ETH zürich

Massive AI Compute. One Efficient Chip.

Designing a PetaFLOP/s chip for AI workloads.

Yichao Zhang[1], Yinrong Li[1], Chenyue Wang[1], Luca Benini[1,2]

[1] Integrated Systems Laboratory, ETH Zurich [2] Department of Electrical, Electronic, and Information Engineering, University of Bologna

01 What AI computing chip needs

AI models: bigger + heterogeneous

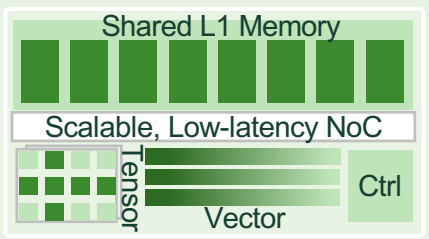
More weights, larger activations, longer context

Matrix, non-linear, recurrent & control computing

Matrix engines
High compute throughput

Large memory
Space for models and working data

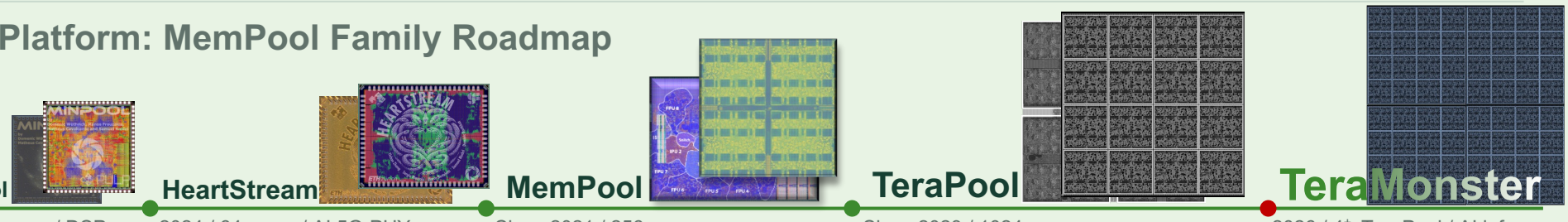
Data movement
Feed compute with less overhead



Scaling from a computing cluster

Large Shared-Memory
Scalable Network-on-Chip
Heterogeneous AI engines

PULP-Platform: MemPool Family Roadmap



02 TERAMONSTER DESIGN

One efficient chip. Working on your AI model.

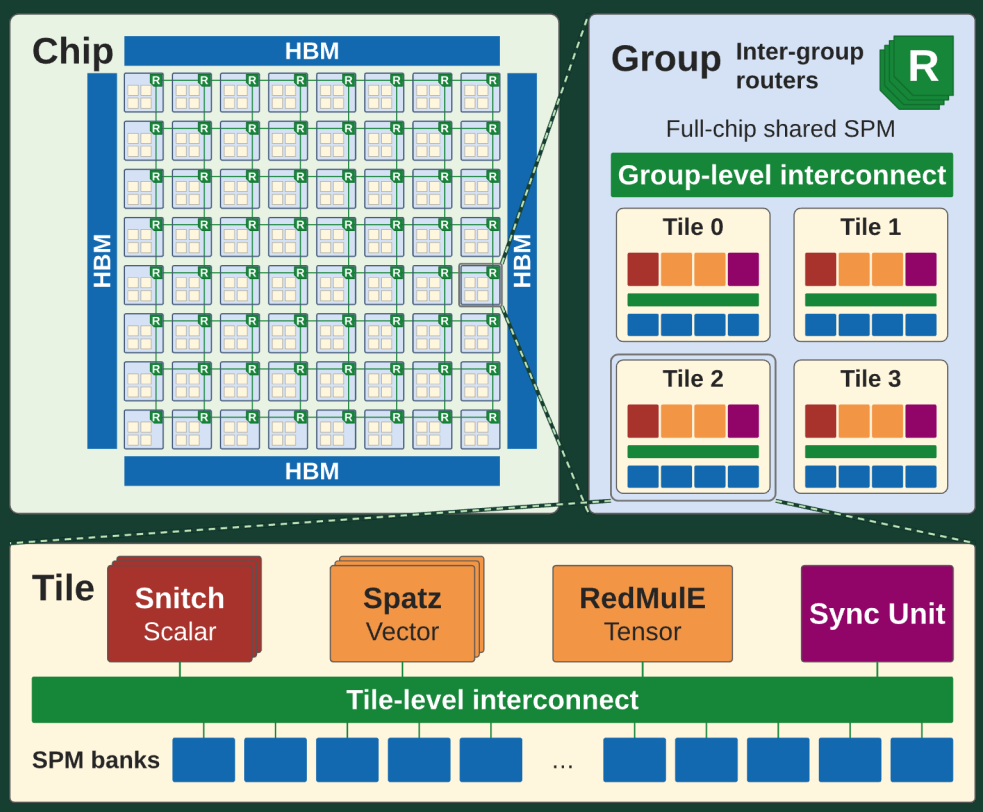
A large heterogeneous cluster for sustained AI inference.

Up to
32×32 Mesh Groups
16384 RISC-V Cores
4096 Tensor Engines
1GiB Shared L1-Mem

Qwen
Qwen3.8-27B[1], FP16

38K Token/s

65ms TTFT Latency



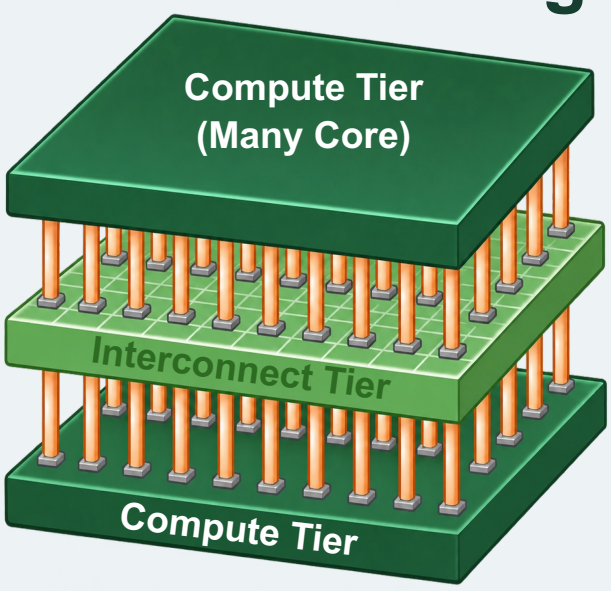
Leading performance and efficiency

	This work*	NVIDIA B200	NVIDIA H100	AMD MI300X	Microsoft Maia 100	Meta MTIA v2
FP16 TFLOPS	2097	2200	989	1307	800	177
Area (mm²)	1366	1600	814	920	820	421
TFLOPS/mm²	1.54	1.38	1.22	1.42	0.98	0.42
TFLOPS/W	2.39	2.20	1.41	1.74	1.14	1.61

* Technology scaling is applied.

03 Beyond 2D scaling

More Compute. Efficient Scaling!



Vertical Compute Scaling
Stack compute units efficiently

Dedicated Interconnection Tier
Physically separate for high complex network topologies

High Bandwidth Vertical Links
Ultra-dense cross-tier links via hybrid-bonding / TSVs

From Concept to Physical Evidence:

The same compute architecture, optimized through 3D physical integration.

-16% Total Die Area

+32% Frequency

(Two-tier 3D-TeraPool folding by low diameter interconnect vs. 2D-TeraPool)

References

[1] Qwen Team, "Qwen3.8-27B model card," https://huggingface.co/Qwen/ Qwen3.8-27B, 2026, accessed Oct. 27, 2026.
[2] Y. Zhang et al., "TeraPool: A Physical Design Aware, 1024 RISC-V Cores Shared-L1-Memory Scaled-up Cluster Design with High Bandwidth Main Memory Link," IEEE Trans. Comput., 2025.
[3] Y. Zhang et al., "TeraNoC: A Multi-Channel 32-bit Fine-Grained, Hybrid Mesh-Crossbar NoC for Efficient Scale-up of 1000+ Core Shared-L1-Memory Clusters," ICCD, 2025.
[4] Y. Zhang et al., "A 410 GFLOP/s, 64 RISC-V Cores, 204.8 GBps Shared-Memory Cluster in 12 nm FinFET with Systolic Execution Support for Efficient B5G/6G AI-Enhanced O-RAN," ESSERC, 2025.