ETH zürich

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

End-to-end Open Source Platforms in the Era of Domain Specialization

from Dream to Reality

Luca Benini lbenini@iis.ee.ethz.ch, luca.benini@unibo.it

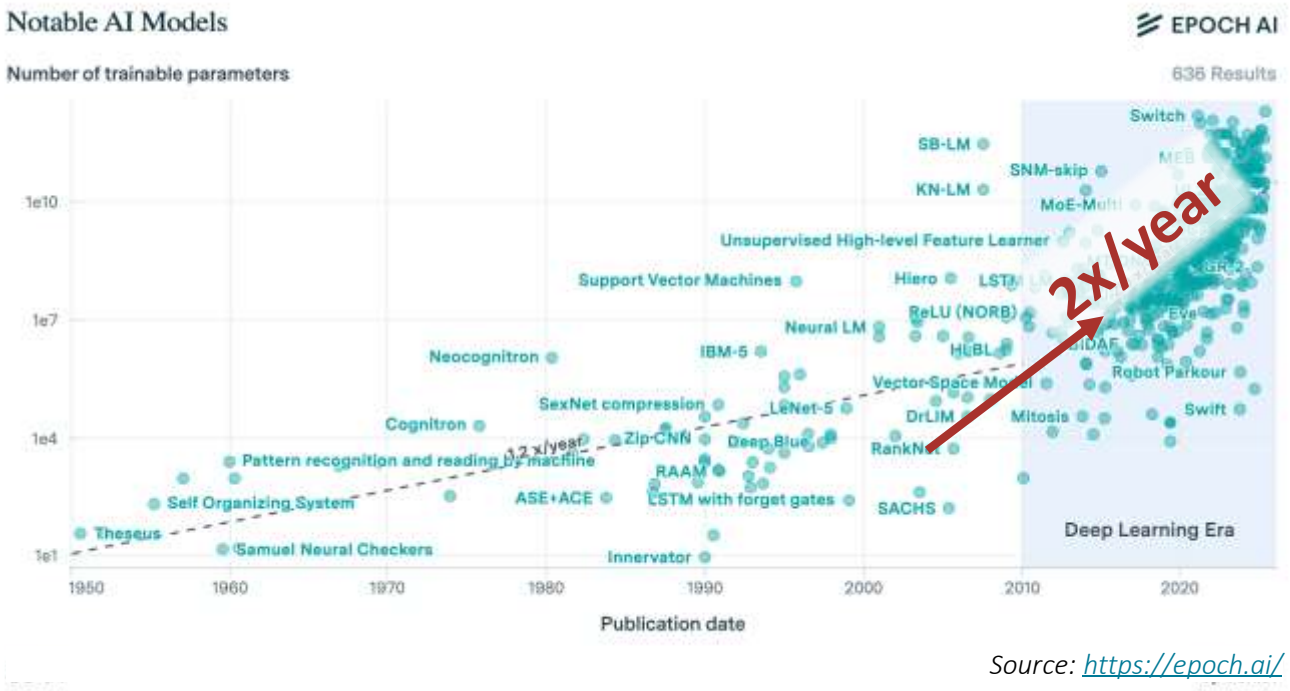
pulp-platform.org
@pulp_platform
company/pulp-platform
youtube.com/pulp_platform

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

Bigger Models, Bigger Bottlenecks

*Deep learning models continue to grow in **scale** and **complexity***

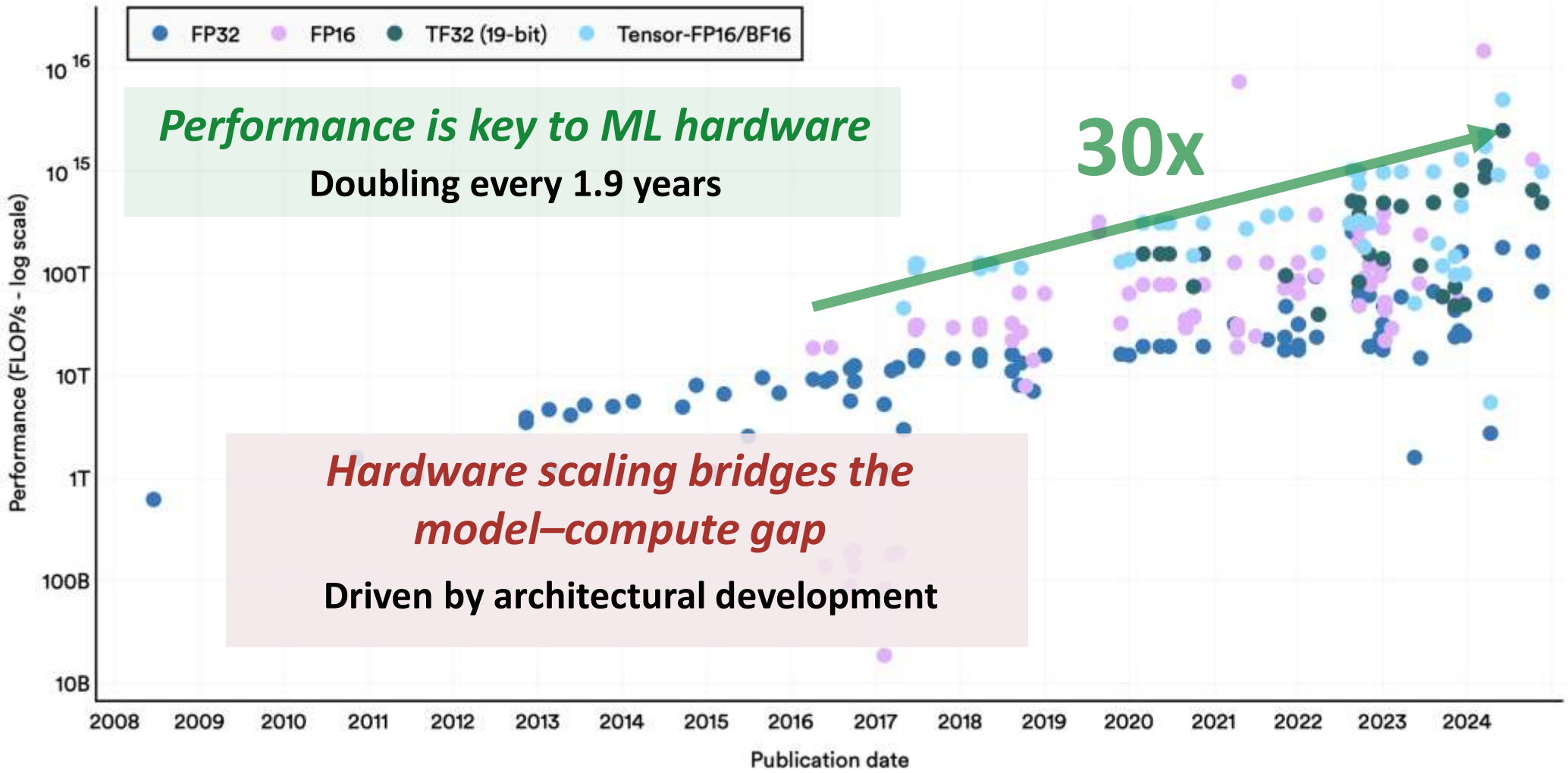
- Growing model sizes demand ever-increasing compute and memory



Source: https://epoch.ai/

Hardware Scaling is key to AI Progress (cloud & edge)

Peak computational throughput of notable ML hardware

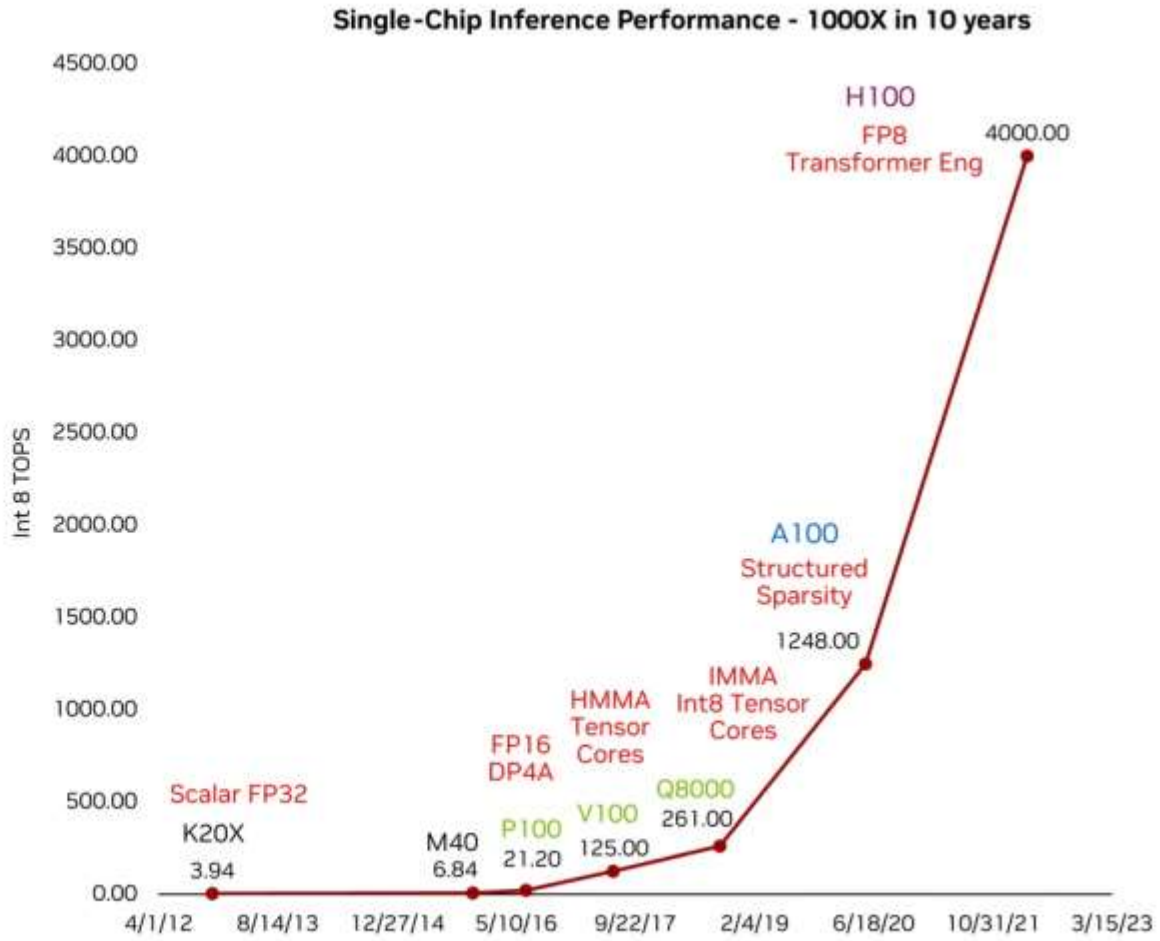


N. Maslej et al., *"Artificial Intelligence Index Report 2025," arXiv Preprint: https://arxiv.org/abs/2504.07139*

How is Industry doing it?

Gains from

- Number Representation
  - FP32, FP16, Int8
  - (TF32, BF16)
  - ~16x
- Complex Instructions
  - DP4, HMMA, IMMA
  - ~12.5x
- Process
  - 28nm, 16nm, 7nm, 5nm
  - ~2.5x
- Sparsity
  - ~2x
- Model efficiency has also improved – overall gain > 1000x



[Dally HotChips 2023]

AI Innovation beyond “NVIDIA Gravity” is Challenging!

It’s the software → **flexibility** key for fast evolution!

Need an **open standard** to counter a monopoly

RISC-V®

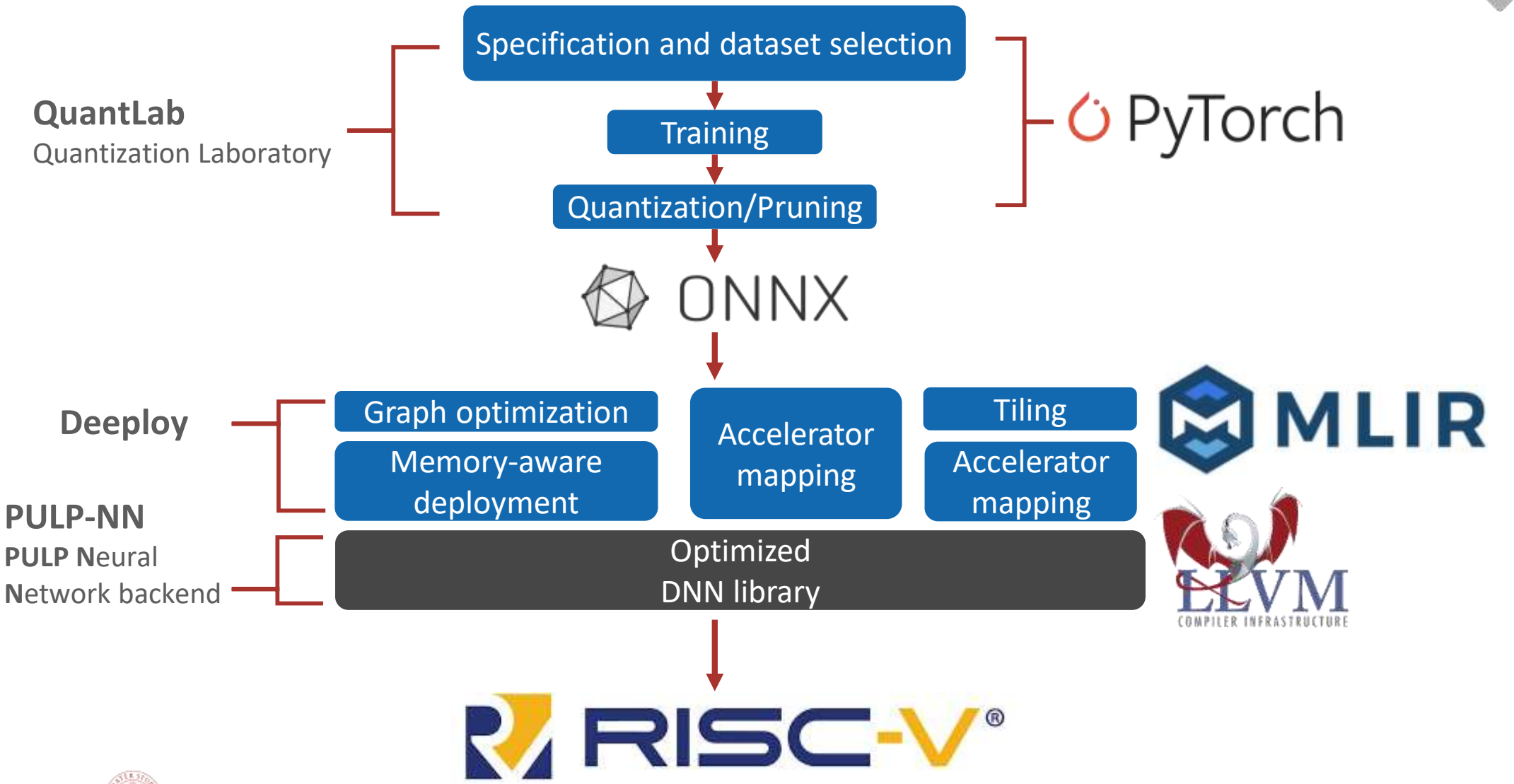
RISC-V: The Free and Open RISC Instruction Set Architecture

Meta

AXELERA

tenstorrent

Fully Open-Source AI SW Stack with RISC-V!



PULP: Open-Source RISC-V Hardware:

RISC-V Cores and Vector Units

RI5CY *CV32E*	Zero R *Ibex*	Snitch	Spatz	Ariane *CVA6*	ARA
RV32	RV32	RV32	RVV	RV64	RVV

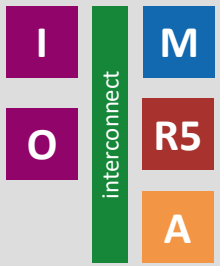
Peripherals

JTAG, SPI, UART, I2S, DMA, GPIO

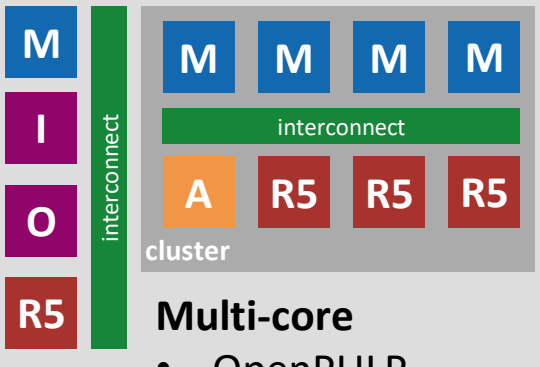
Interconnects

LIC, HCI, APB, FlooNoC, AXI4

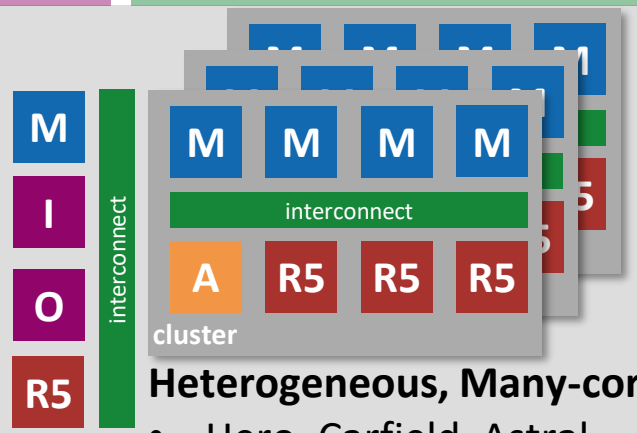
Platforms



Single core
- PULPissimo, Croc
- Cheshire



Multi-core
- OpenPULP
- ControlPULP



Heterogeneous, Many-core
- Hero, Carfield, Astral
- Occamy, Mempool

IOT → **HPC**

Accelerators and ISA extensions

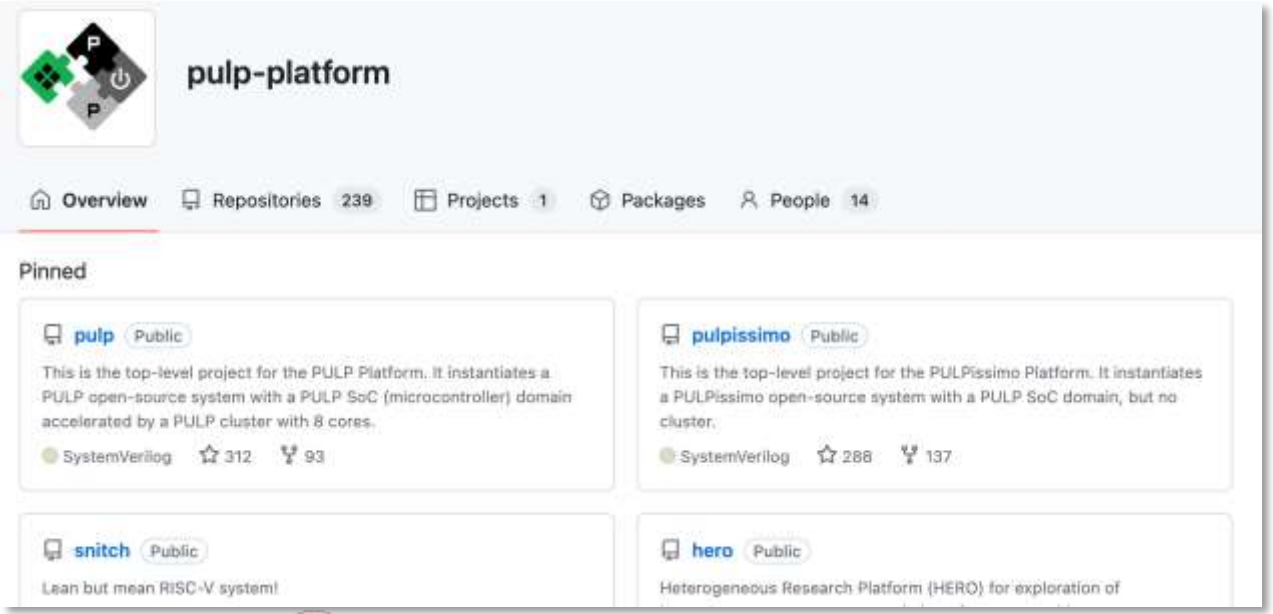
XpulpNN, XpulpTNN	ITA (Transformers)	RBE, NEUREKA (QNNs)	FFT (DSP)	REDMULE (FP-Tensor)

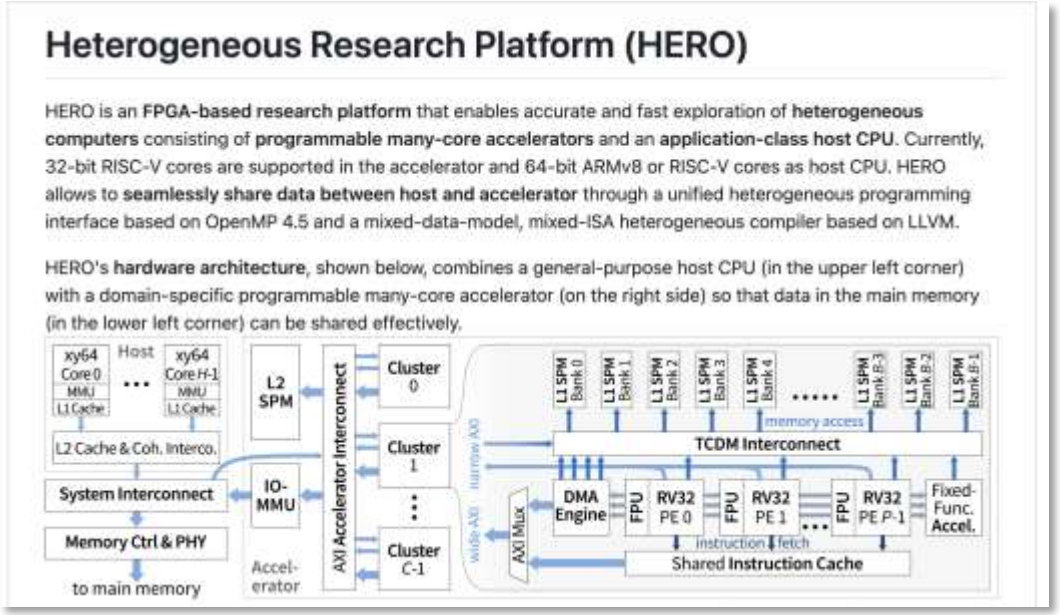
All of our designs are open-source hardware

- **All our development is on GitHub using a permissive license**
  - HDL source code, testbenches, software development kit, virtual platform

https://github.com/pulp-platform

- Allows anyone to use, change, and make products without restrictions.





We have designed over 60 ASICs using open-source HW

All our designs are based on open-source HW published on our GitHub page

- All using a permissive open source license (SolderPad)

https://github.com/pulp-platform

See our chip gallery under: http://asic.ethz.ch/

End-to-end OSHW aims to open **all steps** of IC design

Design

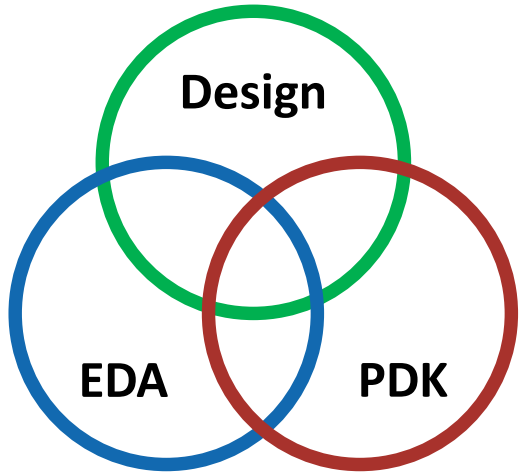
- RTL / HDL descriptions (quite common)
- Schematics / Physical Design (may have dependencies to technology information)

Tools (EDA)

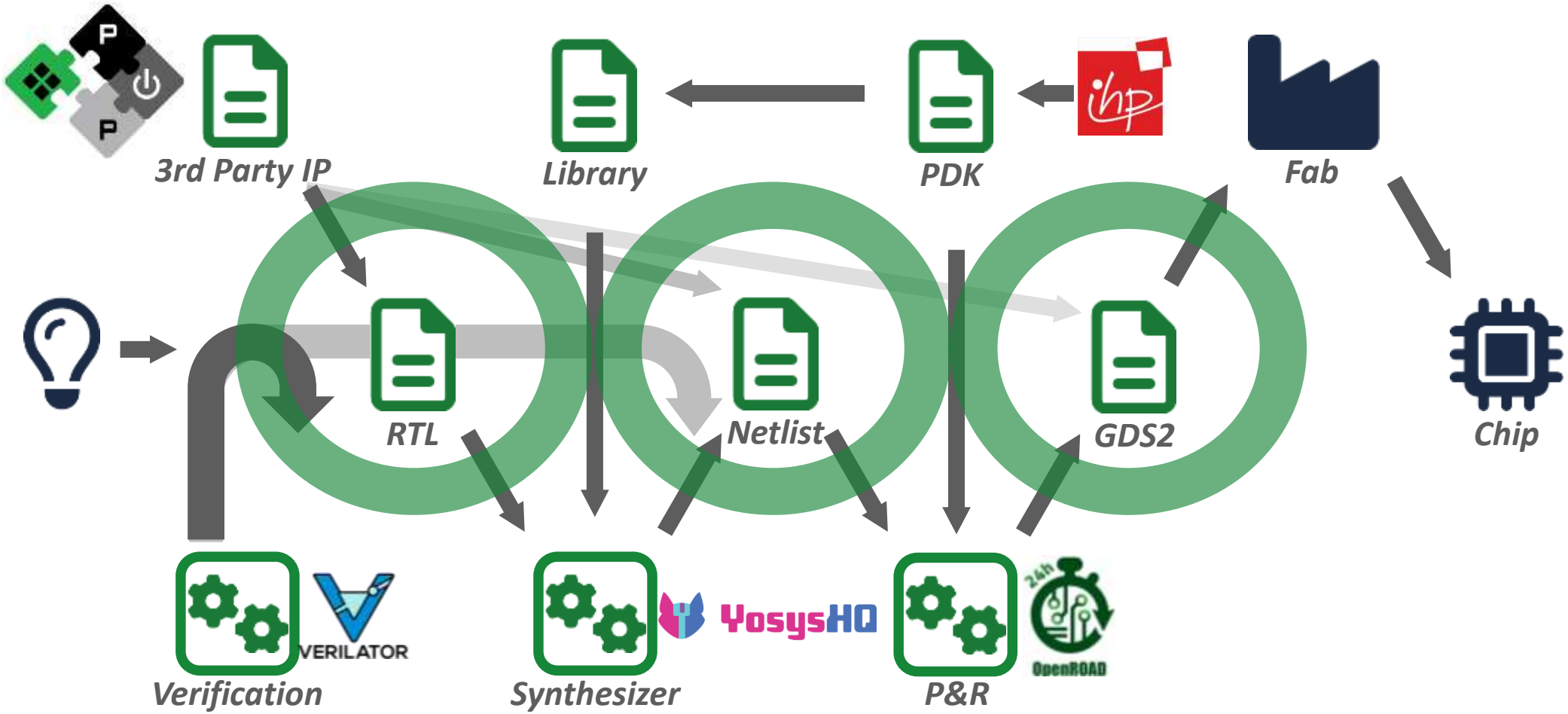
- Front-end tools (Synthesis)
- Back-end tools (Placement and Routing)
- Verification tools (Simulation)

Manufacturing (PDK)

- Design rules for manufacturing (separation, minimum width of metals)
- Layer stack information for parasitics (thickness, dielectric constants..)
- Device models (SPICE parameters) for simulation



End-to-end Open-Source allows sharing of design data



End-to-end Open-Source IC Design is possible today!

Design: from PULP

`github.com/pulp-platform`

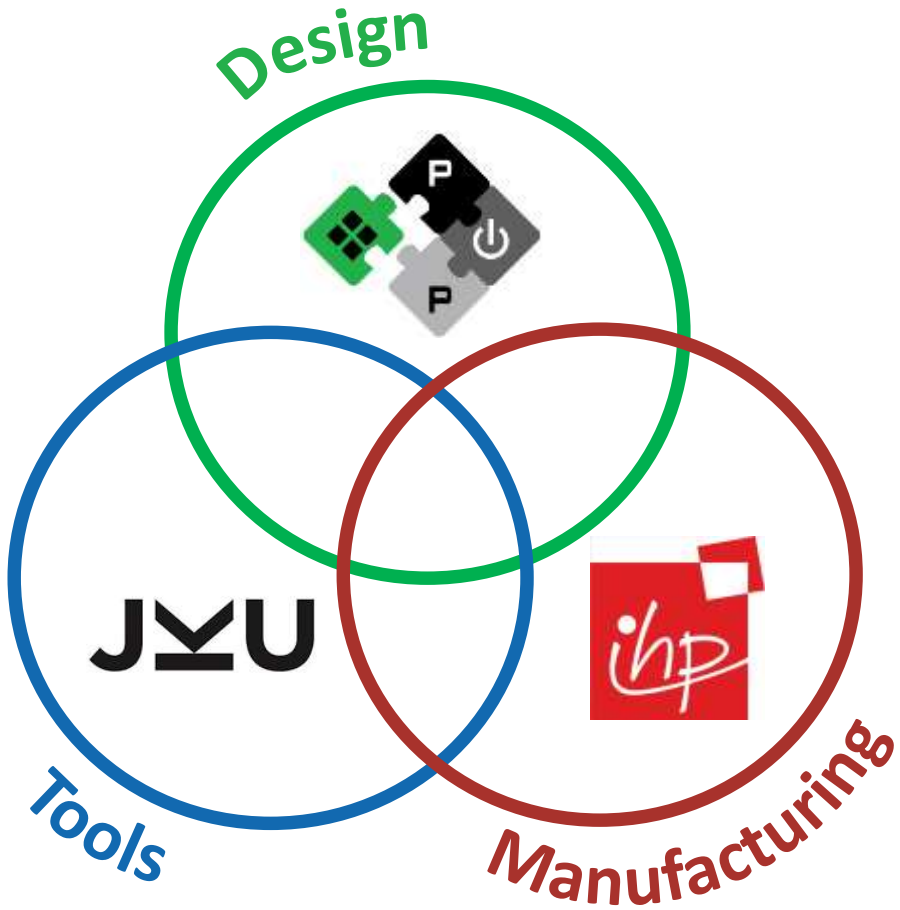
Tools: from Johannes Kepler University (JKU)

Reliable VM with large collection of open-source tools

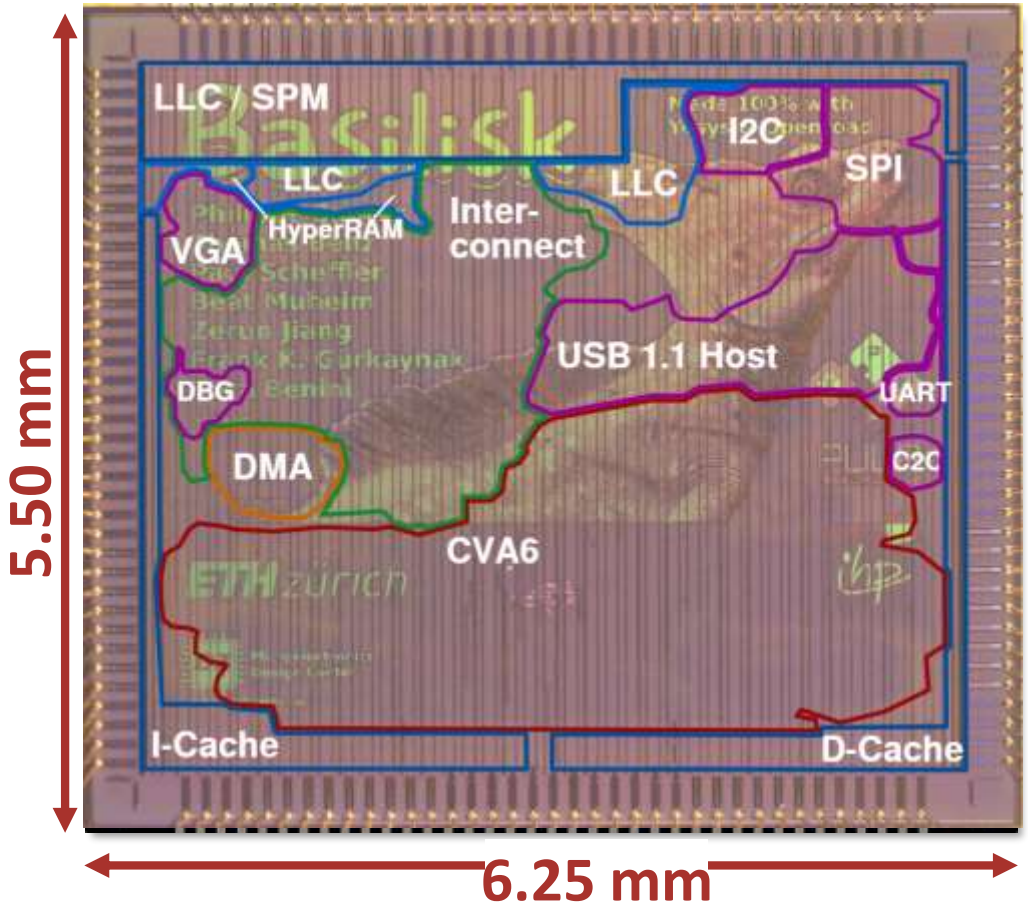
`github.com/iic-jku/IIC-OSIC-TOOLS`

Manufacturing: IHP130nm

`github.com/IHP-GmbH/IHP-Open-PDK`



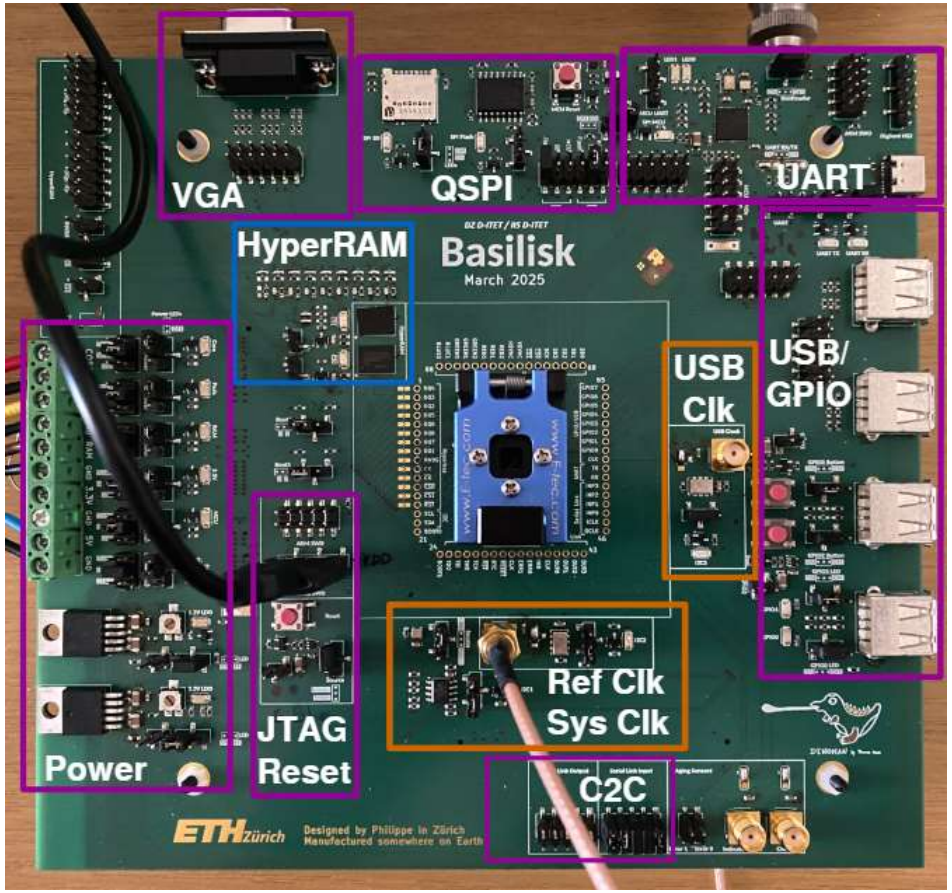
Meet Basilisk: Open **RTL**, Open **EDA**, Open **PDK**



- **Designed in IHP 130nm OpenPDK**
  - **34mm²** (6.25mm x 5.50mm)
  - **~5× larger** than previous end-to-end OS designs
  - 2.7 MGE total, 1.14MGE logic
  - 24 SRAM macros (114 KiB)
  - **62MHz** at nominal voltage (1.2V)
- **RV64GC design runs Linux**
- **Active collaboration with**

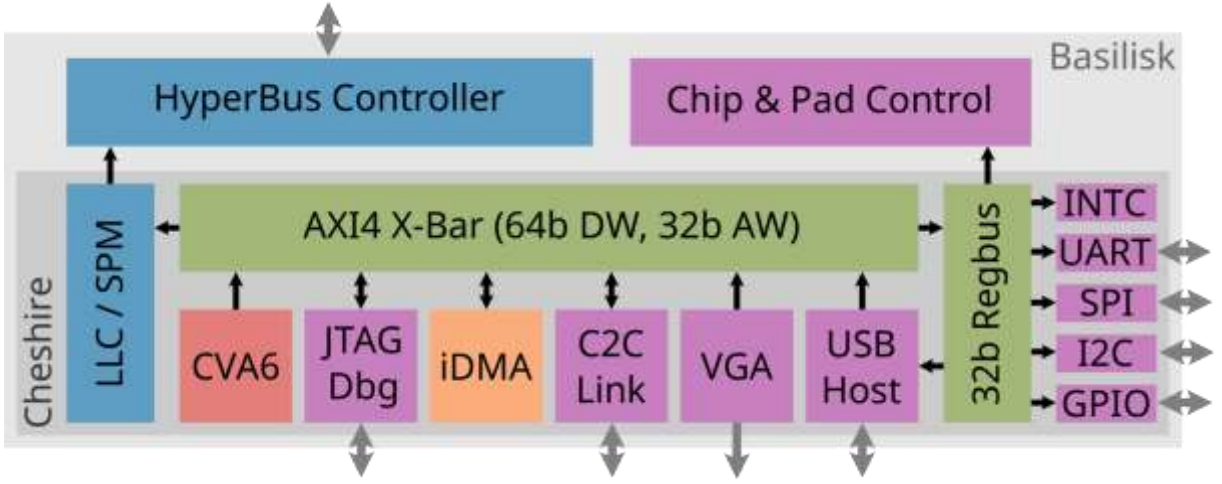
`github.com/pulp-platform/cheshire-ihp130-o`

Basilisk is a complete Linux-capable SoC



- **64-bit RISC-V core**
- **Rich peripherals:**
  - HyperRAM controller **@154MB/s**
  - C2C AXI-Link **@77MB/s**
- **Automatic boot via scratchpad**

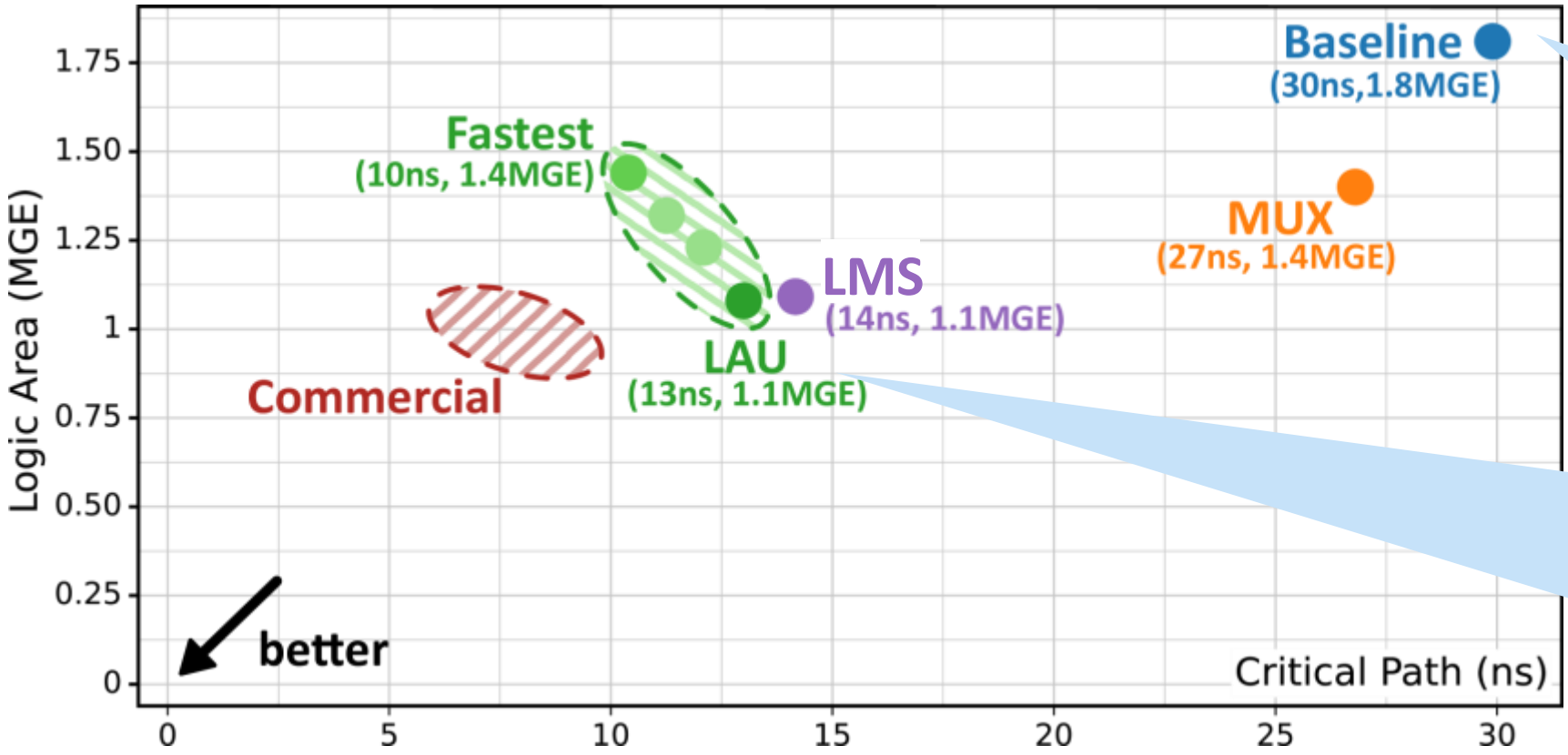
Watch the video

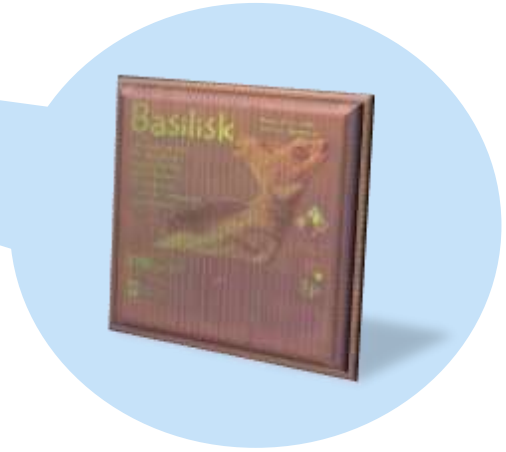


arxiv.org/pdf/2505.10060

Closing the PPA gap to commercial EDA

6



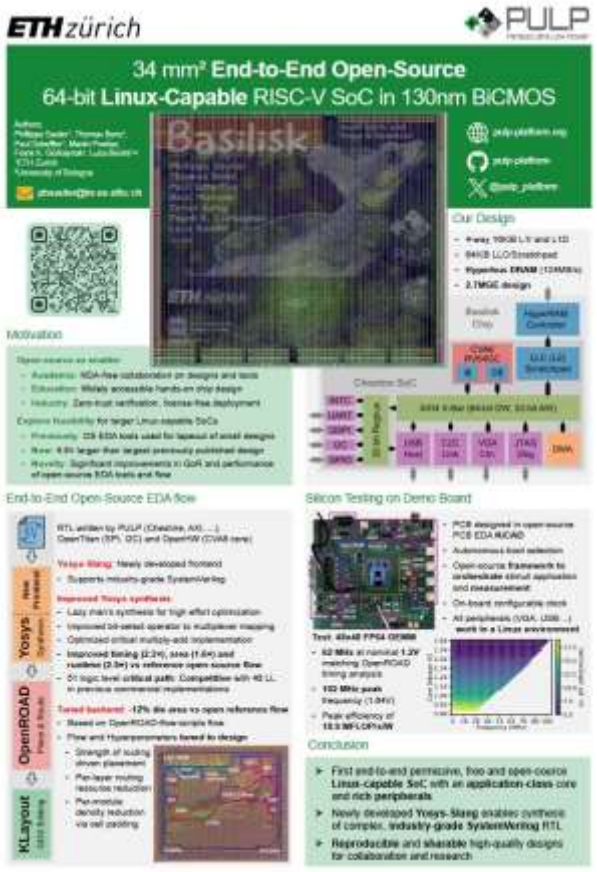


Yosys-slang full Sysverilog Frontend: @ <6sec runtime (from minutes)

Yosys synthesis: 1.1 MGE (1.6×) @ 77 MHz (2.3×), 2.5× less runtime, 2.9× less RAM

OpenROAD P&R: tuning -12% die area, +10% core utilization

We presented Basilisk at HOT CHIPS



Industry Noticed!

"Basilisk at Hot Chips 2025 Presented Ominous Challenge to IP/EDA Status Quo"*

Poster: lnkd.in/daB6HskB

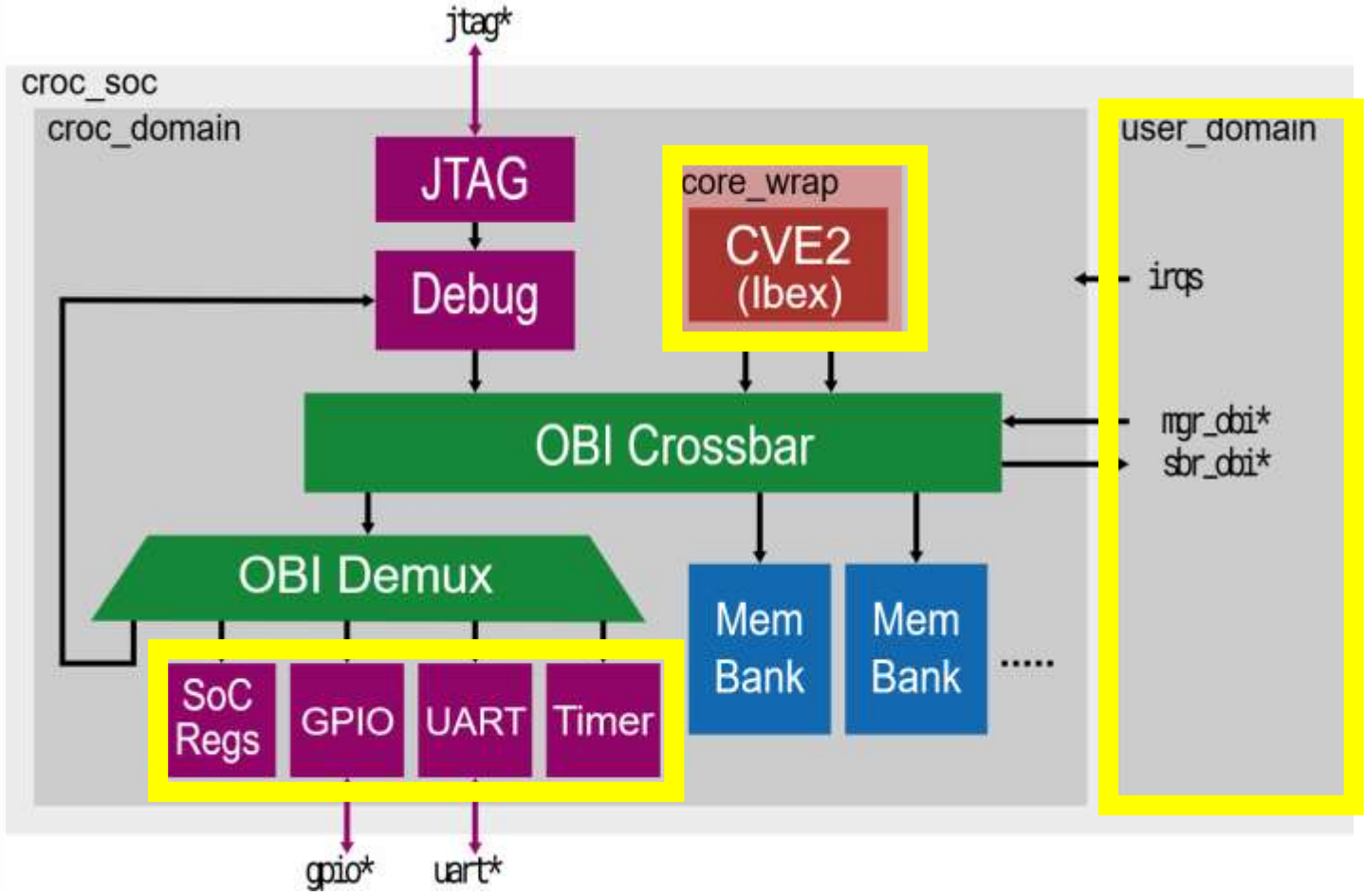
ETH zürich

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

*semiwiki.com/ip/risc-v/361204-basilisk-at-hot-chips-2025-presented-ominous-challenge-to-ip-eda-status-quo/

Croc: a simple SoC for education with PULP IPs

- **32-bit RISC-V core (CVE2)**
- **Options to improve**
  - User domain
  - Adding peripherals
  - Extensions to the core
- **Reference design for VLSI2 lecture and exercises**
- **Pipe-cleaning with two Croc-based tapeouts**
  - Mlem, Koopa (next slide)

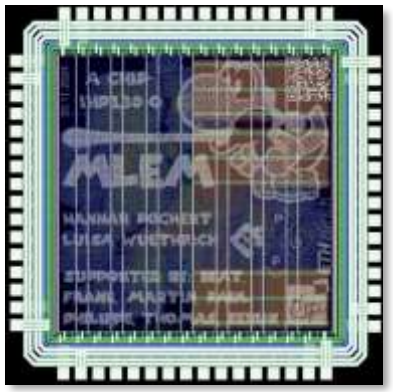


github.com/pulp-platform/croc

At ETH Zürich, IC Design teaching now uses open source HW

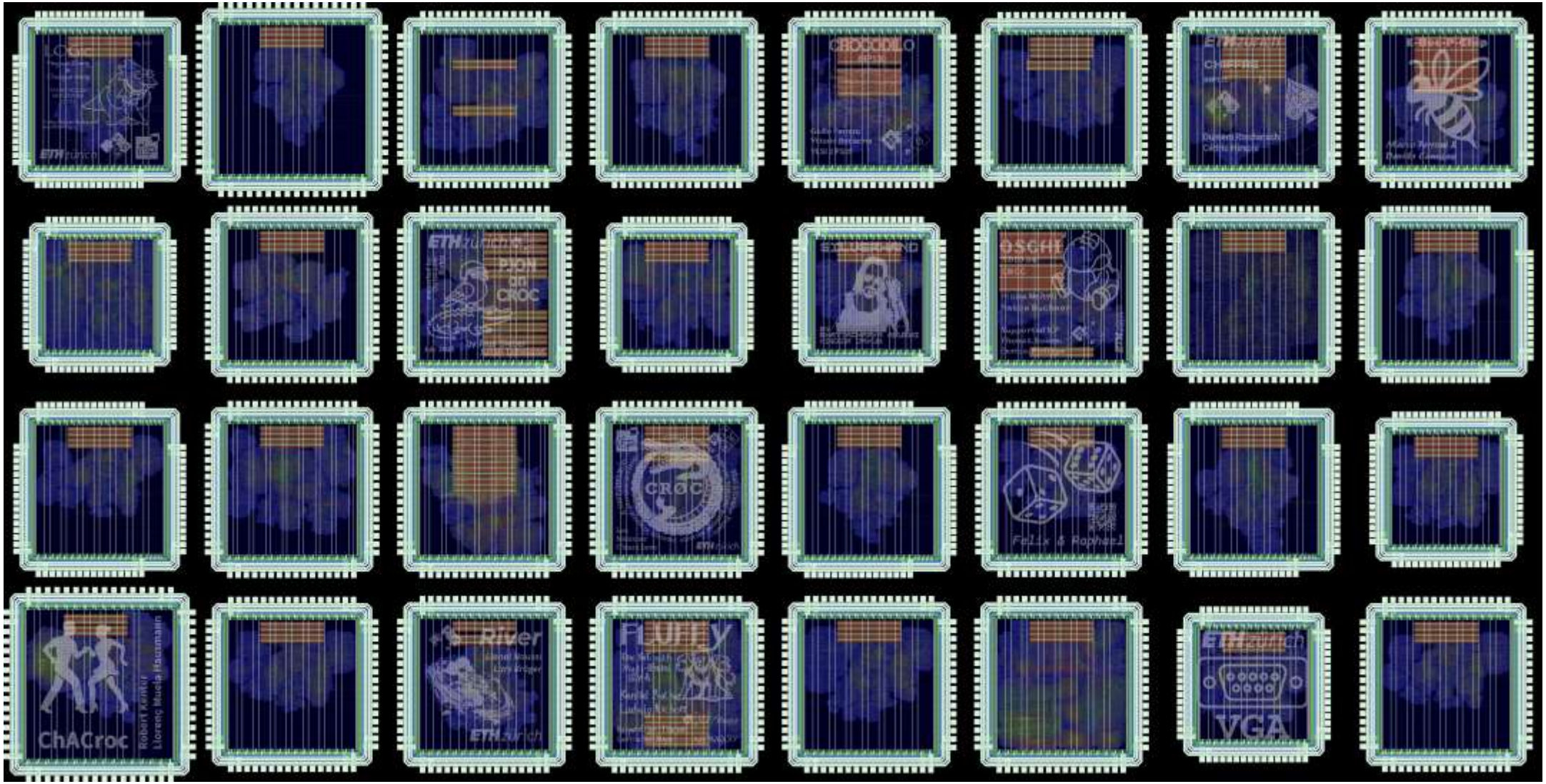
- **In Spring 2025, our IC Design course switched to (mostly) open source**
  - Using IHP 130, Yosys and OpenROAD
    - Parts for backannotated simulation, test pattern generation, DRC/LVS, still use proprietary tools
    - Will be gradually replaced by open tools

https://vlsi.ethz.ch



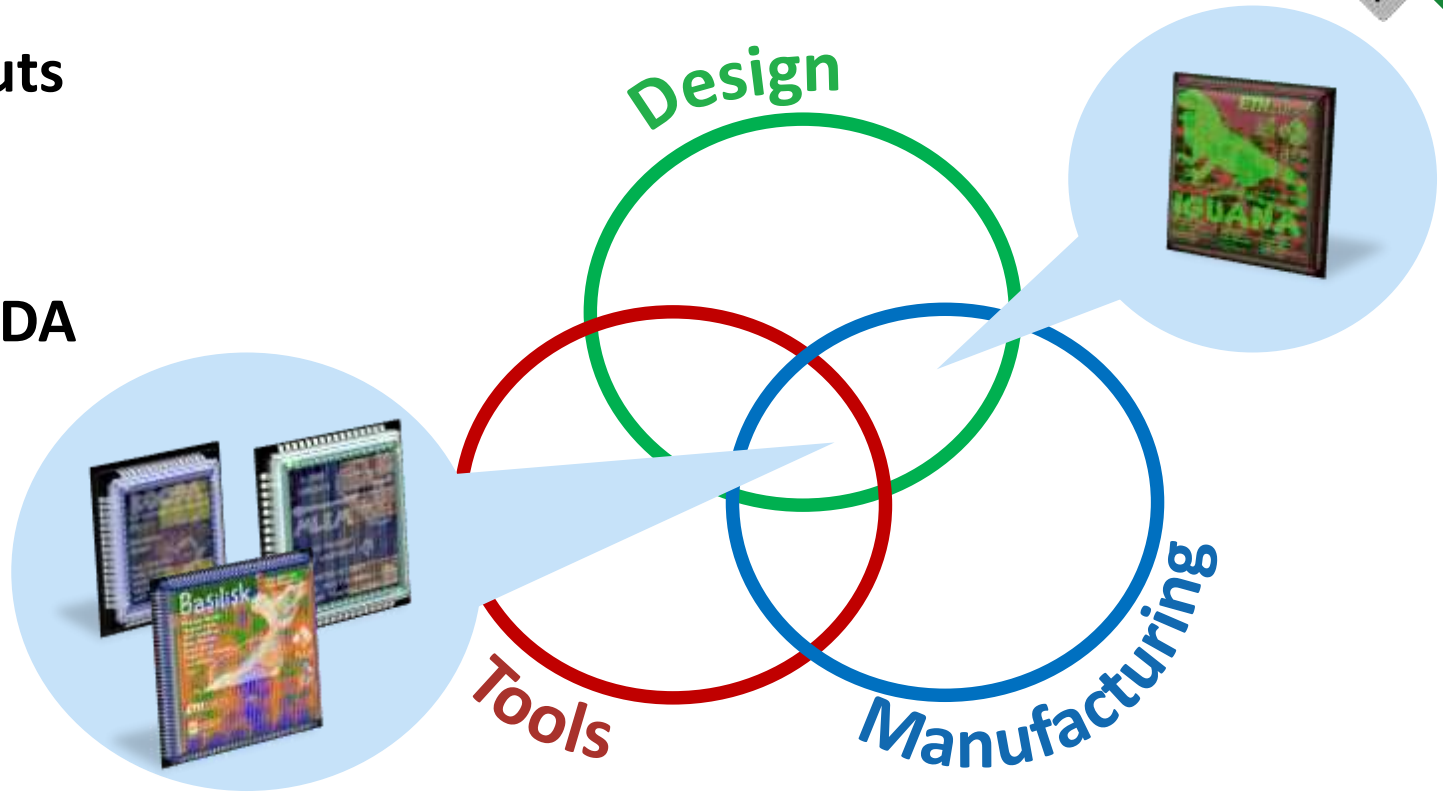
- **Project based grading**
  - Students (in groups of two) will have to modify the Croc reference design
  - Best five designs will be taped-out
- **72 students enrolled**
  - Projects finished in summer
  - Tape-out in IHP130 September

And the students delivered!



Freedom on Tools & Process According to Design Goals

- **10 open-EDAs & PDKs tape-outs with different design choices**
  (+many more coming)
- **Active contribution to open-EDA community**
- **Successful educational goals:**
  - Open-EDA based courses
  - Open-source tape-out student projects



Open-Source Design & Flow for Reproducible SoA Innovation!

End-to-end Open-Source IC Design is already working!

Easier collaboration / sharing

- Need to stand on the shoulder of giants
- Share common parts that all need
- Concentrate work/time where it matters

Open reproducible results

- Everyone can verify performance claims
- Allows us to generate example datasets that can be used to train/improve tools

Reduce entry barriers for all

- You can easily get started with IC Design
- No agreements needed to get started
- Can then decide to stay open or not

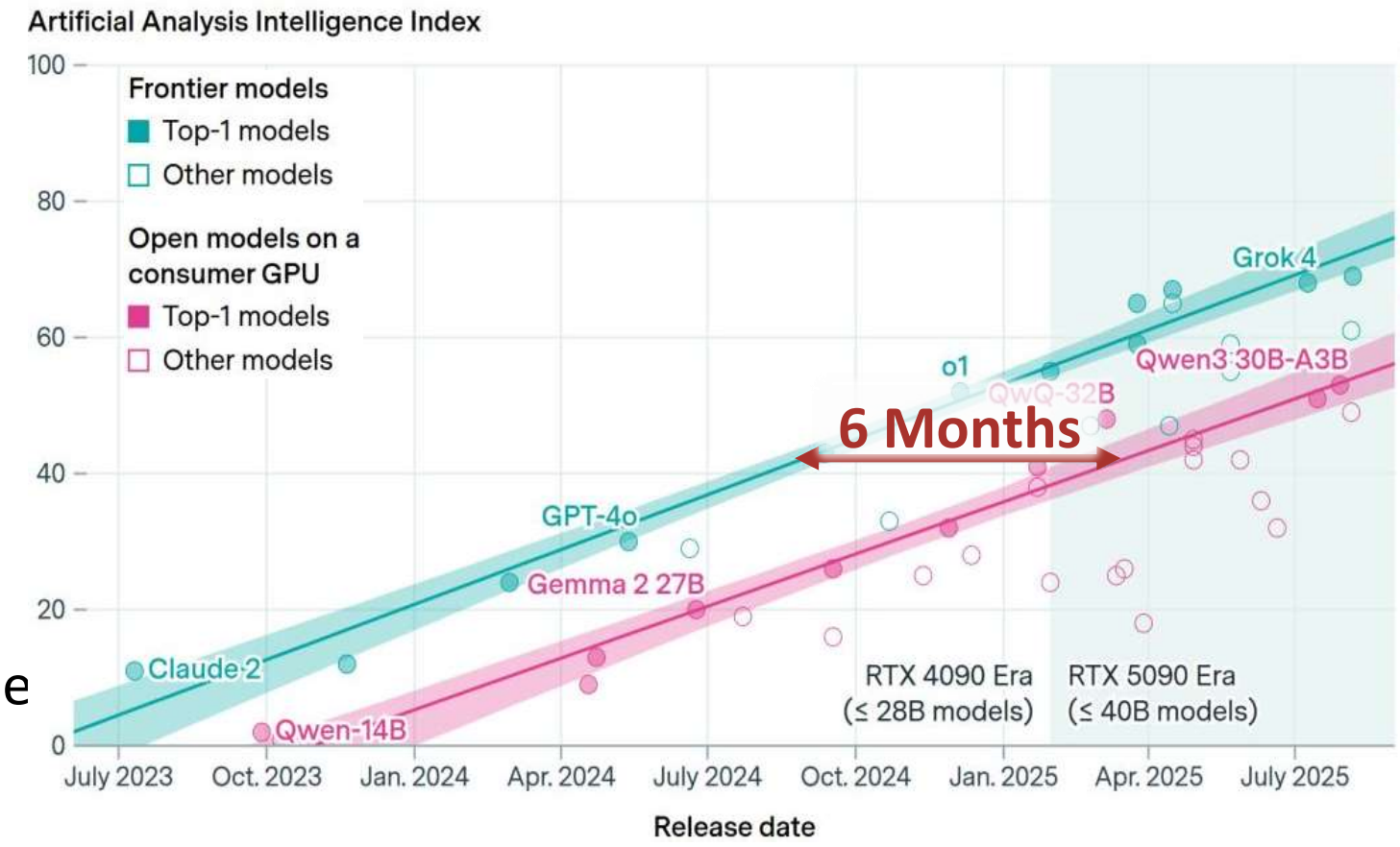
Accessible teaching for all

- Share courses, designs, examples
- Create tutorials, knowledge bases
- Training for industry

Back to AI: Bigger Models, Bottlenecks @ Edge

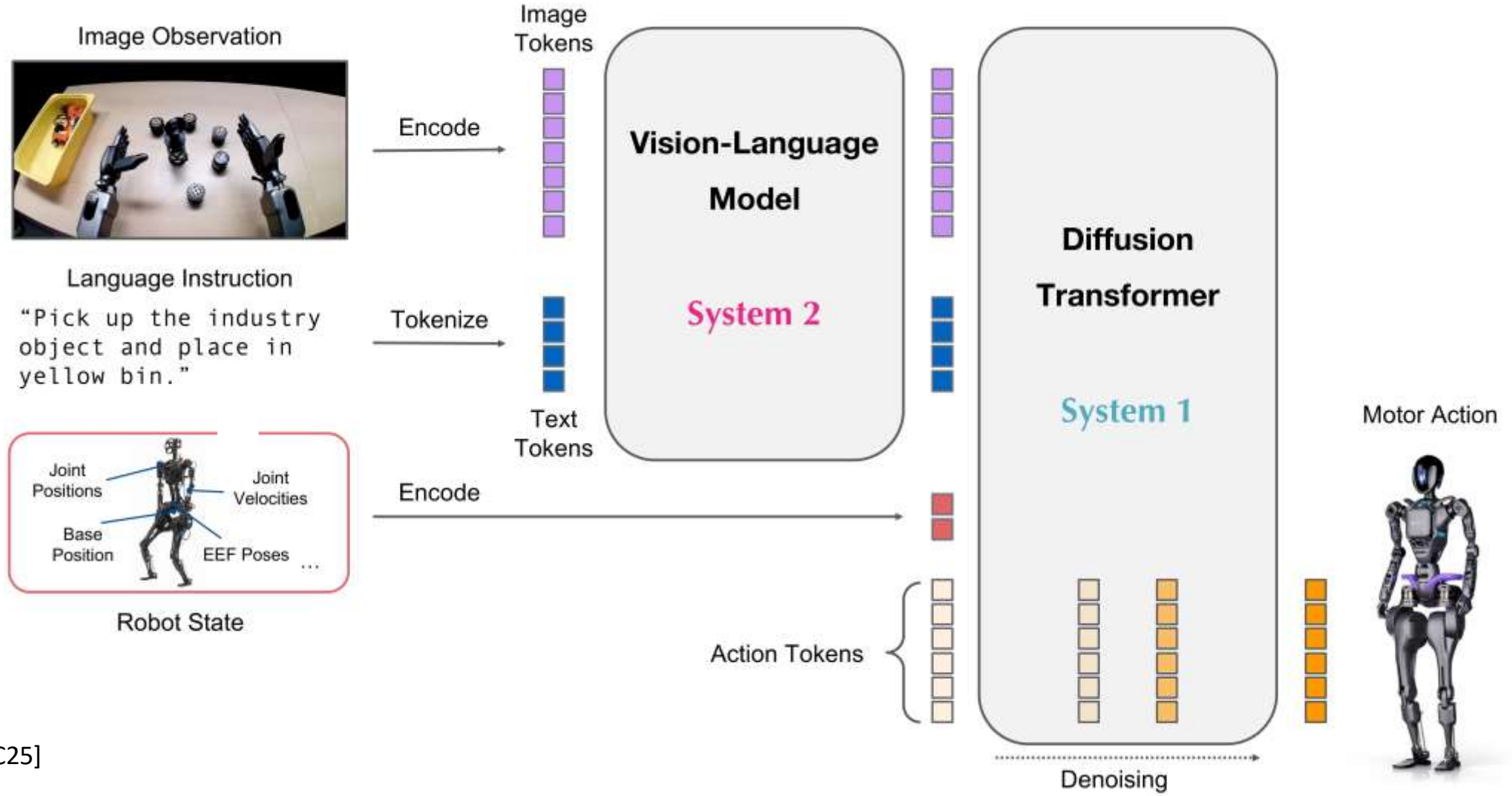
*Deep learning models continue to grow in **scale** and **complexity***

- Growing model sizes demand ever-increasing compute and memory
- Inference compute scale even faster than for training
- Models that fit on a single GPU trail the frontier by less than one year



Source: https://epoch.ai/

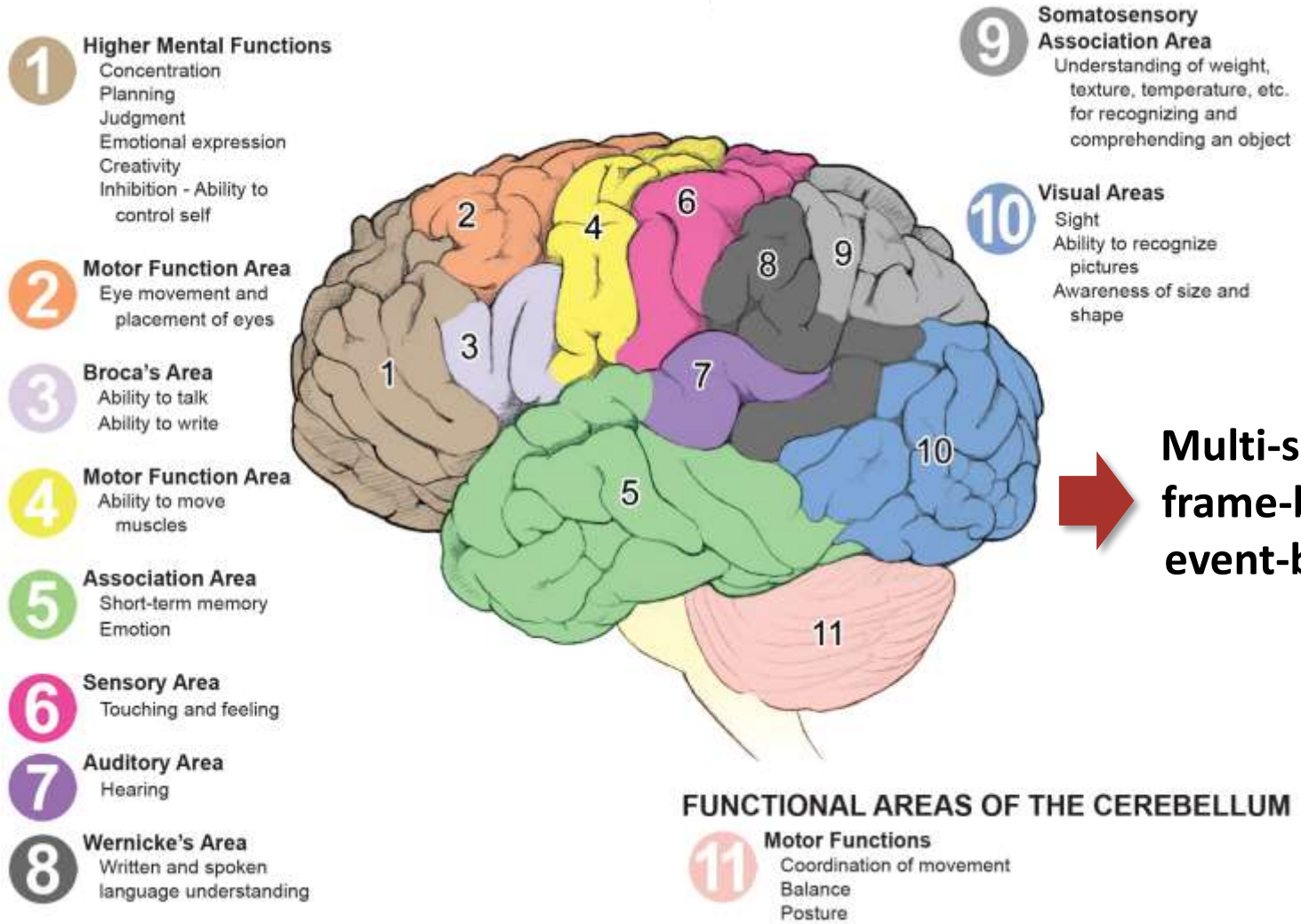
Embodied Gen.AI



[GTC25]

Efficiency through Heterogeneity: Multi-Specialization

Brain-inspired: Multiple areas, different structure different function!



Multi-sensor
frame-based
event-based
→ **Perception** → **Fusion**
Reasoning

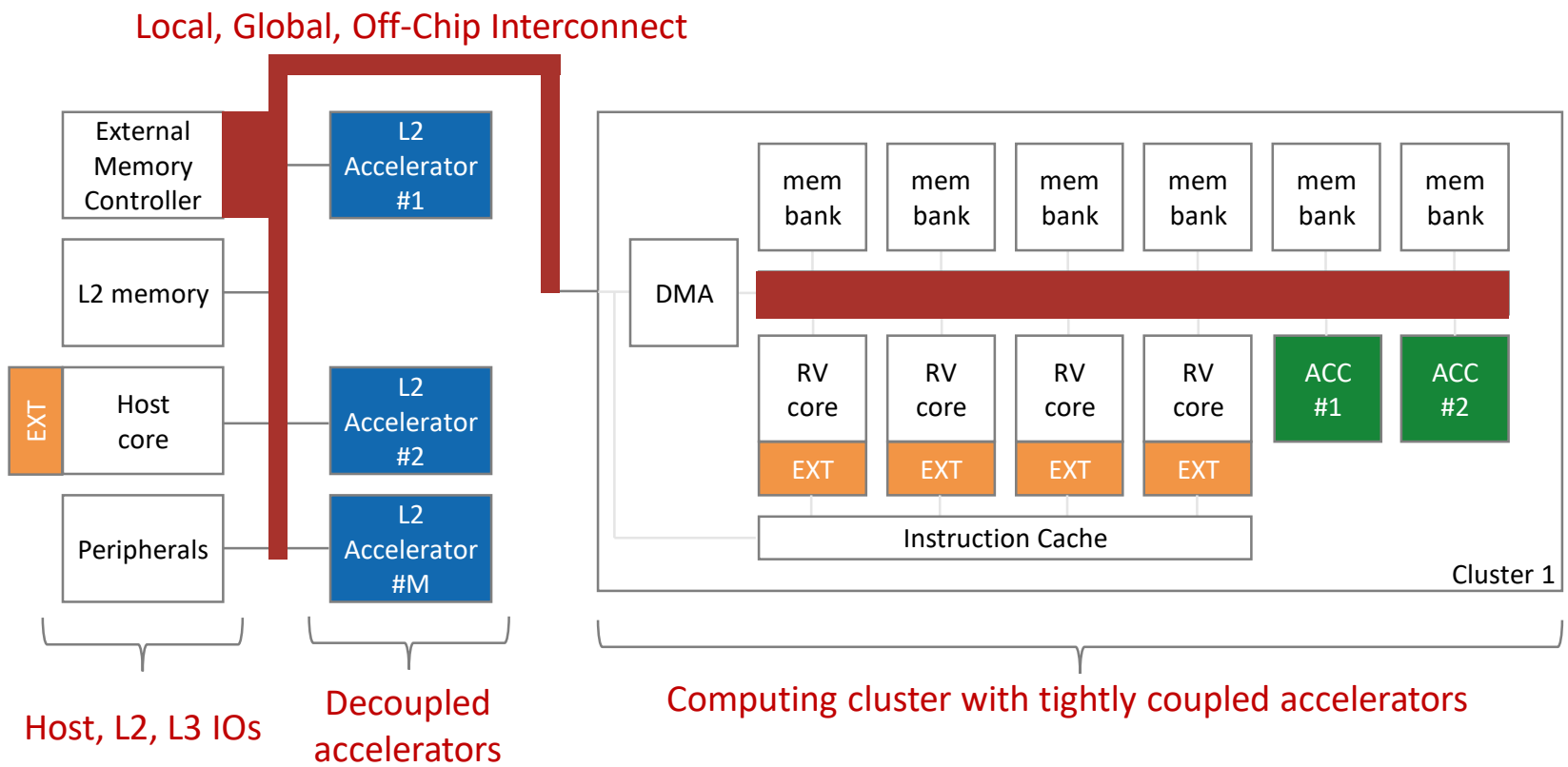
How to Specialize Silicon

Multiple Scales of acceleration

Extensions to processor cores
- Explore new extensions
- Efficient implementations

Shared-memory Accelerators
- Domain specific
- Local memory

Multiple Decoupled Accelerators
- Communication
- Synchronization

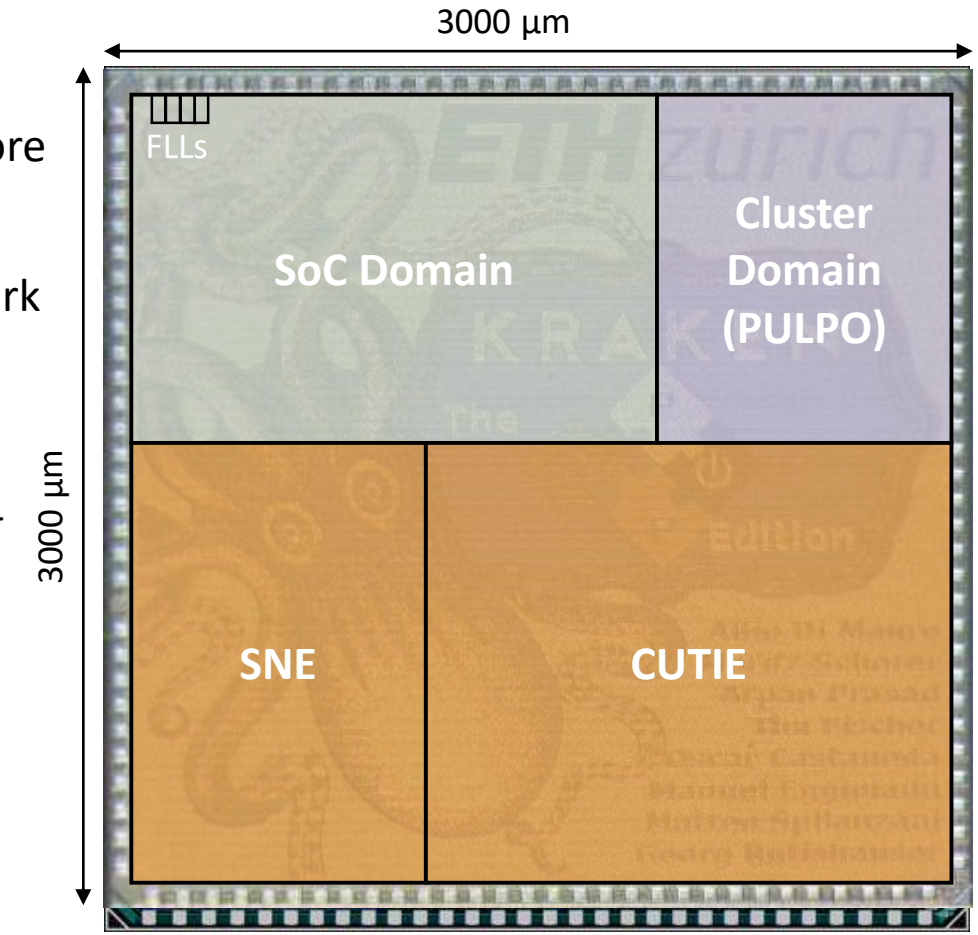


Local, global, package, system → Specialization at scale

Kraken: 22FDX SoC, Multiple Heterogeneous Accelerators

The ***Kraken***: an "Extreme Edge" Brain

- **RISC-V Cluster**
  8 Compute cores +1 DMA core
- **CUTIE**
  Dense ternary-neural-network accelerator
- **SNE**
  Energy-proportional spiking-neural-network accelerator



Technology	22 nm FDSOI
Chip Area	9 mm^2
SRAM SoC	1 MiB
SRAM Cluster	128 KiB
VDD range	0.55 V - **0.8 V**
Cluster Freq	**~370 MHz**
SNE Freq	**~250 MHz**
CUTIE Freq	**~140 MHz**

Specialization in numbers (Joules)

Using 22FDX tech, NT@0.6V, High utilization, minimal IO & overhead

Energy-Efficient RV Core → **20pJ (8bit)**

↓

ISA-based 10-20x →**1pJ (4bit)** → **XPULP**

↓

Configurable DP 10-20x → **100fJ (4bit)** → **RBE**

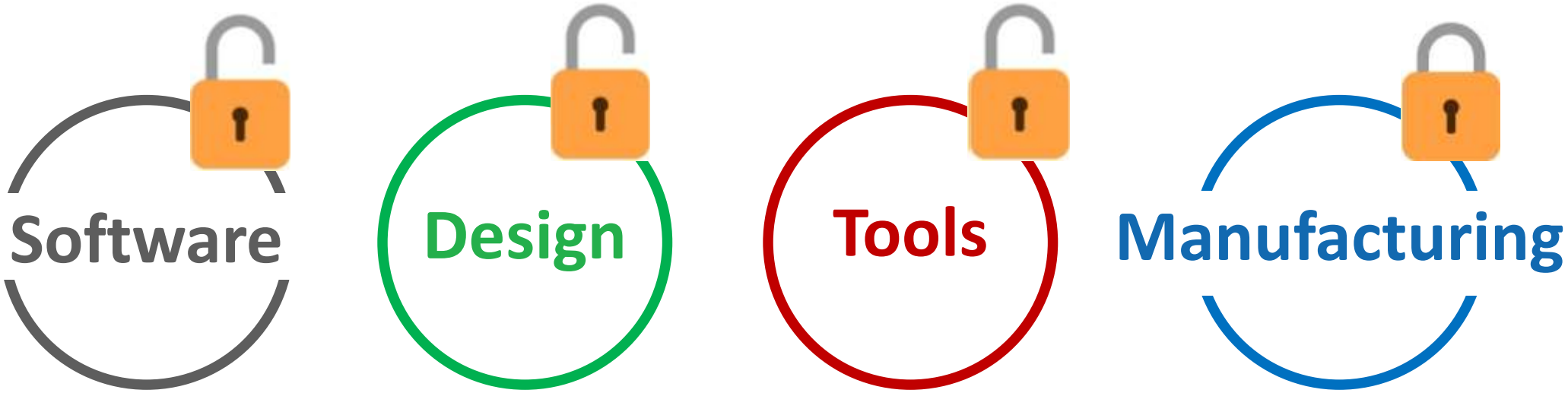
↓

Highly specialized DP 100x →**1fJ (ternary)** → **CUTIE, SNN**

3-4 OoM may not be enough, and we need flexibility!

Open EDAs Heterogeneous Chips in Advanced CMOS?

Extreme Performance + Energy Efficiency is required!



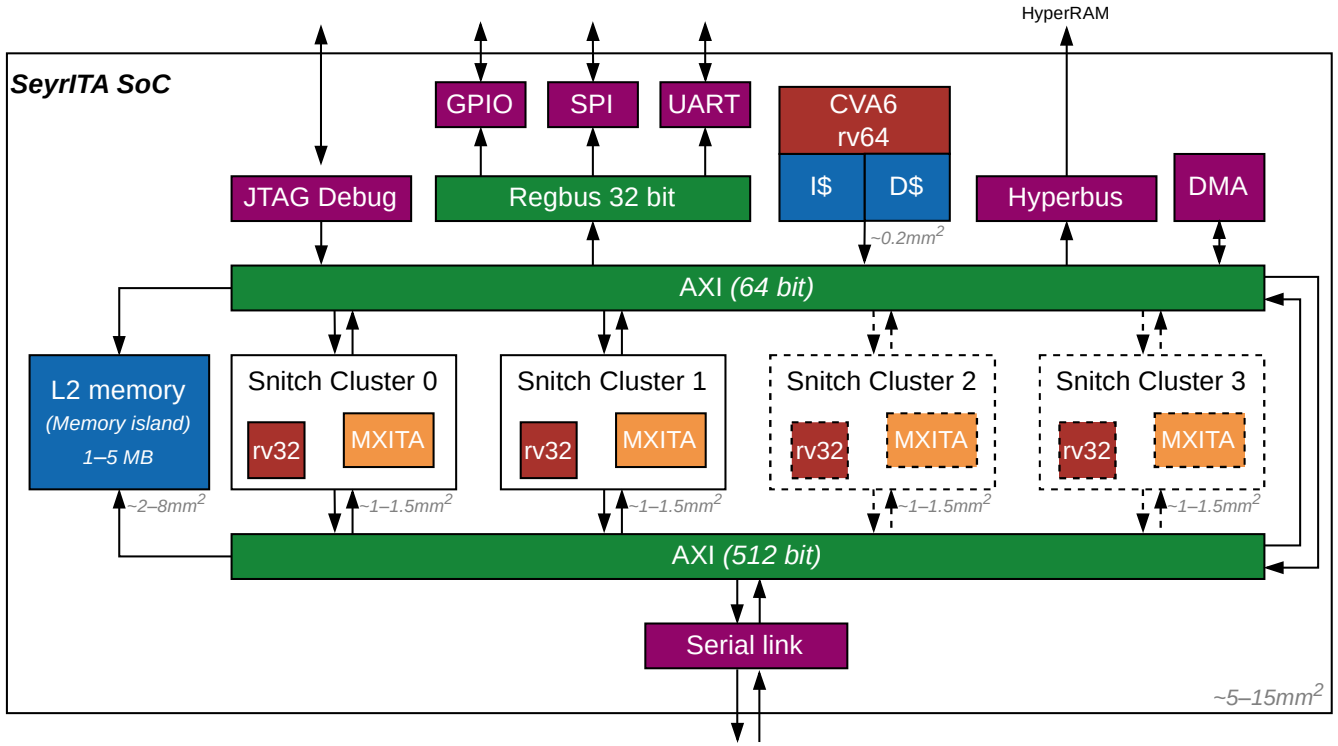
HW-SW Co-Design

HW-SW & Tools Co-Optimization

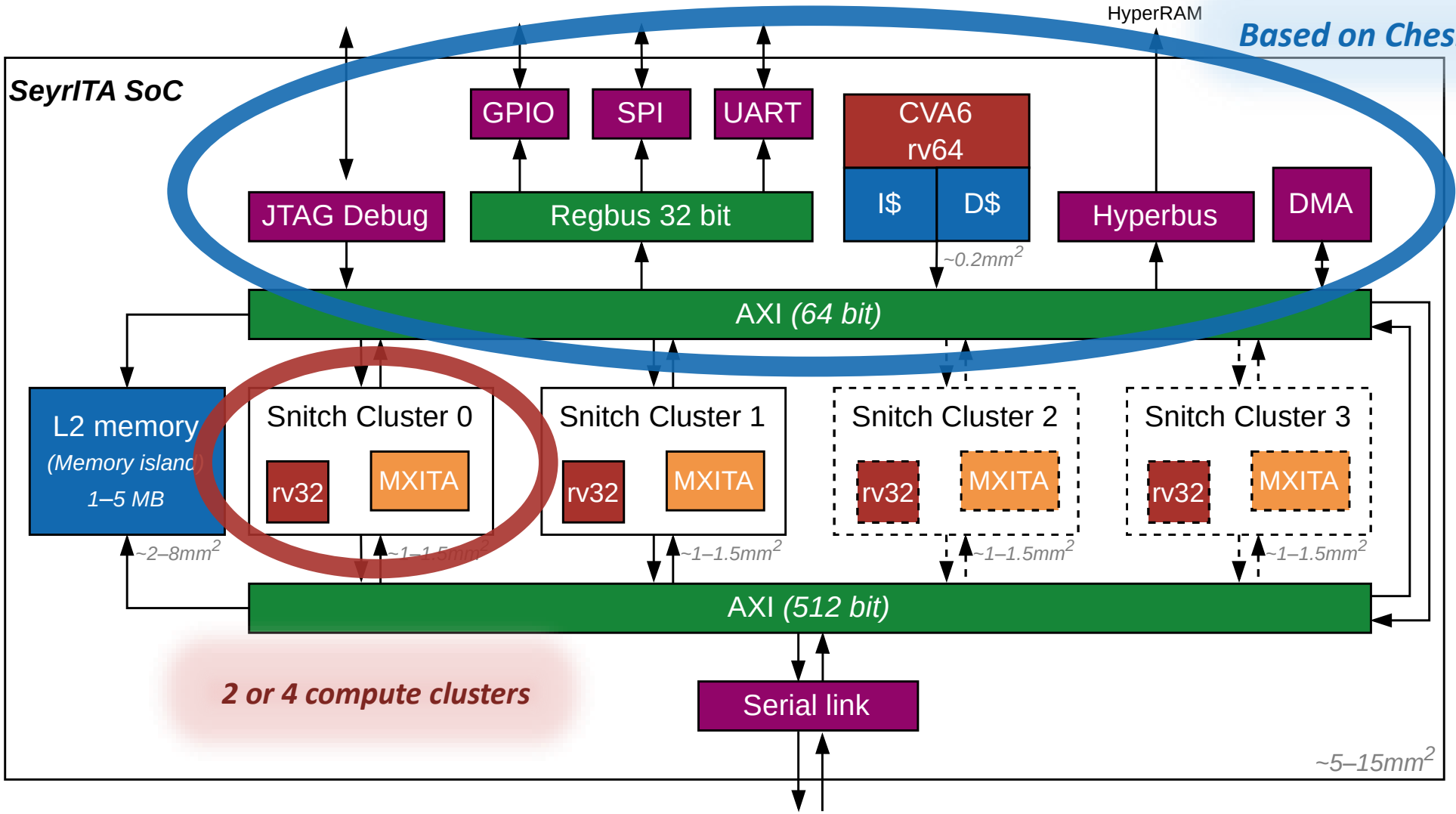
Get the best from advanced nodes!

On the Horizon: SeyrITA – GF22 with Open-Source Tools

- **RISC-V Linux host platform**
- **Transformer accelerator**
  - Targeting BERT, mobileBERT, DEiT-T
  - Leveraging microscaling quantization
  - MXINT and MXFP32 formats
- **10x larger!**
  - 20-40MGE (SeyrITA) vs 2-3MGE (Basilisk)
- **500MHz target frequency**
- **1-2TFLOP/s**

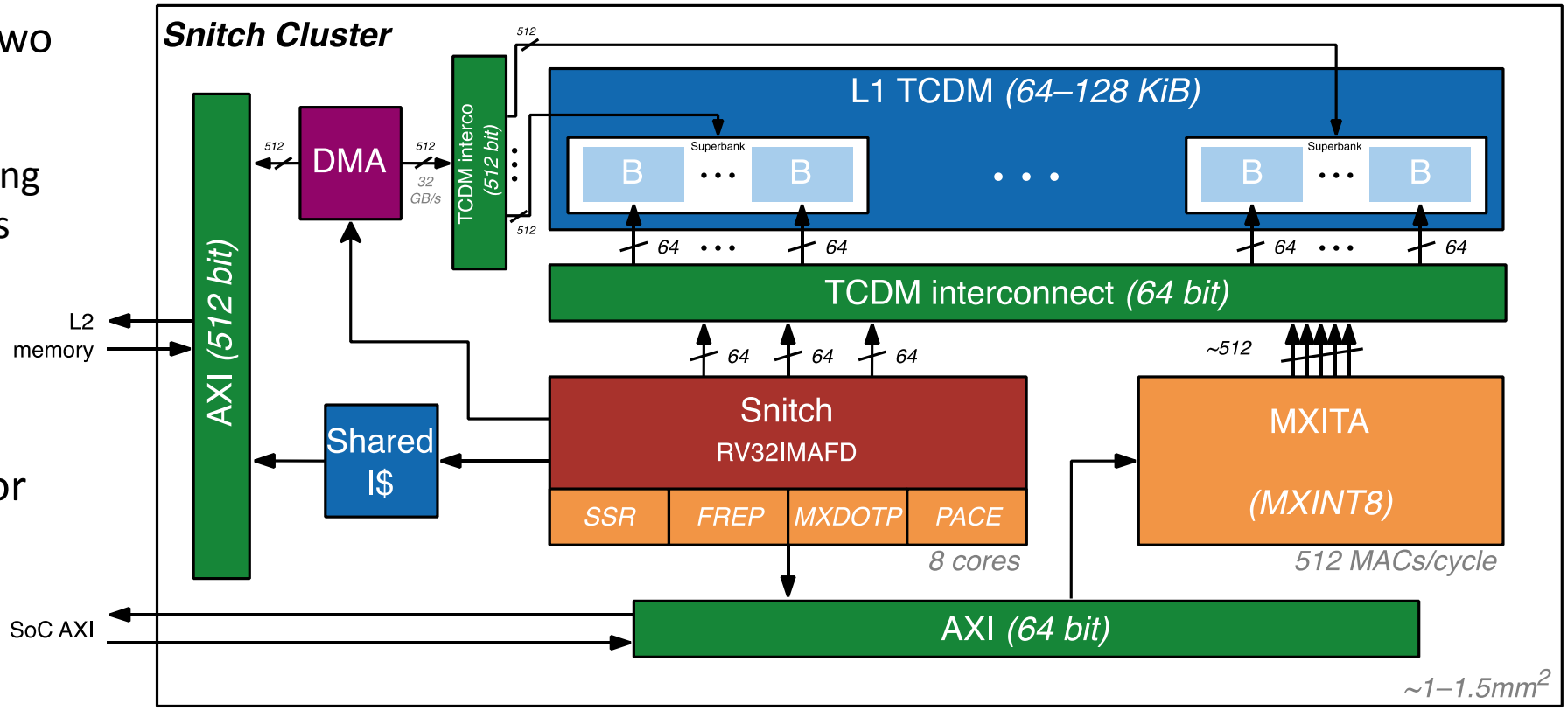


On the Horizon: SeyrITA – Top Level



On the Horizon: SeyrITA – Compute Cluster

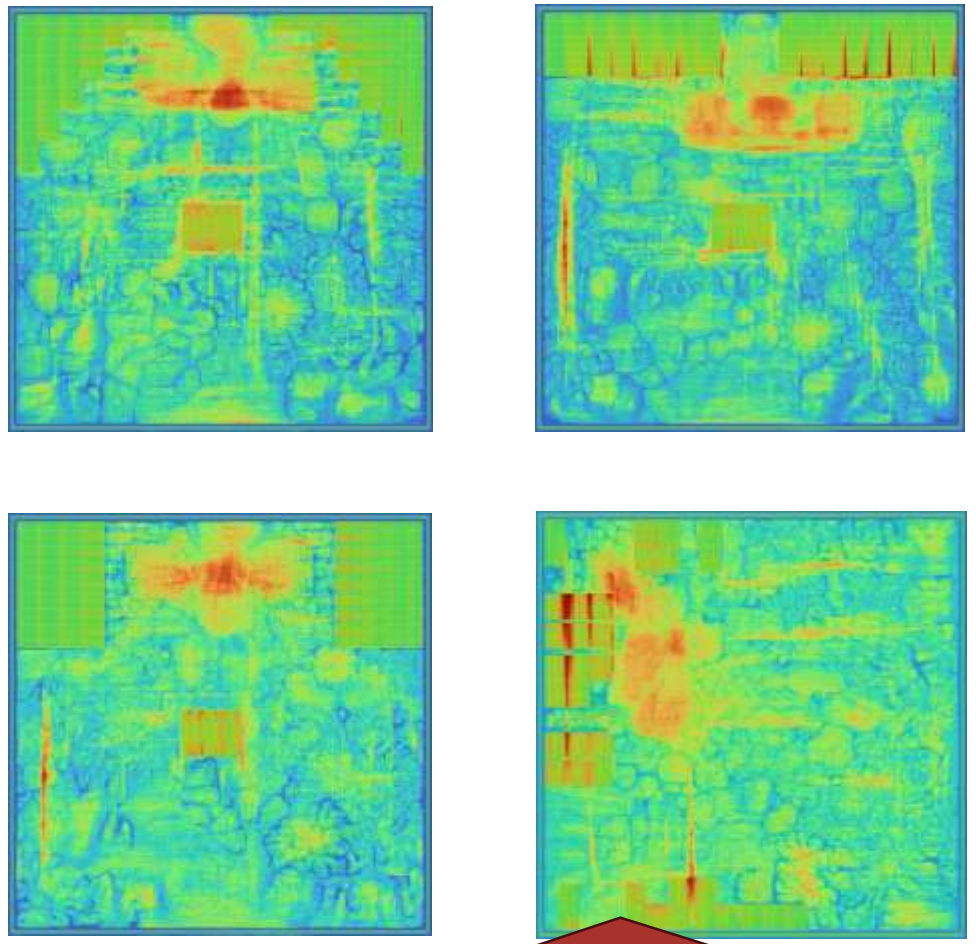
- **8 Snitch cores** with two new ISA extensions:
  - **MXDOTP**: Microscaling (MX) FP dot products
  - **PACE**: Piecewise polynomial approximations
- **MXITA accelerator** for MXINT8 matrix multiplications



One MXITA = 512 MACs/cycle = 1024 FLOP/cycle = 512 GFLOP/s @500MHz

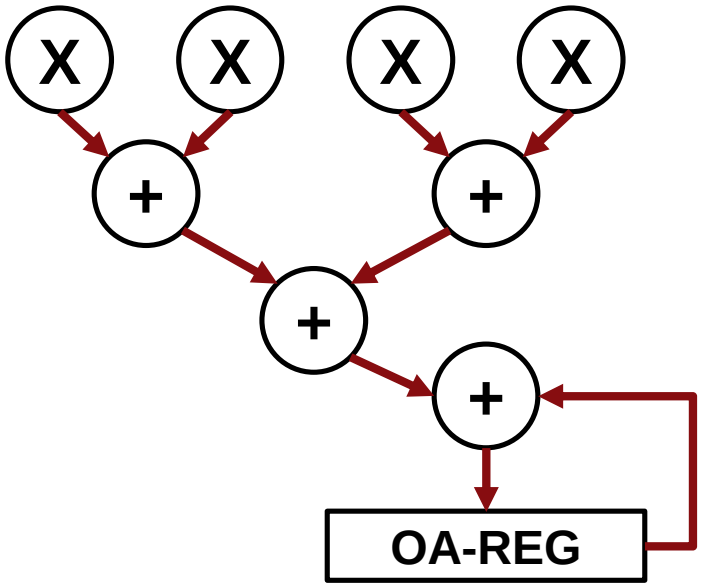
On the Horizon: SeyrITA – Working on the Tapeout

- **Demonstrate** a large **22nm tapeout** with open-source tools
- **Improve tools** and close the **performance** gap
- Identify and **implement missing features** along the way
- **Active Collaboration with**

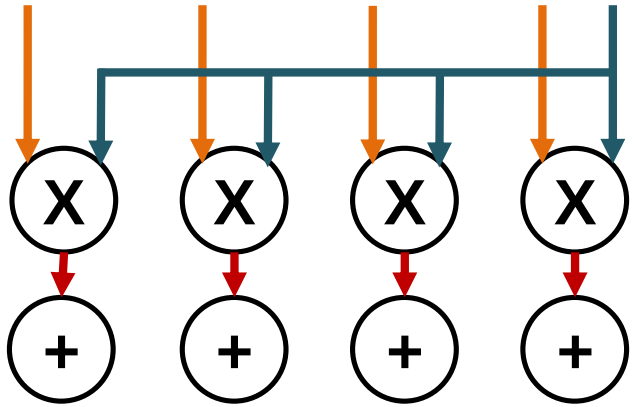


Snitch cluster floorplan exploration

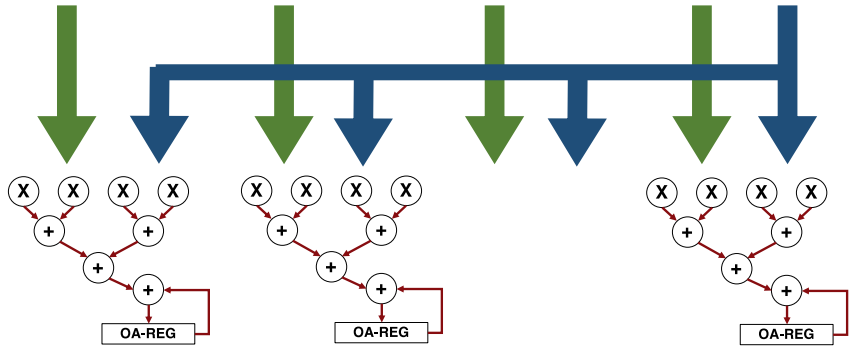
Yes but why? Specialization + EDA multiplicative effect!



Inner Product



Outer Product



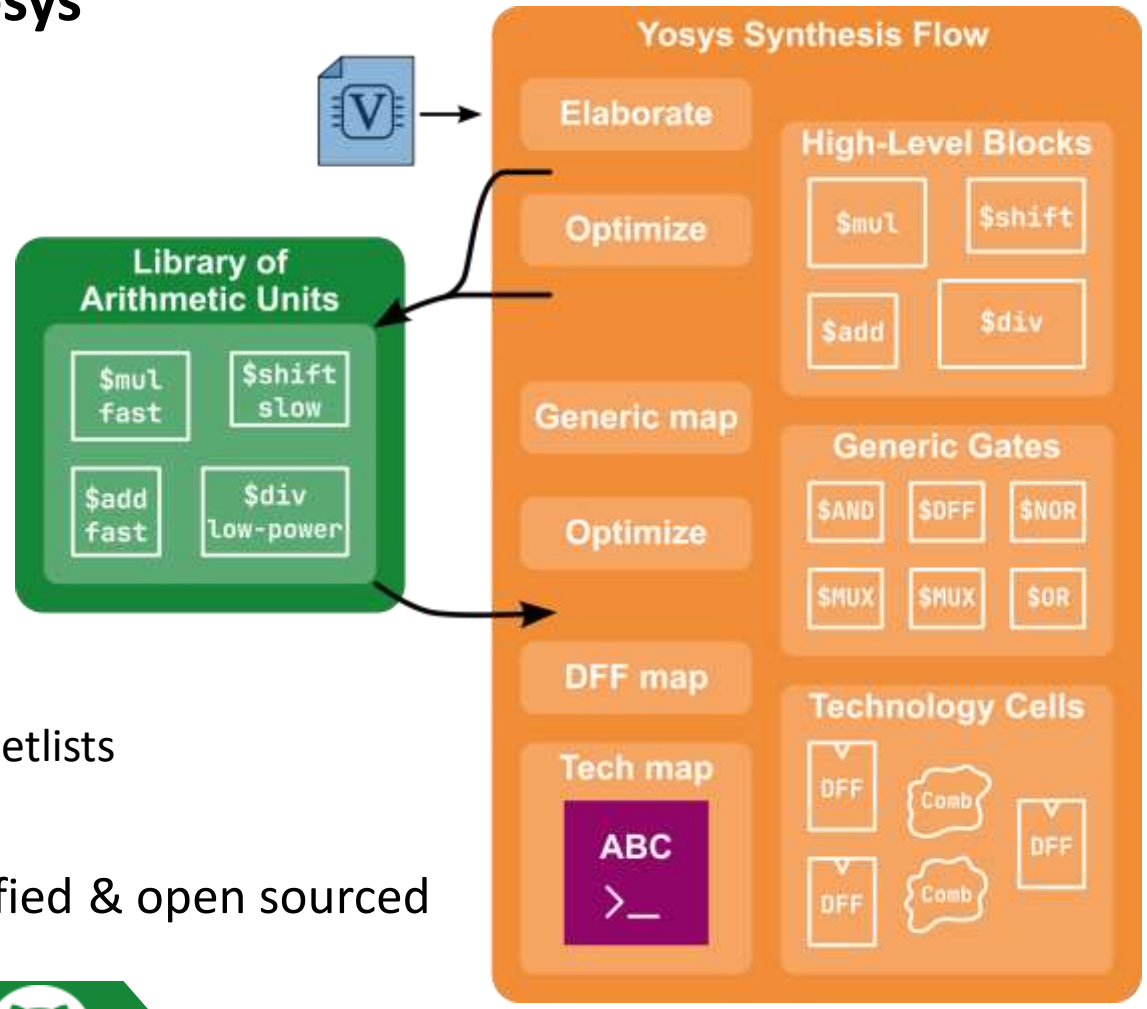
Mixed

Precision tuning – OP/Mem tuning - deep arithmetic optimization – operand network tuning...

Co-Specialize SW, HW, EDA & Technology is the frontier

Library of Arithmetic Unit (LAU)

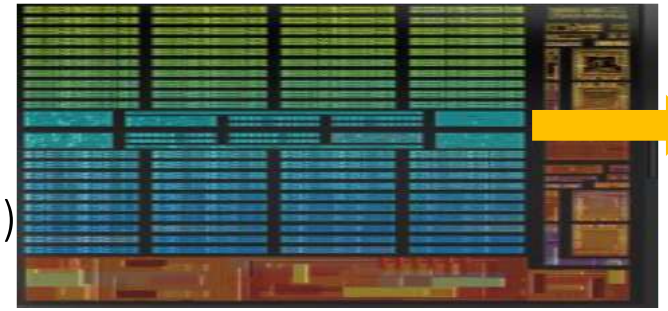
- **Block replacement is implemented in Yosys**
  - Only used in FPGA designs to infer DSP slices
  - Detect and replace arithmetic operators
  - Currently: manual selection
  - ***Next: AI based!***
- **No open-source LAU**
  - Well-optimized library is key to good results
  - A LAU created at IIS as part of a PhD thesis
    - A wide range of arithmetic operations
    - 3 different performance variants of generic gate netlists
    - Thoroughly QoR evaluated and optimized
  - SystemVerilog port: LLM translation, hand-verified & open sourced

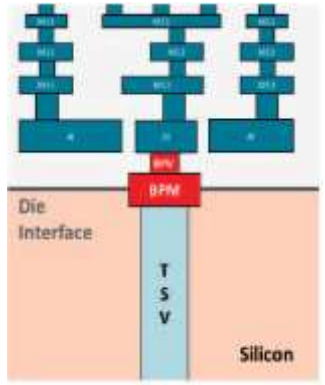


github.com/pulp-platform/elau

What Happens Next?

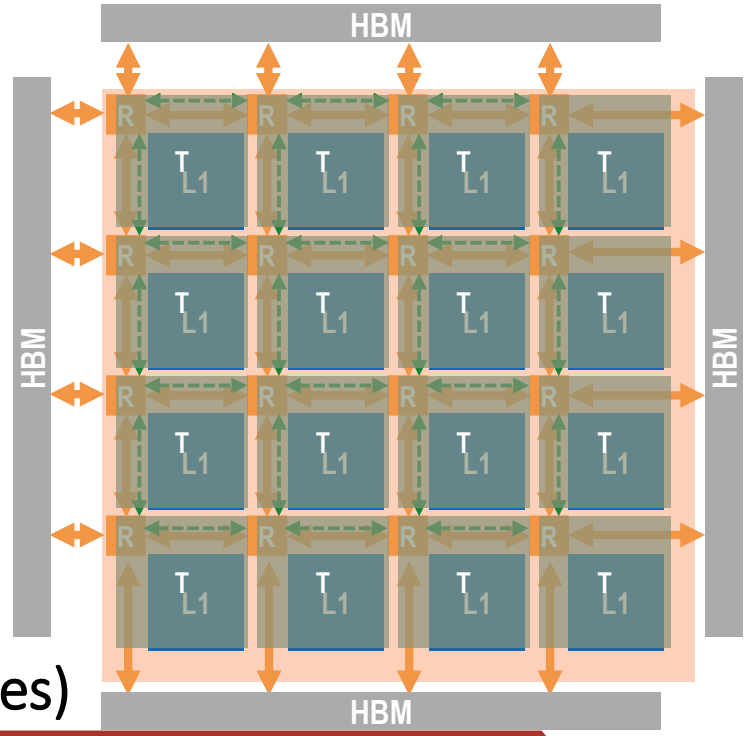
- 3.5D v1
  - 3D stacking on logic + 2.5D HBM (AMD MI300)
  - Face (top) to Back (bottom)
  - Die (top) to Wafer (bottom)



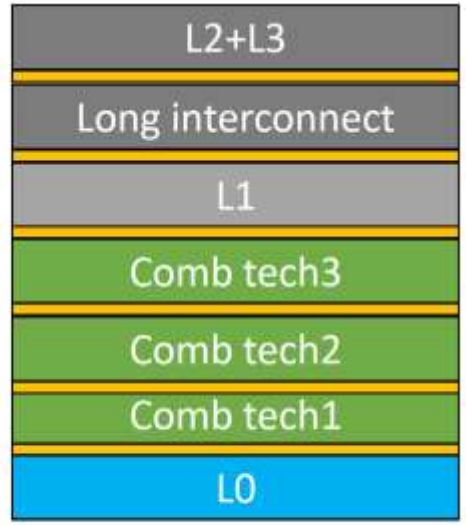


Logic die
Memory + IO die

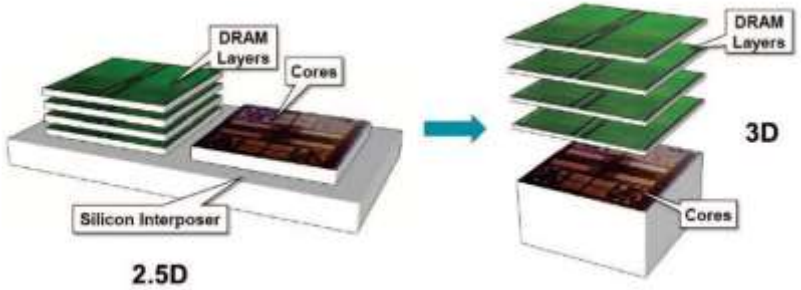
MI300 Instinct™



V1
SRAM+NOC+IO at the bottom



- 3.5D v2?



- Monolithic 3D (CMOS2.0+3D memories)

Technology is going "full 3D", OS-EDA is right there*

*github.com/lamda-bbo/Open3DBench


http://pulp-platform.org
@pulp_platform
The future is bright for open source!