Zhenyu Bai

zhenyu.bai@nus.edu.sg | ORCiD:0000-0003-1143-0762 | Scopus ID:58983402300 | GitHub: FTOD

Education

IRIT lab, University of Toulouse, PhD. in Computer Science — Oct 2019 – May 2023
- Scholarship Founded by French Minister for Higher Education and Research as top Master students.

University of Toulouse, Master in Embbeded Computing Systems — Sep 2017 – Aug 2019

University of Toulouse, BS in Computer Science — Sep 2014 – Aug 2017

Research Experience

Research Fellow (ARTIC Fellow from 2026), School of Computing National University of Singapore (PI: Tulika Mitra) — May 2023 – Present
- Reconfigurable spatial-dataflow architecture and compiler design [1, 2] (collaborations with Tenstorrent & IBM);
- Hardware accelerators for sparse and quantized AI workloads [3, 4, 2];
- Compilers for Coarse Grained Reconfigurable Array (CGRA) [5];
- Dataflow architecture and software co-design for Spiking Neural Networks [6, 7];
- Heterogeneous FPGA-GPU system for AI workloads [8] (collaborations with AMD).

PhD student, IRIT, University of Toulouse, France — Oct 2019 – May 2023
- CPU micro-architcture modeling and program performance analysis for real-time systems [9, 10, 11].

Research Intern, Verimag, Grenoble Alpes University, France — Apr 2019 – Oct 2019
- CPU cache analysis and program analysis for real-time systems [12].

Teaching Experiences

Computer Architecture and VHDL (≈80h), Computer Architecture and ARM assembly(≈50), Compilation Theory(≈60h) , Advanced Compilation (≈10h), Master student project supervisor (3 months/year)

Publications

Name = *Equal contribution.* * *Corresponding author.*

[1] Zhenyu, Bai, Pranav, Dangi, Rohan, Juneja, Zhaoying, Li*, Zhanglu, Yan, Huiying, Lan, and Tulika, Mitra. "A Data-Driven Dynamic Execution Orchestration Architecture". In: *31th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*. 2026.

[2] Bai, Zhenyu, Dangi, Pranav, Li, Huize*, and Mitra, Tulika. "SWAT: Scalable and efficient window attention-based transformers acceleration on FPGAs". In: *Proceedings of the 61st ACM/IEEE Design Automation Conference*. 2024, pp. 1–6.

[3] Dangi, Pranav, Bai, Zhenyu*, Juneja, Rohan, Wijerathne, Dhananjaya, and Mitra, Tulika. "Zed: A generalized accelerator for variably sparse matrix computations in ml". In: *Proceedings of the 2024 International Conference on Parallel Architectures and Compilation Techniques*. 2024, pp. 246–257.

[4] Yin, Chenyang, Bai, Zhenyu*, Venkatram, Pranav, Aggarwal, Shivam, Li, Zhaoying, and Mitra, Tulika. "TerEffic: Highly Efficient Ternary LLM Inference on FPGA". In: *arXiv preprint arXiv:2502.16473* (2025).

[5] Li, Zhaoying, Dangi, Pranav, Yin, Chenyang, Bandara, Thilini Kaushalya, Juneja, Rohan, Tan, Cheng, Bai, Zhenyu*, and Mitra, Tulika. "Enhancing CGRA Efficiency Through Aligned Compute and Communication Provisioning". In: *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1*. 2025, pp. 410–425.

[6] Yan, Zhanglu, Bai, Zhenyu*, Mitra, Tulika, and Wong, Weng-Fai. "SparrowSNN: A Hardware/software Co-design for Energy Efficient ECG Classification". In: *arXiv preprint arXiv:2406.06543(Under Submission for IEEE Transactions on Computers)* (2024).

[7] Yan, Zhanglu, Bai, Zhenyu*, and Wong, Weng-Fai. "Reconsidering the energy efficiency of spiking neural networks". In: *arXiv preprint arXiv:2409.08290* (2024).

[8] Bai, Zhenyu, Wu, Dan, Dangi, Pranav, Wijerathne, Dhananjaya, Miriyala, Venkata Pavan Kumar, and Mitra, Tulika. “Data-aware Dynamic Execution of Irregular Workloads on Heterogeneous Systems”. In: *arXiv preprint arXiv:2502.06304 (Under Submission for IEEE Transactions on Computers)* (2025).

[9] Bai, Zhenyu, Cassé, Hugues, De Michiel, Marianne, Carle, Thomas, and Rochange, Christine. “Improving the Performance of WCET Analysis in the Presence of Variable Latencies”. In: *The 21st ACM SIGPLAN/SIGBED Conference on Languages, Compilers, and Tools for Embedded Systems*. LCTES. London, United Kingdom, 2020. ISBN: 9781450370943. DOI: 10.1145/3372799.3394371. URL: https://doi.org/10.1145/3372799.3394371.

[10] Bai, Zhenyu, Cassé, Hugues, De Michiel, Marianne, Carle, Thomas, and Rochange, Christine. “A Framework for Calculating WCET Based on Execution Decision Diagrams”. In: *ACM Transactions on Embedded Computing Systems*. 2022. DOI: 10.1145/3476879. URL: https://doi.org/10.1145/3476879.

[11] Bai, Zhenyu*, Cassé, Hugues, Carle, Thomas, and Rochange, Christine. “Computing Execution Times With Execution Decision Diagrams in the Presence of Out-of-Order Resources”. In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 42.11 (2023), pp. 3665–3678. DOI: 10.1109/TCAD.2023.3258752.

[12] Bai, Zhenyu, Monniaux, David, and Maïza, Claire. “PLRU cache analysis”. In: *Proceedings of the 13th Junior Researcher Workshop on Real-Time Computing*. JRWRTC 2019. 2019.

[13] Li, Wei, Bai, Zhenyu, Wang, Heru, Dangi, Pranav, Zhang, Zhiqiang, Tan, Cheng, Lan, Huiying, Wong, Weng-Fai, and Mitra, Tulika. “TL: Automatic End-to-End Compiler of Tile-Based Languages for Spatial Dataflow Architectures”. In: *arXiv preprint arXiv:2512.22168* (2025).

[14] Li, Huize, Li, Zhaoying, Bai, Zhenyu, and Mitra, Tulika. “ASADI: Accelerating sparse attention using diagonal-based in-situ computing”. In: *2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE. 2024, pp. 774–787.

[15] Bai, Zhenyu. “Modélisation du comportement temporel du pipeline pour le calcul de WCET”. PhD thesis. Université Paul Sabatier-Toulouse III, 2023.

[16] Bai, Zhenyu, Cassé, Hugues, Michiel, Marianne de, Carle, Thomas, and Rochange, Christine. “Déterminer le WCET d’applications temps-réel en présence de latences d’exécution variables”. In: *Conférence francophone d’informatique en Parallélisme, Architecture et Système (COMPAS 2021)*. 2021.