ZHEYU YAN

PhD Candidate
Department of Computer Science and Engineering
University of Notre Dame

zyan2@nd.edu **574-210-6016** **Notre Dame, IN** **sites.nd.edu/zheyu-yan/** **zyyan**
Address: 222 Cushing Hall of Engineering, Notre Dame, IN 46556

RESEARCH INTERESTS AND SPECIALTIES

- **Cross-layer design of Compute-in-Memory (CiM) deep neural network (DNN) accelerators:** we are the first to link the DNN properties with device variations in CiM accelerators so as to provide device, circuit, architecture and algorithm solution to the device variation issue[1]–[3], [5]–[8].
- **Co-Design of Neural Architecture and Hardware Accelerators:** we are the first to propose co-design frameworks to explore neural architectures and FPGAs, ASICs, and CiM platforms[4], [5], [8], [10], [12], [13].

EMPLOYMENT

Amazon Lab 126 — June 2022 – Sept. 2022

- **ASIC Engineer Intern.** Focus on designing full integer accelerators for transformer models.

Univerisity of Notre Dame — Aug. 2019 – May. 2022

- **Teaching Assistant** for courses: (1) CSE60321 Advanced Computer Architecture, (2) CSE30151 Theory of Computing, and (3) CSE10102 Elements of Computing II.

EDUCATION

Univerisity of Notre Dame — Aug. 2019 – May 2024 (Expected)

- **Ph.D. in Computer Engineering (Advisor Dr. Yiyu Shi and Dr. Xiaobo Sharon Hu).** Focus on software-hardware co-design of compute-in-memory deep neural network accelerators.

Zhejiang University — Aug. 2015 – June 2019

- **B.S. in Electronic Engineering (Advisor Dr. Cheng Zhuo).** 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
- Computer Science and Engineering Department, CSE-Select Fellowship — Aug. 2019
- Design Automation Conference Young Fellow — July 2020

PROPOSAL WRITING

- PI: Yiyu Shi, "SPX: Collaborative Research: Scalable Neural Network Paradigms to Address Variability in Emerging Device based Platforms for Large Scale Neuromorphic Computing," National Science Foundation, 9/1/19 – 8/31/24, $360,132
- PI: Yiyu Shi, "Collaborative Research: DESC: Type II: REFRESH: Revisiting Expanding FPGA Real-estate for Environmentally Sustainability Heterogeneous-Systems," National Science Foundation, 11/01/23-10/31/26, $500,000
- PI: Yiyu Shi, "FuSe-TG: Cross-layer Co-Design for Self-Evolving Implantable Devices," National Science Foundation, 06/01/23-05/31/25, $300,000
- PI: Yiyu Shi, "AI-Based EDA Tools for Co-Design of ReRAM/FeFET and CMOS based Heterogeneous Systems," AI Chip Center for Emerging Smart Systems Limited, 09/01/2022-08/31/2026, $1,200,000

TEACHING EXPERIENCE

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

MENTOR OF STUDENTS

- Carl Xu, June-Sept. 2022 (Summer research intern from Penn High School) [16]
- Xiaoting Yu, June-Sept. 2023 (Summer research intern from Southern University of Science and Technology) [17]

SELECTED PUBLICATIONS

[1] **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.

[2] **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.

[3] **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.

[4] **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.

[5] **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.

[6] **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.

[7] **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.

[8] **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.

[9] **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.

[10] 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.

[11] 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*, vol. 71, no. 11, pp. 2717–2727, 2022.

[12] 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.

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

[14] 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)*,

[15] 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)*.

Selected Under Review and Work-in-Progress Papers

[16] **Z. Yan**, C. Xu, 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)*, 2023.

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

[18] **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)*, 2023.

INVITED TALKS

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

PROFESSIONAL SERVICES

Journal Reviewer

- Nature Reviews Electrical Engineering
- 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)

Conference Reviewer

- IEEE/ACM Design Automation Conference 2020 - 2023
- IEEE/ACM International Conference on Computer-Aided Design 2019-2020
- AAAI Conference on Artificial Intelligence 2022