DAC Young Fellows

DESIGN AUTOMATION CONFERENCE
FROM CHIPS TO SYSTEMS — LEARN TODAY, CREATE TOMORROW

Advanced NoCs for Scalable AI Systems: Toward Smart and Energy-Efficient Interconnects in 3D Architectures

Lorenzo Leone[1] Fellow ID 682, Frank K. Gürkaynak[1], Luca Benini[1,2]

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

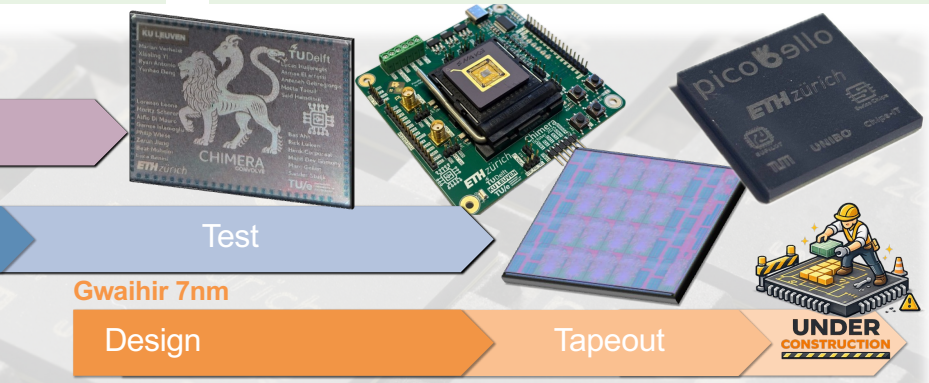
1. Motivation & Challenges

- **Attention-heavy inference** stresses dataflow and memory traffic
- Technology scaling enables **thousands of PEs on-chip**
- Large AI chips need scalable fabrics beyond **XBAR-based interconnects**
- Large tiled systems amplify pressure on the **NoC fabric** and **off-chip** traffic

2. Solutions

- **NoC-based scaleout** to connect many PEs
- Enable on-chip **multicast, barriers**, and **reductions**
- **CMOS2.0**[1]**:** Drive NoC evolution beyond 3D
- Preserve **energy-efficiency, scalability**, and **physical implementability**

3. From Idea to Silicon



Chimera 22nm: Design → Tapeout → Test

Picobello 7nm: Design → Tapeout → Test

Gwaihir 7nm: Design → Tapeout

July24 — July25 — July 26 (now) — 2027

3. Low-Power AI-MCU

Flexible and Heterogeneous AI-MCUs for real-time transformer inference

Chimera[2] (GF22)

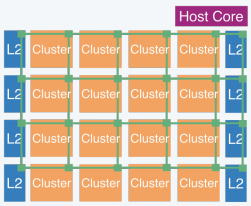
- 5 heterogeneous clusters
- Transformer Acceleration Cluster
- Low-latency XBAR
- High-bandwidth shared-L2

Up to **3.1 TOPS/W** on Transformer Acceleration Cluster

1.8x area efficiency on SoA accelerators

4. Scalable AI System

NoC-based scaleout of compute clusters for ML based application



Picobello (TSMC7)

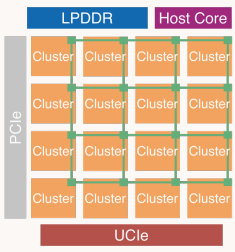
- 144 scalar PEs
- 16 FP Tensor Units
- FlooNoC[3] mesh
- Wide Physical Links

FlooNoC-based SoC

Peak Performance **4.1 TFLOPS**

5. Smart NoC

A Lightweight Collective-Capable NoC for Large-Scale ML Acelerators[4]



Gwaihir (TSMC7)

- 144 vector PEs
- 16 MX Tensor Units
- NoC with collectives
- 2x Wide Physical Links
- High-Speed I/O

5.3x and **2.8x** speedup on **broadcast** and **reduction**

Expected peak performance **32 TFLOPS**

CONVOLVE — THE EUPILOT — GlobalFoundries — tsmc — SwissChips

1. J. Ryckaert and S. B. Samavedam, "The CMOS 2.0 Revolution," Nature Reviews Electrical Engineering, vol. 1, 2024.
2. L. Leone, P. Wiese, et al., "CHIMERA: A Flexible and Scalable 3.1 TOPS/W AI-MCU with Transformer Accelerator and 563 Gb/s Shared-L2 Memory Subsystem with QoS Guarantees," arXiv:2606.02358, 2026.
3. T. Fischer, et al., "FlooNoC: A 645-Gb/s/link 0.15-pJ/B/hop Open-Source NoC With Wide Physical Links and End-to-End AXI4 Parallel Multistream Support," IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol. 33, no. 4, 2025.
4. L. Colagrande, L. Leone, *et al.*, "A Lightweight High-Throughput Collective-Capable NoC for Large-Scale ML Accelerators," arXiv:2603.26438, 2026.

ETH zürich — ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA — PULP Parallel Ultra Low Power