# GABRIELE TOMBESI

New York, NY | Ph.D. Candidate, Computer Science (May 2027)
gtombesi@cs.columbia.edu | LinkedIn | GitHub | Scholar | Website

## EDUCATION

**Columbia University** *Expected May 2027*
Ph.D. in Computer Science (GPA 3.97/4.00), New York, NY.

**Polytechnic of Turin / Grenoble INP / EPFL** *2020*
M.S. in Micro and Nano Technologies for Integrated Systems.

**Polytechnic of Turin** *2018*
B.S. in Physical Engineering.

## PROFILE

Ph.D. researcher in computer architecture and hardware/software co-design for AI accelerators, developing performance models, configurable tiled architectures, automated workload-mapping frameworks, HLS/RTL implementation and FPGA prototyping. Experience spans NoC/data movement, FPGA prototyping, silicon validation, and low-level ML kernels, with industry experience at Tenstorrent and NVIDIA Research.

## TECHNICAL SKILLS

**Programming & modeling:** C/C++, Python, SystemC, analytical/performance modeling, design-space exploration, profiling, Bash, Tcl, Git.
**Architecture & HW/SW co-design:** Tiled/systolic/vector accelerators, NoC, DMA/P2P/multicast data movement, memory hierarchies, multicore scheduling, workload mapping.
**ML acceleration:** GEMM/GEMV, Transformers, attention/FlashAttention, speculative decoding, tensor sharding, tensor parallelism, pipeline parallelism.
**RTL, HLS & prototyping:** SystemVerilog/Verilog, VHDL, Catapult/Stratus HLS, MatchLib, Vivado, Questa, FPGA prototyping, post-silicon validation.

## RESEARCH & ENGINEERING EXPERIENCE

**Columbia University — System-Level Design Group** *Jan. 2021 – Present*
*Ph.D. Researcher | Advisor: Prof. Luca P. Carloni | New York, NY*

- **Architecture exploration & performance modeling:** Built analytical and SystemC models of tiled AI accelerators to identify compute, memory-bandwidth, and NoC bottlenecks; explored dataflows, memory reuse, tile allocation, tensor sharding, and resource mappings to guide architectural decisions.
- **SPECTRA — workload-driven accelerator reconfiguration:** Developed a runtime-reconfigurable architecture for speculative LLM decoding, adapting tile dataflows and M/N/K sharding across draft decoding and batched verification. On a 20-tile FPGA prototype, achieved up to 2.09× speedup from tile-level reconfiguration and a further 1.25× from system-level adaptation.
- **FLIP2M / OASIS — automated workload mapping**: Developed an open-source optimization framework for a 49-tile heterogeneous SoC, jointly exploring accelerator allocation, tensor parallelism, memory bandwidth, and pipeline parallelism; achieved up to 1.94× latency improvement and 2.59× lower energy-delay product.
- **ESP — accelerator integration & data movement**: Core contributor to the open-source heterogeneous SoC platform; integrated SystemC/Catapult HLS and MatchLib flows and designed SystemVerilog accelerator-socket extensions for flexible peer-to-peer transfers between NoC-connected compute tiles.

**Tenstorrent** *May – Aug. 2026*
*Hardware/Software Co-Design Intern | Zurich, Switzerland*

- **Low-level accelerator kernels:** Implemented and optimized matrix-multiplication kernels on Tensix NEO cores, exploring 1D/2D partitioning, multicore mappings, on-chip data movement, and synchronization overheads.
- **Attention execution:** Implemented non-causal FlashAttention for prefill and optimized FPU/SFPU scheduling and data movement to overlap matrix operations, softmax processing, and communication across execution resources.

- **Performance characterization:** Profiled compute, communication, and synchronization bottlenecks across multicore configurations and used measurements to refine kernel scheduling, sharding, and dataflow choices.

**NVIDIA Research** *May – Aug. 2024*
*Summer Research Intern | Santa Clara, CA*

- Developed a parameterized SystemC/HLS accelerator generator for sliding-window vision workloads, enabling rapid exploration of architectural parallelism, dataflows, and implementation tradeoffs.
- Designed reusable SystemC/MatchLib data-movement and memory-orchestration mechanisms; mapped OpenCV kernels to assess throughput, memory reuse, hardware utilization, and resource cost.

## ADDITIONAL EXPERIENCE

**Columbia University — Master's Thesis Research** *Mar. – Oct. 2020*
Developed a compositional post-silicon validation methodology and modular VHDL debug infrastructure for heterogeneous NoC-based SoCs; integrated the debug hardware into a full-system platform subsequently fabricated in 12 nm.

**Columbia University — Teaching Assistant, System-on-Chip Platforms** *Fall 2022, 2023*
Designed and delivered labs spanning SystemC/HLS accelerator modeling, GEMM/Conv2D design, RTL integration, NoC/DMA interfaces, software bring-up, and FPGA deployment; led the course's adoption of Catapult HLS.

**IBM Research — Summer Research Intern** *Jun. – Sep. 2019*
*San Jose, CA.* Characterized metal-oxide semiconductor materials for gas sensing through experimental measurements and electronic-transport analysis.

## SELECTED PUBLICATIONS

[1] **SPECTRA: Adaptive Execution of Speculative Decoding on a Runtime-Reconfigurable Tiled Architecture.** G. Tombesi et al. *IEEE/ACM ICCAD, 2026.* First author.

[2] **FLIP2M: Flexible Intra-layer Parallelism and Inter-layer Pipelining for Multi-model AR/VR Workloads.** G. Tombesi et al. *ACM TECS (CASES/ESWEEK), 2025.* First author.

[3] **SoCProbe: Compositional Post-Silicon Validation of Heterogeneous NoC-Based SoCs.** G. Tombesi et al. *IEEE Design & Test, 2023.* First author; Best Paper Award.

[4] **Teaching Collaborative SoC Engineering with HLS-Based Accelerator Design and Agile System-Level Integration.** G. Tombesi et al. *WCAE, 2025.* First author.

[5] **RECONFORMER: A Multi-Level Run-Time Reconfigurable System-on-Chip for Accelerating Transformers.** J. Yang, G. Tombesi et al. *FPL, 2025.*

[6] **A Coherence-Aware Runtime Reconfigurable System-on-Chip for Efficient Language Model Inference.** J. Yang, G. Wallace, J. Zuckerman, G. Tombesi, L. Carloni. *IEEE Micro, 2026.*

[7] **A 12 nm Linux-SMP-Capable RISC-V SoC with 14 Accelerator Types, Distributed Hardware Power Management and Flexible NoC-Based Data Orchestration.** M. Cassel dos Santos et al. *ISSCC, 2024.*

[8] **EPOCHS-1: A 12 nm Highly Heterogeneous Open-Source SoC with Distributed Power Management and Integrated Hybrid Voltage Regulation.** J. Zuckerman et al. *IEEE JSSC, 2025.*