Fabian Peddinghaus

fabianpedd@gmail.com | linkedin.com/in/fabianpeddinghaus | github.com/fpedd | fabianpeddinghaus.com

Experience

Senior Software Developer Apr. 2023 – Present
Axelera AI, ML Compilation Group *Zurich, Switzerland*

- Joined Axelera's compiler team as one of its first engineers; built core components of the production AI compiler across three silicon generations, from early silicon bring-up to large-scale customer deployment.
- Built memory, scheduling, and tiling optimization passes targeting Axelera's in-memory compute architecture, delivering multi-fold inference speedups and substantial peak-memory reductions on production networks, outperforming comparable NVIDIA and Hailo accelerators on equivalent workloads, see public benchmark.
- Architected the compiler's concurrency and synchronization model, enabling sharding across multiple cores and chips; built graph-level rewrite passes for the compiler frontend to ingest ONNX and PyTorch models.
- Optimized attention and matmul kernels for transformer inference in an internal kernel DSL, significantly accelerating prefill and decode for Llama, Qwen, and similar architectures on Axelera silicon.
- Initiated the compiler performance dashboard and on-silicon measurement infrastructure, enabling same-day triage of regressions; designed and open-sourced OmniMalloc, a static memory allocation framework.

Visiting Student Researcher Sep. 2022 – Jan. 2023
Stanford University, Department of Electrical Engineering *Stanford, United States*

- Co-developed a novel systolic array-based AI accelerator for Transformers and CNNs, utilizing on-chip Resistive RAM, presented at the VLSI Symposium and open-sourced project code.
- Engineered the power-aware ML compiler featuring graph optimization and spatiotemporal power-gating, reducing overall power consumption by 40% while maintaining performance.

Graduate Research Assistant Oct. 2021 – Sep. 2022
TU Munich, Department of Electronic Design Automation *Munich, Germany*

- Created muRISCV-NN, an open-source AI kernel library for RISC-V architectures, achieving 60% higher performance than LLVM's auto-vectorizer with optimized kernels for the Vector and Packed extensions.
- Integrated library with TensorFlow Lite and microTVM backends, enabling seamless deployment of ML models on RISC-V platforms with ARM CMSIS-NN compatible interfaces.
- Mentored 50+ students as TA for two graduate-level courses in embedded systems and ML.

Undergraduate Research Assistant Oct. 2018 – Mar. 2021
TU Berlin, Department of Flight Mechanics and Flight Control *Berlin, Germany*

- Designed and published a distributed wireless control system, achieving 2× lower latency and higher reliability over a centralized baseline in real-world flight conditions.

Education

Technical University of Munich Apr. 2021 – Apr. 2023
M.Sc. Electrical and Computer Engineering, 4.0/4.0 GPA, top 2% *Munich, Germany*
Master's Thesis on power-aware optimization and compilation for edge AI accelerators.

Technical University of Berlin Oct. 2017 – Mar. 2021
B.Sc. Computer Engineering, 3.7/4.0 GPA, ranked 1st in cohort *Berlin, Germany*
Bachelor's Thesis on spectrum sensing algorithms for nanosatellites in LEO.

German Academic Scholarship Foundation Feb. 2020 – Apr. 2023
Merit-based scholarship awarded to top 0.5% of students in Germany.

Technologies

Programming Languages: Python, C, C++, Assembly (RISC-V/RVV, ARM), SystemC
ML Frameworks: TVM/TensorIR, MLIR (torch-mlir, Linalg), ONNX, PyTorch, TensorFlow
Scientific Computing: NumPy, SciPy, Z3-Solver, SymPy, Matplotlib, Jupyter, Pandas
Development Tools: Git, CMake, Make, Docker, Pytest, GDB, Valgrind, Linux/Unix
Languages: German (native), English (fluent), Spanish (basic)