ZHEYU YAN

Assistant Professor - ZJUY100 Young Researcher
College of Integrated Circuits
Zhejiang University

@ zyan2@zju.edu.cn   Hangzhou, Zhejiang   sites.nd.edu/zheyu-yan/   zyyan

EMPLOYMENT

Zhejiang University

Jan 2025 – Present

Assistant Professor
Research interests: efficient on-device learning, machine learning for EDA.

Univerisity of Notre Dame

Feb. 2024 – Dec. 2024

Postdoctoral Researcher
Mentor: Dr. Yiyu Shi.
Research interests: AI accelerator design and Compute-in-Memory (CiM) systems robustness improvement.

EDUCATION

Univerisity of Notre Dame

Aug. 2019 – May 2024

Ph.D. in Computer Engineering
Dissertation: Software-Hardware Co-Design of Neural Network Accelerators using Emerging Technologies.

Zhejiang University

Aug. 2015 – June 2019

B.S. in Electronic Engineering
Thesis: DNN quantization for digital accelerators.

AWARDS

• Best Paper Award, IEEE/ACM International Conference on Computer-Aided Design	Nov. 2023
• Best Demonstration (First Place), University Demo, IEEE/ACM Design Automation Conference	Dec. 2021
• Best Paper Award Nomination, IEEE/ACM International Conference on Computer-Aided Design	Oct. 2024
• CSE-Select Fellowship, Computer Science and Engineering Department	Aug. 2019
• Young Fellow, IEEE/ACM Design Automation Conference	July 2020

RESEARCH GRANTS

- “Embedding sensors hardware-software co-design,” Chengxi Great Walkway, 09/01/2025-08/31/2026, ¥400,000 (¥100,000)

PROFESSIONAL SERVICES

Conference Program Committee

- Design, Automation and Test in Europe Conference and Exhibition (DATE) 2024
- IEEE/ACM International Conference on Computer-Aided Design (ICCAD) 2024
- IEEE Computer Society Annual Symposium on VLSI (ISVLSI) 2024 - 2025
- Great Lakes Symposium on VLSI (GLSVLSI) 2024

Journal Reviewer

- Nature Reviews Electrical Engineering
- IEEE Transactions on Computers (TC)
- IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)
- IEEE Transactions on Emerging Topics in Computing (TETC)
- IEEE Transactions on Neural Networks and Learning Systems (TNNLS)
- IEEE Transactions on Very Large Scale Integration Systems (TVLSI)
- IEEE Transactions on Circuits and Systems for Artificial Intelligence (TCAS-AI)
- IEEE Transactions on Circuits and Systems I: Regular Papers (TCAS-I)
- IEEE Transactions on Circuits and Systems II: Express Briefs (TCAS-II)

- ACM Transactions on Design Automation of Electronic Systems (TODAES)
- ACM Transactions on Architecture and Code Optimization (TACO)
- ACM Transactions on Embedded Computing Systems (TECS)

PUBLICATIONS

Book Chapters

[B1] **Z. Yan**, Q. Lu, W. Jiang, *et al.*, Hardware–software co-design of deep neural architectures: From fpgas and asics to computing-in-memories, in *Embedded Machine Learning for Cyber-Physical, IoT, and Edge Computing: Software Optimizations and Hardware/Software Codesign*, Springer, 2023, pp. 271–301.

[B2] **Z. Yan**, X. S. Hu, and Y. Shi, On the reliability of computing-in-memory accelerators for deep neural networks, in *System Dependability and Analytics: Approaching System Dependability from Data, System and Analytics Perspectives*, Springer, 2022, pp. 167–190.

Journal Articles

[J1] Y. Guo, **Z. Yan**, X. Yu, *et al.*, Hardware design and the fairness of a neural network, *Nature Electronics*, ***Impact Factor 34.3***, *[Equal contrib. **first author**]*, 2024.

[J2] Z. Jia, T. Zhou, **Z. Yan**, J. Hu, and S. Yiyu, Personalized meta-federated learning for iot-enabled health monitoring, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), Impact Factor 2.9*, 2024.

[J3] Y. Jiang, K. Ni, T. Kämpfe, C. Zhuo, **Z. Yan**, and X. Yin, Csa-cim: Enhancing multi-functional computing-in-memory with configurable sense amplifiers, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2024.

[J4] **Z. Yan**, X. S. Hu, and Y. Shi, Compute-in-memory based neural network accelerators for safety-critical systems: Worst-case scenarios and protections, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, ***Impact Factor 2.9***, 2024.

[J5] **Z. Yan**, X. S. Hu, and Y. Shi, U-swim: Universal selective write-verify for computing-in-memory neural accelerators, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, ***Impact Factor 2.9***, 2024.

[J6] T. Wang, J. Zhang, J. Xiong, S. Bian, **Z. Yan**, *et al.*, Visualnet: An end-to-end human visual system inspired framework to reduce inference latency of deep neural networks, *IEEE Transactions on Computers*, *Impact Factor 3.4*, vol. 71, no. 11, pp. 2717–2727, 2022.

[J7] W. Jiang, Q. Lou, **Z. Yan**, *et al.*, Device-circuit-architecture co-exploration for computing-in-memory neural accelerators, *IEEE Transactions on Computers*, *Impact Factor 3.4*, 2020.

Conference Papers

[C1] H. Du, C. Wen, Z. Chen, L. Zhang, Q. Sun, **Z. Yan**, and C. Zhuo, Algorithm-hardware co-design of a unified accelerator for non-linear functions in transformers, in *2025 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, IEEE, 2025.

[C2] R. Qin, P. Ren, **Z. Yan**, *et al.*, Nvcim-pt: An nvcim-assisted prompt tuning framework for edge llms, in *2025 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, IEEE, 2025.

[C3] Y. Qin, Z. Jia, **Z. Yan**, *et al.*, A 10.60 μW 150 gops mixed-bit-width sparse cnn accelerator for life-threatening ventricular arrhythmia detection, *2025 30th Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2025.

[C4] R. Qin, Y. Hu, **Z. Yan**, J. Xiong, A. Abbasi, and Y. Shi, Towards fairness of neural architecture search via llms, in *2024 29th Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2024.

[C5] R. Qin, **Z. Yan**, D. Zeng, *et al.*, Robust implementation of retrieval-augmented generation on edge-based computing-in-memory architectures, *2024 International Conference on Computer-Aided Design (ICCAD)*, 2024.

[C6] Y. Qin, **Z. Yan**, Z. Pan, W. Wen, X. S. Hu, and Y. Shi, Tsb: Tiny shared block for efficient dnn deployment on nvcim accelerators, *2024 International Conference on Computer-Aided Design (ICCAD)*, 2024.

[C7] Y. Qin, **Z. Yan**, W. Wen, X. S. Hu, and Y. Shi, Special session: Sustainable deployment of deep neural networks on non-volatile compute-in-memory accelerators, in *2024 International Conference on Hardware/Software Codesign and System Synthesis (CODES+ ISSS)*, IEEE, 2024, pp. 1–9.

[C8] X. Wang, **Z. Yan**, M. Chang, Y. Shi, and W. Qian, Dasals: Differentiable architecture search-driven approximate logic synthesis, *2023 International Conference on Computer-Aided Design (ICCAD)*, 2023.

[C9] **Z. Yan**, Y. Qin, X. S. Hu, and Y. Shi, Improving realistic worst-case performance of nvcim dnn accelerators through training with right-censored gaussian noise, *2023 International Conference on Computer-Aided Design (ICCAD)* ***Acceptance rate 22.9%***, 2023.

[C10] **Z. Yan**, Y. Qin, X. S. Hu, and Y. Shi, On the viability of using llms for sw/hw co-design: An example in designing cim dnn accelerators, in *Proceedings of the 36th IEEE International System-on-chip Conference*, 2023.

[C11] B. Lu, **Z. Yan**, Y. Shi, and S. Ren, A semi-decoupled approach to fast and optimal hardware-software co-design of neural accelerators, *TinyML Summit*, 2022.

[C12] **Z. Yan**, X. S. Hu, and Y. Shi, Computing in memory neural network accelerators for safety-critical systems: Can small device variations be disastrous? *2022 International Conference on Computer-Aided Design (ICCAD)* ***Acceptance rate 22.0%***, 2022.

[C13] **Z. Yan**, X. S. Hu, and Y. Shi, Swim: Selective write-verify for computing-in-memory neural accelerators, *2022 59th ACM/IEEE Design Automation Conference (DAC)*, ***Acceptance rate 22.6%***, 2022.

[C14] **Z. Yan**, W. Jiang, X. S. Hu, and Y. Shi, Radars: Memory efficient reinforcement learning aided differentiable neural architecture search, in *2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC)*, IEEE, 2022, pp. 128–133.

[C15] **Z. Yan**, D.-C. Juan, X. S. Hu, and Y. Shi, Uncertainty modeling of emerging device based computing-in-memory neural accelerators with application to neural architecture search, in *2021 26th Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2021.

[C16] L. Yang, **Z. Yan**, M. Li, *et al.*, Co-exploration of neural architectures and heterogeneous asic accelerator designs targeting multiple tasks, in *2020 57th ACM/IEEE Design Automation Conference (DAC)*, IEEE, 2020.

[C17] **Z. Yan**, Y. Shi, W. Liao, M. Hashimoto, X. Zhou, and C. Zhuo, When single event upset meets deep neural networks: Observations, explorations, and remedies, in *2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC)*, IEEE, 2020.

INVITED TALKS

• Kyoto University, Kyoto, Japan (host: Prof. Hashimoto)	Oct. 2024
• AI Chip Center for Emerging Smart Systems, HongKong, China (host: Dr. Luhong Liang)	Oct. 2023
• Zhejiang University, Hangzhou, China (host: Prof. Cheng Zhuo)	June 2023
• Technical University of Munich, Munich, Germany (host: Prof. Ulf Schlichtmann)	Dec. 2022

STUDENTS MENTORED

Ph.D. Student

Master Student

Undergraduate Student

- Carl Xu, June-Sept. 2022 (Summer research intern from Penn High School) [J5]
- Xiaoting Yu, June-Sept. 2023 (Summer research intern from Southern University of Science and Technology) [J1]

COURSE TAUGHT

• Advanced Computer Architecture (TA CSE60321)	Spring 2021 & Fall 2022
• Theory of Computing (TA CSE30151)	Fall 2019
• Elements of Computing II (TA CSE10102)	Spring 2020