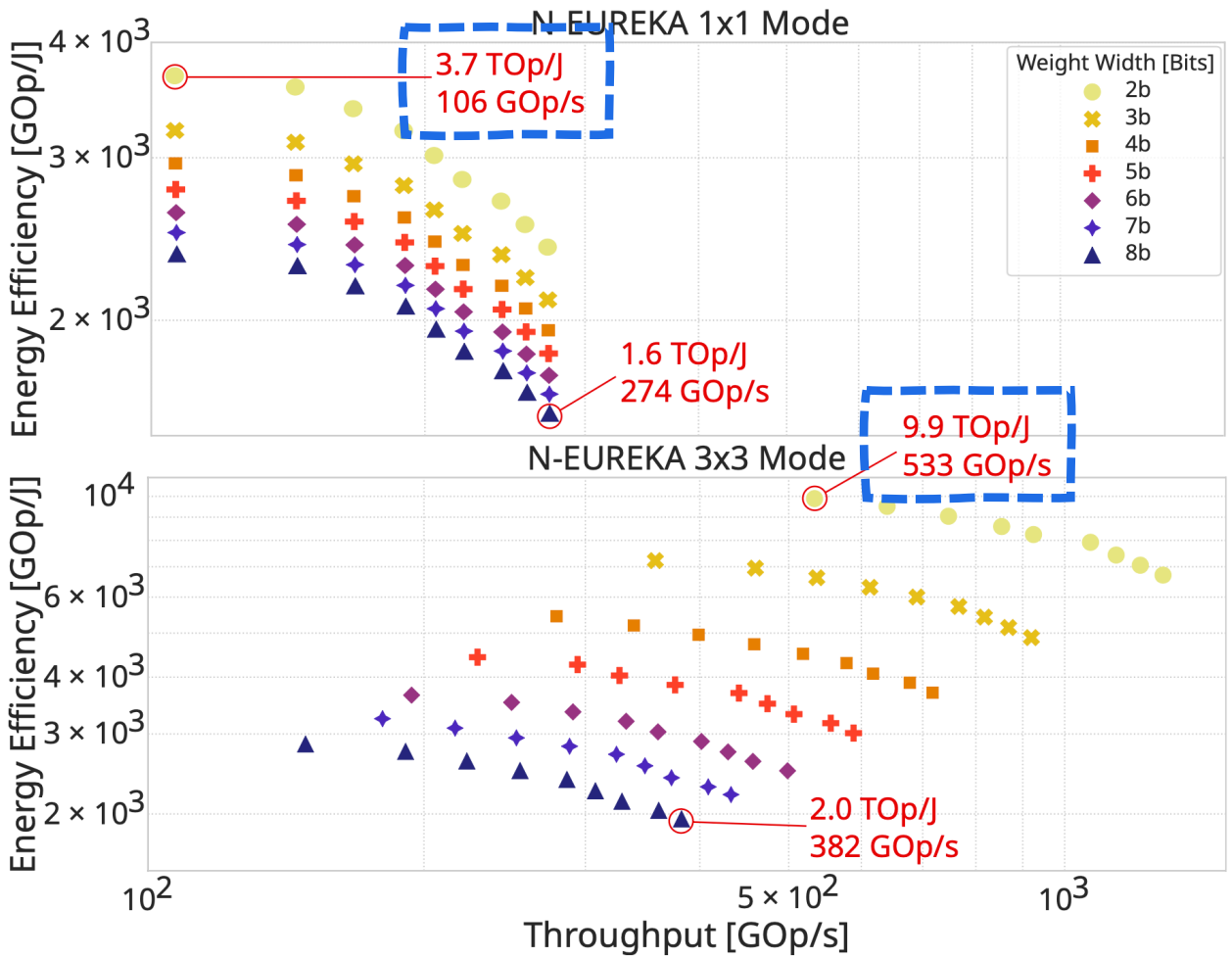


Siracusa Performance and Efficiency

RISC-V cores (GP, DSP, ...)



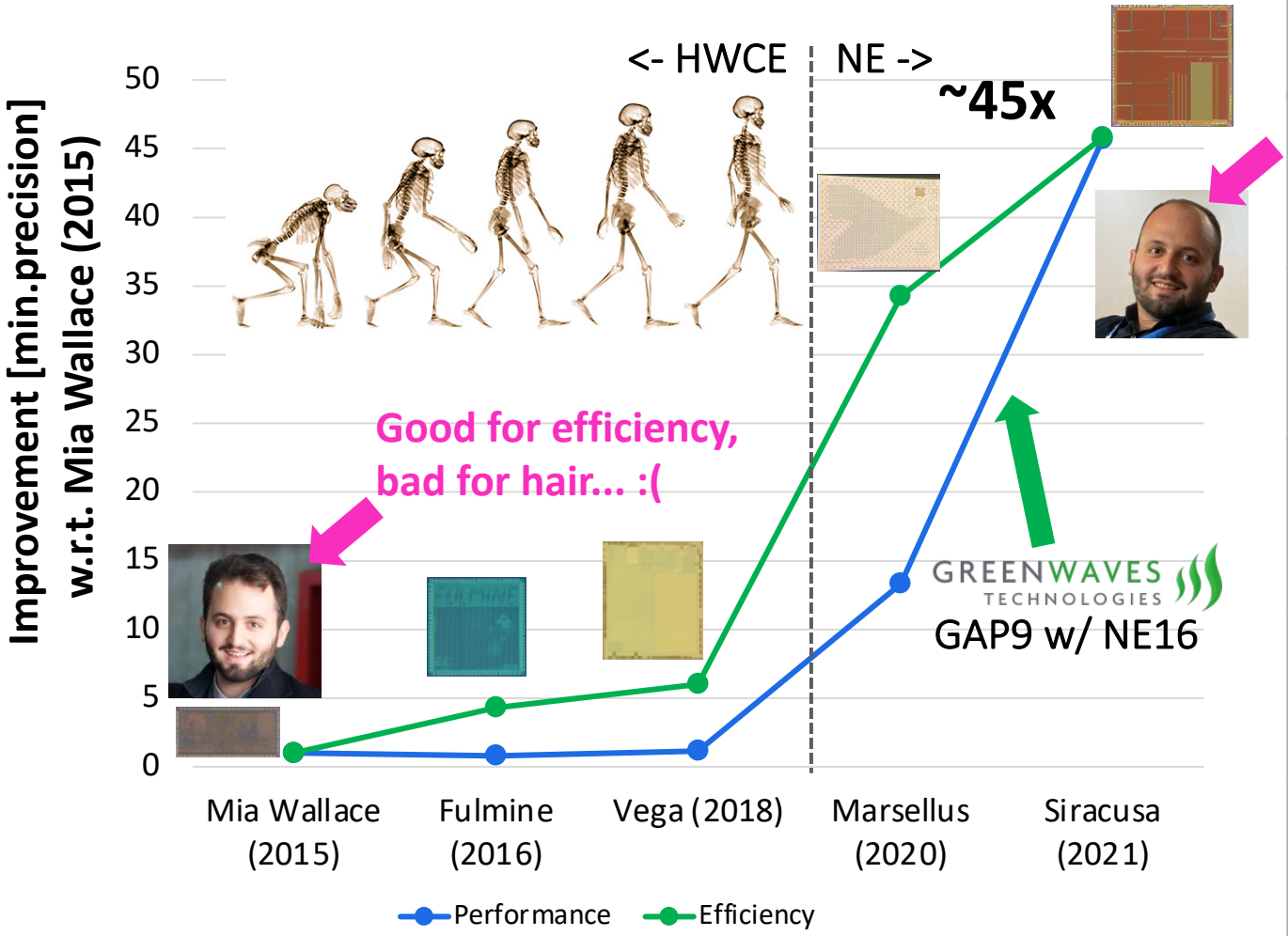
NEureka



Very near to out target

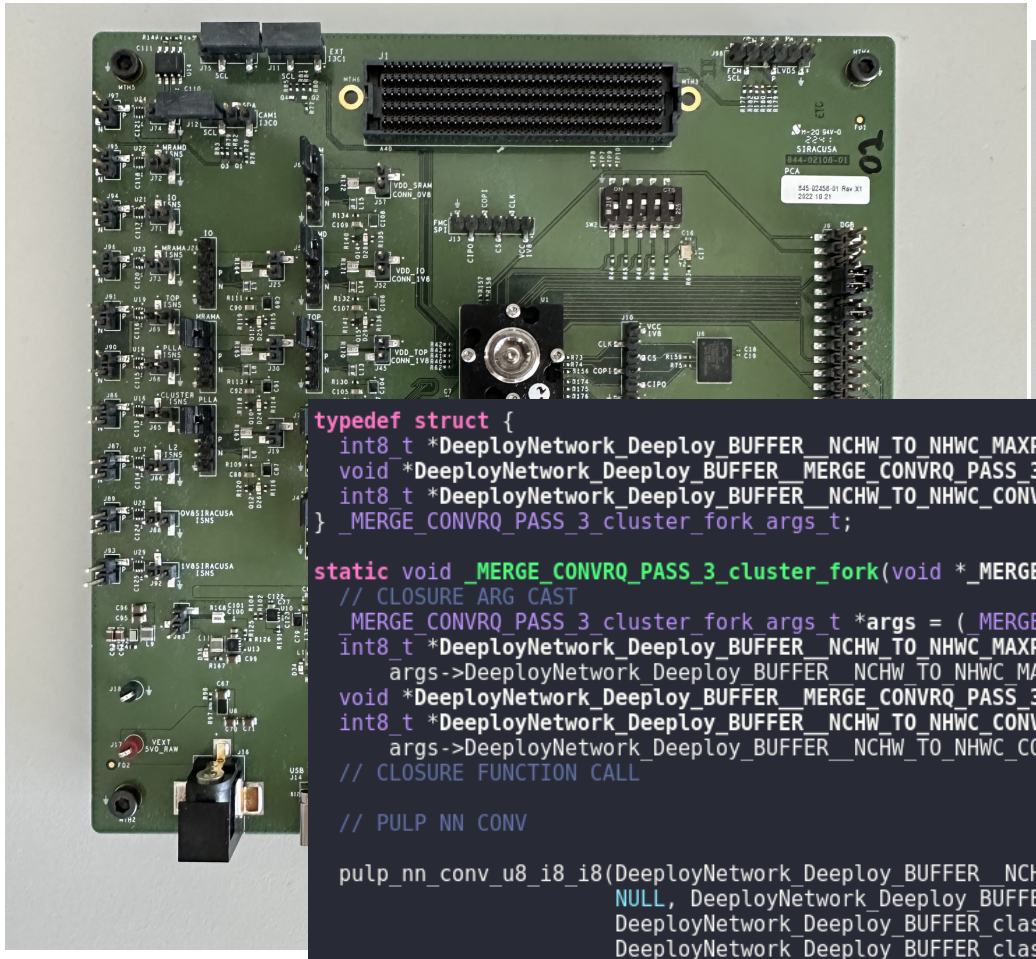
- hundreds of GOPS
- ~10 TOPS/W

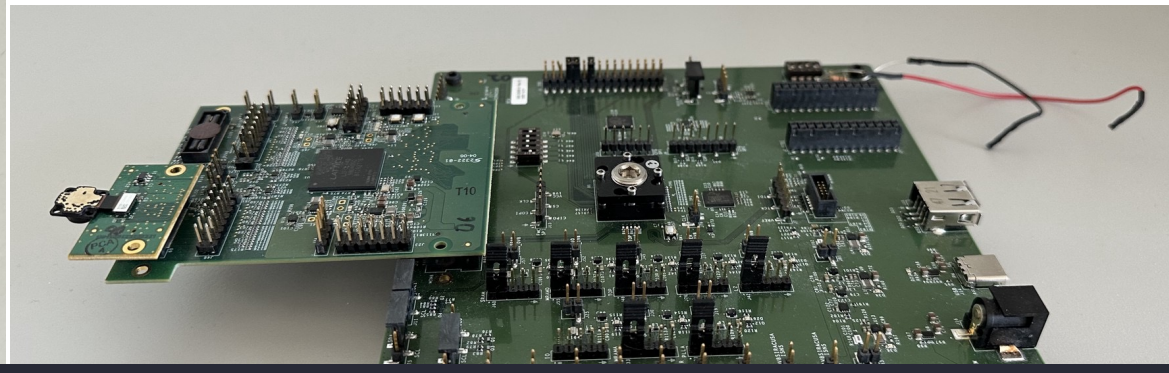
The Evolution of the Accelerator Species



Chip (year)	Tech (nm)	HW Accelerator
Mia Wallace (2015)	65	**HWCE** (gen. 1) Conv 5x5 16x16b
Fulmine (2016)	65	**HWCE** (gen. 2) Conv 5x5 4x16b
Vega (2018)	22	**HWCE** (gen. 4) Conv 3x3 8x8b
Marsellus (2020)	22	**RBE** (NE gen. 1) Conv 3x3, 1x1 2x8b
Siracusa (2021)	16	**NEureka** (NE gen. 2.5) Conv 3x3, 1x1, DWConv 3x3 2x8b

The real challenge: using it!





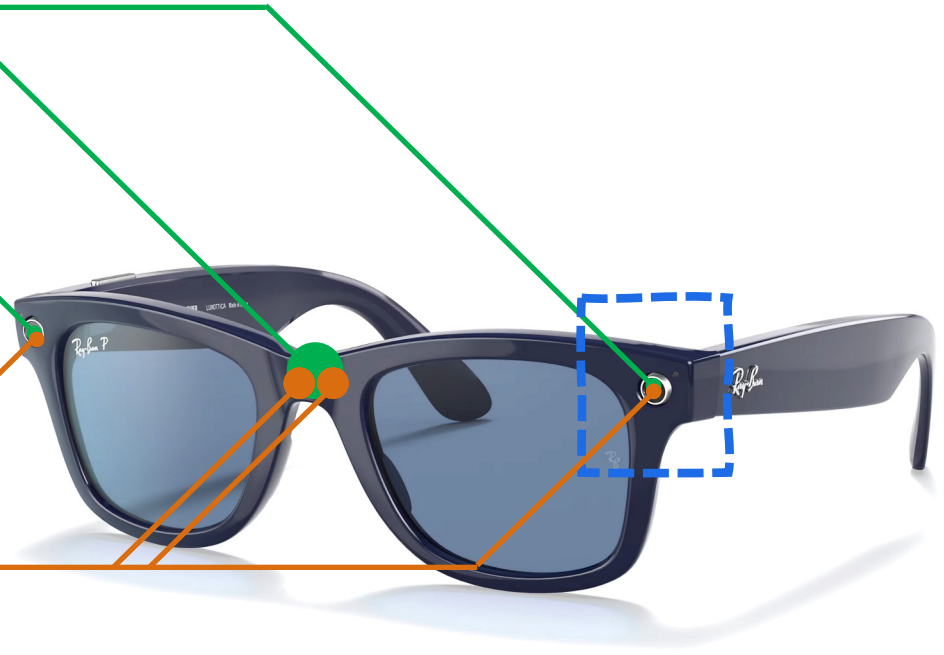
The need for automated Deployment tools: see Moritz+Alessio's presentation tomorrow!

```
typedef struct {
  int8_t *DeeployNetwork_Deeploy_BUFFER__NCHW_TO_NHWC_MAXPOOL_PASS_2_TransposeOut_Out;
  void *DeeployNetwork_Deeploy_BUFFER__MERGE_CONVRQ_PASS_3_buffer;
  int8_t *DeeployNetwork_Deeploy_BUFFER__NCHW_TO_NHWC_CONV_PASS_3_TransposeOut_Out;
} _MERGE_CONVRQ_PASS_3_cluster_fork_args_t;

static void _MERGE_CONVRQ_PASS_3_cluster_fork(void *_MERGE_CONVRQ_PASS_3_cluster_fork_args) {
  // CLOSURE ARG CAST
  _MERGE_CONVRQ_PASS_3_cluster_fork_args_t *args = (_MERGE_CONVRQ_PASS_3_cluster_fork_args_t *)_MERGE_CONVRQ_PASS_3_cluster_fork_args;
  int8_t *DeeployNetwork_Deeploy_BUFFER__NCHW_TO_NHWC_MAXPOOL_PASS_2_TransposeOut_Out =
      args->DeeployNetwork_Deeploy_BUFFER__NCHW_TO_NHWC_MAXPOOL_PASS_2_TransposeOut_Out;
  void *DeeployNetwork_Deeploy_BUFFER__MERGE_CONVRQ_PASS_3_buffer = args->DeeployNetwork_Deeploy_BUFFER__MERGE_CONVRQ_PASS_3_buffer;
  int8_t *DeeployNetwork_Deeploy_BUFFER__NCHW_TO_NHWC_CONV_PASS_3_TransposeOut_Out =
      args->DeeployNetwork_Deeploy_BUFFER__NCHW_TO_NHWC_CONV_PASS_3_TransposeOut_Out;
  // CLOSURE FUNCTION CALL

  // PULP NN CONV

  pulp_nn_conv_u8_i8_i8(DeeployNetwork_Deeploy_BUFFER__NCHW_TO_NHWC_MAXPOOL_PASS_2_TransposeOut_Out, DeeployNetwork_Deeploy_BUFFER__MERGE_CONVRQ_PASS_3_buffer,
                        NULL, DeeployNetwork_Deeploy_BUFFER__NCHW_TO_NHWC_CONV_PASS_3_TransposeOut_Out,
                        DeeployNetwork_Deeploy_BUFFER_classifier__QL_REPLACED__INTEGERIZE_PACT_CONV2D_PASS_3_weight,
                        DeeployNetwork_Deeploy_BUFFER_classifier__QL_REPLACED__INTEGERIZE_SIGNED_ACT_PASS_0_mul,
                        DeeployNetwork_Deeploy_BUFFER_classifier__QL_REPLACED__INTEGERIZE_SIGNED_ACT_PASS_0_add, 1, 18, 4, 4, 64, 4, 4, 128, 3, 3, 1, 1, 1, 1,
                        1, 1, 1, 1);
}
```

Back to the vision!

Distributed, on-sensor computing

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



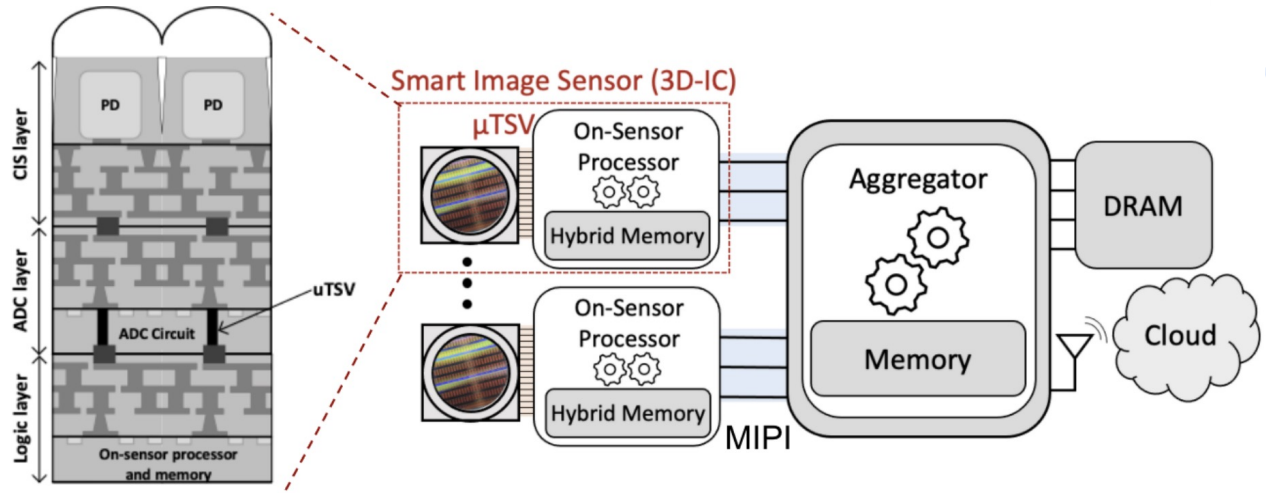
Acceleration

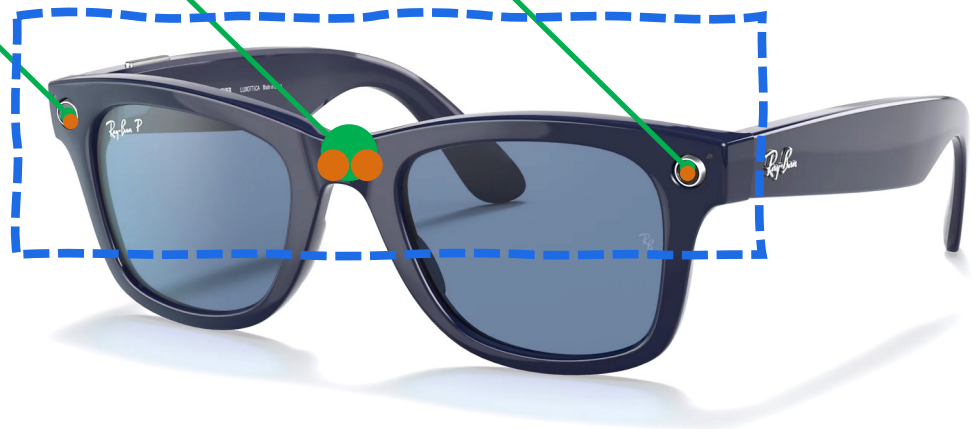
- On-chip **NVM** for DNN weights
- L1 **HW acceleration** for DNNs
- L0 **acceleration** for diverse processing

Back to the vision!

Distributed, on-sensor computing

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

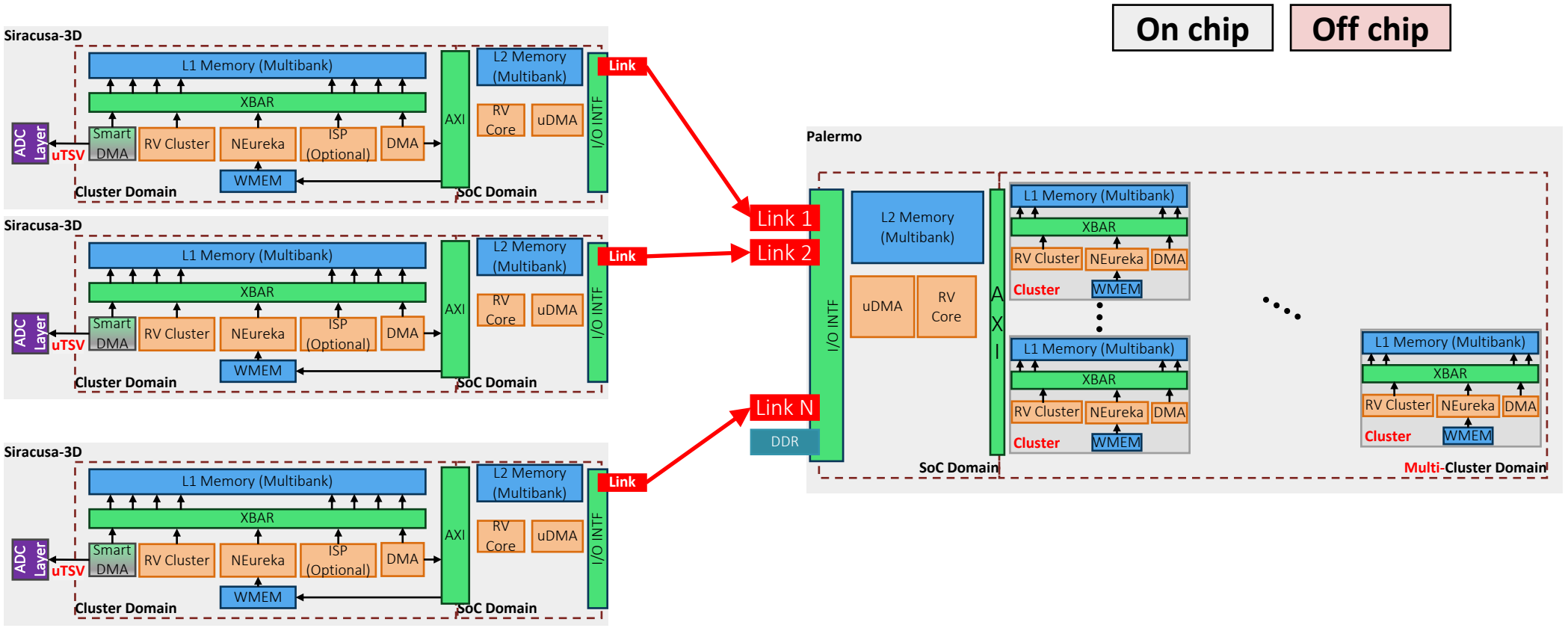




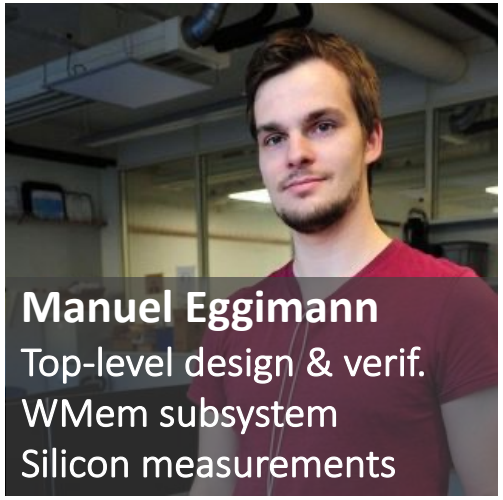
From focus on single node towards distributed network of on-sensor nodes

[J. Gomez et al., Distributed On-Sensor Compute System for AR/VR Devices: A Semi-Analytical Simulation Framework for Power Estimation]

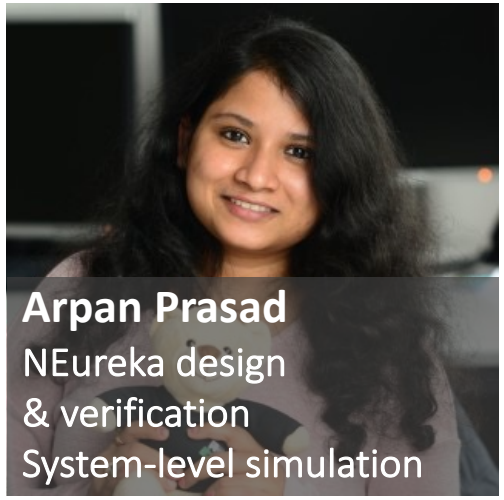
Distributed on-sensor computing simulation with GVSOC



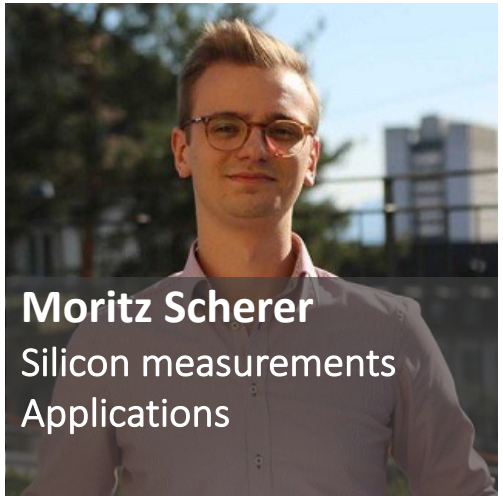
Siracusa Team @ ETH / UNIBO



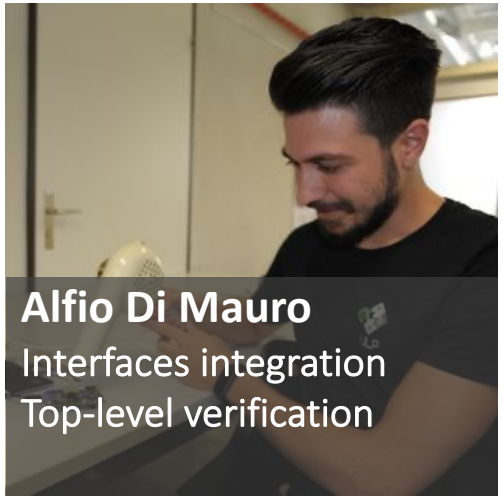
Manuel Eggimann
Top-level design & verif.
WMem subsystem
Silicon measurements



Arpan Prasad
NEureka design
& verification
System-level simulation



Moritz Scherer
Silicon measurements
Applications



Alfio Di Mauro
Interfaces integration
Top-level verification

Davide Rossi
Siracusa architecture

Francesco Conti
NEureka architecture
Siracusa architecture

Luca Benini
Siracusa architecture
Project lead

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.
- [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] 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.

ETH zürich

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA