ETH zürich ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

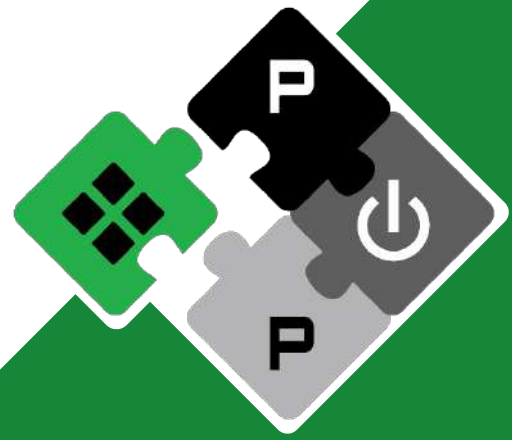
Designing ASICs for the Physical AI Era

Insights and Perspectives for the Swiss ASIC Community

Luca Benini

lbenini@iis.ee.ethz.ch

luca.benini@unibo.it



@pulp_platform

pulp-platform.org

youtube.com/pulp_platform

PULP Platform

Open Source Hardware, the way it should be!

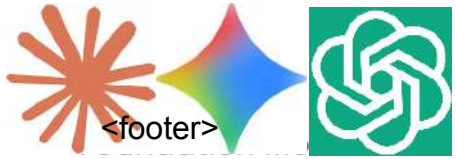
The Race for Physical AI is On





Perception AI





Generative AI





Agentic AI



Physical AI

Reference from Nvdiai Keynote 2025

Physical AI: Artificial Intelligence Everywhere

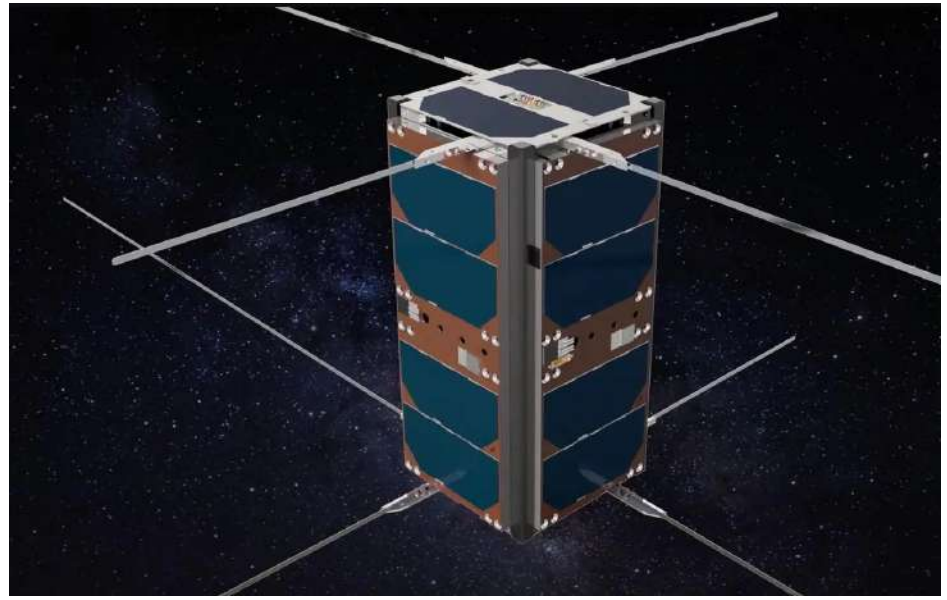
Humanoid



Interceptor Drone



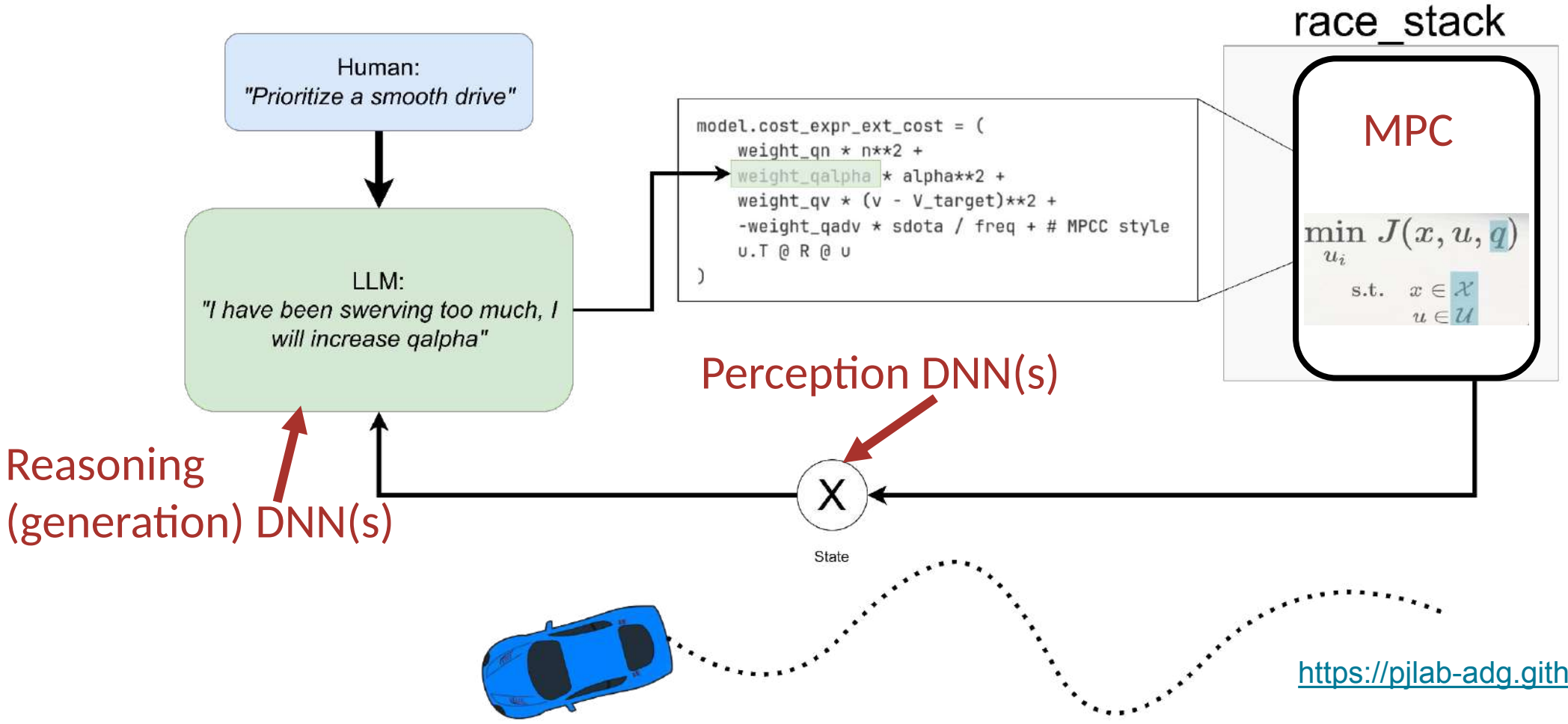
Cubesat



Targeting longer tasks under constant reasoning power constraints

Perception and Agentic Reasoning in Physical AI

LLM Reasoning on Human Commands & Robot Observations

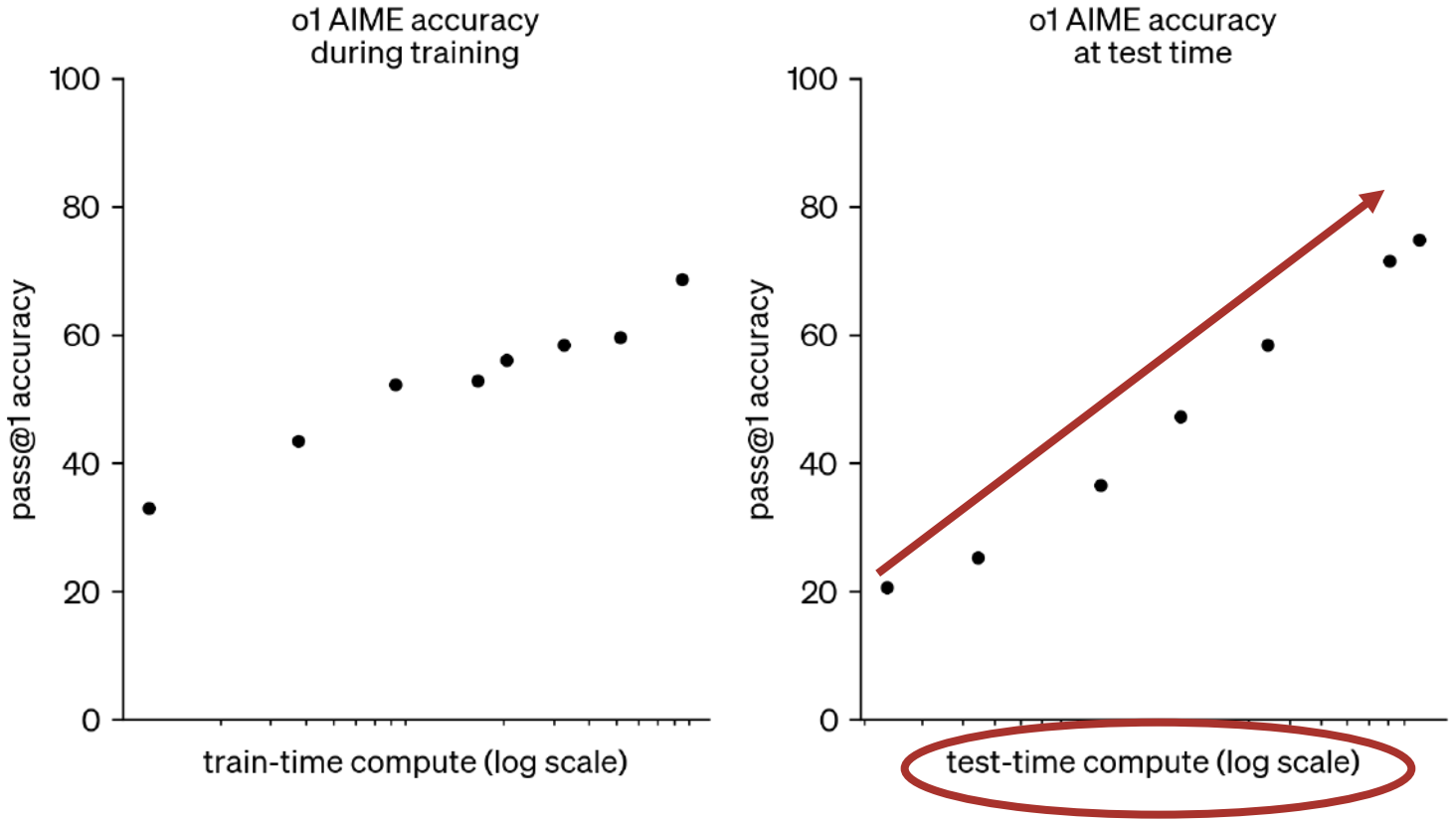


https://pjlab-adg.github.io/DiLu/

Bigger Models, Bigger Bottlenecks @ **Inference**

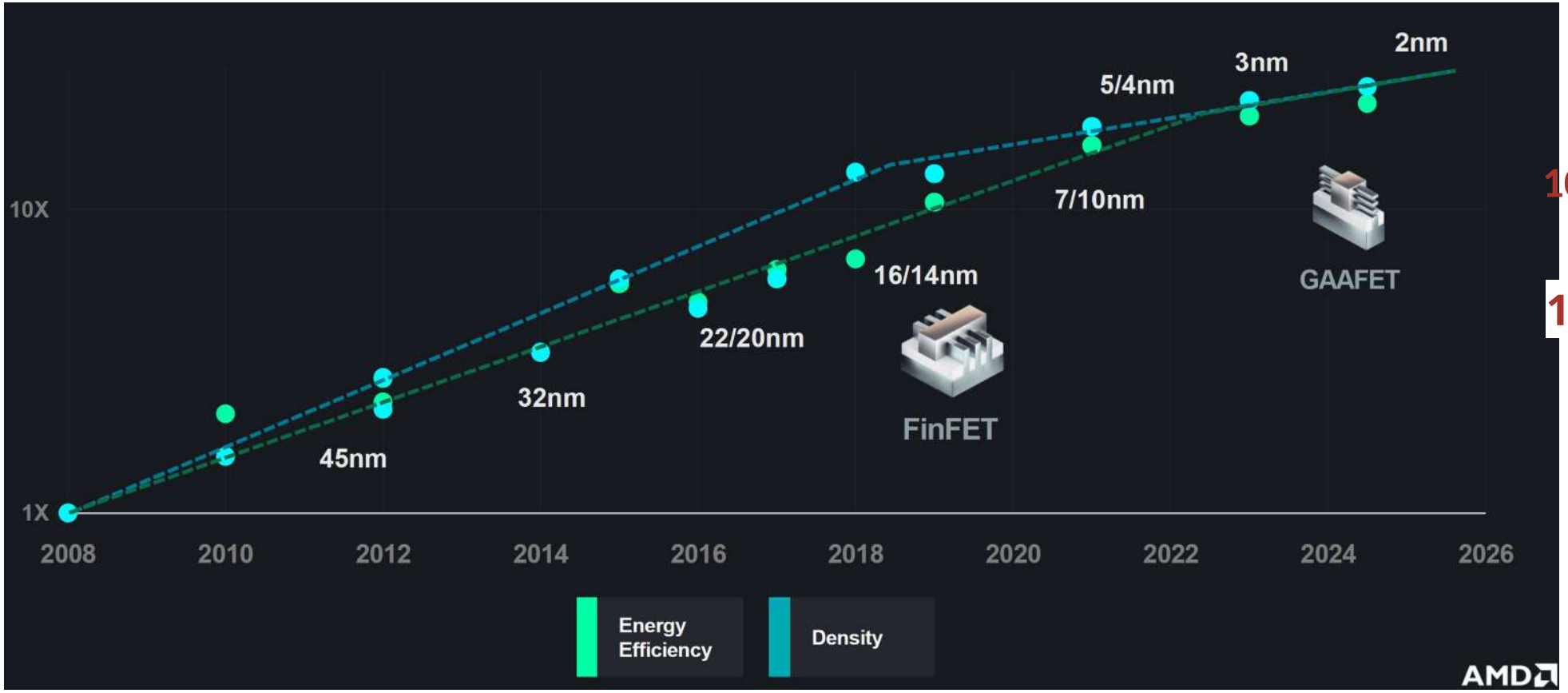
*Deep learning models continue to grow in **scale** and **complexity***

- Growing model sizes demand ever-increasing compute and memory
- Agentic inference compute scales faster than training



[openai.com/index/learning-to-reason-with-llms/]

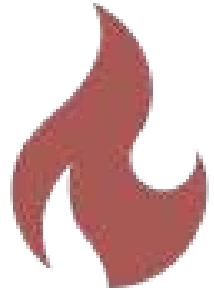
Efficiency Challenge: Moore is not Enough!



[AMD HotChips24]

Model complexity
10× every ~2.5 years

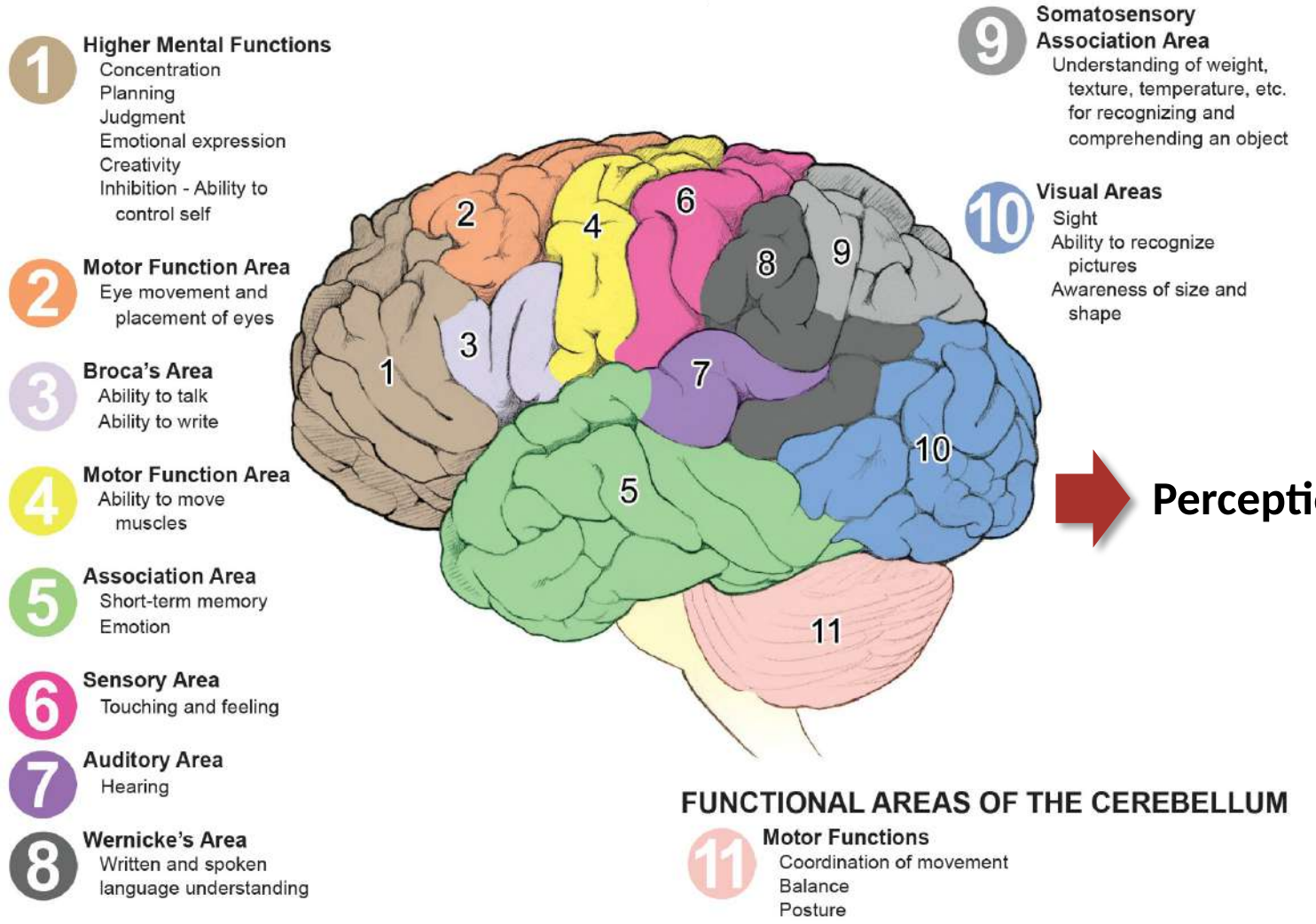
Moore's Law
10x every 12 years!



Algorithm, Architecture, Design

Efficiency via Domain Specialization

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

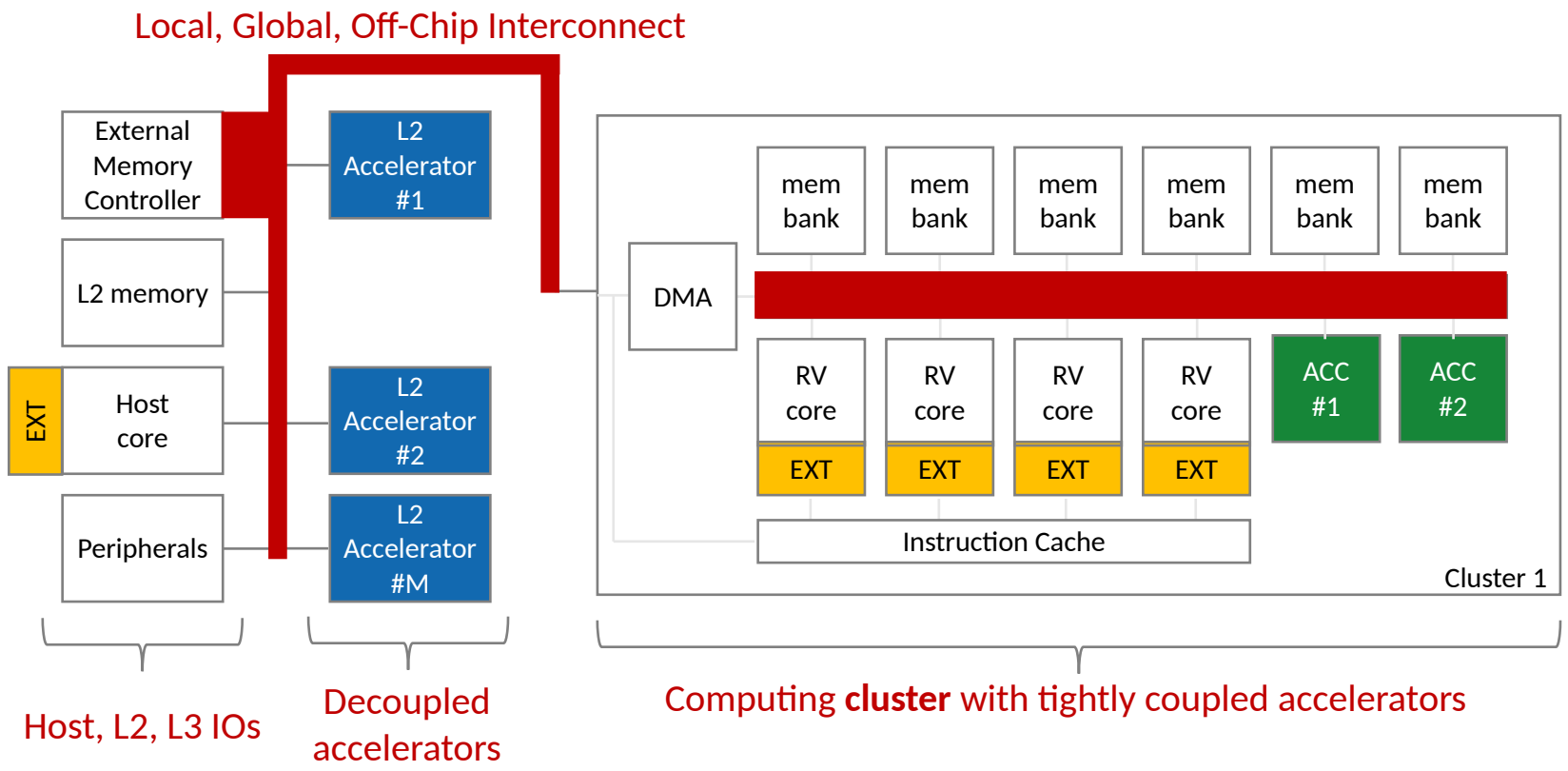


Perception → **Multi-modal Agentic reasoning**

Flexible Domain-Specific Architecture (DSA)

Multiple dimensions for Specialization

- Extensions to processor cores
  - ISA extension
  - Same L0 (RF)
- Tensor Engines
  - ISA extension or fast offload
  - Decoupled L0
- Decoupled Accelerators
  - Explicit Coarse Grained offoad
  - Decoupled L1
- Specialized Mem and Interco
  - Multicast/reductions/compression
  - Compute near/in mem & NoC



Specialization in Perspective: Amdahl's Law

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

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

↓

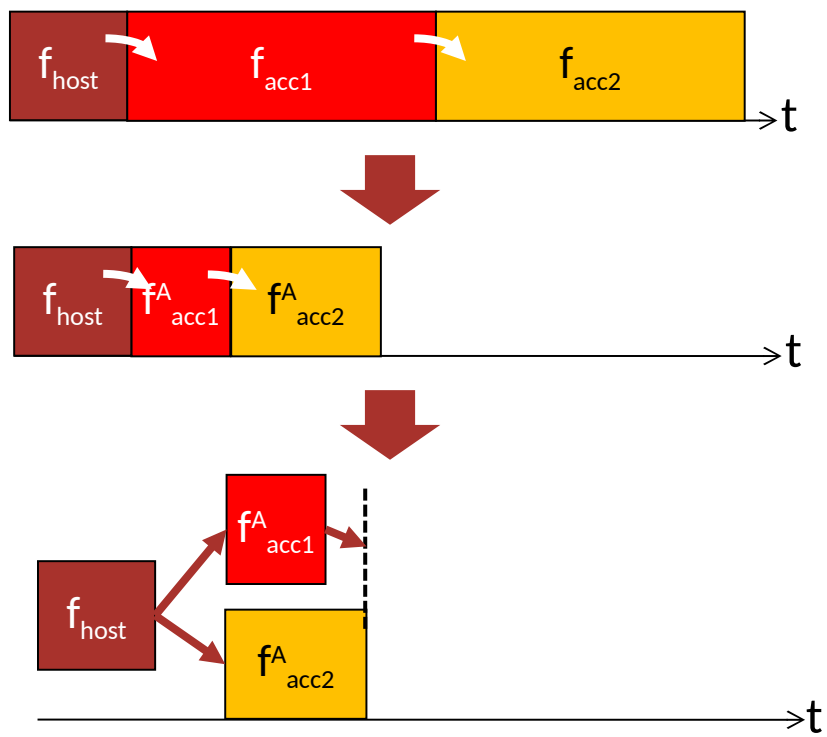
ISA-based 10-20x → **1pJ (4bit)**

↓

Tightly coupled 10-20x → **100fJ (4bit)**

↓

Decoupled 100x → **1fJ (ternary)**

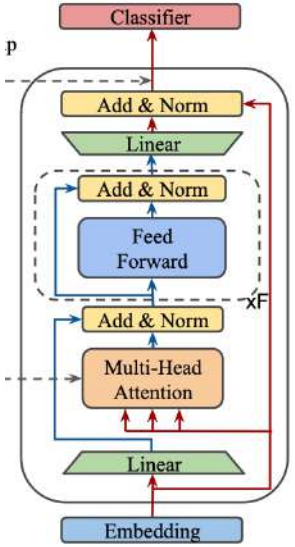


Specialization game: efficiency vs. utilization Tradeoff

A Fast-Evolving Model Zoo

[Z. Sun et al.] MobileBERT

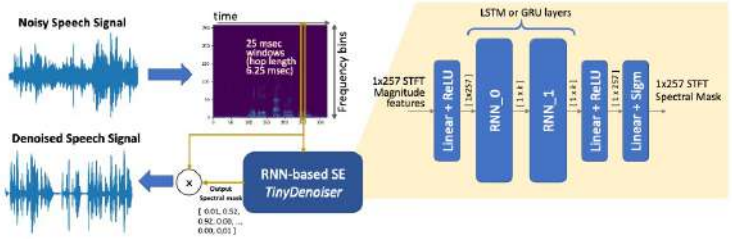
Encoder Transformer





[M. Oquab et al.]

Multimodal Transformer

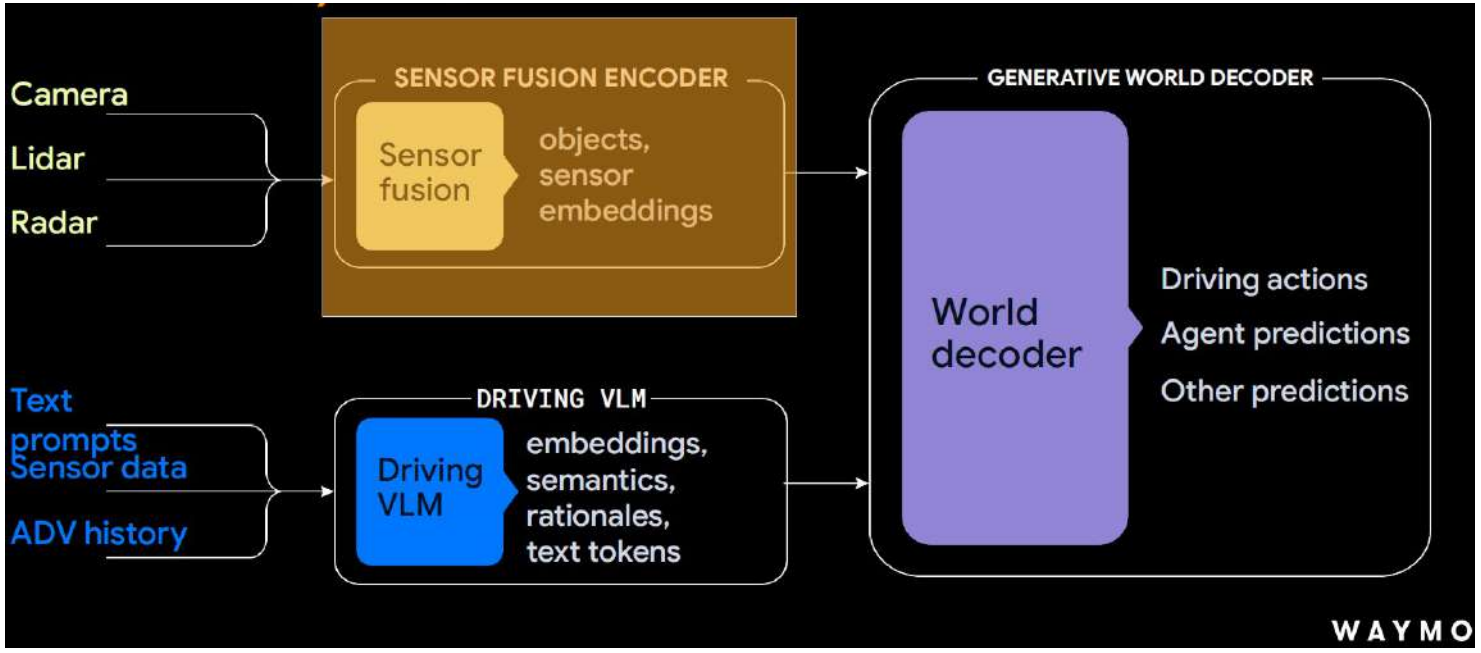


[M. Rusci et al.] TinyDenoiser

Recurrent models

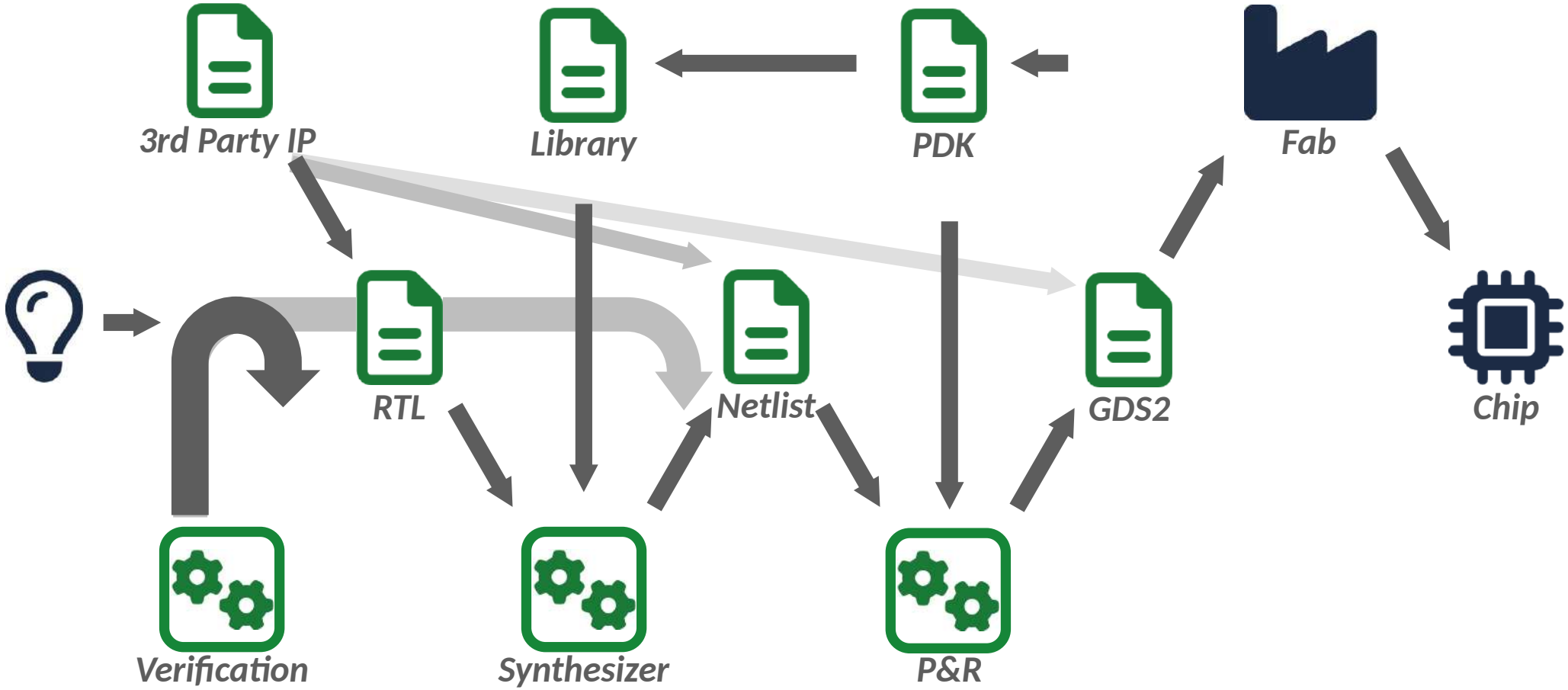
State based models

[Waymo HC26] **Multi-stage Multi-modal**

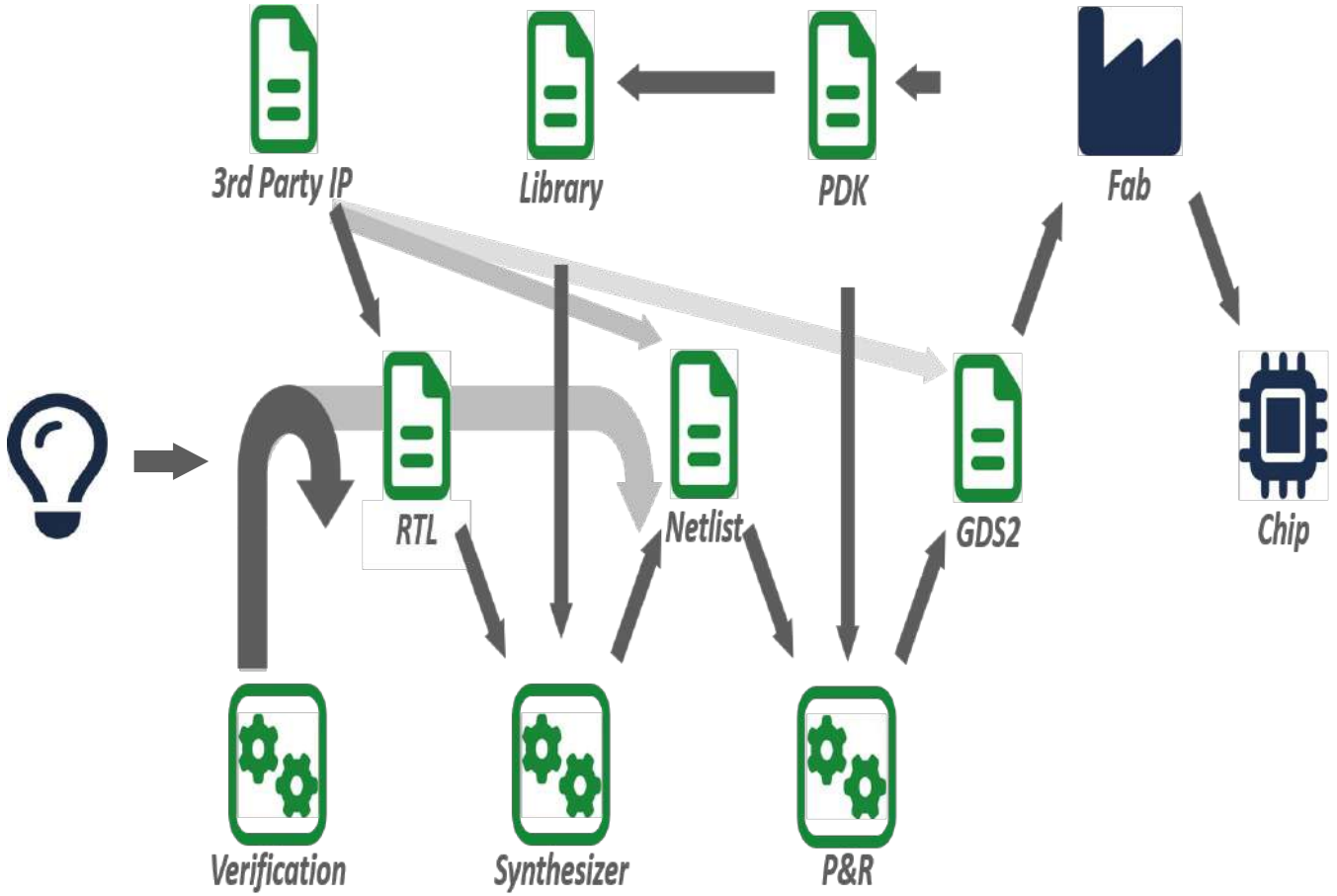


Must specialize fast!

How to Accelerate Chip Design?



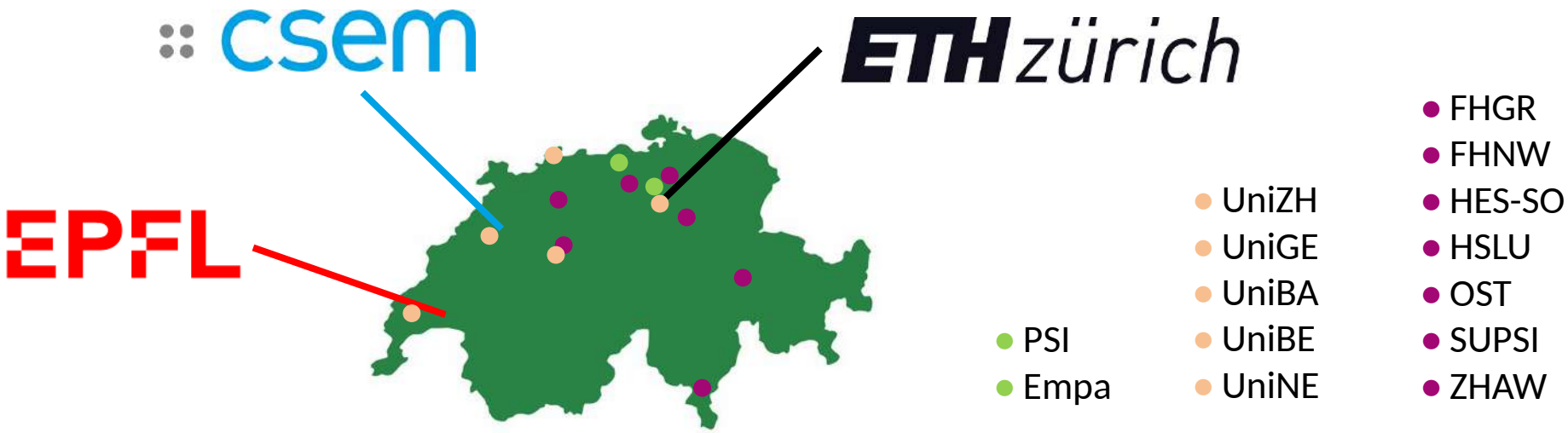
How to Accelerate Chip Design?



Three Pillars

1. People & Skills
2. Open-source
3. AI for Chip design

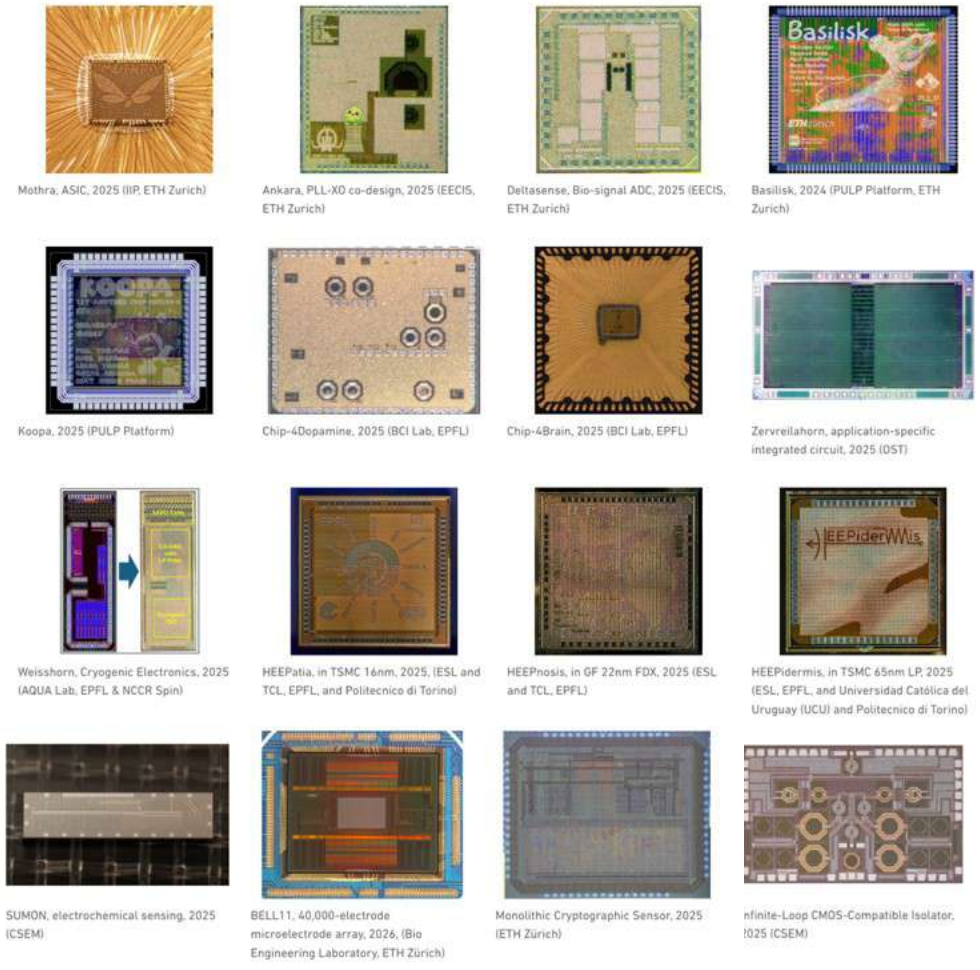
People and Skills: SwissChips Initiative



- Three-year transitional measure launched on February 1, 2024
- CHF 26 million from SERI and CHF 7.8 million in contributions from CSEM, EPFL, and ETH Zurich
- Goals: Secure shared IC design infrastructure, fund research projects, train specialists, and strengthen the Swiss semiconductor ecosystem and international cooperation

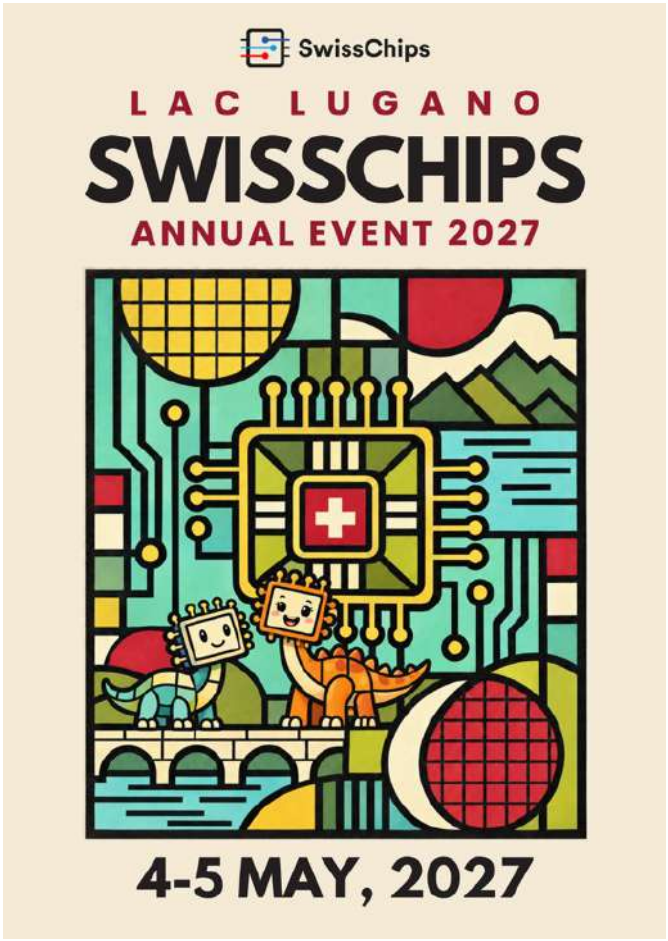
SwissChips in numbers: The first two years

Metric	Number
Scientific publications	**138**
Fabricated chip prototypes	**70**
Externally supported chip prototypes	**19**
Released IP blocks	14
Doctoral students	**123**
Postdocs	30
Professors / subproject leaders	39
Technical and administrative staff	20
International awards	7
Workshops and summer schools	9



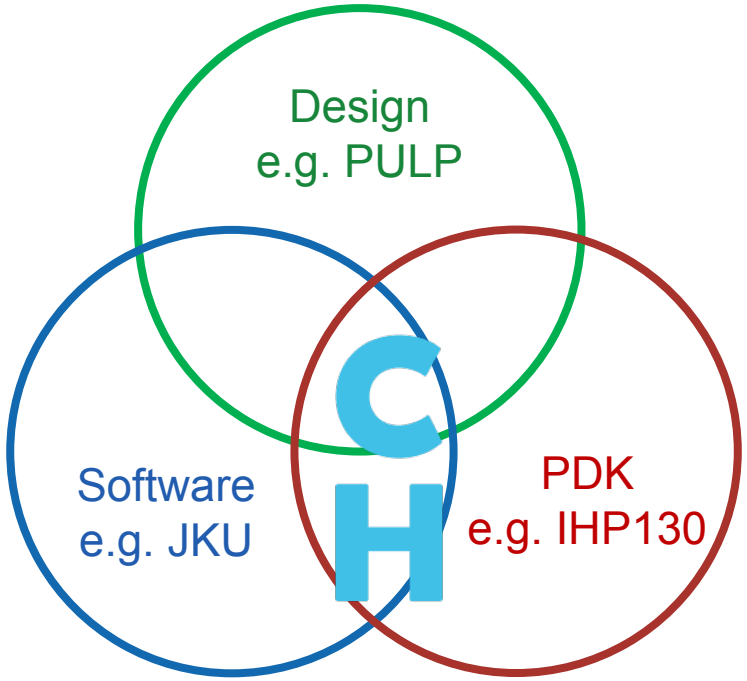
Upcoming SwissChips events





https://swisschips.ethz.ch/

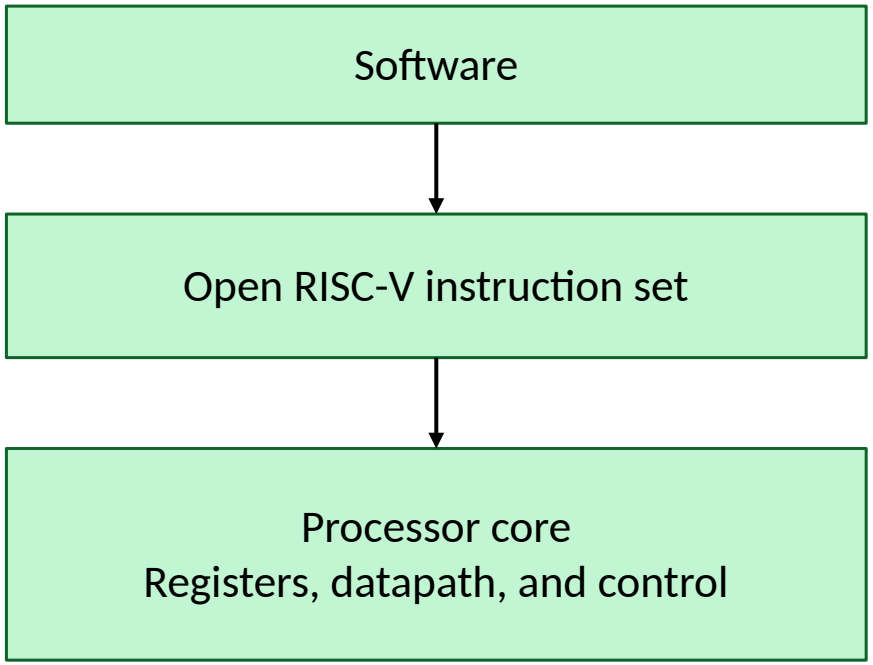
Open Chip design from source code to layout (end-to-end)



- Open designs: Processor and system descriptions
- Open tools: Software for simulation, synthesis, and layout
- Open PDKs: e.g. IHP 130 nm with public rules and models. Fabrication through prototype runs

Open designs: Architectures such as RISC-V

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



- An instruction set defines which instructions a processor understands
- The RISC-V instruction set is open and carries no architecture licensing fees
- Custom processor cores and extensions can be built on it
- Example: ETH Zurich's PULP platform provides open cores to inspect and adapt

Reliable collection of open source tools is essential!

- **Most open source developers have their own idea about the environment**
  - Getting several independent tools to work at the same time is very demanding work
  - Preparing/sharing lectures require you to be able to point to a reliable image

`https://github.com/iic-jku/IIC-OSIC-TOOLS` to the

README Apache-2.0 license

IIC-OSIC-TOOLS

DOI 10.5281/zenodo.15044726

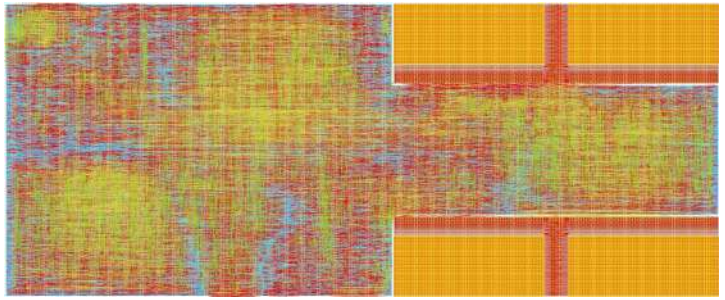
This environment is based on the efabless.com FOSS-ASIC-TOOLS.

IIC-OSIC-TOOLS (Integrated Infrastructure for Collaborative Open Source IC Tools) is an all-in-one Docker container for open-source-based integrated circuit designs for analog and digital circuit flows. The CPU architectures `x86_64/amd64` and `aarch64/arm64` are natively supported based on Ubuntu 24.04 LTS (since release `2025.01`). This collection of tools is curated by the Department for Integrated Circuits (DIC), Johannes Kepler University (JKU).

Thanks Harald

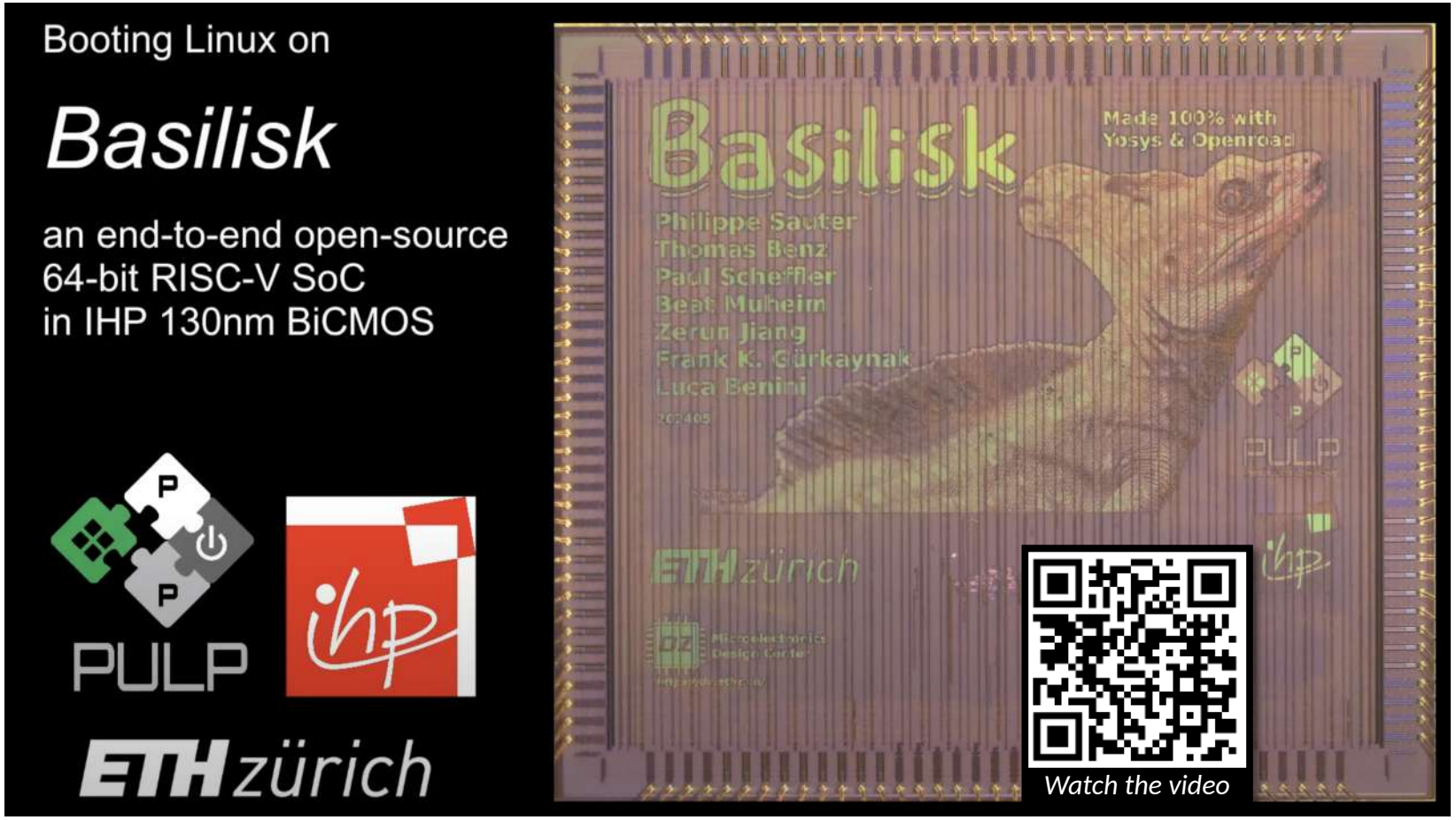
End-to-end open-source flows **REQUIRE** open PDKs

- **At the moment, the choice is limited**
  - Skywater 130nm
  - IHP 130nm → Swisschip Tiny-tapeout (99.4% utilized, 200+ projects)
  - Globalfoundries 180MCU (will be abandoned soon)
- **Some are sort of available**
  - IC Sprout 55nm (taped out our first chip)
- **We need more open PDKs to design more relevant designs**
  - Something in the 65nm – 28nm range would be a **game changer**
  - **And we will soon have alternatives (to be announced in 2027, this is a done deal!)**





We can make **today** large end-to-end open-source chips

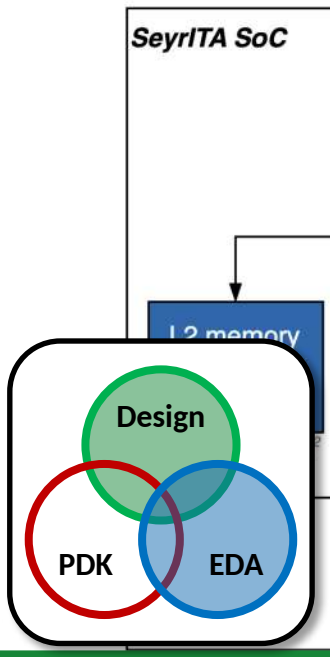


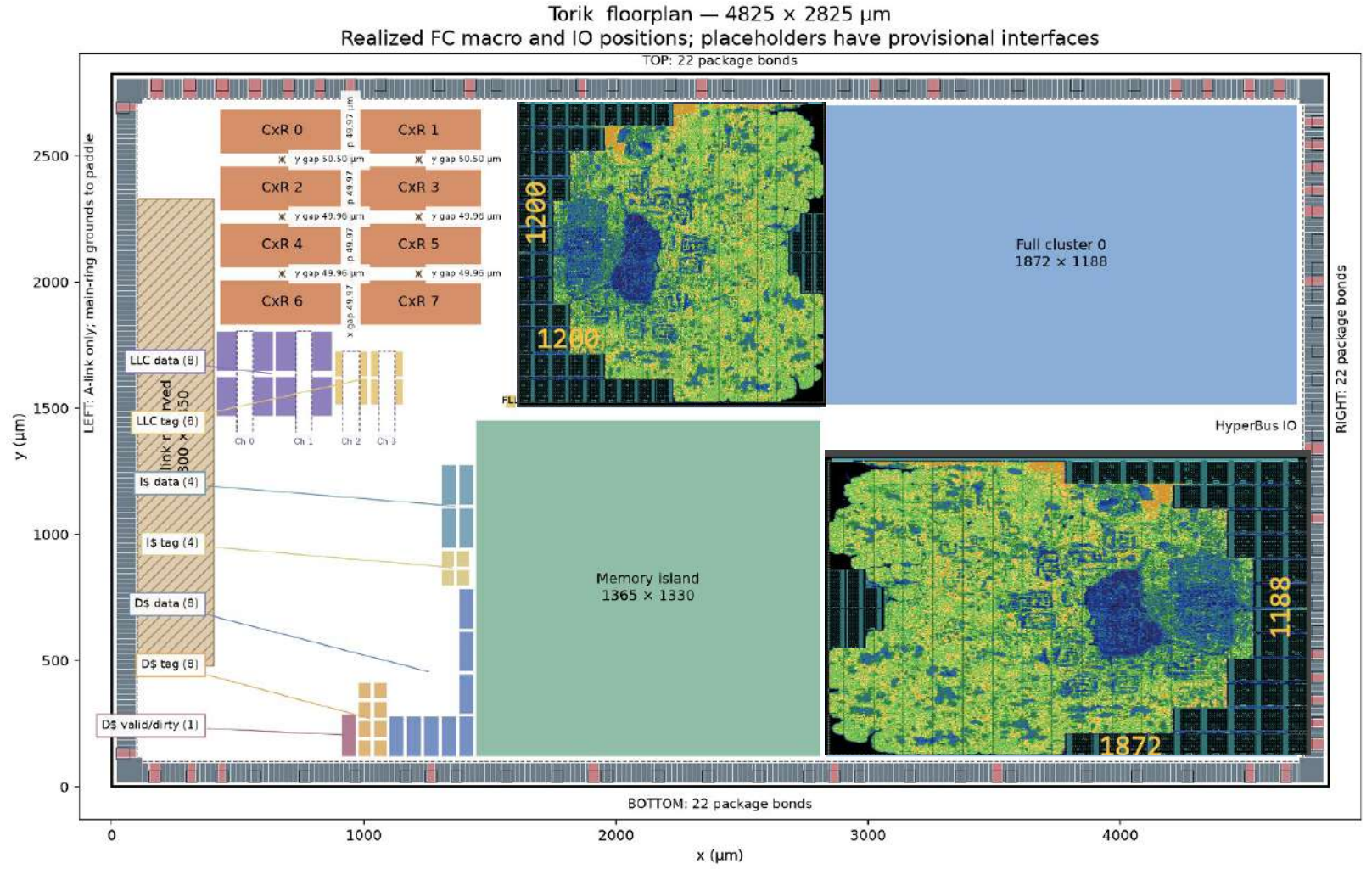
- **Open Design**
  - Cheshire with CVA6
- **Open EDA**
  - Yosys (with –slang)
  - OpenROAD
  - KiCAD (for board)
- **Open PDK**
  - IHP 130
  - Open standard cells
  - Memory macros

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

Torik: our most ambitious project for Physical AI

- **Designi**
  - State of
- **Collabo**





(ITA)

ng work

gn,
echnology
nter problems daily
solve them
em at a time
we will get it done

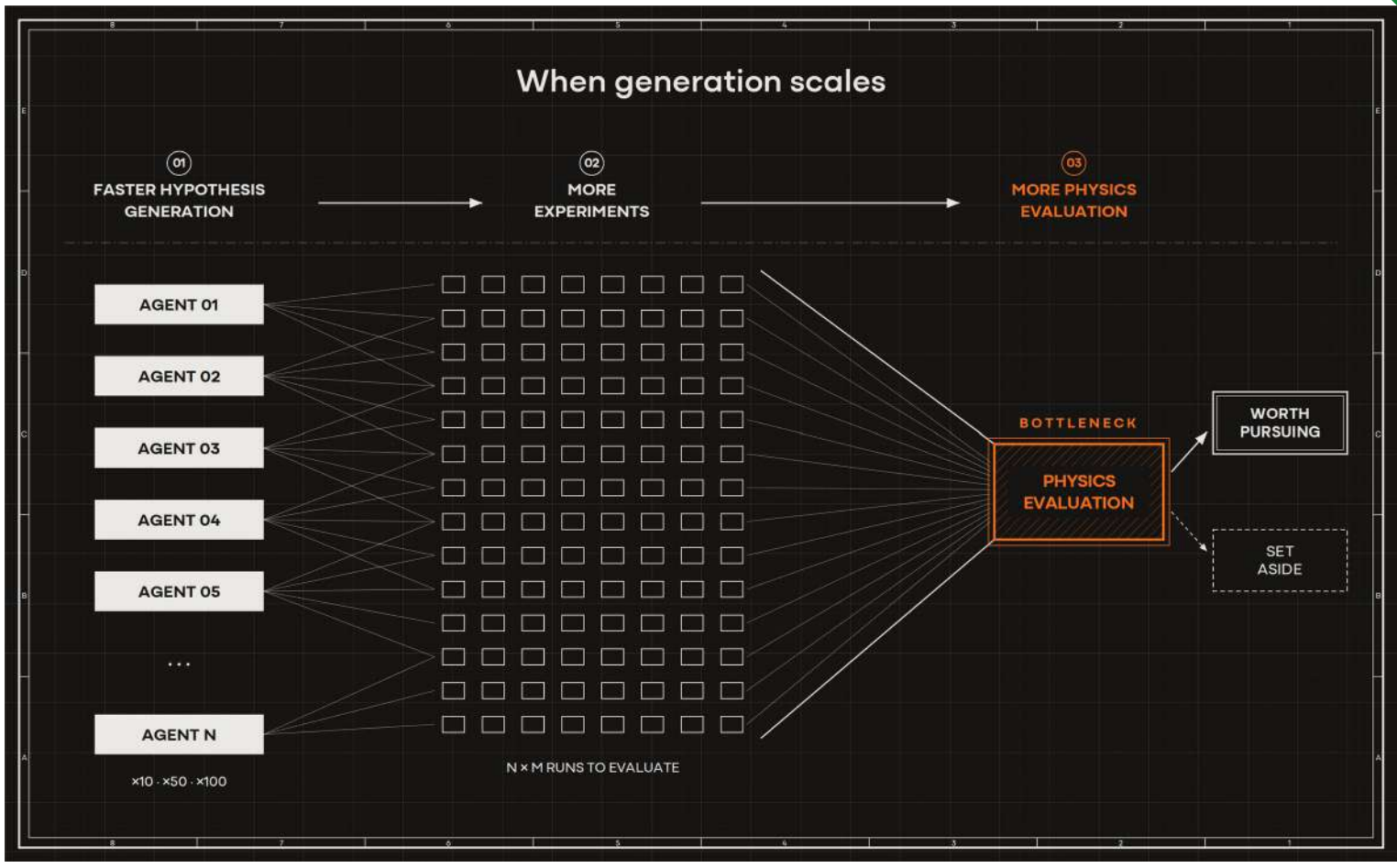
e-out

https:/

AI for Chip design: Towards Agentic Optimization

- **Design Space is exploding**
  - Domain adaptation is required at the month time scale
  - Architecture and implementation: scale-up, scale-out, 3D
- **How to explore it?**
  - Hand-crafting is not viable at the speed required
  - Need speed and Intelligence → AI agents

https://chipagents.ai/blogs/ai-native-flow-for-chip-design



ACTERO: Agentic NPU Optimizer

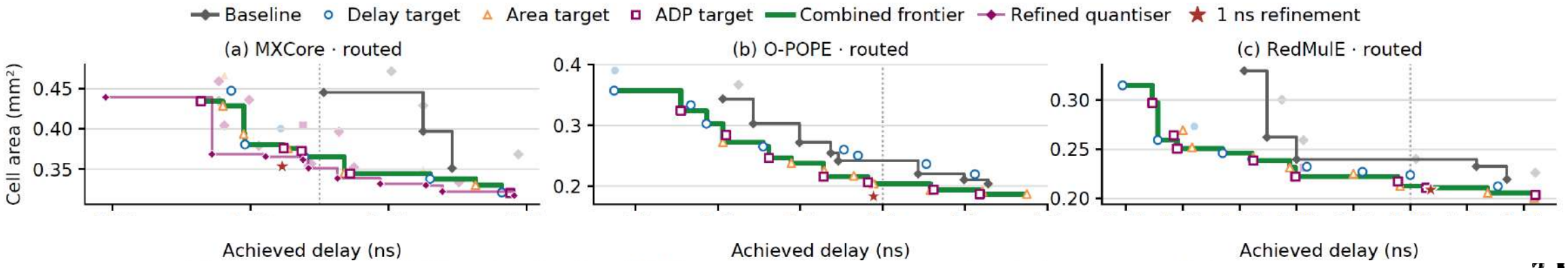
Arithmetic-Centric Tensor Engine RTL Optimisation with Multi-Fidelity Feedback and Agentic Search

Tensor engine

Baseline: N × U

Control · buffers

MXCore
32 × MXDOTP
Fixed 3 / flex ≤5
O-POPE
64 × SDOTP
4 pipes (fixed)
RedMulE
96 × FMA
2 pipes (fixed)



The Future is AI-Boosted Extreme Co-design

Flexibility

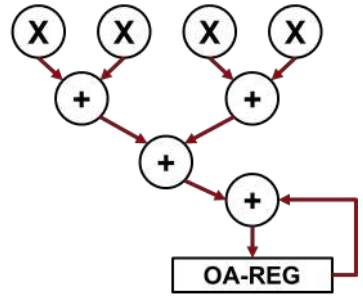
Energy-Efficient RV Core → **20pJ (8bit)**

ISA-based 10-20x → **1pJ (4bit)**

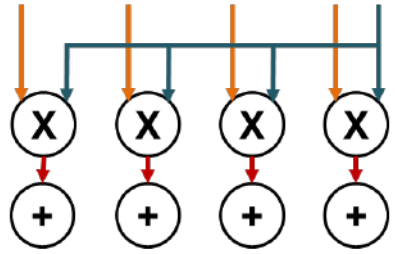
Configurable DP 10-20x → **100fJ (4bit)**

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

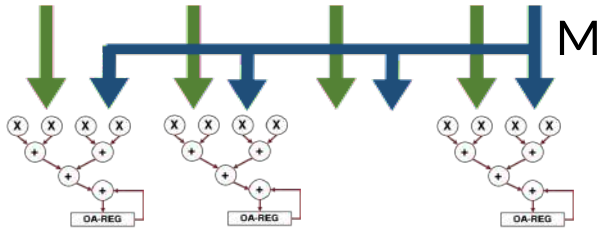
AI- boosted EDA



Inner Product



Outer Product



Mixed

Precision tuning – OP/Mem tuning - deep arithmetic optimization – 3D placement…

Co-Specialize SW, HW, EDA & Technology is the frontier & Open Platforms are Key!



Thank You!