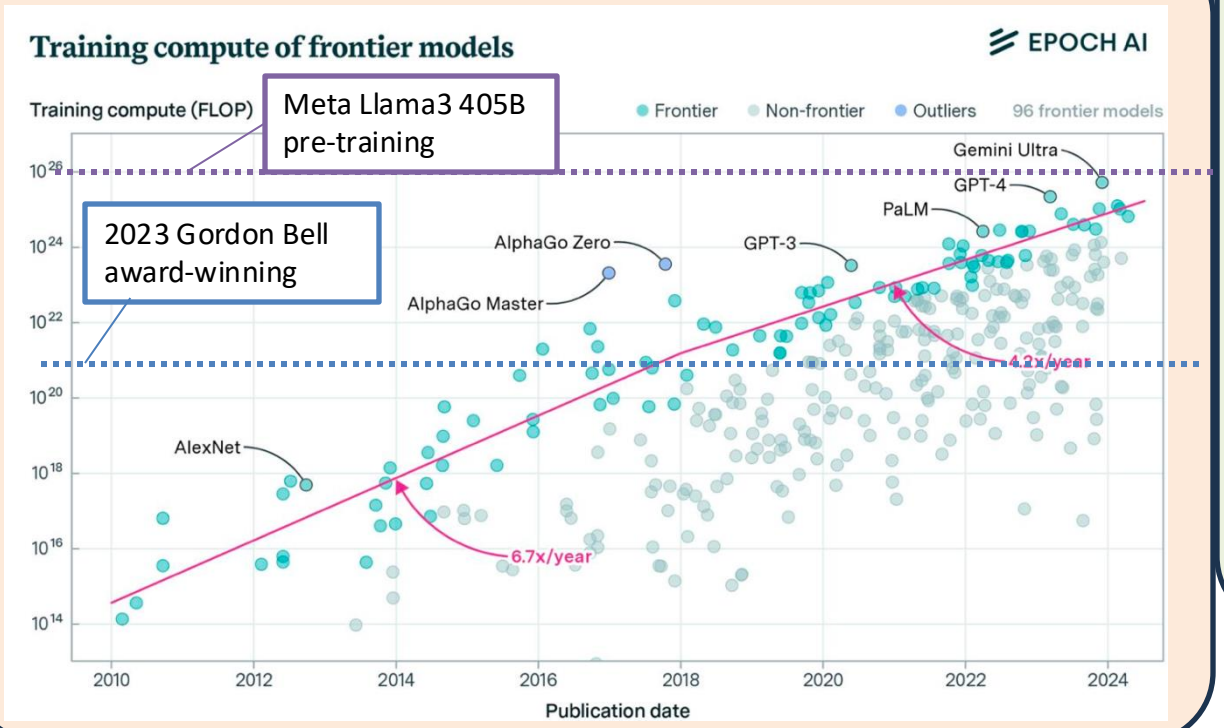
New Computing Architecture for AI – The European Ecosystem

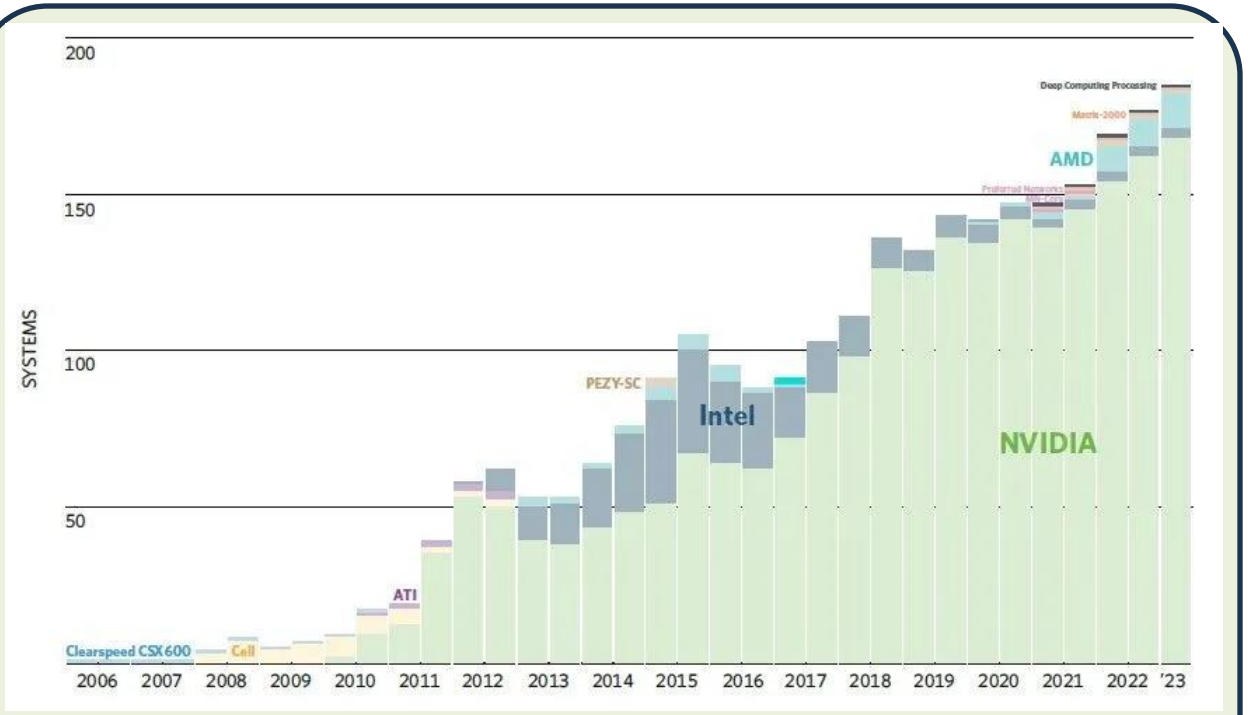
Prof. Andrea Bartolini

The Department of Electrical, Electronic and Information Engineering (**DEI**) – <a.bartolini@unibo.it>

AI Architectures some trends…

Training Compute of Frontier AI Models Grows by 4-5x per Year, Sevilla and Roldán (2024)





Bommasani, Rishi, et al. "**On the Opportunities and Risks of Foundation Models**." *Center for Research on Foundation Models (CRFM), Stanford Institute for Human-Centered Artificial Intelligence (HAI).*

Meta

The Llama 3 Herd of Models

Llama Team, AI @ Meta[1]

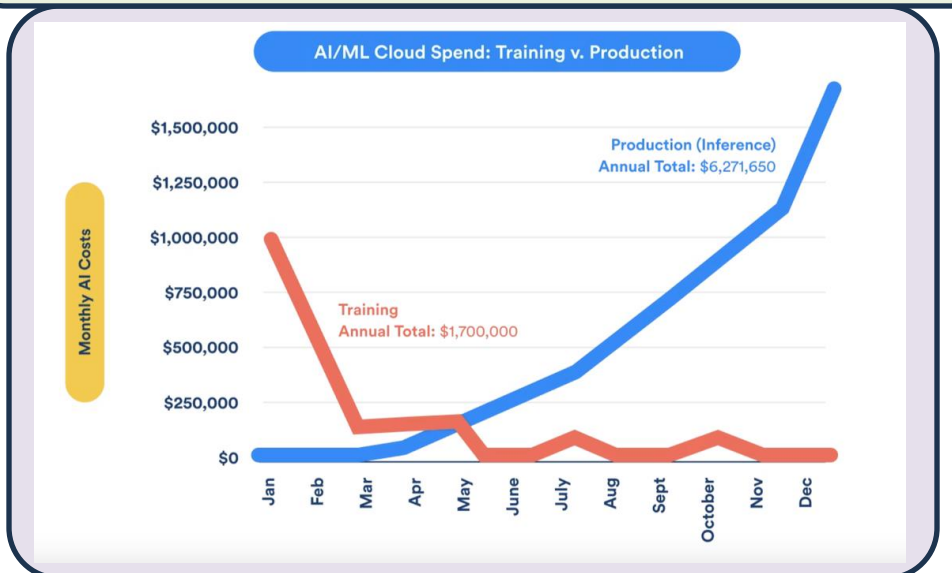
[1]A detailed contributor list can be found in the appendix of this paper.

Impact of environmental factor on training performance at scale – diurnal 1-2% throughput variation

GPUs	TP	CP	PP	DP	Seq. Len.	Batch size/DP	Tokens/Batch	TFLOPs/GPU	BF16 MFU
8,192	8	1	16	64	8,192	32	16M	430	43%
16,384	8	1	16	128	8,192	16	16M	400	41%
16,384	8	16	16	4	131,072	16	16M	380	38%

Table 4 Scaling configurations and MFU for each stage of Llama 3 405B pre-training. See text and Figure 5 for descriptions of each type of parallelism.

https://arxiv.org/abs/2407.21783



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

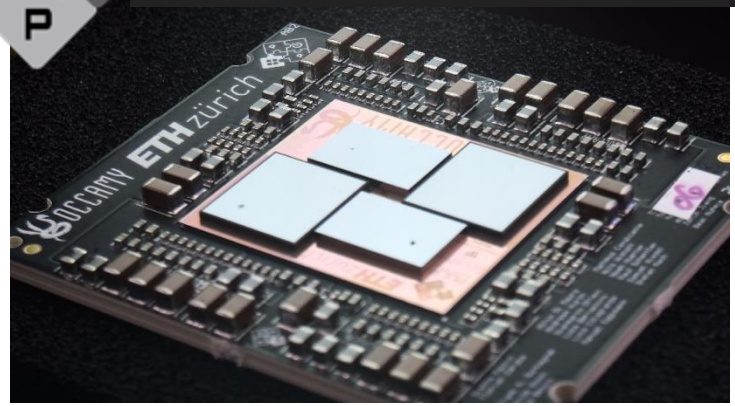
European Ecosystem of AI Platforms

European Processor Initiative

PULP Platform

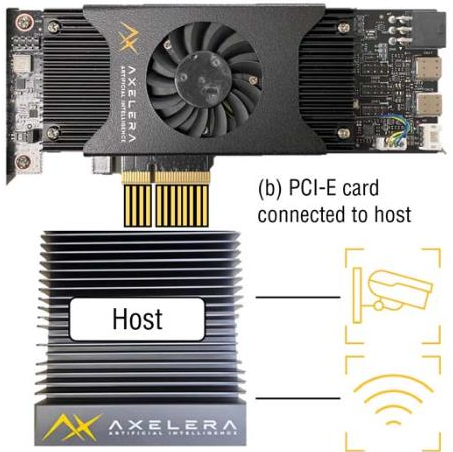
Open Source Hardware, the way it should be!

https://github.com/pulp-platform/occamy



Occamy:
70mm2@GF12
Up to **686 GFLOPS**
89% FPU util.
40 GFLOPS/W

RISC-V®



Metis AIPU
209.6 TOPS
15 to 82 TOPS/W

AXELERA
ARTIFICIAL INTELLIGENCE

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

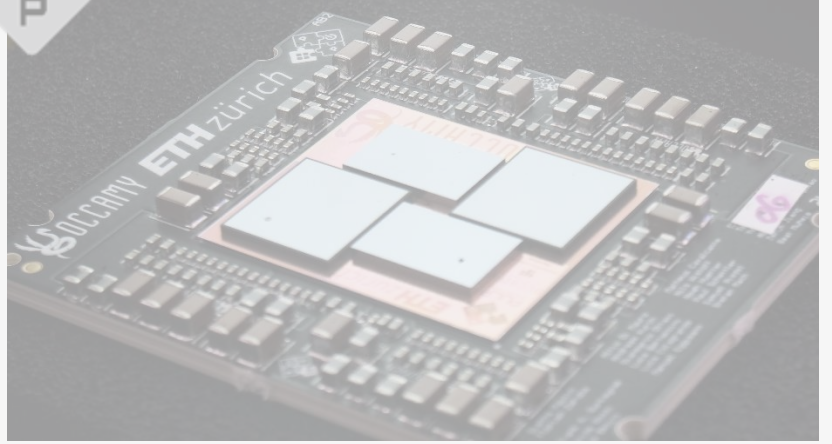
European Ecosystem of AI Platforms

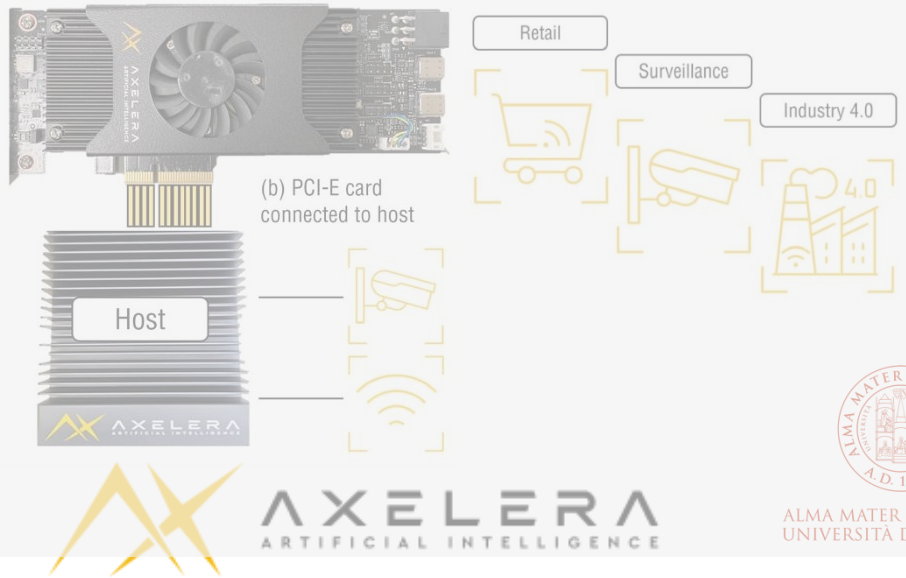
European Processor Initiative

RISC-V®

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

https://github.com/pulp-platform/occamy

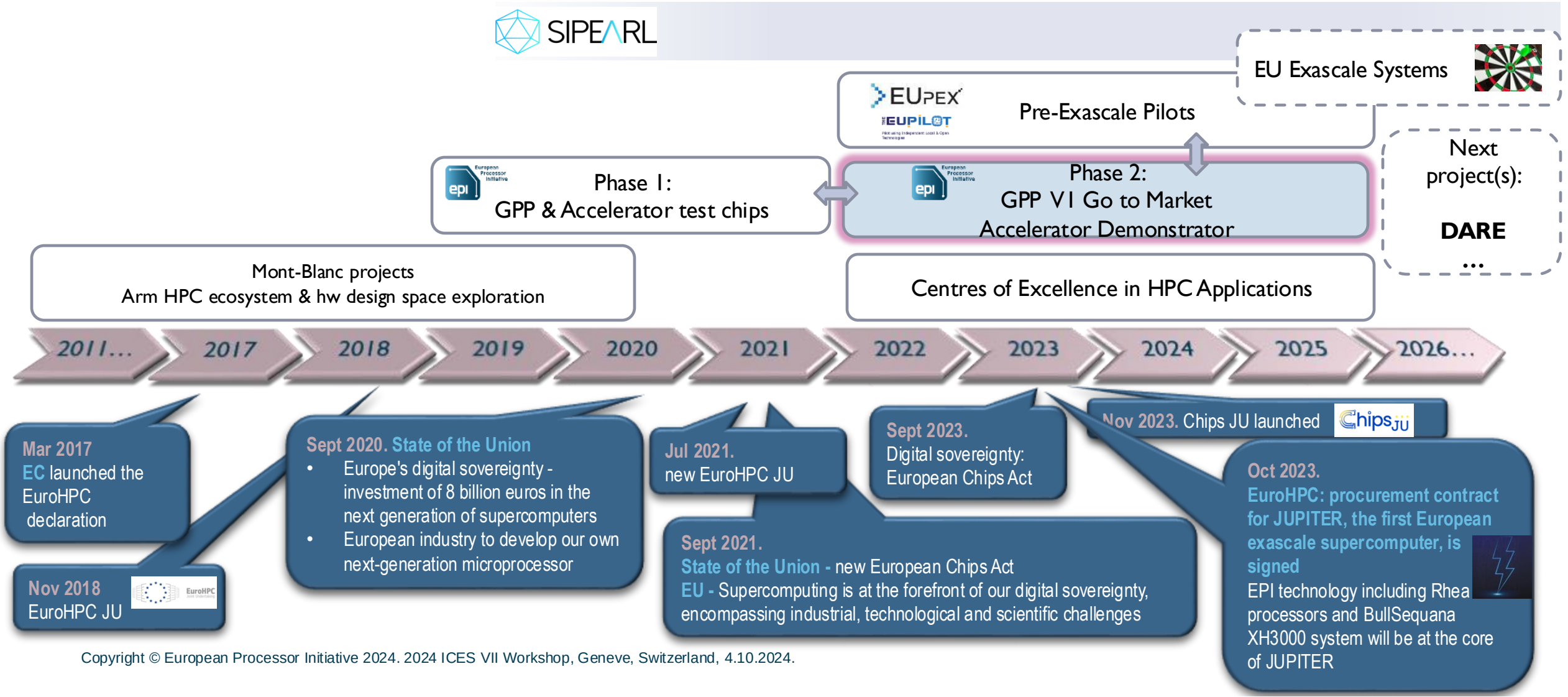




AXELERA
ARTIFICIAL INTELLIGENCE

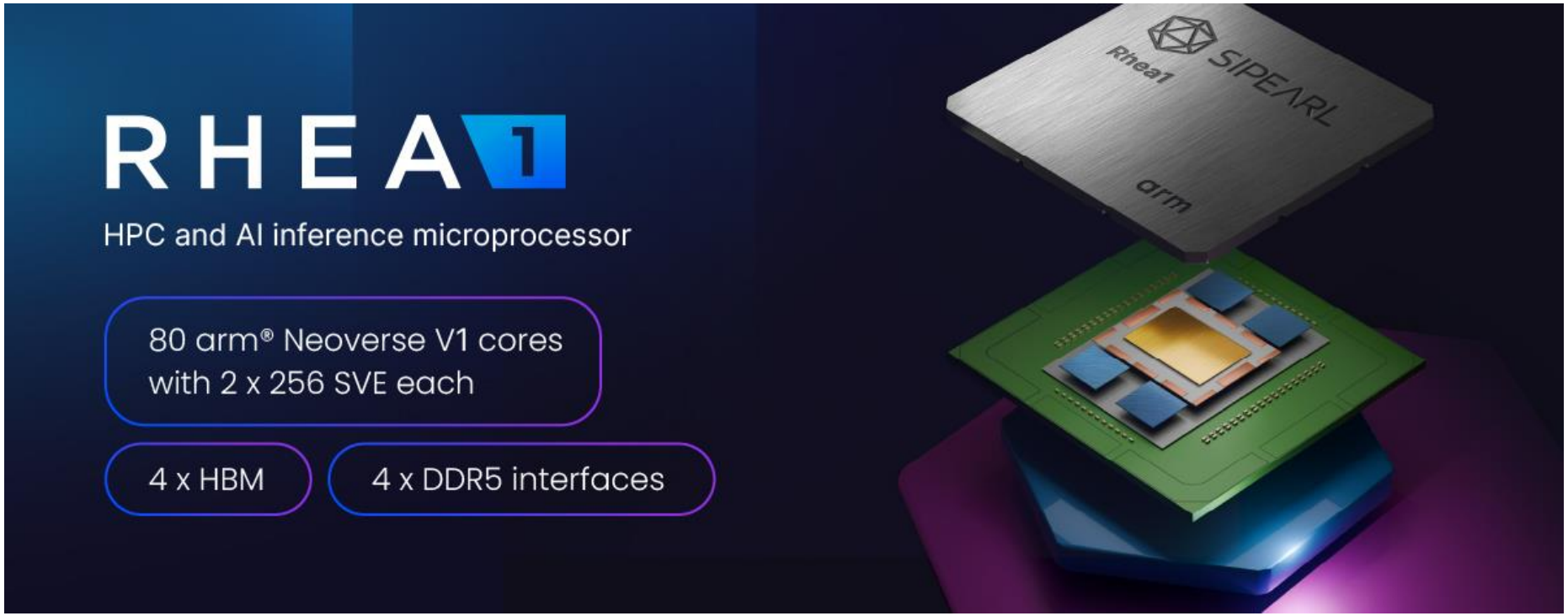
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

THE EU HPC & EPI TIMELINE

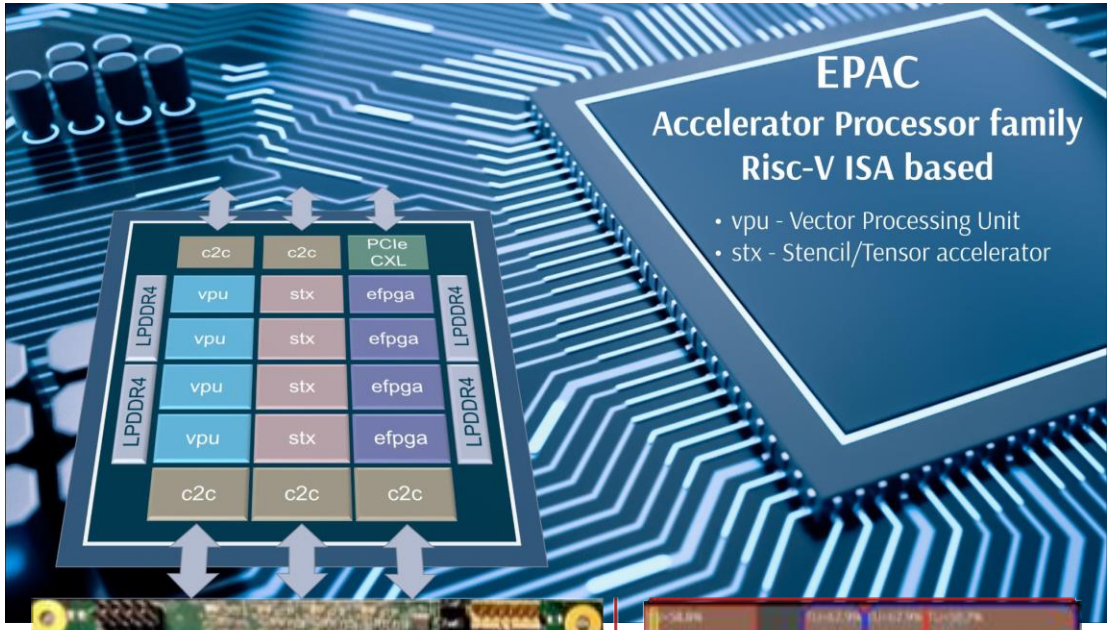


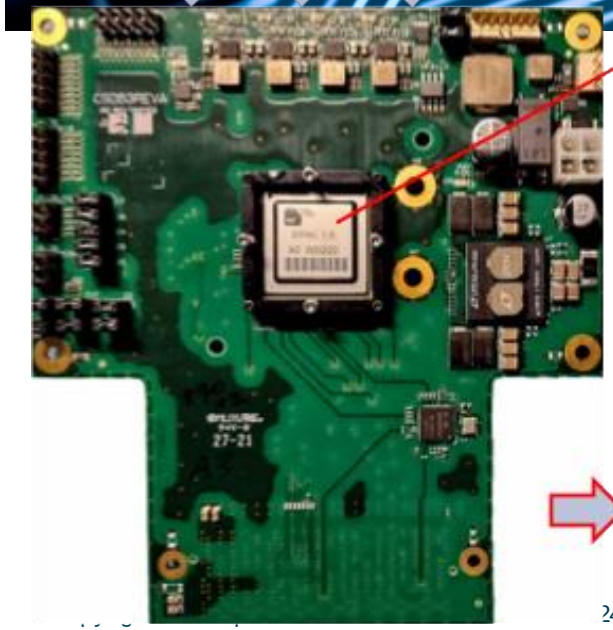
EPI PROJECT FACTSHEET

- Currently in Phase 2 (2022-2025)
- Consortium of 27 strategically chosen key European academic and industrial partners
- Total budget: 70 M€
- Funded by EuroHPC JU (50%)
  - and co-funded by Croatia, France, Germany, Greece, Italy, the Netherlands, Portugal, Spain, Sweden and Switzerland



EPAC VISION AND CONTRIBUTIONS





- **VEC -** Self-hosted RISC-V CPU + wide VPU (256 double elements) supporting RVV 0.7.1 / 1.0
- **STX -** RISC-V CPU + specific cores for stencil and neural network computation
- **VRP -** RISC-V CPU with support for variable precision arithmetic (data size up to 512 bit)
- **eFPGA -** On-chip reconfigurable logic
- **Ziptillion -** IP compressing/decompressing data to/from the main memory
- **KVX -** FPGA demonstrator of the Kalray RISC-V CPU targeting HPC and ML

EPAC 1.5 test chip back, brought-up and running well

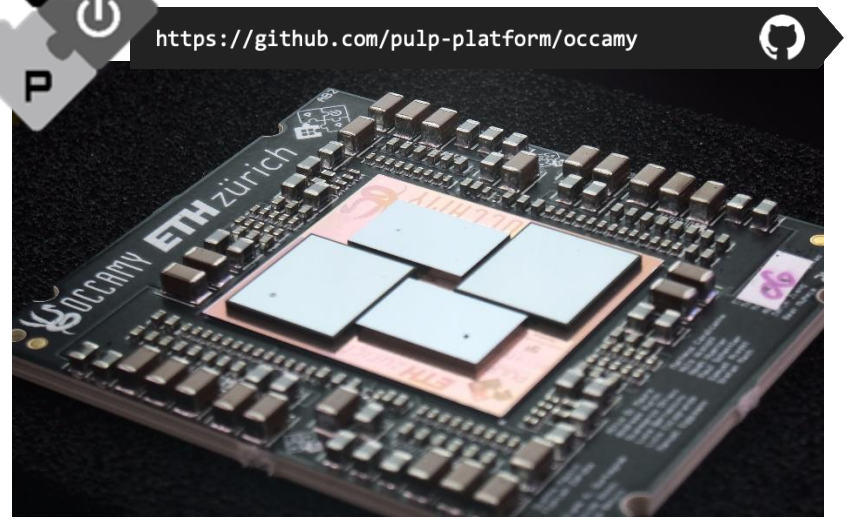
European Ecosystem of AI platforms

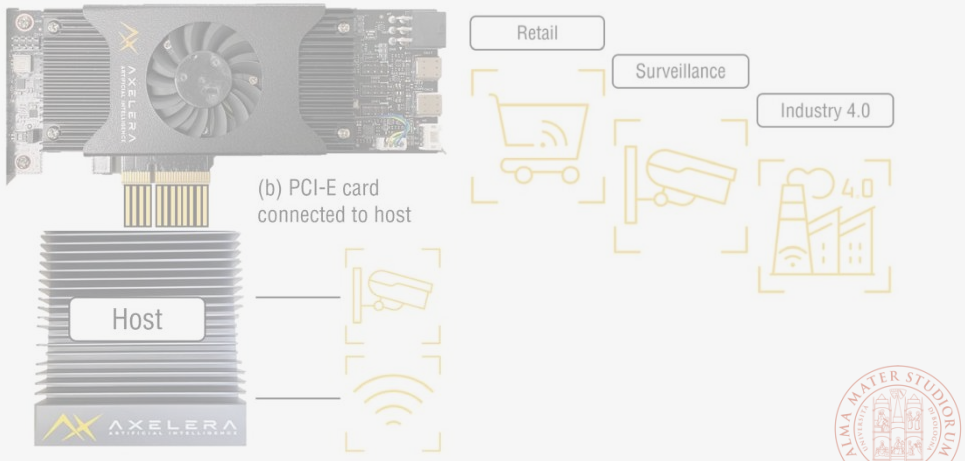
European Processor Initiative

RISC-V®

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

https://github.com/pulp-platform/occamy



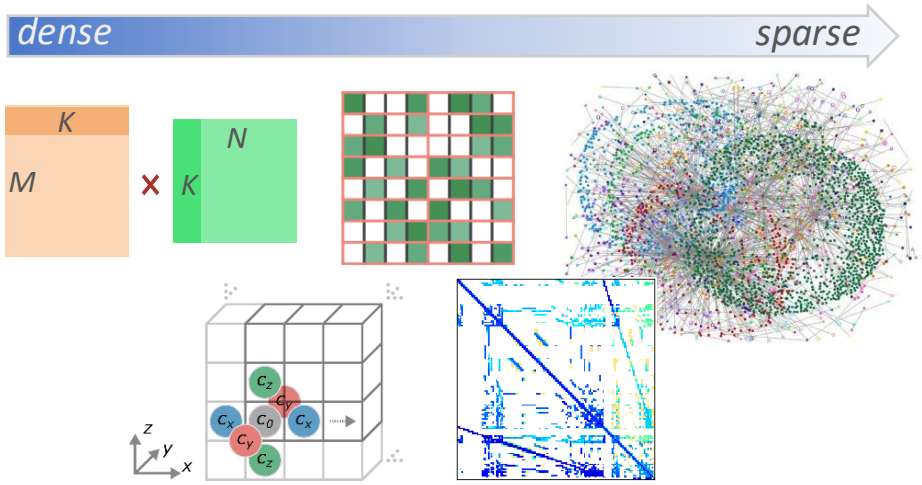


AXELERA
ARTIFICIAL INTELLIGENCE

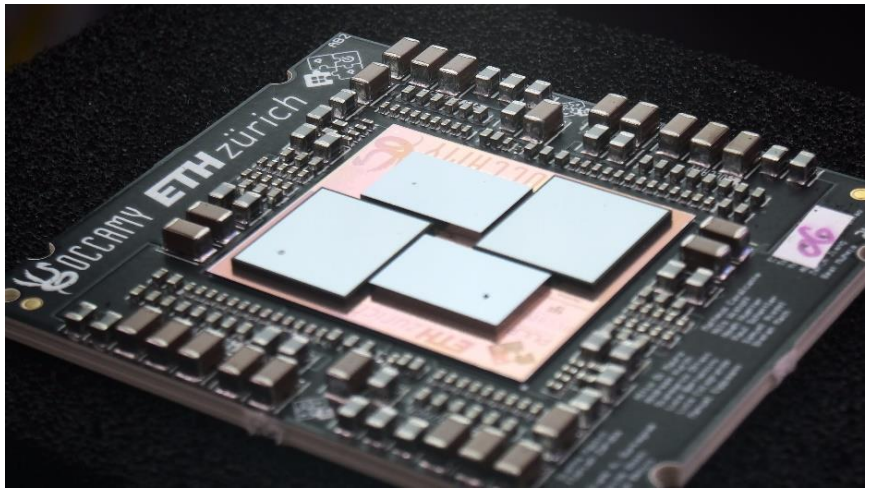
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Towards A *Flexible* Manycore RV Accelerator

- **Gen-AI workloads increasingly mix *dense* and *sparse* computations**
  - Dense: stencils, encoding…
  - Sparse: weight and activation sparsification in CNNs, DNNs, LLMs, as well as graph NNs



➢ **Next-generation systems must handle *both* compute types efficiently**
  - Need flexibility in quantization and sparsification
  - Accelerators focus on only one or are too *specialized* / *inflexible* to be future-proof

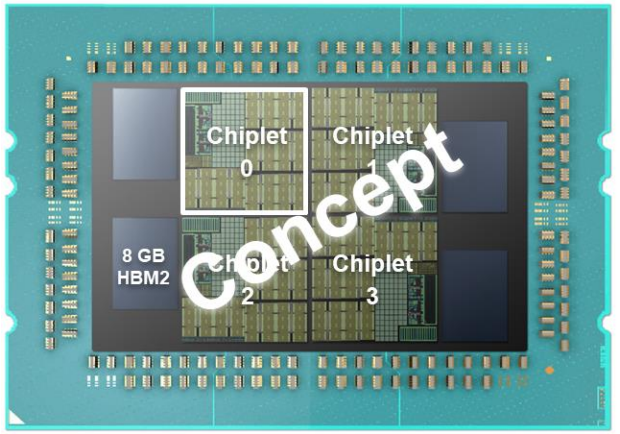


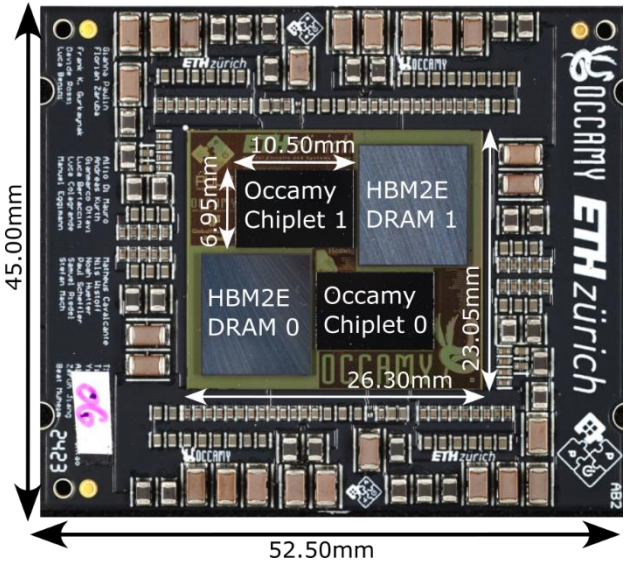
- **Occamy, a *flexible* RV chiplet scalable manycore for efficient *sparse and dense* Gen-AI**

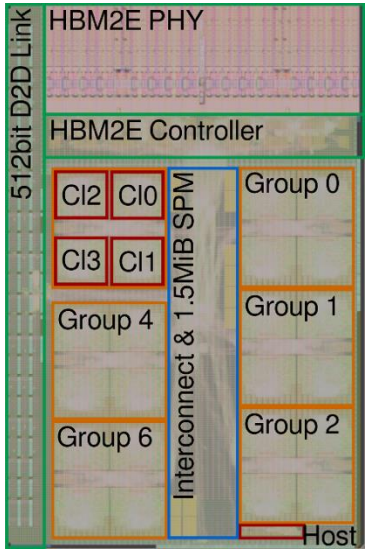
Paulin et al., «Occamy: A 432-Core 28.1 DP-GFLOP/s/W 83% FPU Utilization Dual-Chiplet, Dual-HBM2E RISC-V-based Accelerator for Stencil and Sparse Linear Algebra Computations with 8-to-64-bit Floating-Point Support in 12nm FinFET» https://arxiv.org/pdf/2406.15068

Occamy: Inception and key Figures

- **From concept to tapeout in 15 months**
  - Manticore concept at Hot Chips 2020 → GF challenge: take concept to prototype (multi-chiplet in GF12)
  - Kickoff in April 2021, ≈25 people, up to 10 full-time
  - Tapeout July 2022 (GF12) and September 2022 (GF65)
  - GlobalFoundries, Synopsys, Rambus, Micron, Avery
- **2.5D assembly by IZM (Fraunhofer)**
  - February to September 2023
  - Received two *early* samples in August 2023
- **High-complexity multi-chiplet prototype**
  - 73mm^2, 600 MGE each chiplet
  - 606mm^2 65nm passive interposer
  - RO4350B, low-CTE, high stability 12-layer PCB





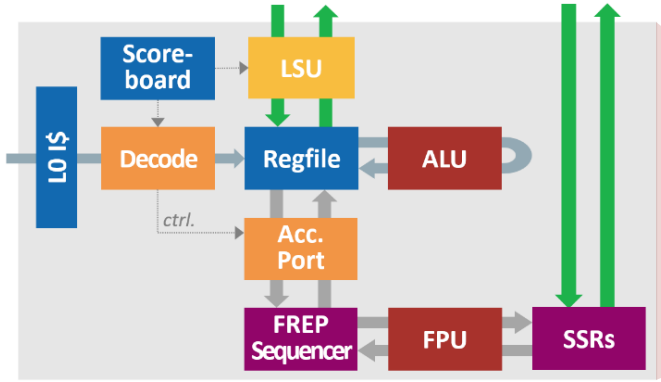


ETH zürich ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

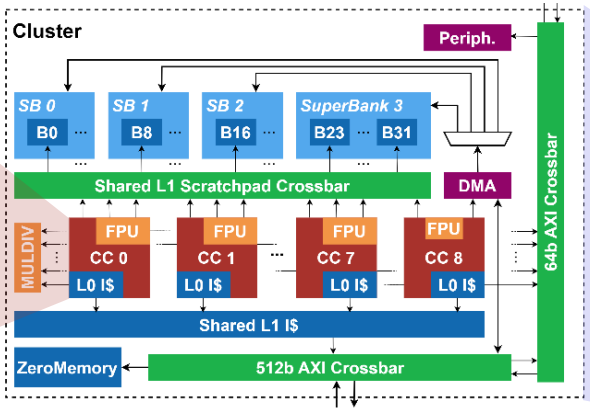
Paulin et al., «Occamy: A 432-Core 28.1 DP-GFLOP/s/W 83% FPU Utilization Dual-Chiplet, Dual-HBM2E RISC-V-based Accelerator for Stencil and Sparse Linear Algebra Computations with 8-to-64-bit Floating-Point Support in 12nm FinFET» https://arxiv.org/pdf/2406.15068

Achieving Scale through Hierarchical Design

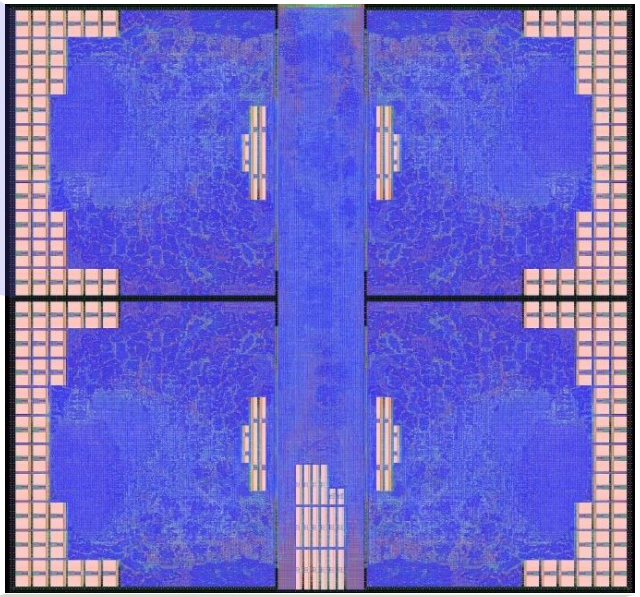
Snitch Core



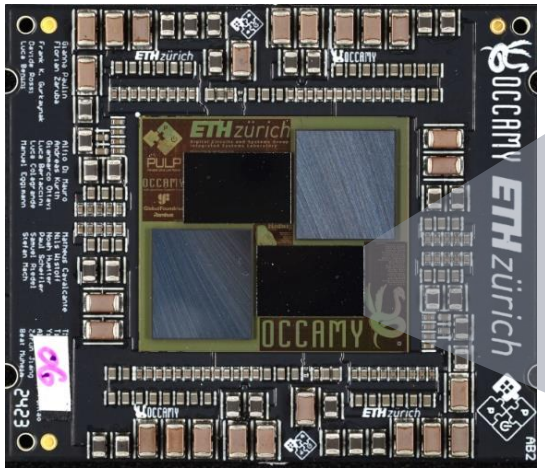
Snitch Cluster



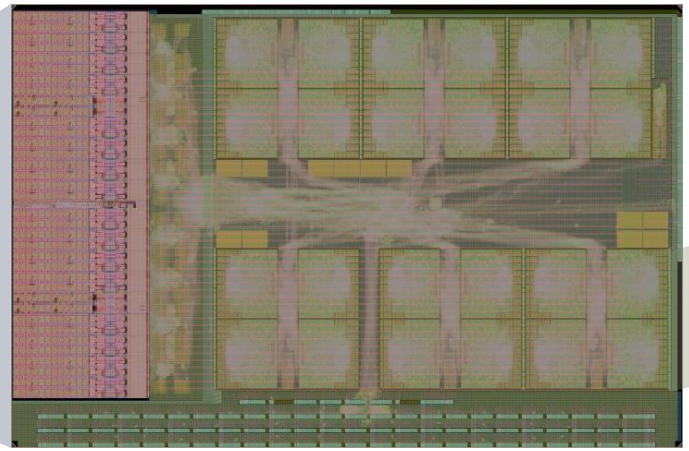
Occamy Group



Occamy System



Occamy Chiplet



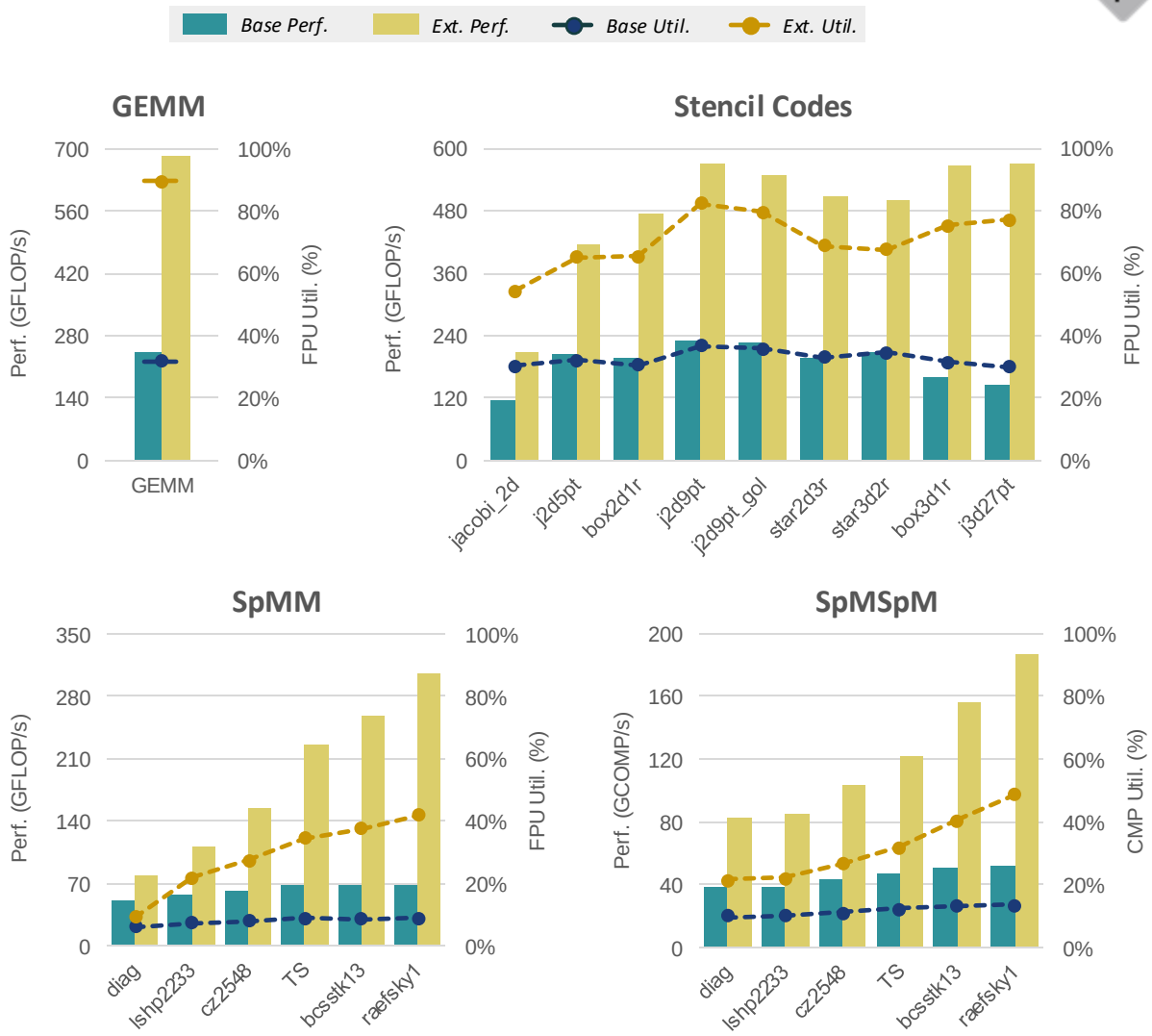
Paulin et al., «Occamy: A 432-Core 28.1 DP-GFLOP/s/W 83% FPU Utilization Dual-Chiplet, Dual-HBM2E RISC-V-based Accelerator for Stencil and Sparse Linear Algebra Computations with 8-to-64-bit Floating-Point Support in 12nm FinFET» https://arxiv.org/pdf/2406.15068

Key challenge: tolerate memory latency at local and global level →never wait for memory!

Occamy Performance and FPU Utilization

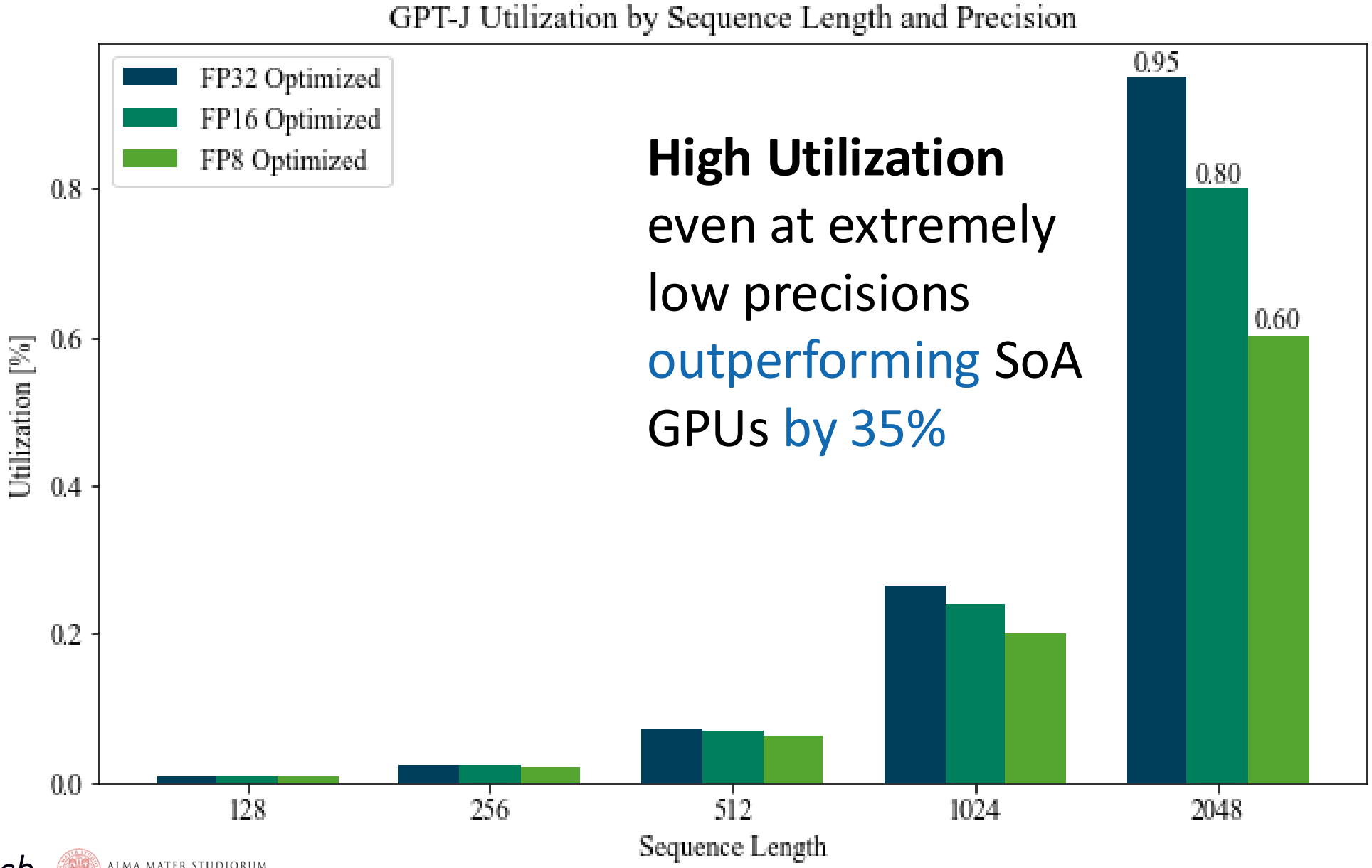
- **Mixed workloads @ 1GHz, 0.8 V, 25°C**
  - RV32G baseline vs. code using ISA extensions
  - All workloads use FP64 data, int16 indices
  - Sparse LHS real-world matrices, RHS 1% density

➢ ***Near-ideal* dense, *leading* sparse perf.**
  - ***GEMM***: **686 GFLOP/s – 40 GFLOP/s/W**, **89%** FPU util. competitive with GPUs
  - ***Stencils:*** Up to **571 GFLOP/s - 28 GFLOP/s/W**, **83%** FPU util. (**≥15%** more than GPU code gens)
  - ***SpMM:*** Up to **307 GFLOP/s - 16 GFLOP/s/W**, **42%** FPU util. (**≥1.6×** more than sd. LA on GPUs)
  - ***SpMSpM:*** Up to **187 GCOMP/s - 17 GCOMP/s/W**, **49%** index comparator utilization



Paulin et al., «Occamy: A 432-Core 28.1 DP-GFLOP/s/W 83% FPU Utilization Dual-Chiplet, Dual-HBM2E RISC-V-based Accelerator for Stencil and Sparse Linear Algebra Computations with 8-to-64-bit Floating-Point Support in 12nm FinFET» https://arxiv.org/pdf/2406.15068

Large Language Inference and Training



High Utilization even at extremely low precisions outperforming SoA GPUs by 35%

European Ecosystem of AI platforms

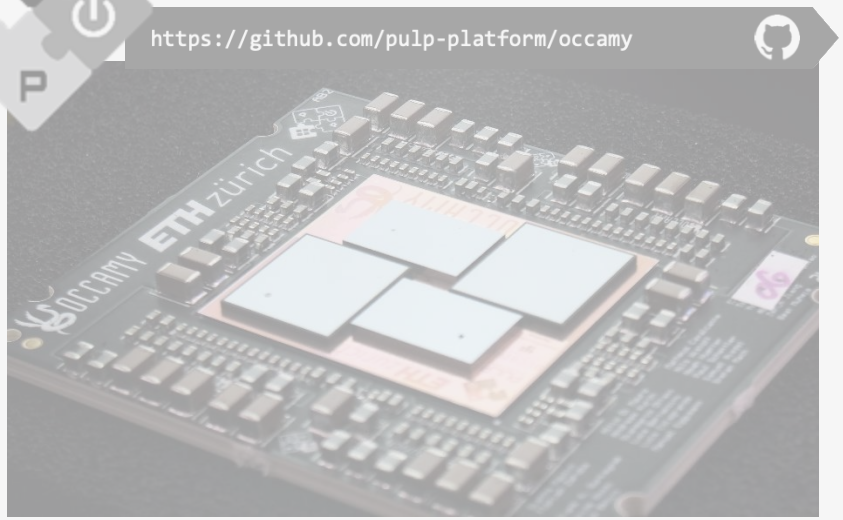
European Processor Initiative

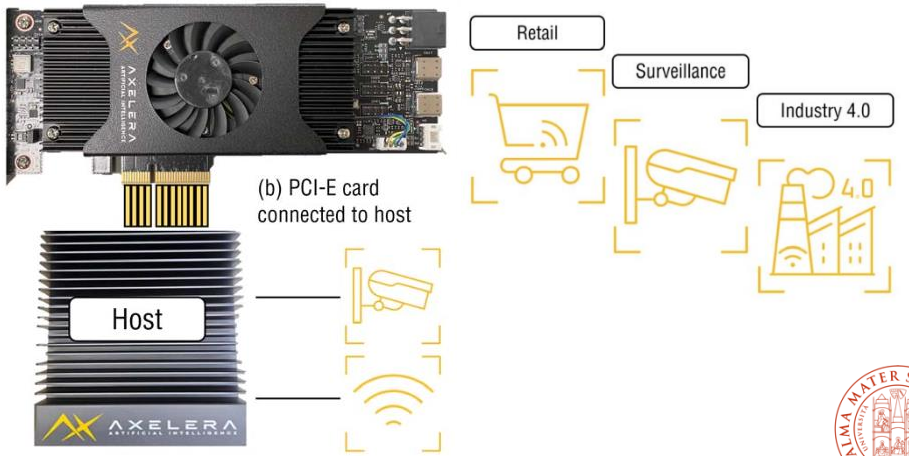
RISC-V®

PULP Platform

Open Source Hardware, the way it should be!

https://github.com/pulp-platform/occamy





AXELERA ARTIFICIAL INTELLIGENCE

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

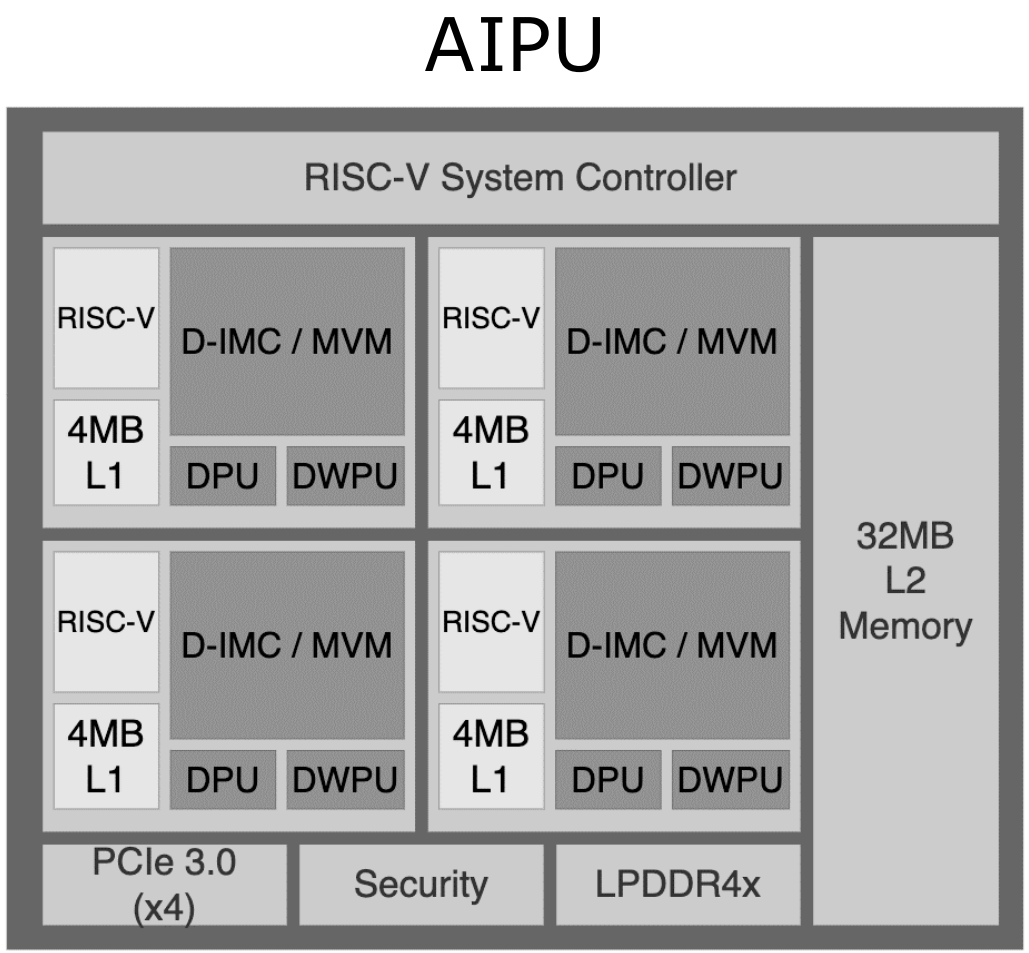
Metis - AI Platform

- AI Edge inference accelerator
  - M.2 module or PCIe card
- Metis AIPU executes all tasks of an AI workload
  - Offload complete network(s)
  - Not just individual layers
- Easy-to-use software stack
  - Voyager SDK combining compilation and quantization flow

PCIe card connected to host

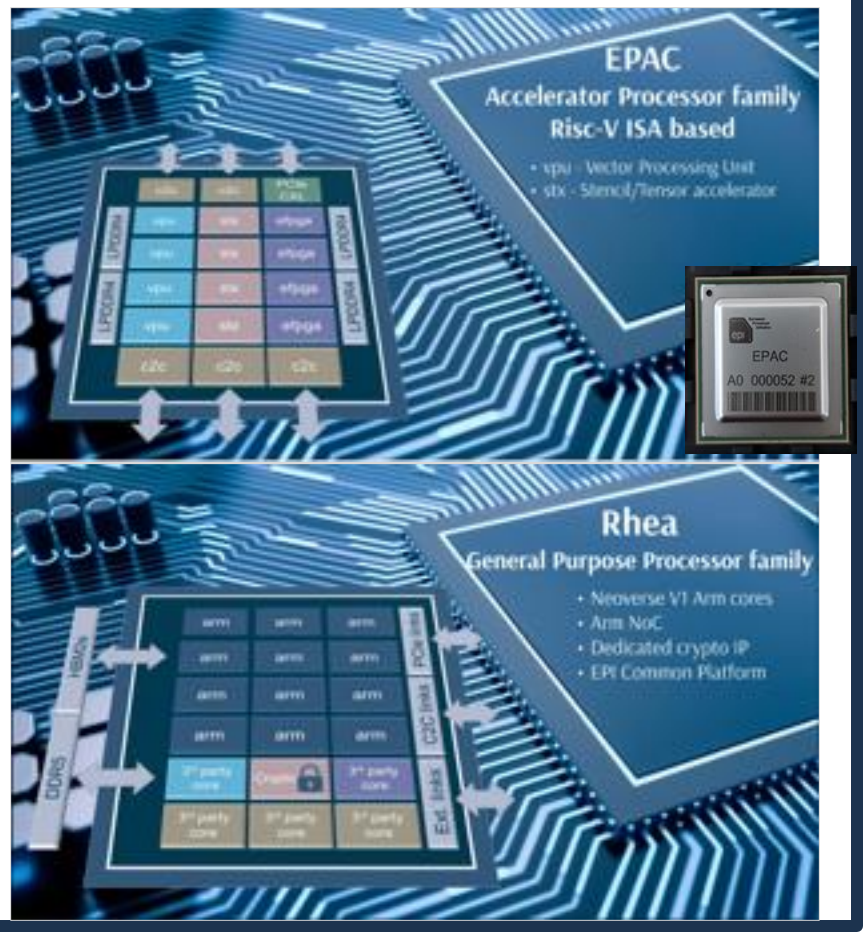
Metis AI Processing Unit (AIPU)

- **Quad-core System-on-Chip**
  - PCIe 3.0 4x link to host
  - LPDDR4x
  - RISC-V controlled
  - 48 MiByte on-chip SRAM
    - 4 MiByte L1 per AI core
    - 32 MiByte L2 shared
  - 4 MiByte D-IMC (Digital in-Memory Computing)
    - 1M 8-bit D-IMC weights per AI core
    - 52.4 TOPS per AI core



European Ecosystem of AI platforms

European Processor Initiative

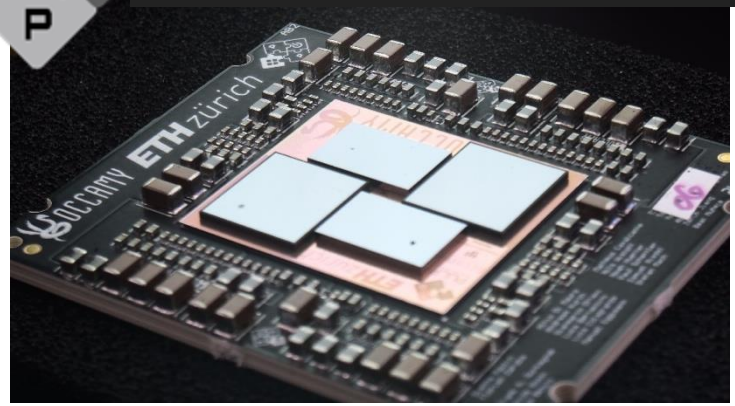


RISC-V®

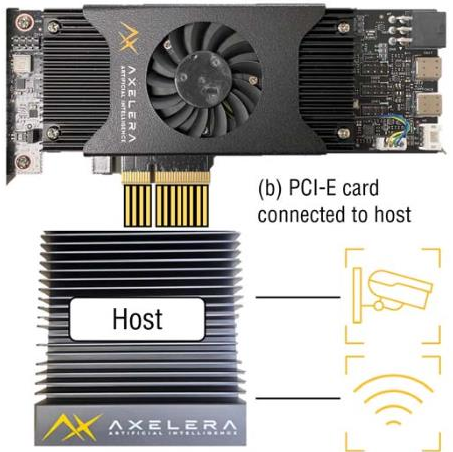
PULP Platform

Open Source Hardware, the way it should be!

https://github.com/pulp-platform/occamy



Occamy:
70mm2@GF12
Up to **686 GFLOPS**
89% FPU util.
40 GFLOPS/W



Metis AIPU
209.6 TOPS
15 to 82 TOPS/W

AXELERA ARTIFICIAL INTELLIGENCE

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Credits:

Andrea Bartolini