ETH zürich

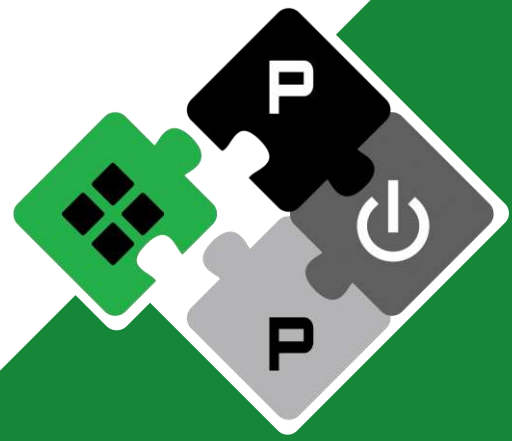
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

IEEE/ACM ICCAD
2025 INTERNATIONAL CONFERENCE ON COMPUTER-AIDED DESIGN
44th Edition

End-to-end Open Source Platforms in the era of Domain-Specific Design Automation

From Dream to Reality

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



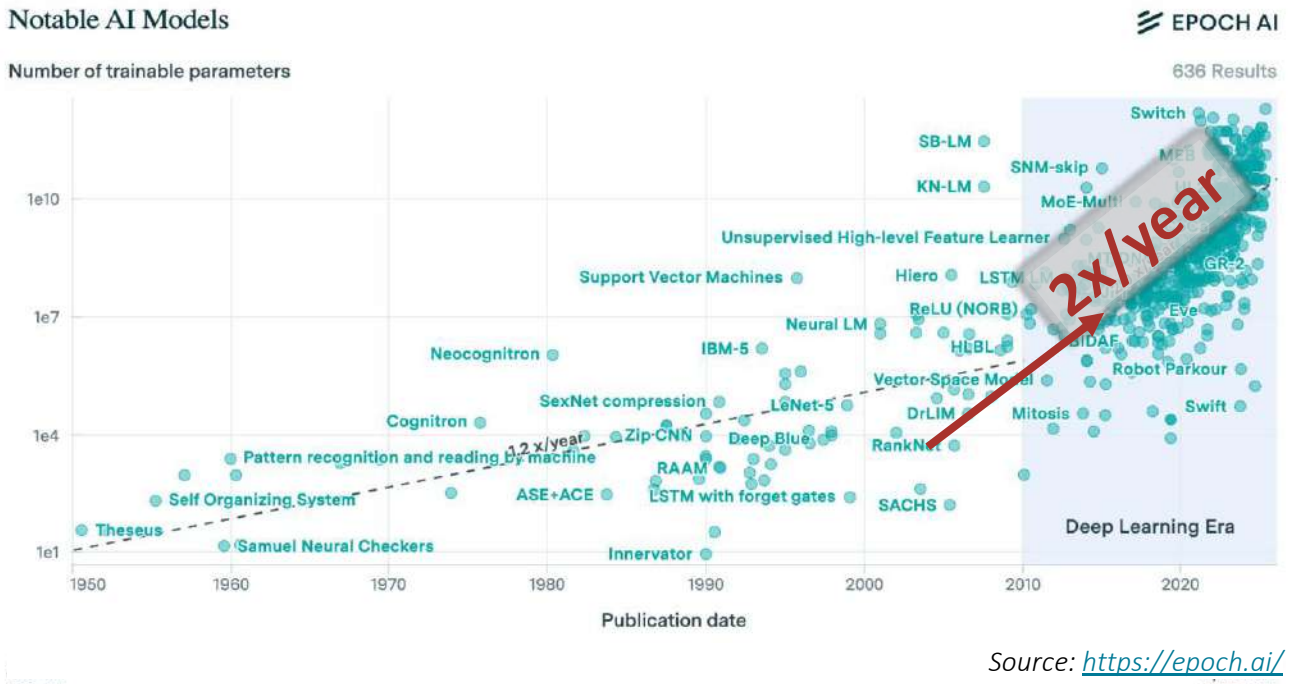
pulp-platform.org
@pulp_platform
company/pulp-platform
youtube.com/pulp_platform

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

AI Hardware Platforms: Model Scaling Laws

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

- Growing model sizes demand ever-increasing compute and memory

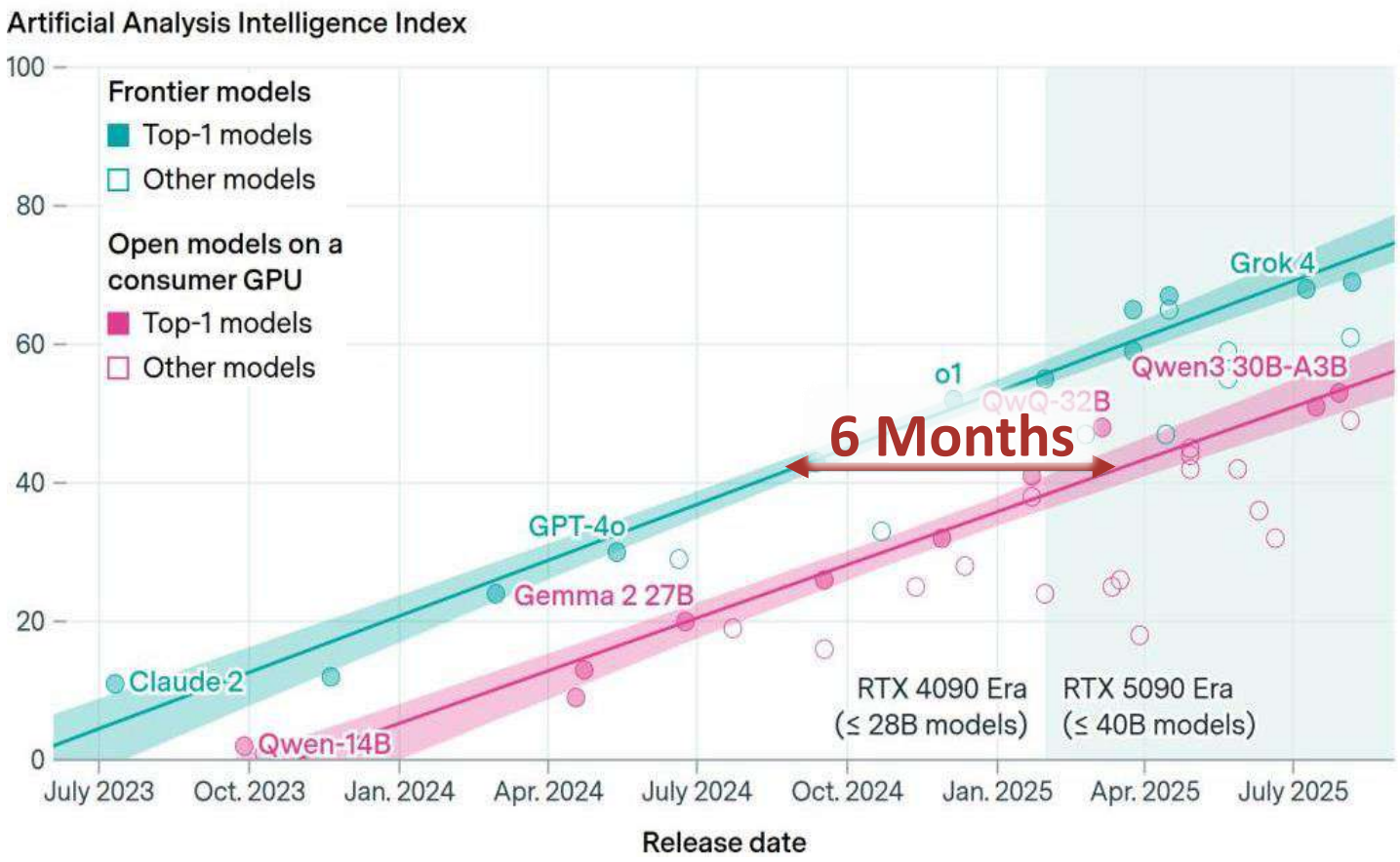


Source: https://epoch.ai/

AI Hardware Platforms: Scaling @Inference, @Edge

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

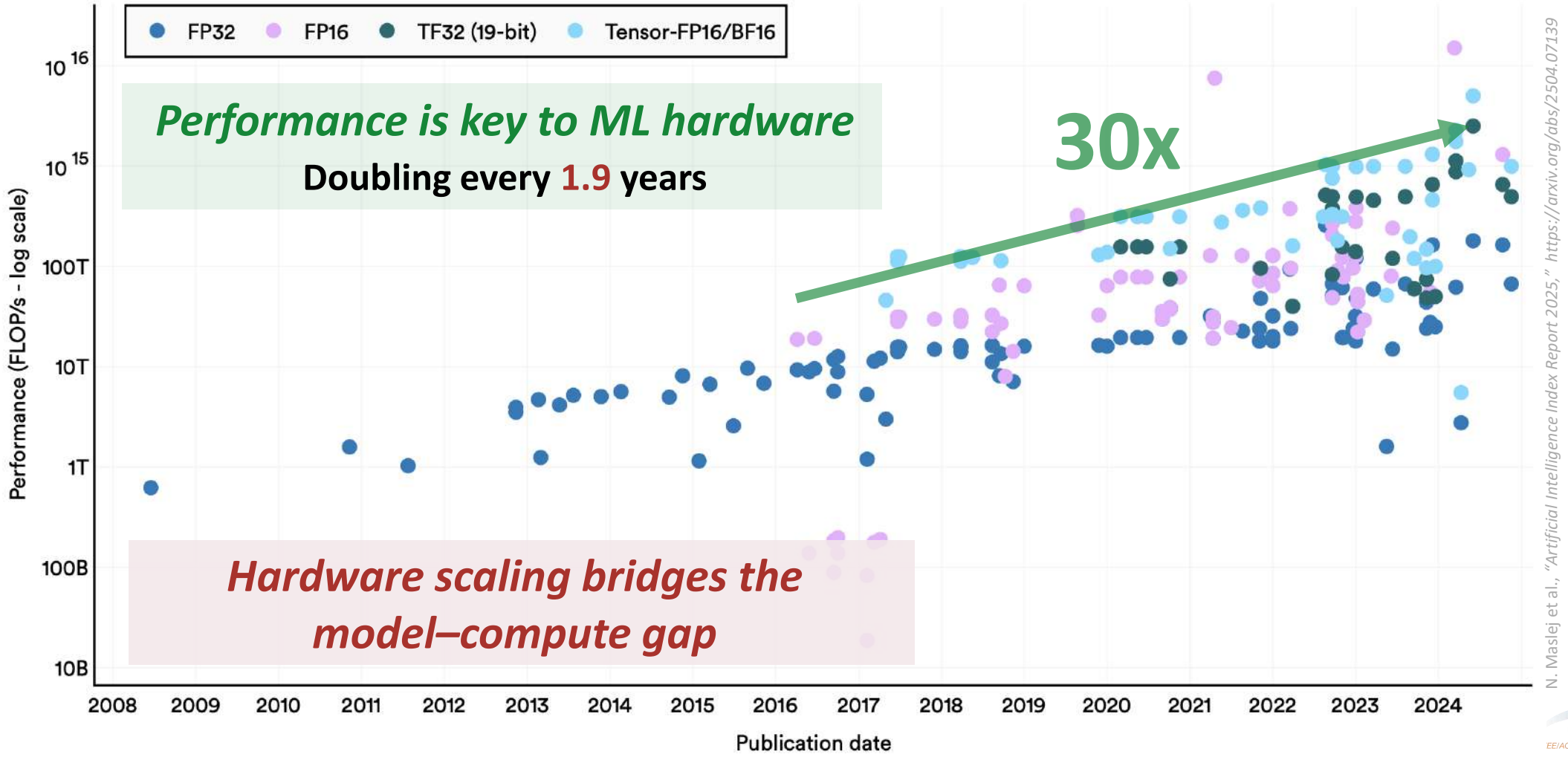
- Growing model sizes demand ever-increasing compute and memory
- ***Inference*** compute scales even faster than for training
- ***Edge*** models that fit on a single GPU trail the frontier by less than one year



Source: https://epoch.ai/

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

Peak computational throughput of notable ML hardware (energy efficiency must track!)

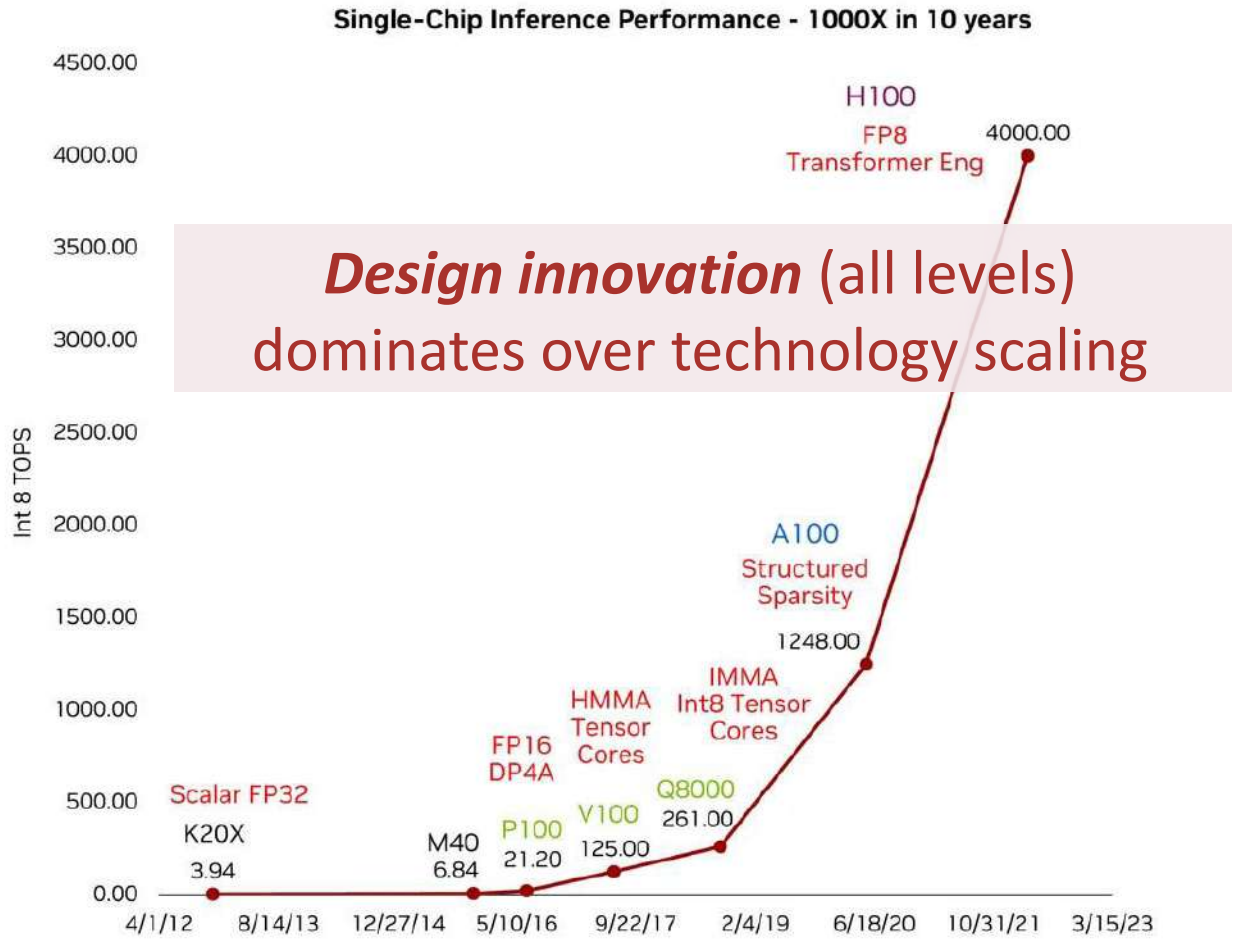


N. Maslej et al., *"Artificial Intelligence Index Report 2025,"* https://arxiv.org/abs/2504.07139

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



Design innovation (all levels) dominates over technology scaling

[Dally HotChips 2023]

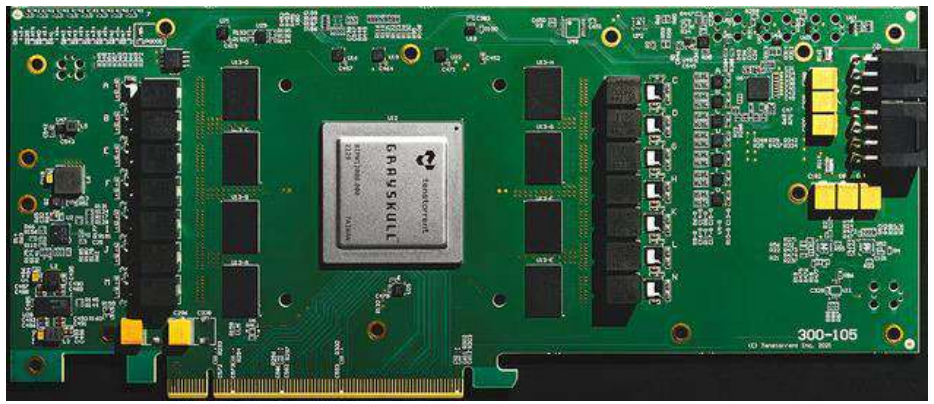
Innovation beyond “NVIDIA Gravity” is Challenging!

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

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

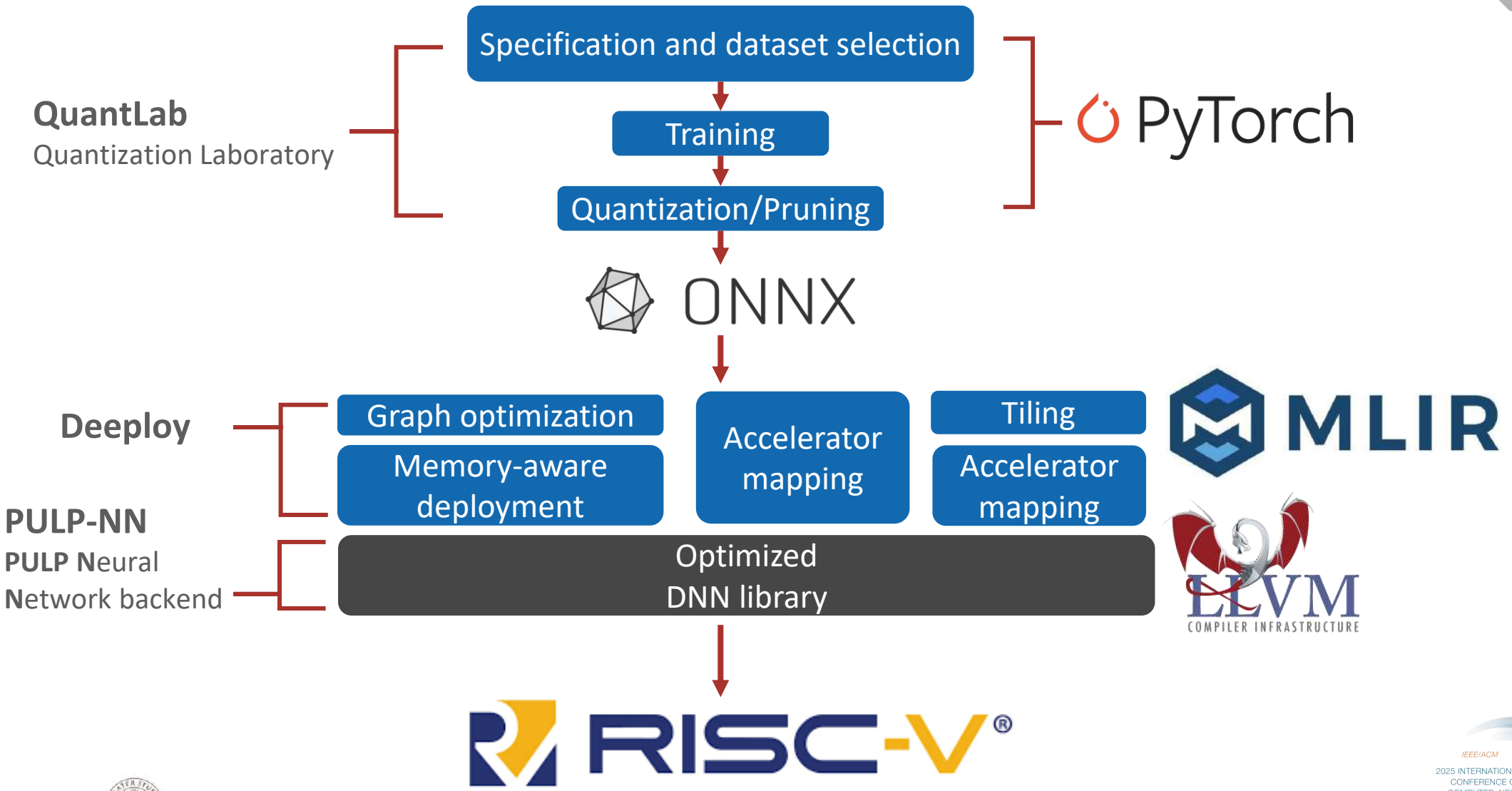
RISC-V®

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



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

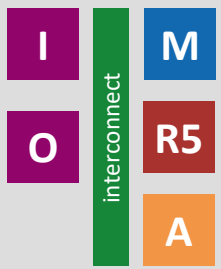
Peripherals

JTAG, SPI, UART, I2S, DMA, GPIO

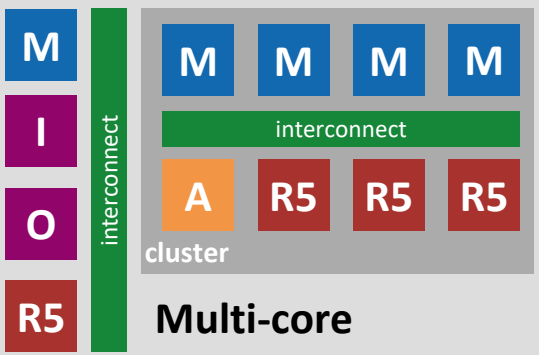
Interconnects

LIC, HCI, APB, FlooNoC, AXI4

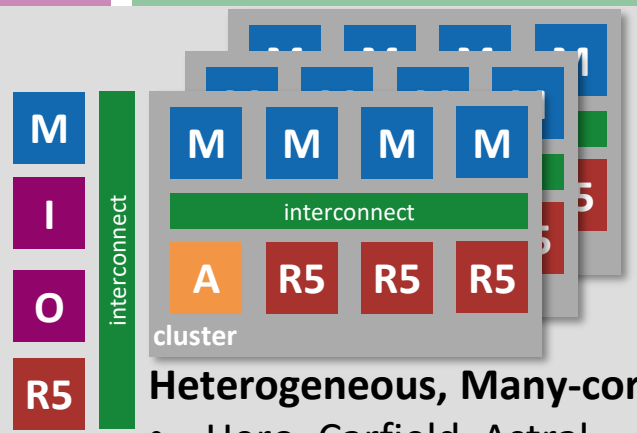
Platforms



Single core
- PULPissimo, Croc
- Cheshire



Multi-core
- OpenPULP
- ControlPULP



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

IOT → **HPC**

Accelerators and ISA extensions

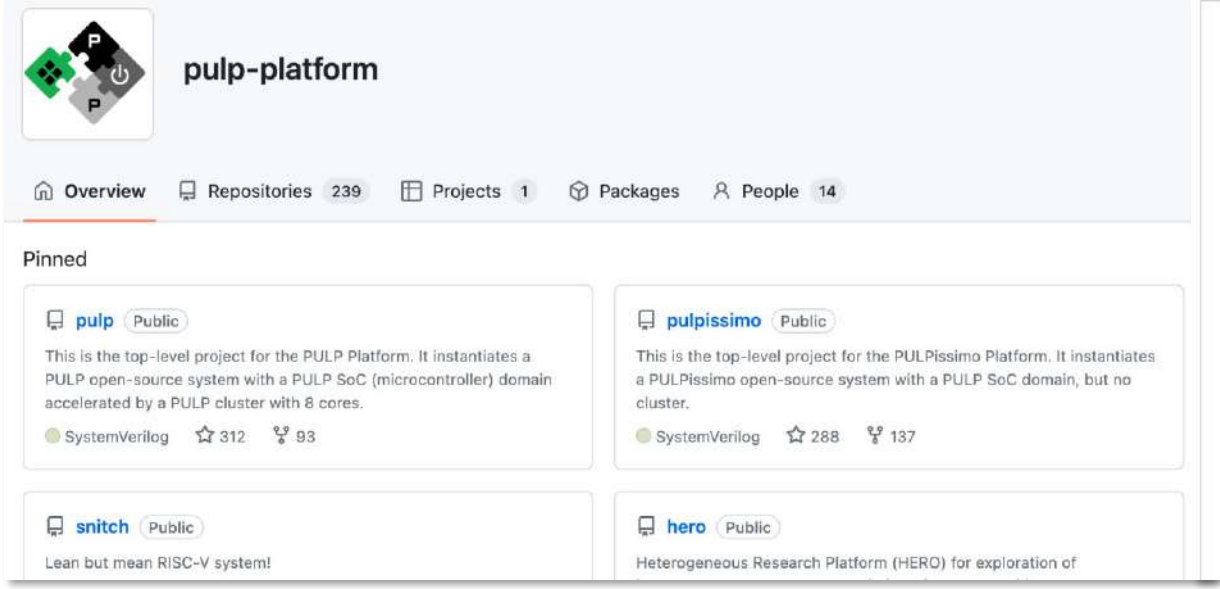
XpulpNN, XpulpTNN	ITA (Transformers)	RBE, NEUREKA (QNNs)	FFT (DSP)	REDMULE (FP-Tensor)

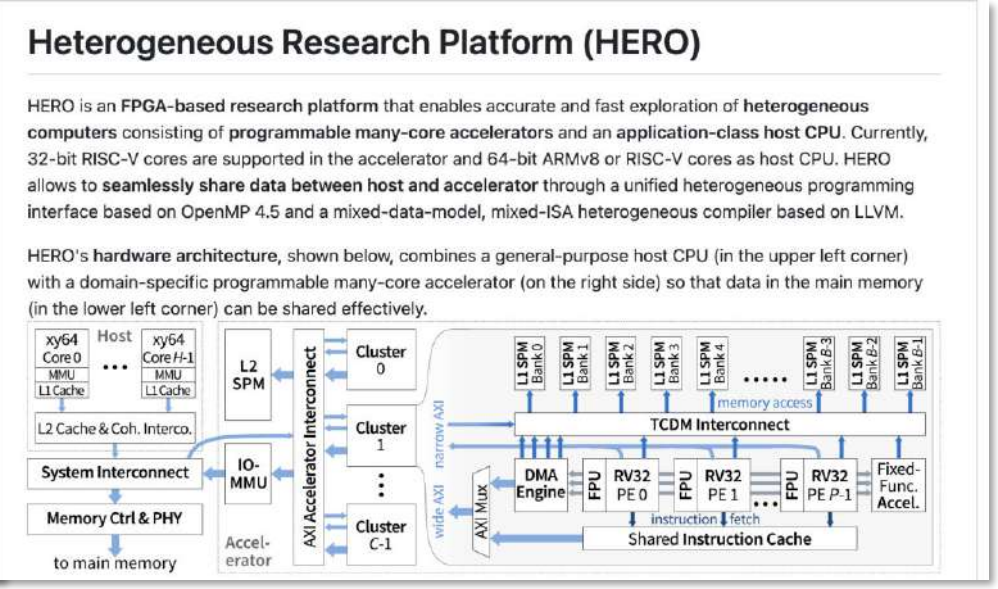
All of our designs are open-source hardware

- **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

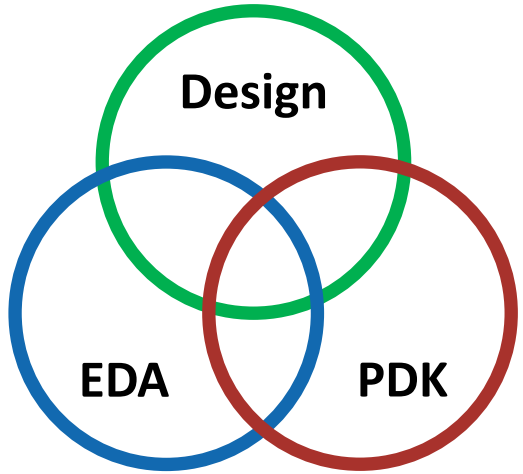
- RTL / HDL descriptions (quite common)
- Schematics / Physical Design (may have dependencies to technology information)

Tools (EDA)

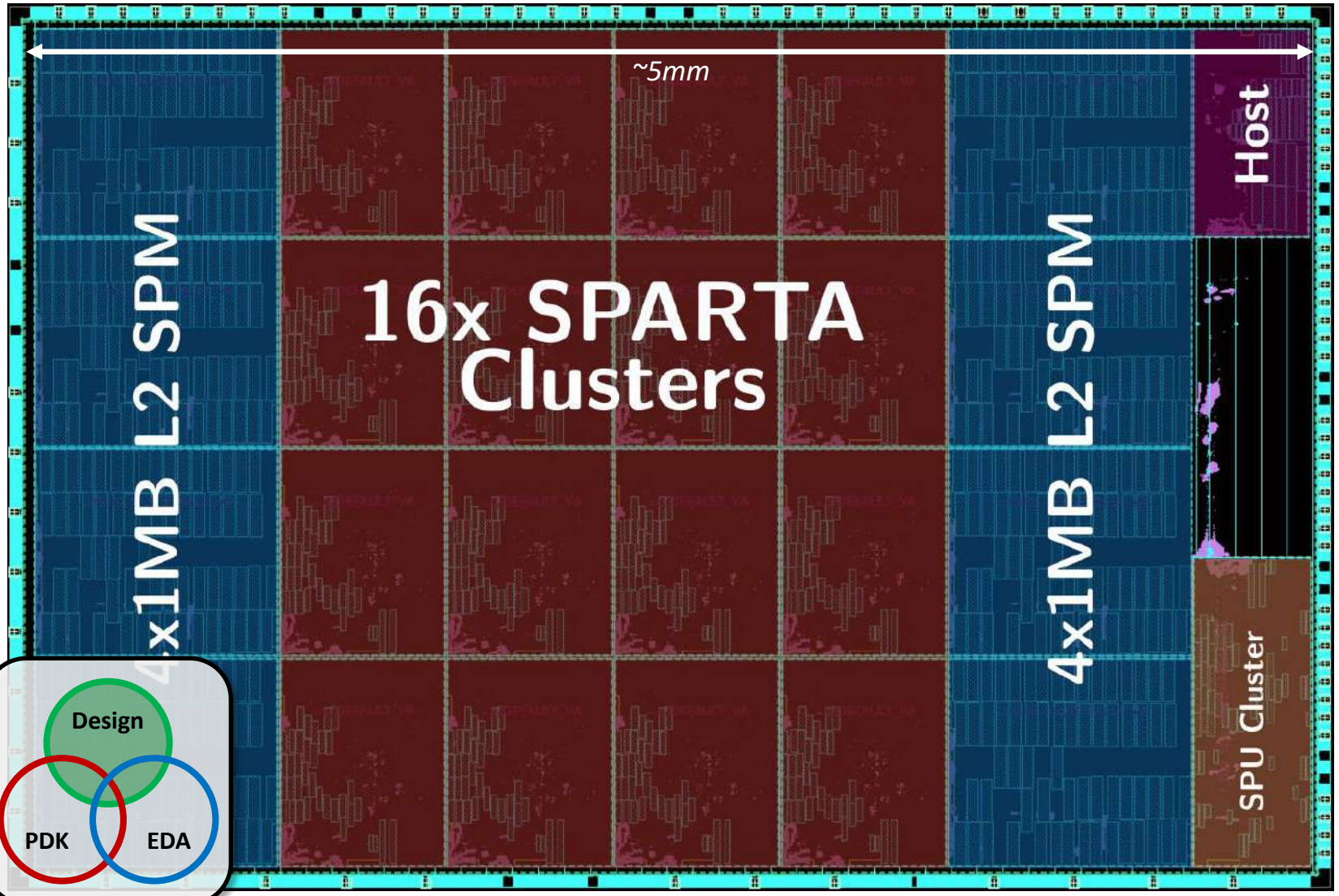
- Front-end tools (Synthesis)
- Back-end tools (Placement and Routing)
- Verification tools (Simulation)

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



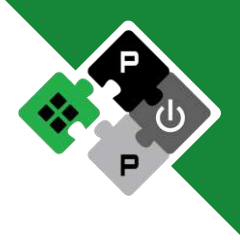
Here is Picobello: our latest design in TSMC 7nm



- 16 SPARTA clusters totaling 144x RISC-V cores with FP8-FP64-bit support
- 8x 1MB of on-chip L2
- Linux capable CVA6 Host
- Peripherals (JTAG, SPI, I2C)
- Running at 1+ GHz (WC),
  > 256 GFLOP/s (FP64)
  > 2 TFLOP/s (FP8)
- Tape-out August 2025
- Part of the EU Pilot project

THE EUPILOT

You need help to make large modern SoCs



- **There are many innovative parts in Picobello**
  - Hopefully you will read about it in publications starting in 2026
- **But you can not afford to design all parts of an SoC from scratch by yourself**
  - **It builds on successful designs from the past**
    - CVA6 core : `https://github.com/openhwgroup/cva6`
    - Cheshire platform & peripherals : `https://github.com/pulp-platform/cheshire`
    - Snitch clusters : `https://github.com/pulp-platform/snitch_cluster`
    - FlooNoc : `https://github.com/pulp-platform/floonoc`
    - AXI : `https://github.com/pulp-platform/axi`
  - **Needs collaborations with experienced teams**
    - University of Bologna, TU-Munich

Supported by open-source designs in HDL

What about technology specific IPs? FLL, DDR,… ?

Sharing our FLL with others that need it

- **When you start designing in more modern technologies clock rates increase**
  - For 28nm and less clock rates of 500MHz – 2GHz are easily possible
  - It is difficult to bring such clocks externally, an internal clock generator could help
  - Good luck getting access to a low-cost clocking IP ☺
- **We designed an FLL (Frequency Locked Loop) back in 2016**
  - D. E. Bellasi and L. Benini, "Smart Energy-Efficient Clock Synthesizer for Duty-Cycled Sensor SoCs in 65 nm/28nm CMOS," in *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 64, no. 9, pp. 2322-2333, Sept. 2017, doi: 10.1109/TCSI.2017.2694322.
  - Ported and taped-out in many technologies: UMC65, TSCM65, GF22, GF12, TSMC7…

People have asked us repeatedly if we could share our FLL

.. and we really want to share our FLL, we know how much it helped us

Issues of sharing technology specific IP (like our FLL)

- **Assuming you have no objections from your own institution to share your IP**

The design requires technology data that is under NDA

- You can not share anything without getting permission from the technology provider
- Usually this requires a multi-party NDA

Your design may contain standard cells that come from a different provider

- You will have to contact the IP provider to ask for permission
- Our FLL for GF22 uses INVECAS standard cell libraries now owned by Synopsys
- In theory you can send a design WITHOUT the cells only references, and add the cells locally

You have used EDA tools with academic licenses for your design

- Most academic license agreements would not allow you to transfer the output to others
- You need to contact the EDA vendor and ask for their permission to share

Issues of sharing technology specific IP (like our FLL)

- Assuming you have no objections from your own institution to share your IP

The design requires technology data that is under NDA

- You can not share anything without getting permission from the technology provider
- Usually this requires a multi-party NDA

Your design may contain standard cells that are from a different provider

- You will have to contact the IP owner to ask for permission
- Our FLL for GF22 uses INVECAS standard cells now owned by Synopsys
- In theory you can send the GDS WITHOUT the cells, only references, and add the cells locally

You have used EDA tools with academic licenses for your design

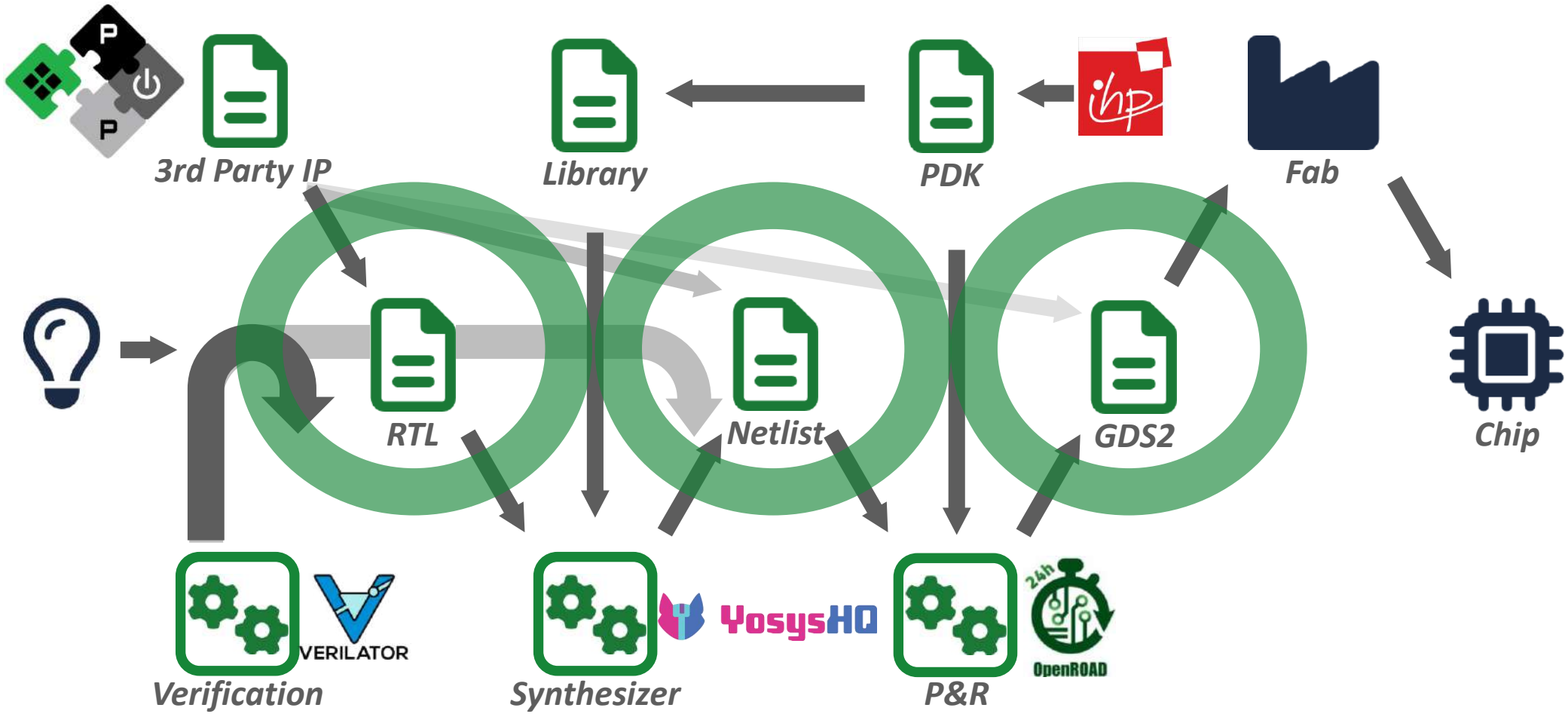
- Most academic license agreements would not allow you to transfer the output to others

These things take a loooong time and are very tedious!

ON average we spend

6 months to put these in place

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`

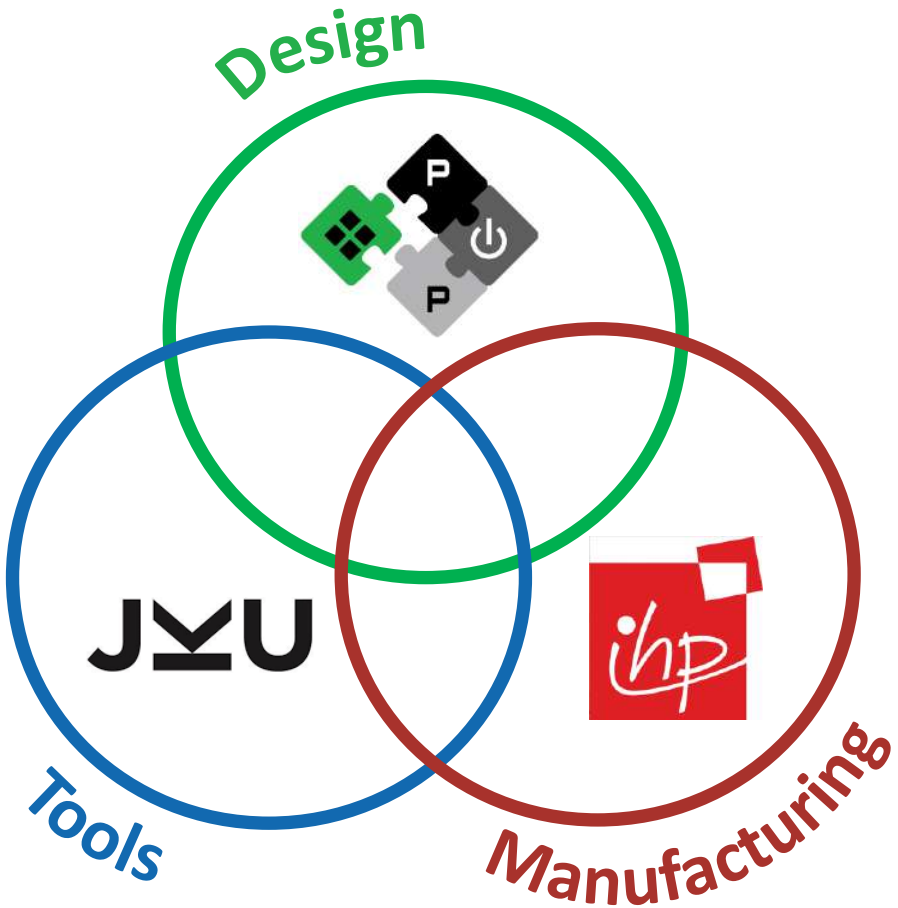
Tools: from Johannes Kepler University (JKU)

Reliable VM with large collection of open-source tools

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

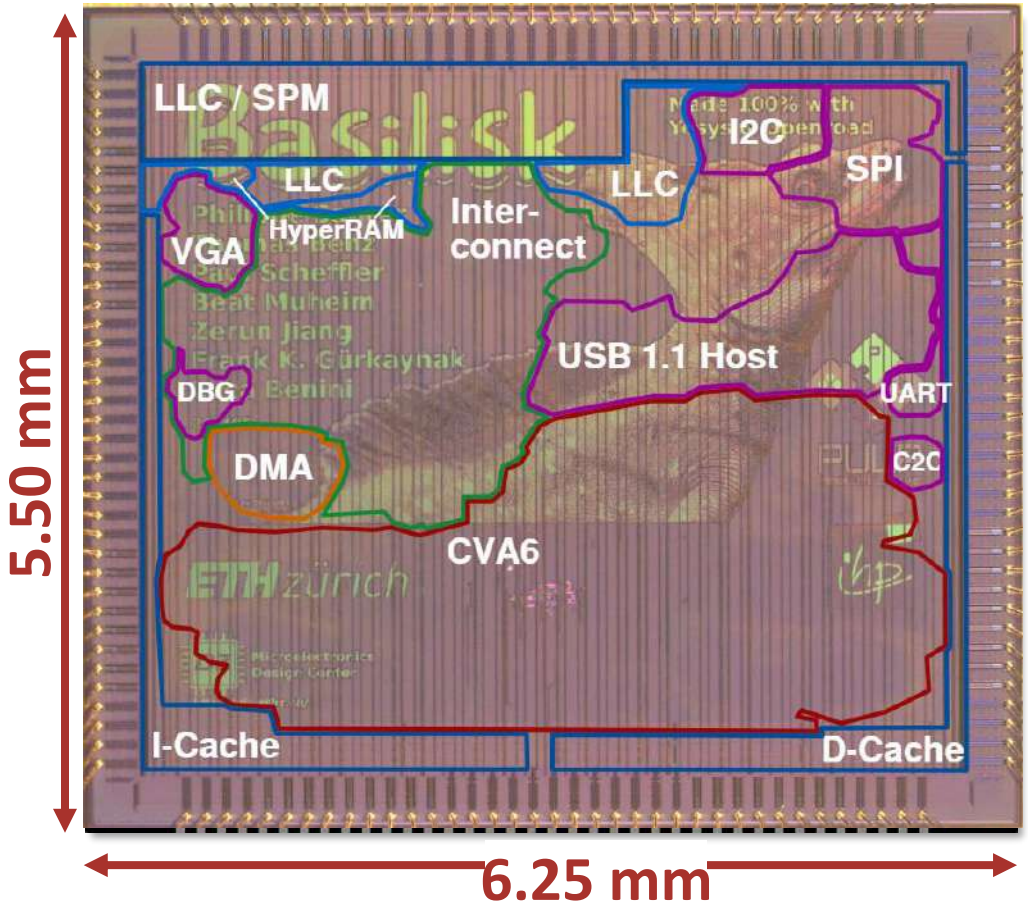
Manufacturing: IHP130nm

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

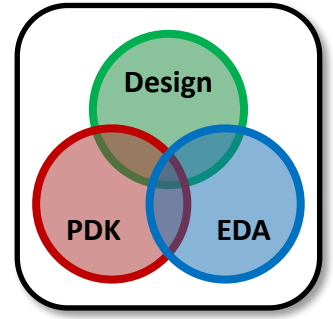


Is it practical? Can we really use this for large SoCs tapeouts?

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

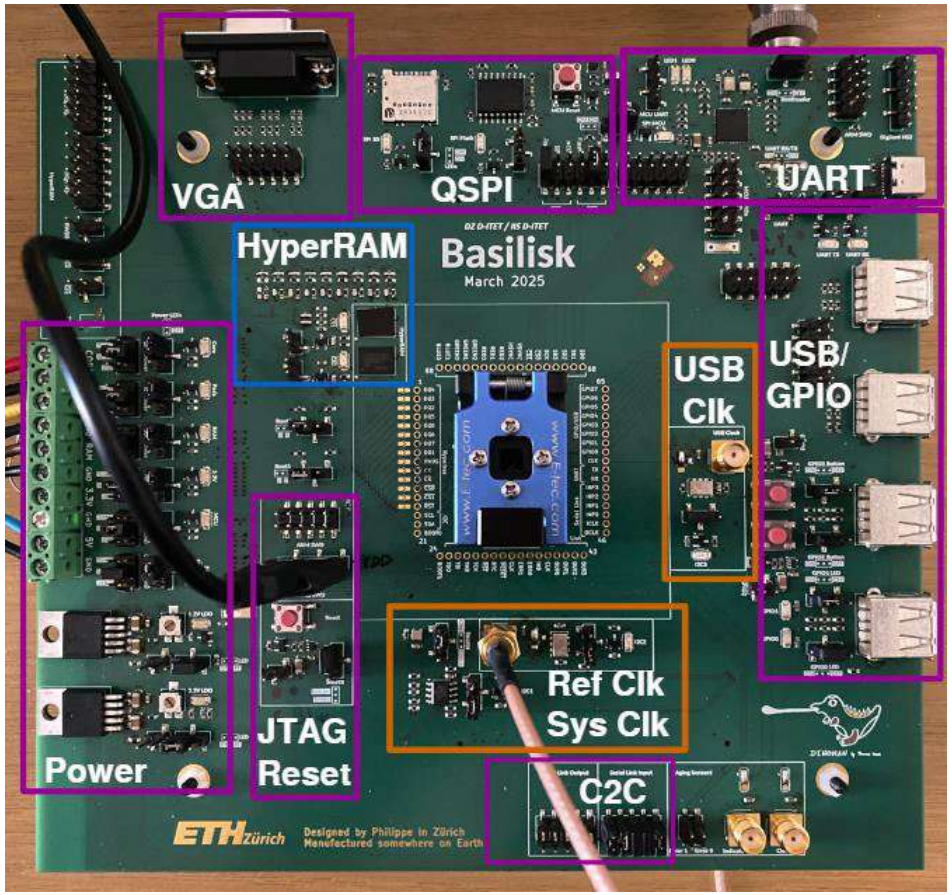


- **Designed in IHP 130nm OpenPDK**
  - **34mm²** (6.25mm x 5.50mm)
  - **~5× larger** than previous end-to-end OS designs
  - 2.7 MGE total, 1.14MGE logic
  - 24 SRAM macros (114 KiB)
  - **62MHz** at nominal voltage (1.2V)
- **RV64GC design runs Linux**
- **Active collaboration with**



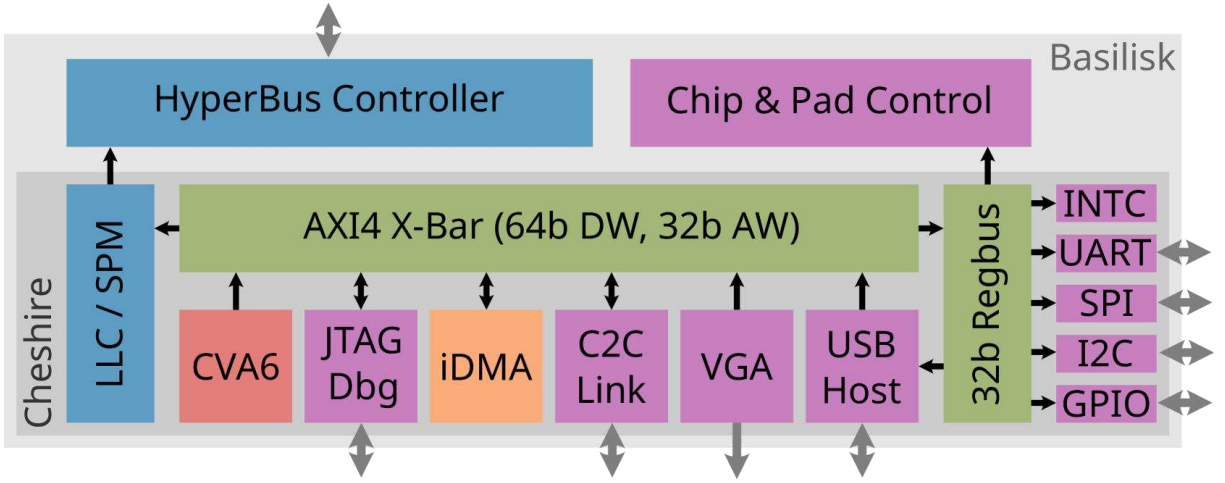
github.com/pulp-platform/cheshire-ihp130-o

Basilisk is a complete Linux-capable SoC



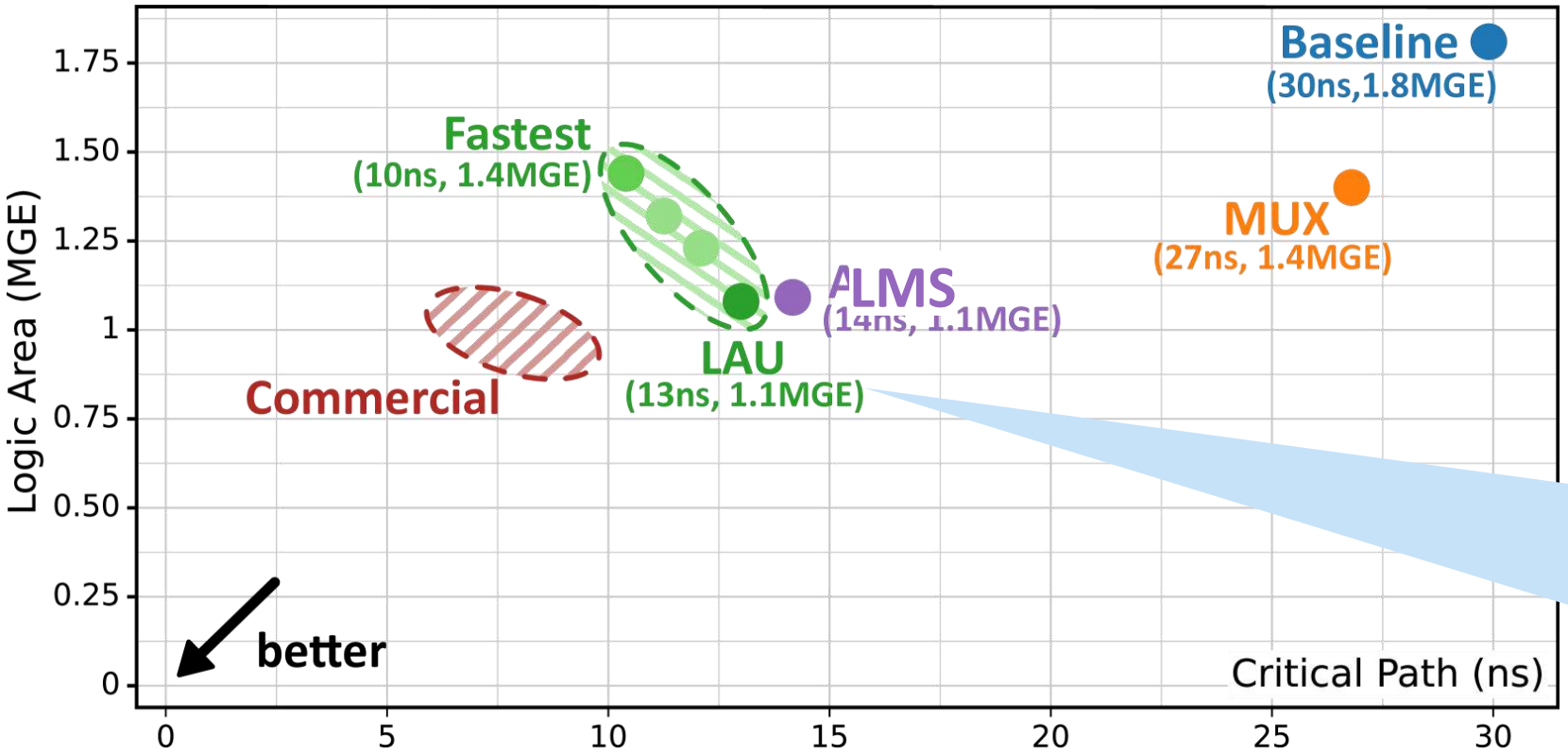
- **64-bit RISC-V core**
- **Rich peripherals:**
  - HyperRAM controller **@154MB/s**
  - C2C AXI-Link **@77MB/s**
- **Automatic boot via scratchpad**

Watch the video



arxiv.org/pdf/2505.10060

What about PPA gap wrt commercial EDA?



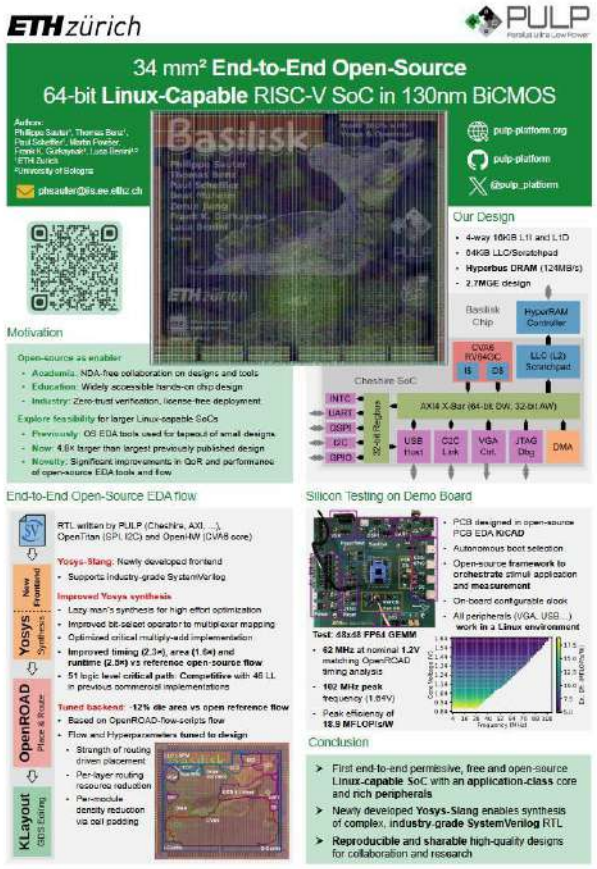
Yosys-slang full Sysverilog Frontend: @ <6sec runtime (from minutes)

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

Commercial EDA leads, but OS-EDA IS usable, now!

We presented Basilisk at HOT CHIPS



Industry Noticed!

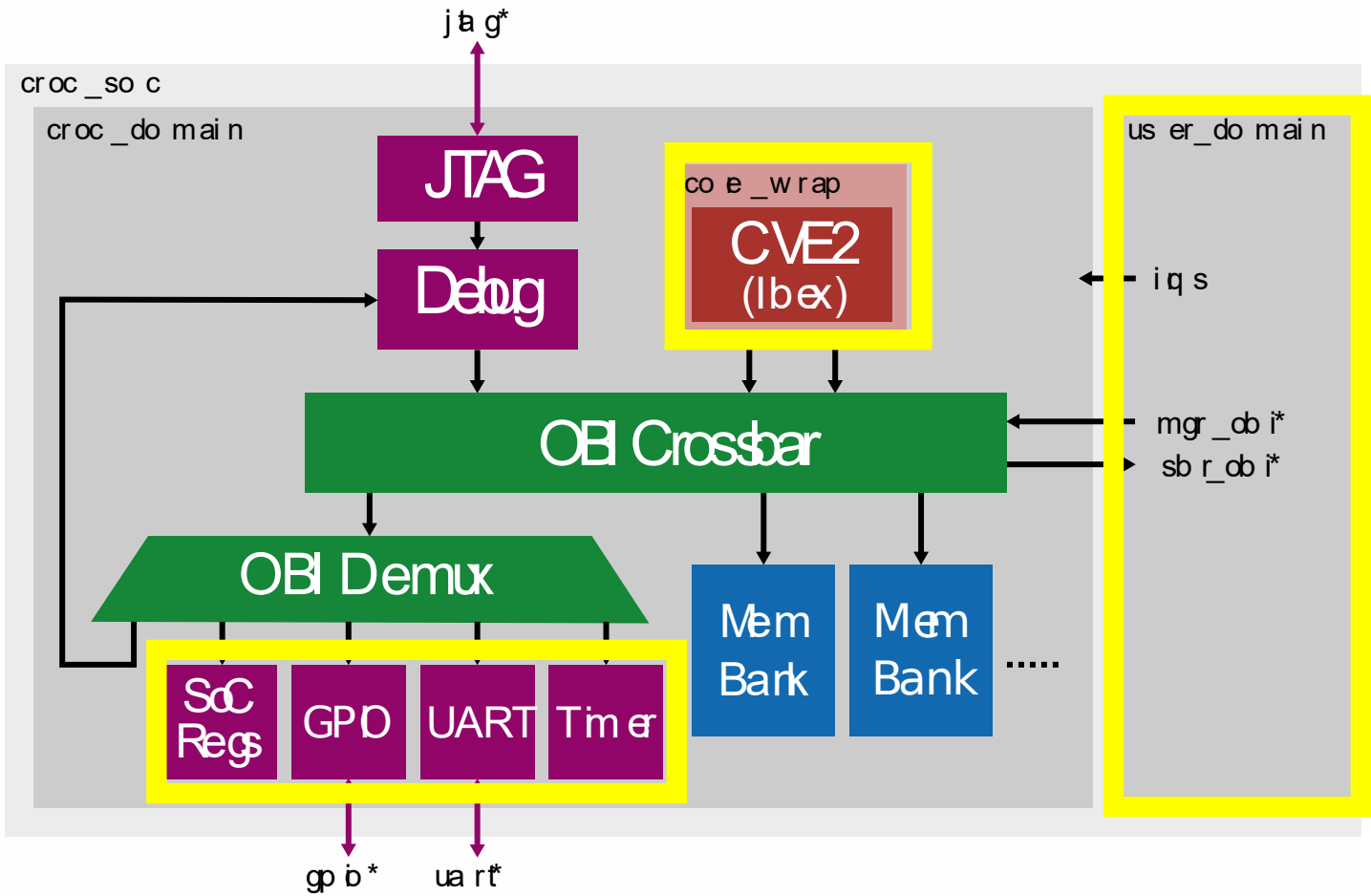
"Basilisk at Hot Chips 2025 Presented Ominous Challenge to IP/EDA Status Quo"*

Poster: lnkd.in/daB6HskB

*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)

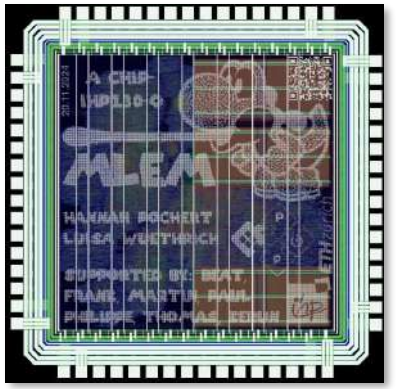


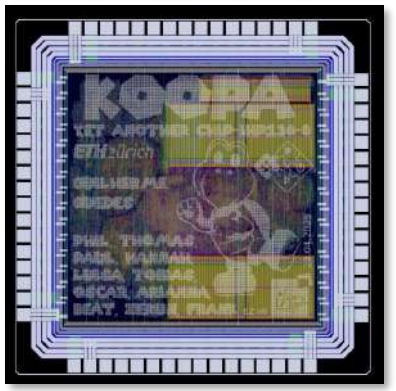
github.com/pulp-platform/croc

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

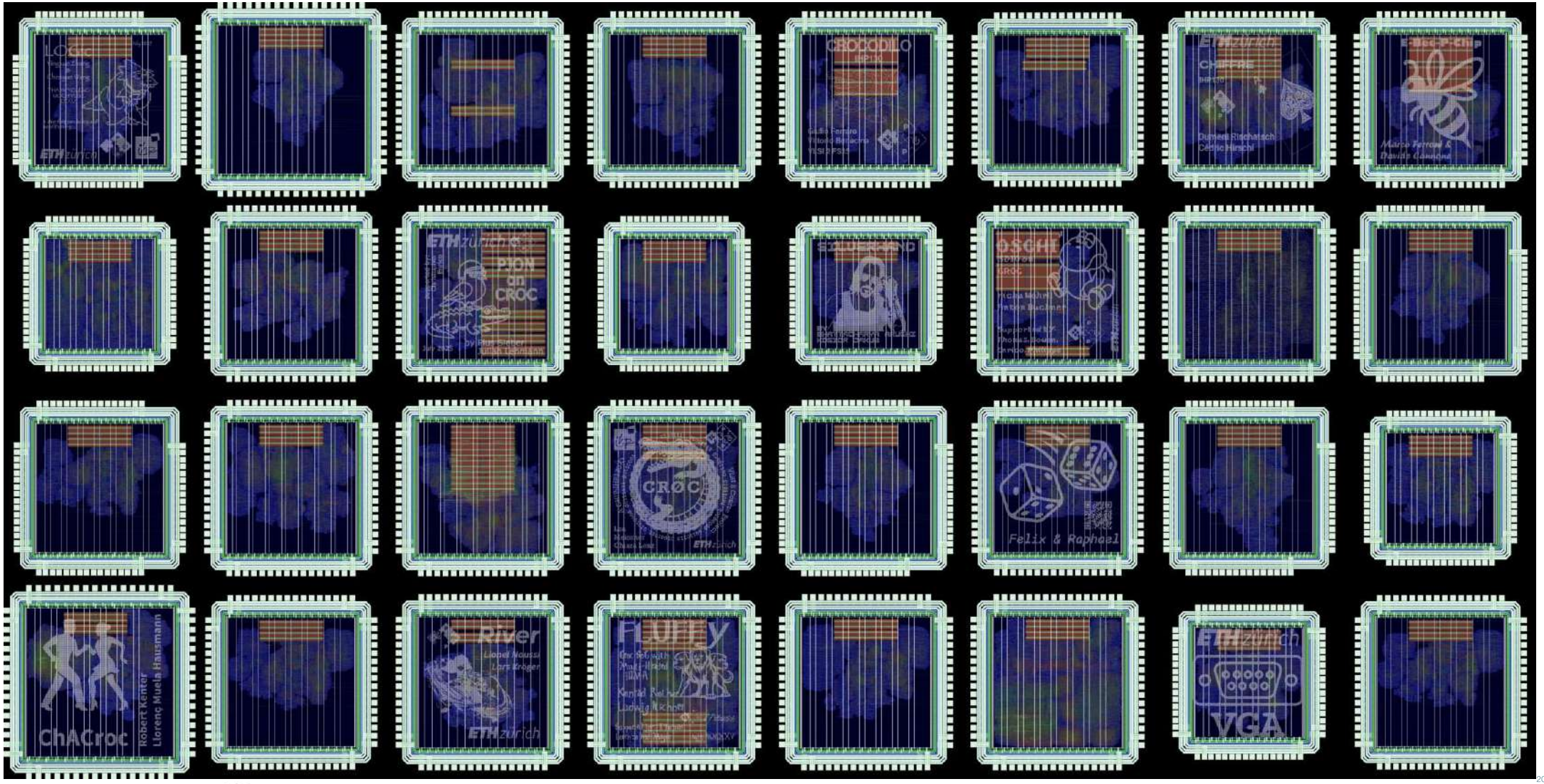
https://vlsi.ethz.ch





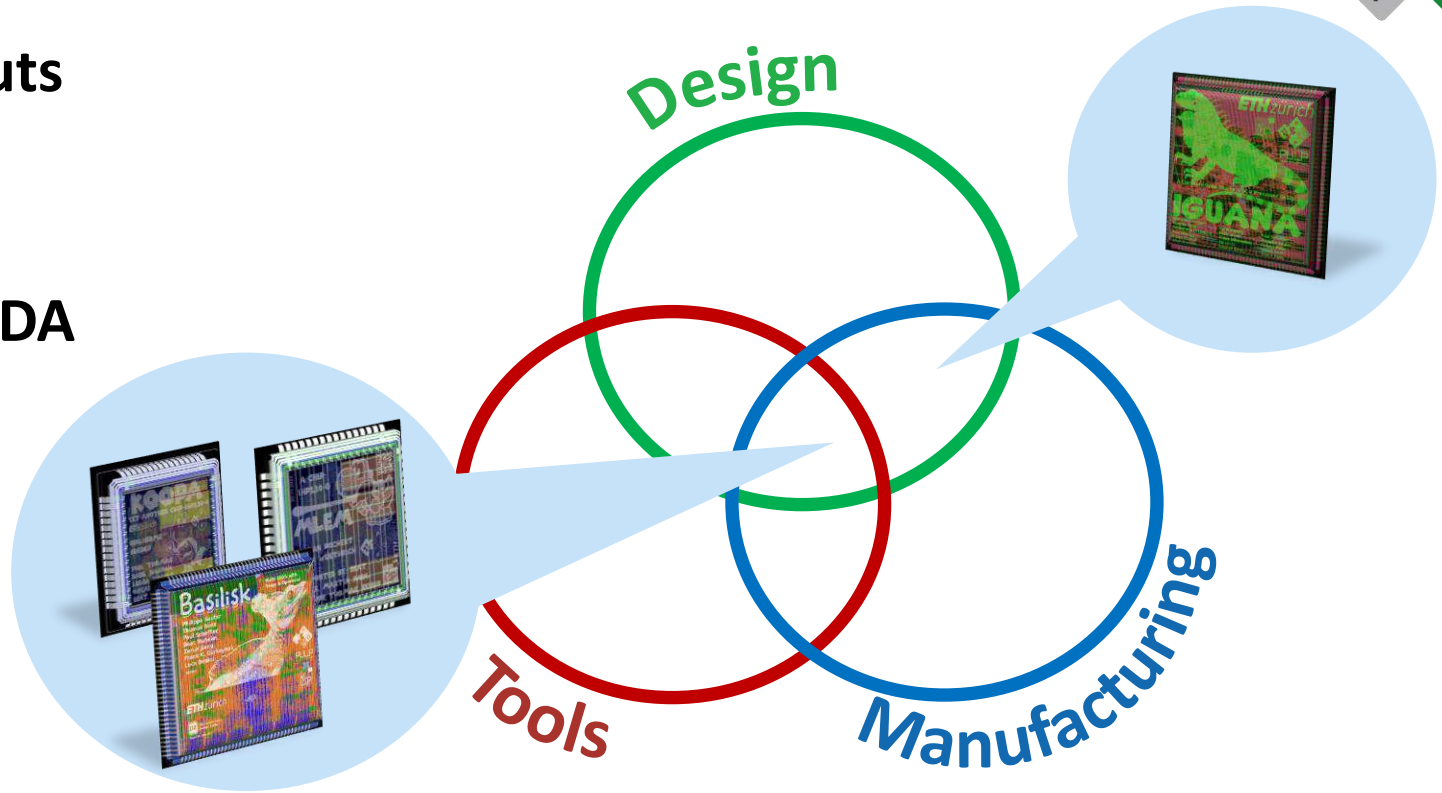
- **Project based grading**
  - Students (in groups of two) will have to modify the Croc reference design
  - Best five designs will be taped-out
- **72 students enrolled**
  - Projects finished in summer
  - Tape-out in IHP130 September

...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 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

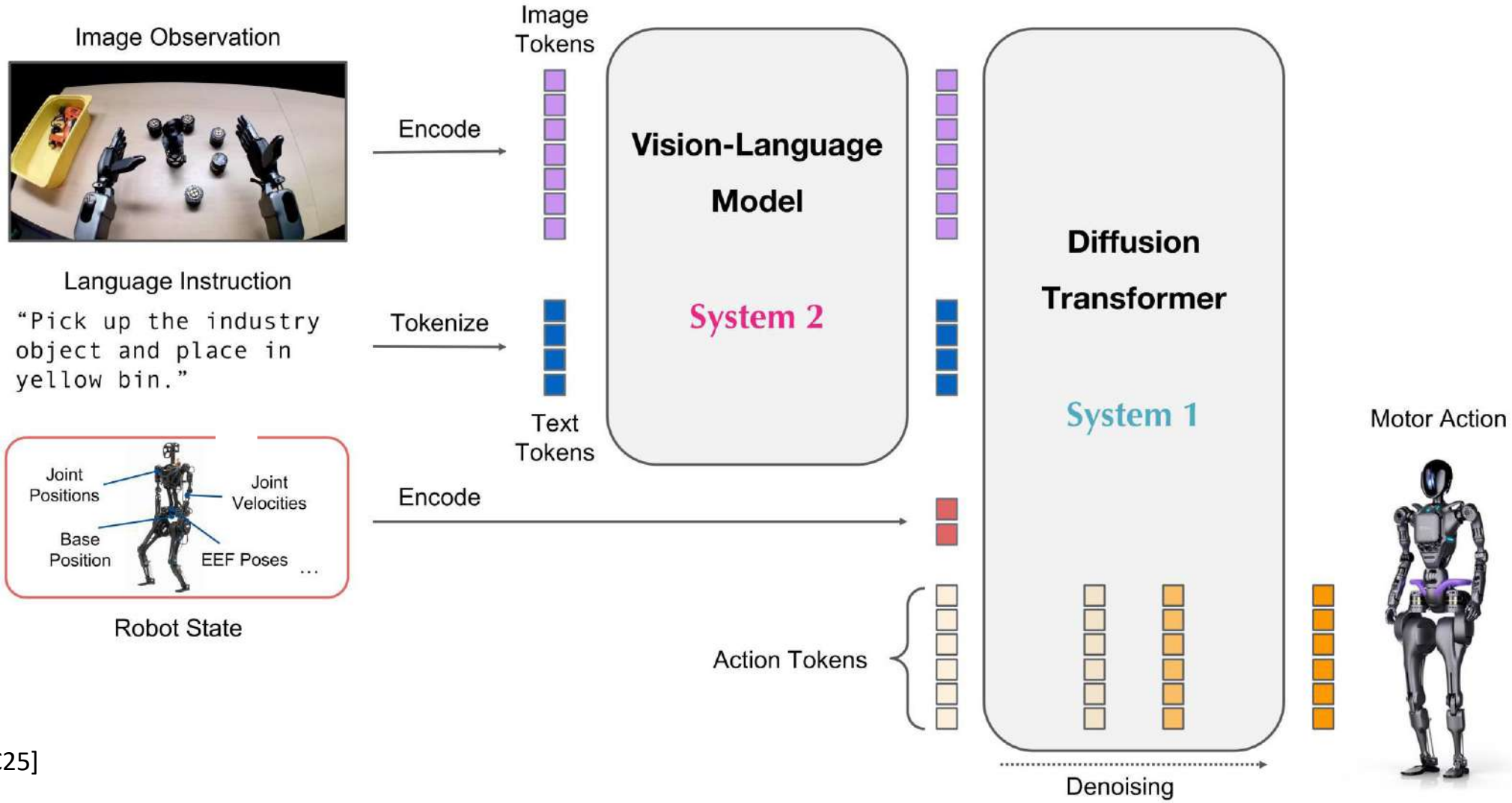
Reduce entry barriers for all

- 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: Embodied Gen.AI → Scale & Efficiency



[GTC25]

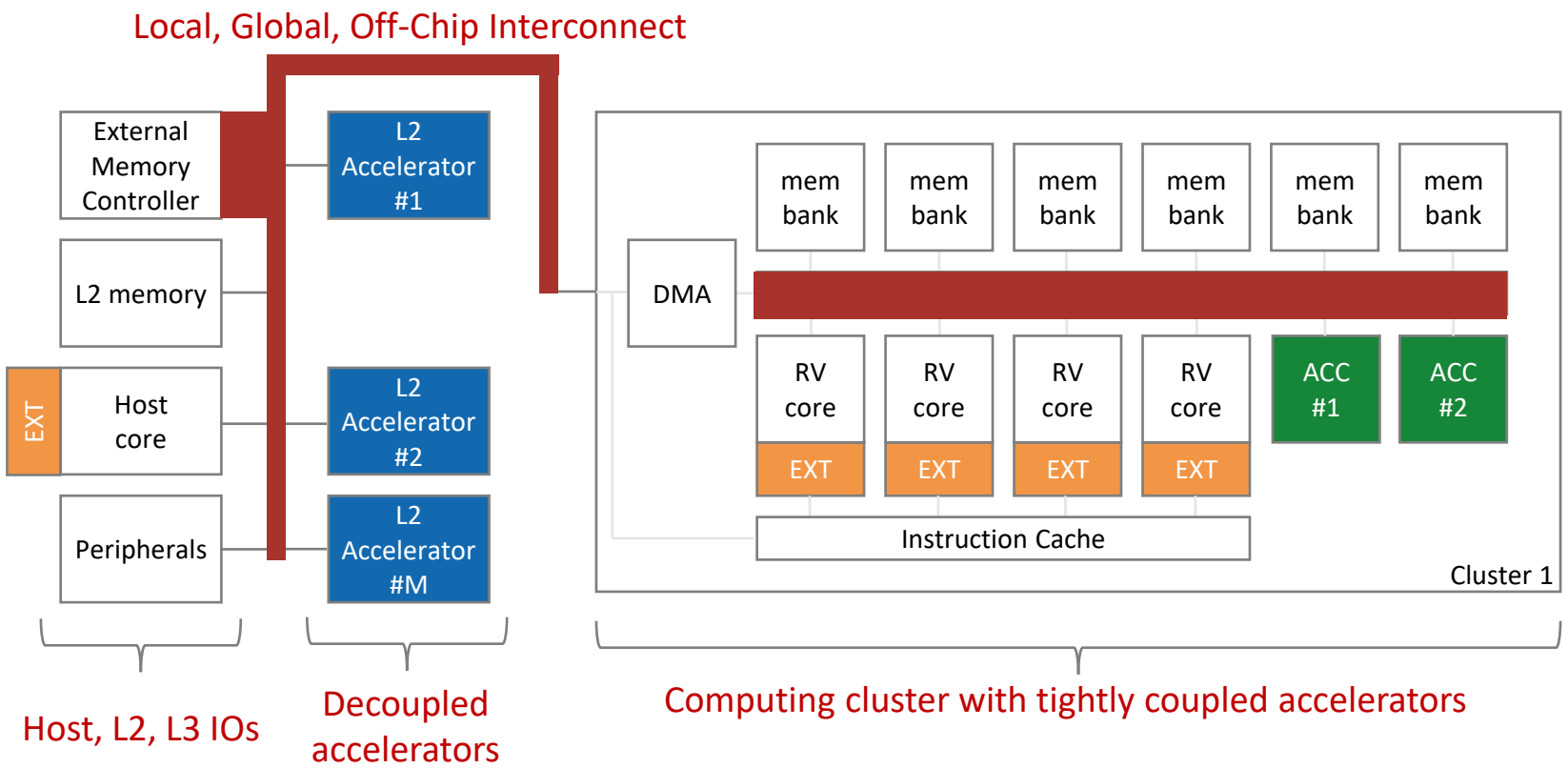
Back to AI: Heterogeneous Accelerated HW Platform

Multiple Scales of acceleration

Extensions to processor cores
- Explore new extensions
- Efficient implementations

Shared-memory Accelerators
- Domain specific
- Local memory

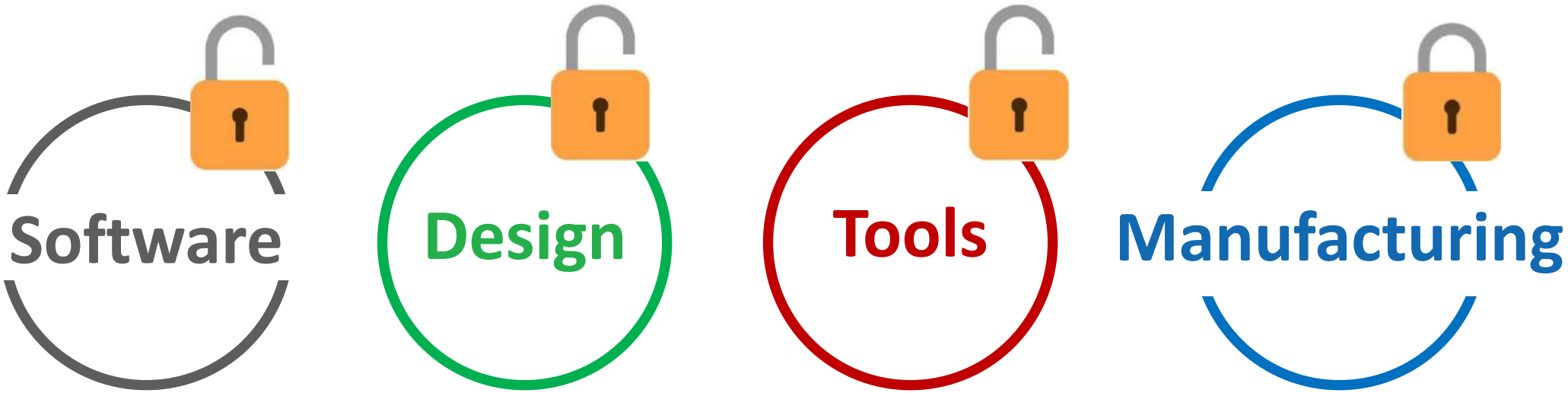
Multiple Decoupled Accelerators
- Communication
- Synchronization



Local, global, package, system → Specialization at scale

Open EDAs Heterogeneous Chips in Advanced CMOS?

Extreme Performance + Energy Efficiency is required!



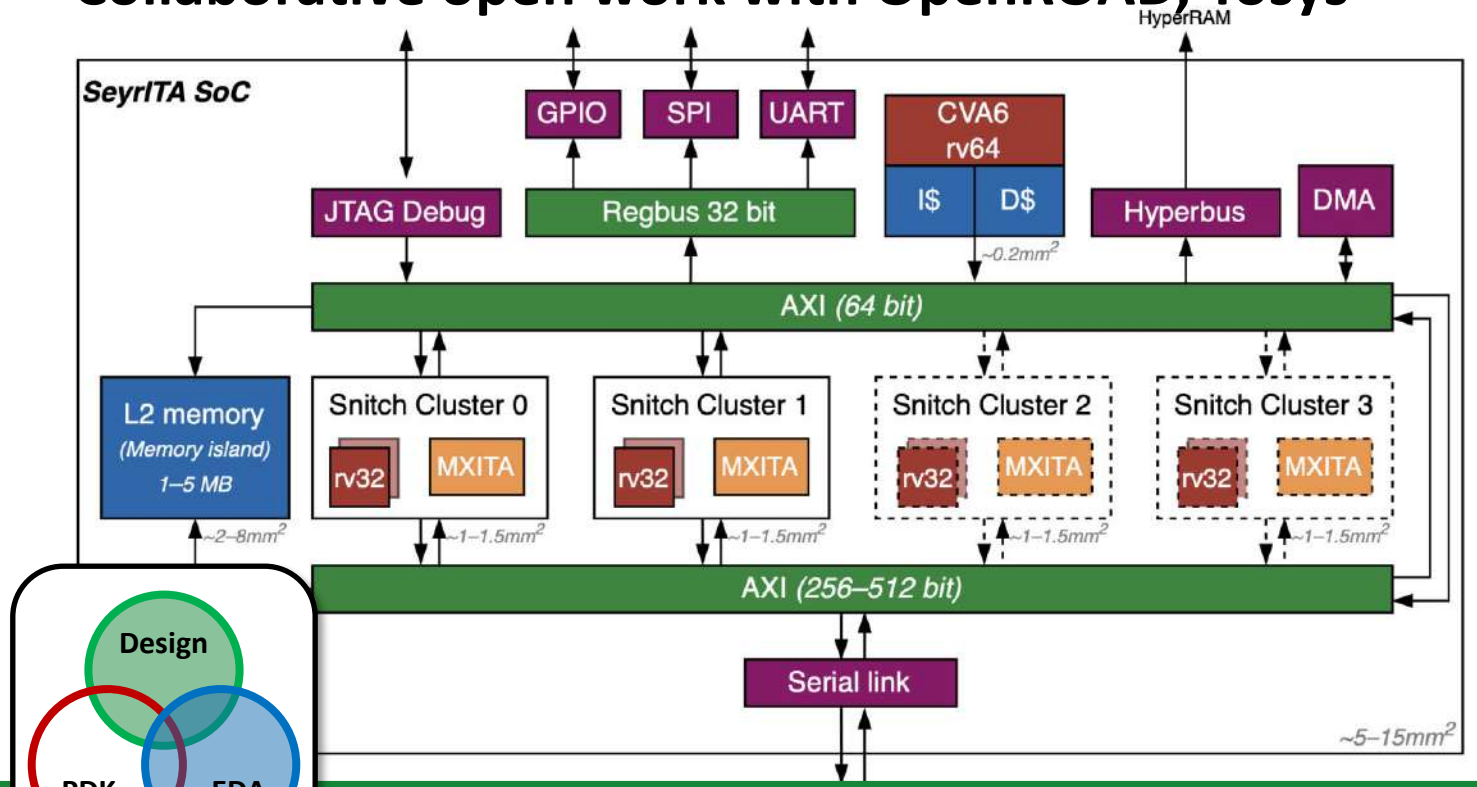
HW-SW Co-Design

HW-SW & Tools Co-Optimization

Get the best from advanced nodes!

SeyrITA, our most ambitious project for embodied AI

- **Designing a research relevant SoC using open source EDA in GF22**
  - State of the art accelerator for embodied AI, using Integer Transformer Accelerator (ITA)
- **Collaborative open work with OpenROAD, Yosys**



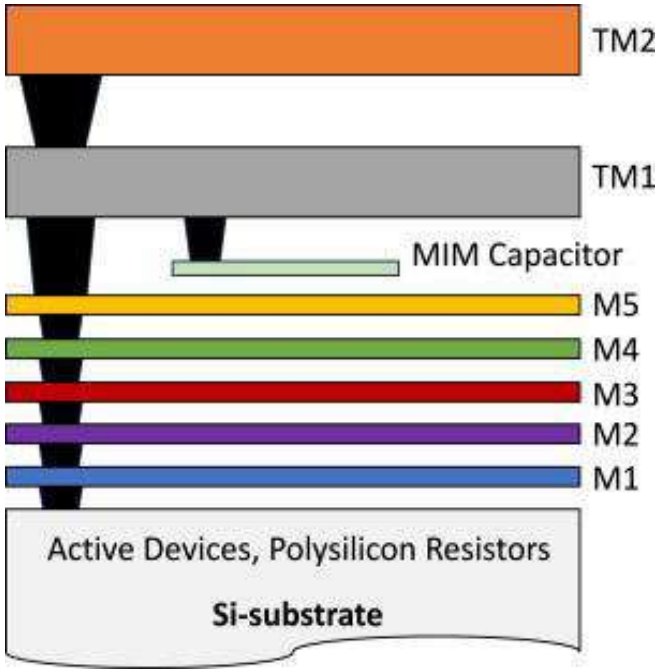
https://github.com/pulp-platform/ita

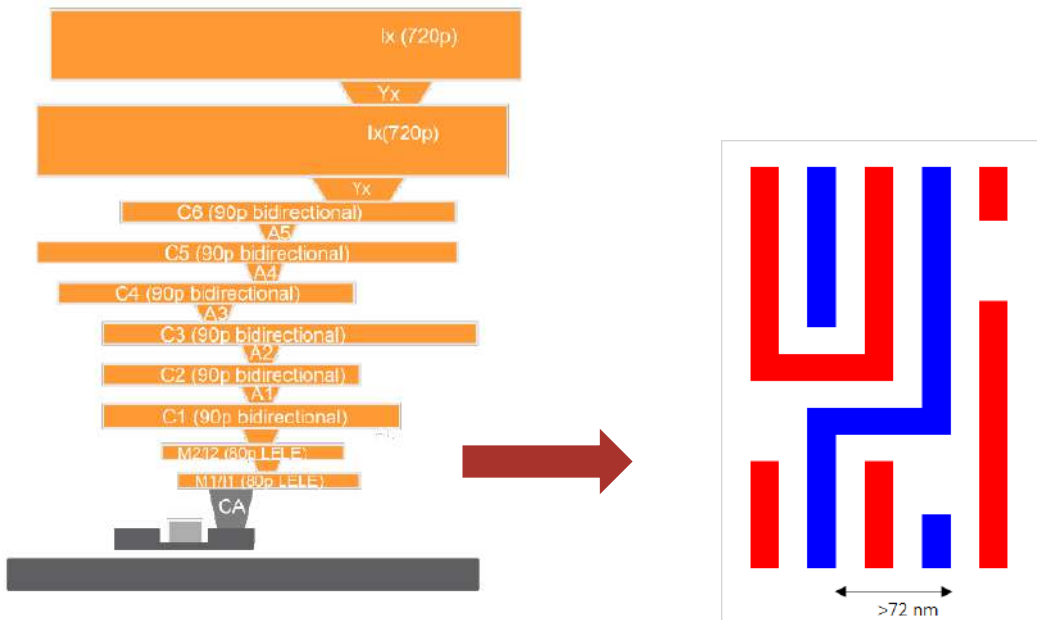
GlobalFoundries™

- **Challenging work**
  - Large design, modern technology
  - We encounter problems daily
  - We try to solve them one problem at a time
  - Confident we will get it done
- **Target tape-out**

Early 2026

Old Versus Modern Nodes: Metal Stacks

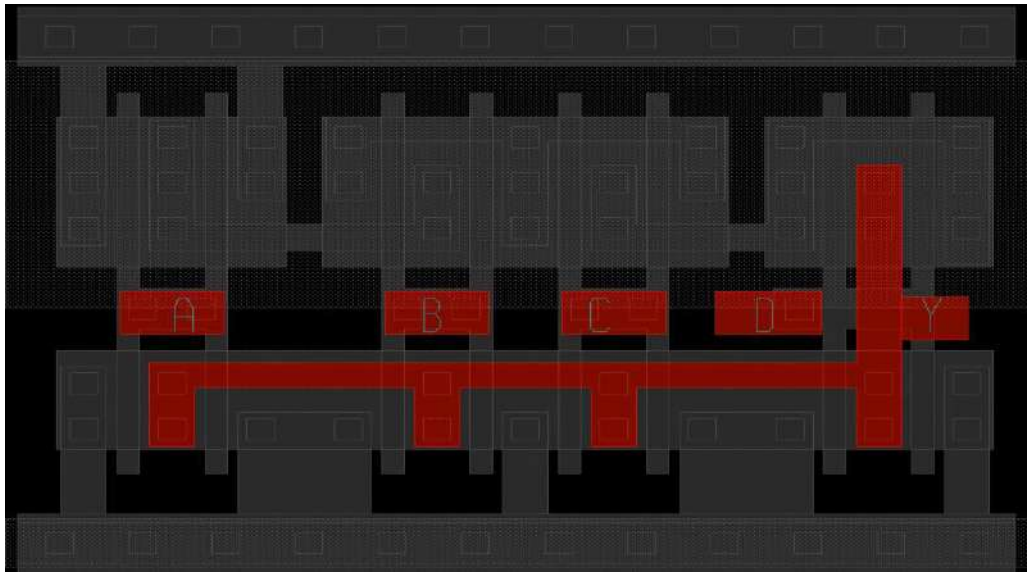




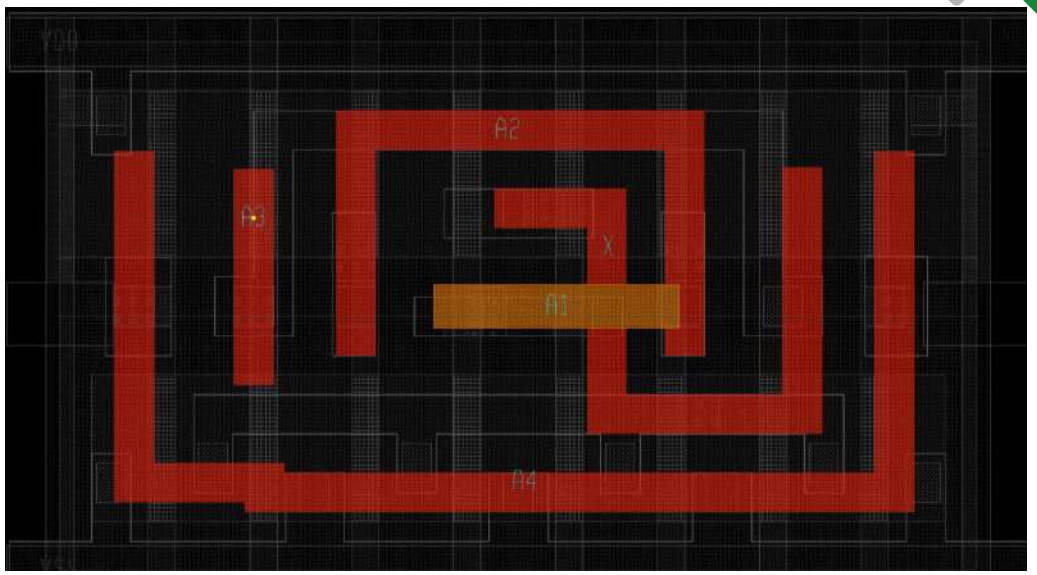
- **IHP130 metal stack**
  - 7 metal layers with two width groups
  - 420nm M1 pitch

- **GF22FDX metal stack**
  - 10 metal layers with 3 width groups
  - 80nm M1 pitch
  - M1, M2 need double-patterning →colored routing

Old Versus Modern Nodes: Standard Cells



IHP130 NOR4 Cell 5.76µm x 4.22µm



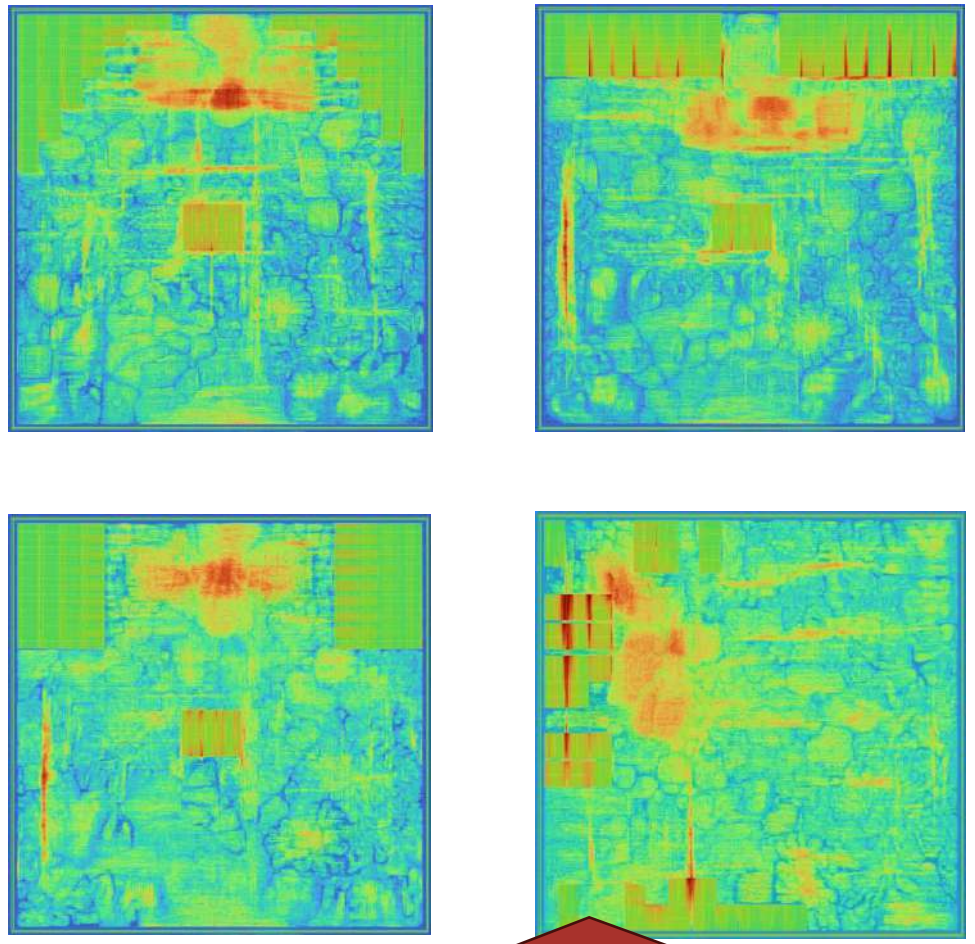
GF22 NOR4 Cell YYµm x ZZµm

- **IHP130 Cells**
  - Larger, with lower density
  - Simple pin shapes
- **GF22FDX Cells**
  - Smaller and denser
  - Pins on multiple layers
  - Irregular pin shapes

Much more complex Synthesis and P&R tooling!

Working on SeyrITA Tapeout

- **First** 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**

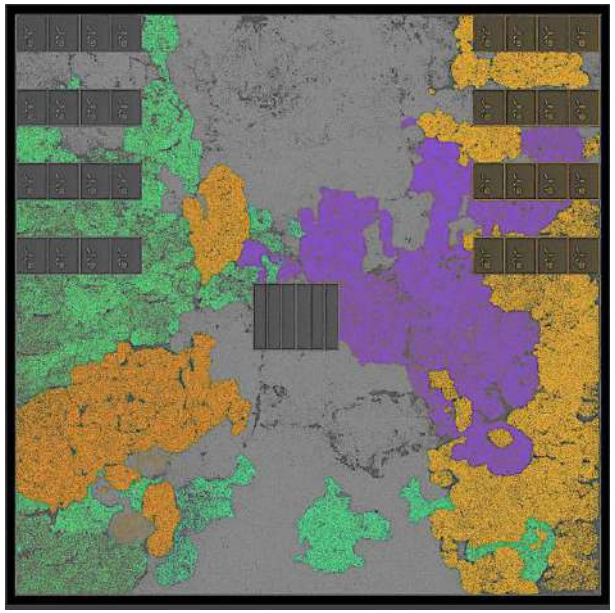


Cluster floorplan exploration

Significant Improvements in QoR

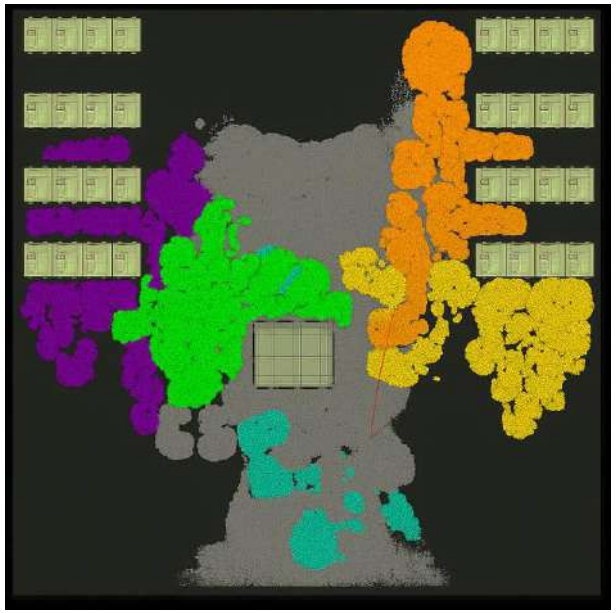
- In addition to tool fixes and improvements, **aggressive hyperparameter tuning**
- All leading to **56% higher frequency** and **42% area reduction!**

Baseline



231 MHz – 7.7 MGE

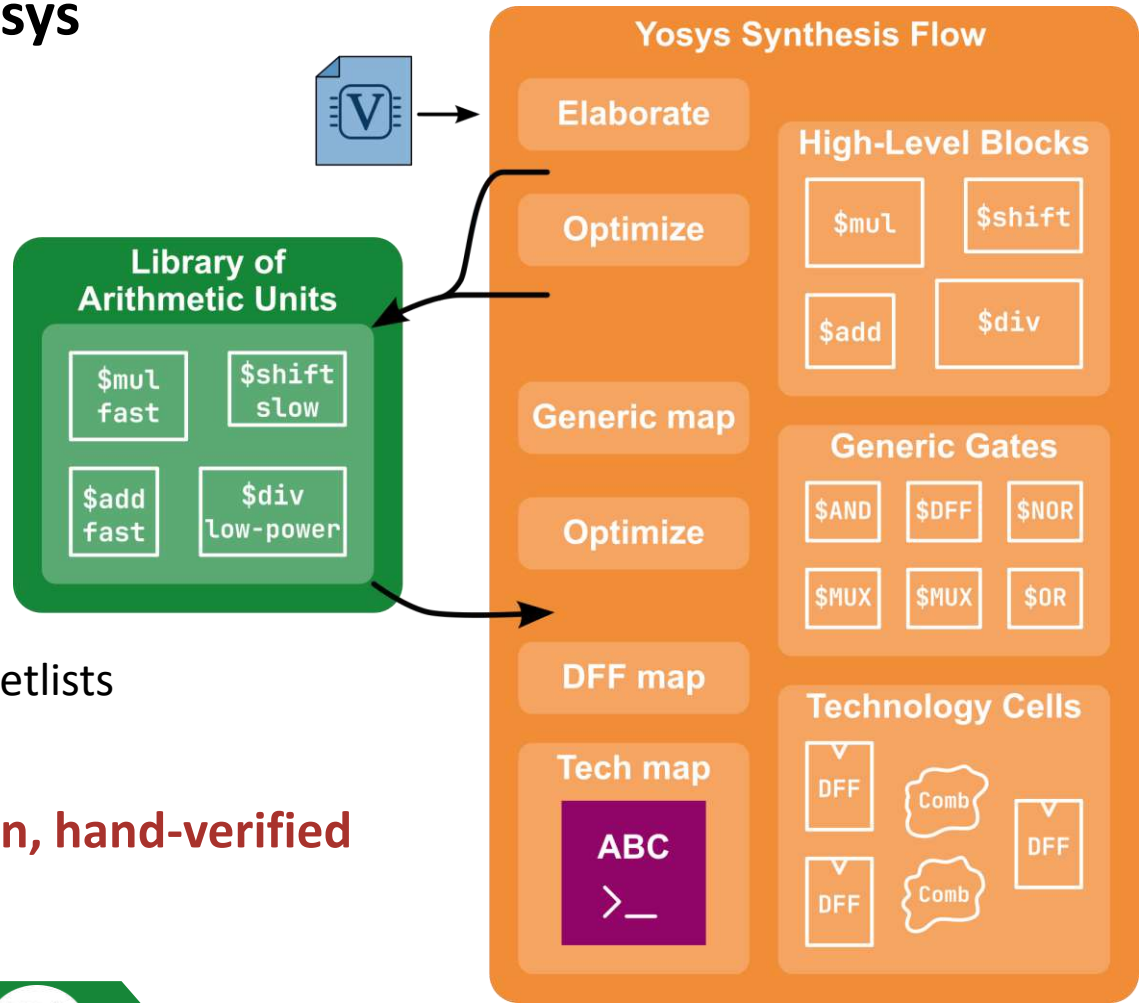
Optimized



360 MHz – 4.5 MGE

Not only Backed! Library of Arithmetic Unit (LAU)

- **Block replacement is implemented in Yosys**
  - Detect and replace arithmetic operators
  - Currently: manual selection
  - **Next: algorithmic, AI based!**
- **No open-source LAU**
  - Rich, optimizd library is key to good results
  - We built it
    - A wide range of arithmetic operations
    - 3 different performance variants of generic gate netlists
    - Thoroughly QoR evaluated and optimized
  - **SystemVerilog port from VHDL: LLM translation, hand-verified**



github.com/pulp-platform/elau

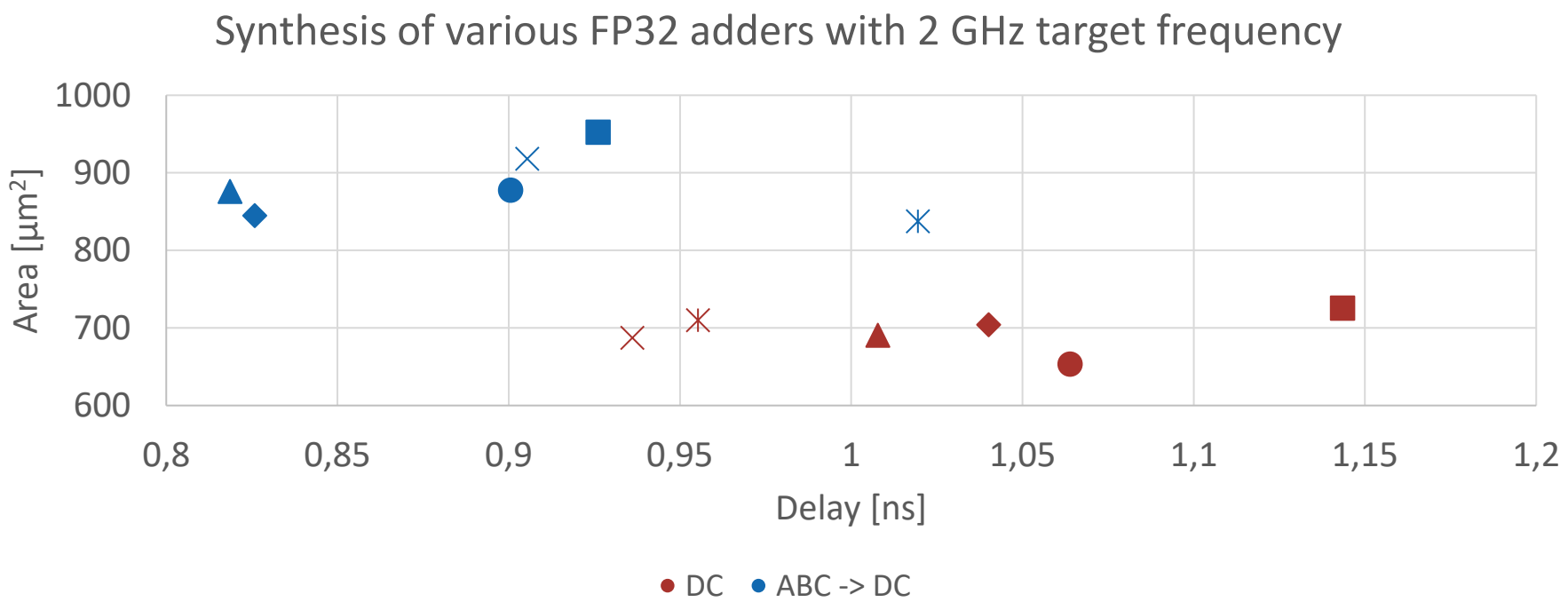
Yosys+ABC Synthesis

- Explored alternative adder topologies in Yosys
- 32-bit FMA with 1 pipeline stage:

Implementation	Area (µm²)	Frequency (MHz)
Double tree (BK)	1183	820
Low fanout tree (KS)	1169	851
Binary (SK)	1386	944
Condensed dbl tree (HC)	1375	968

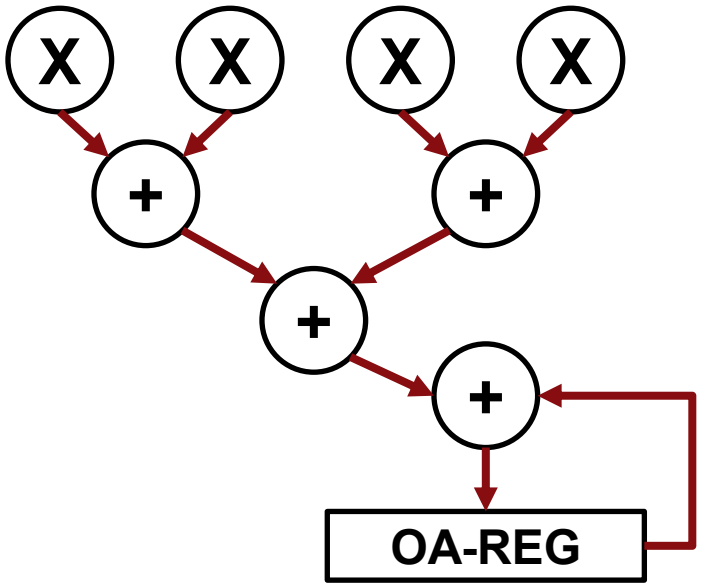
There is still room for improvement in Synthesis!

- Explored various FP32 adders:
  - Applied ABC logic optimizations before Synopsys Design Compiler synthesis, leading to higher frequency Pareto points for several designs.

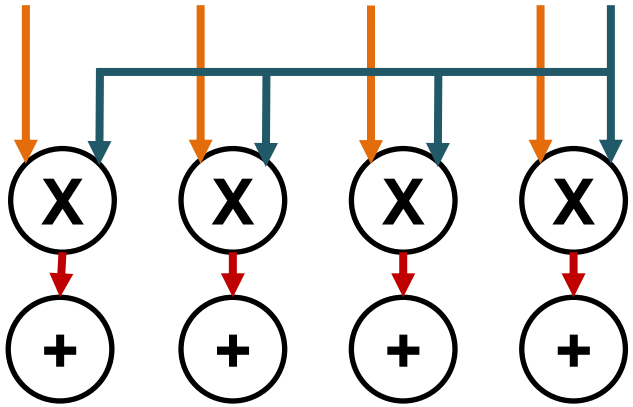


Each marker represents a different design

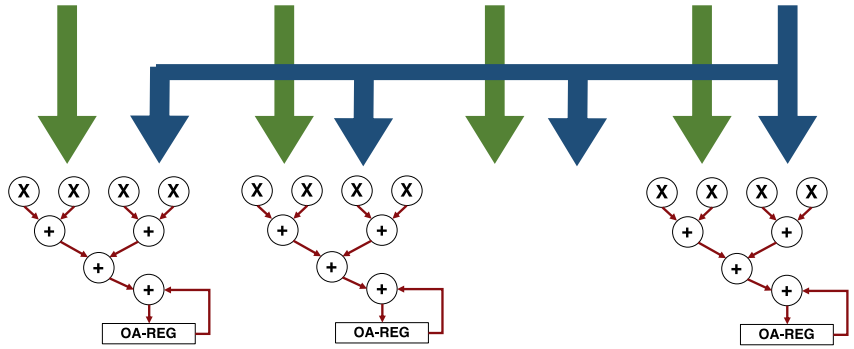
Specialization + EDA multiplicative effect



Inner Product



Outer Product



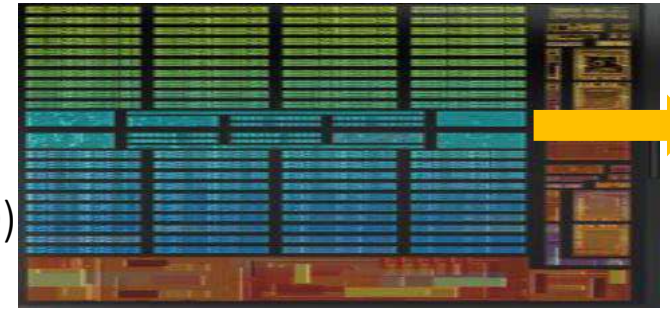
Mixed

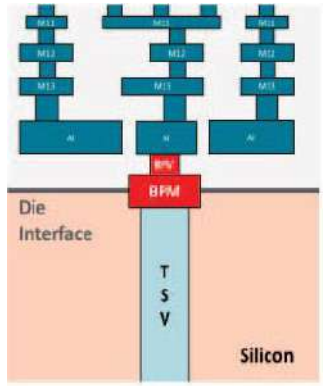
Precision tuning – OP/Mem tuning - deep arithmetic optimization – operand network tuning...

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

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)



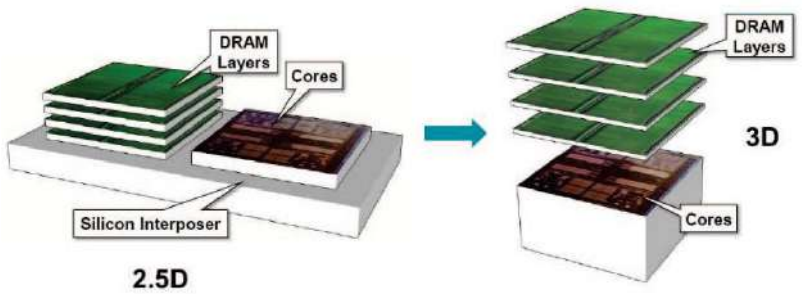


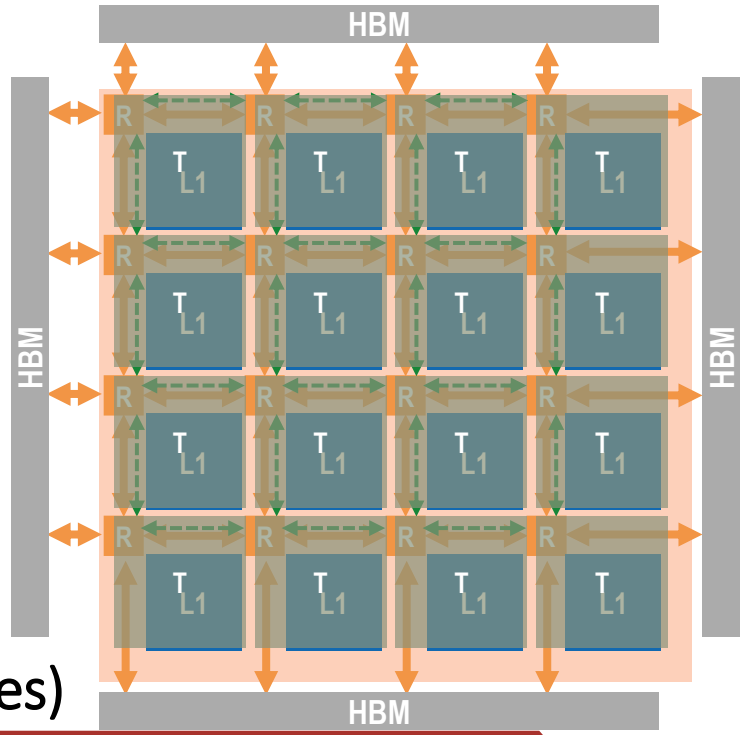
MI300 Instinct™

Logic die

Memory + IO die

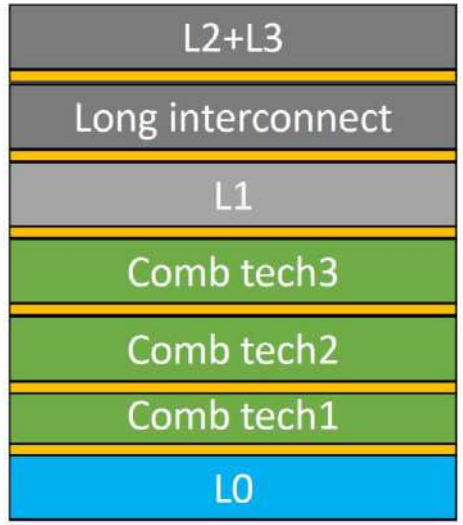
- 3.5D v2?





V1

SRAM+NOC+IO at the bottom



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

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

*github.com/lamda-bbo/Open3DBench

40

Wrapping up…

- **Lowering barriers OS EDA allows you to experiment before investing**
  - You can see if your company can venture into IC Design by experimenting at your own pace
  - Try out how an IC design process fits your company, identify the gaps, judge the benefits
- **Open source EDA is also good for commercial EDA**
  - “Develops” people skilled in EDA – help filling expertise gap
  - Allows EDA companies to concentrate their efforts on parts that make a difference (i.e. no need to develop waveform viewers)
  - More SMEs that venture into IC design the more EDA licenses will eventually be sold.
- **Not limited by license costs but by available CPUs gives opportunities**
  - Makes public (or cross partner) CI flows much easier
  - Being able to run many iterations (with small variations) opens new possibilities
  - Early evaluations of technical choices do not require signoff accuracy

There are still many challenges, our work is not done!

- **We need more open PDKs**
  - Something in the 65nm – 28nm range would be a **game changer**
  - Drafted an open letter to raise awareness with 300+ signatures

https://open-source-chips.eu/

- **Parts of open EDA already in good shape, but there are gaps**
  - Many independent groups are working on tools, need to support inter-operability
  - Opportunity to go beyond what standard tools are able to deliver, reduce the PPA gap!
- **End-to-end open source design requires set of IPs that others can use**
  - Memories, Serial I/O, Data converters, Memory controllers, PHYs for common protocols
  - List is long, we will have enough work to do ☺
  - Need a larger community and good (better) coodination

Open Source EDA for Europe: **ODE4EC**

- **HORIZON-JU-CHIPS-2025-IA-EDA-two-stage proposal**
  - 20MEUR funding from EU, total project **50MEUR**
  - Organized in three sub projects: Digital, Analog, Productivity
  - PO stage passed with flying colors, FPP submitted September
  - If successful start in April/May 2026
  - Mirrors (and extends) similar efforts in USA and China
- **Large consortium**
  - 24 partners (DIG), 27 partner (AMS), 24 partners (PIV)
  - From 14 Countries (AT, DK, FI, FR, DE, GR, HU, IT, LT, PT, SI, ES, SE, SE, CH, UK)
  - Includes broad participation from most open source contributors in EU
- **Great opportunity to make a difference**


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