Yongan (Luke) Zhang

832-566-7124 | yzhang919@gatech.edu | Personal Site

Research Interest

- AI-enabled Hardware Design Automation
- Software/Hardware Co-design for Efficient AI

Education

- **Georgia Tech** Atlanta, GA, USA
  PhD, Computer Science, Advisor: Prof. Yingyan Lin *Aug. 2023 – May. 2025*
- **Rice University** Houston, TX, USA
  MS, Electrical and Computer Engineering, Advisor: Prof. Yingyan Lin *Jan. 2020 – May 2023*
- **Rice University** Houston, TX, USA
  BS, Electrical and Computer Engineering *Aug. 2015 – May 2019*

Experiences

- **Research Intern, Meta** Mentor: Dr. Yuecheng Li
  Adaptive Once-for-all model compression *May 2024 - Jan 2025*
  - Designed a once-for-all AI model compression framework, enabling fine-tuning the model once and flexibly pruning the model for different accuracy and efficiency tradeoff without retraining
  - Profiled the flexibly pruned models for improved hardware efficiency on existing VR hardware
  - Explored potential domain specific acceleration opportunities for runtime adaptive model compression
- **Research Intern, Meta** Mentor: Dr. Yuecheng Li
  Reconfigurable hardware acceleration for VR mobile telepresence pipeline *May 2022 - Dec 2022*
  - Designed the run-time reconfigurable architecture for improved hardware resource efficiency
  - Designed the fine-grained operation scheduling for model-to-hardware mapping
  - Designed RTL-verified performance modeling for flexible DSE
  - Constructed design automation flow to auto generate the arch design and scheduling given Pytorch models
  - Worked with a hybrid of Catapult HLS, Vivado, RTL, C++ and Python for the whole flow
- **Ph.D. Intern, PNNL** Mentor: Dr. Ang Li
  Multi-FPGA acceleration for scalable Graph Neural Networks implementation *Jan 2022 – May 2022*
  - Designed the multi-FPGA architecture for large GNN acceleration
  - Implemented from arch design to final board deployment (fixed model-to-hardware mapping)
  - Worked with Xilinx HLS and Vivado for arch, and Pynq for deployment

Publications

1. **Y. Zhang**, Z. Yu, Z. Ye, S. Li, C. Li, Y. Lin,“SLAVA: Scalable LLM-Driven Verilog Design With an Assertion-Guided Automated Self-Refinement Loop”, *Under Review, 2025.*
2. **Y. Zhang**, Y. Li, S. Sarwar, H. Sumbul, Y. Fu, H. You, C. Wan, Y. Lin,“Re-CATA: Real-Time and Flexible Accelerator Design Framework for On-device Codec Avatars”, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), 2025.*
3. Z. Ye, Y. Fu, J. Zhang, L. Li, **Y. Zhang**, S. Li, C. Wan, C. Li, S. Prathipati, Y. Lin,“Gaussian Blending Unit: An Edge GPU Plug-in for Real-Time Gaussian-Based Rendering in AR/VR”, *IEEE International Symposium on High Performance Computer Architecture (HPCA), 2025.*
4. **Y. Zhang**, Z. Yu, Y. Fu, C. Wan, Y. Lin, “MG-Verilog: Multi-grained Dataset Towards Enhanced LLM-assisted Verilog Generation”, *1st IEEE International Workshop on LLM-Aided Design (LAD), 2024, Best Paper.*
5. **Y. Zhang**, X. Zhang, P. Xu, Y. Zhao, C. Hao, D. Chen, Y. Lin, “AutoAI2C: An Automated Hardware Generator for DNN Acceleration On Both FPGA and ASIC”, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), 2024.*
6. **Y. Zhang**, Y. Fu, Z. Yu, K. Zhao, C. Wan, C. Li, Y. Lin, “INVITED: Data4AIGChip: An Automated Data Generation and Validation Flow for LLM-assisted Hardware Design”, *The 58th Design Automation Conference (DAC), 2024.*

7. Y. Fu, Z. Yu, J. Li, J. Qian, **Y. Zhang**, X. Yuan, D. Shi, R. Yakunin, Y. Lin,“AmoebaLLM: Constructing Any-Shape Large Language Models for Efficient and Instant Deployment”, *The 38th Conference on Neural Information Processing Systems (NeurIPS), 2024.*
8. Y. Fu*, **Y. Zhang***, Z. Yu*, S. Li, Z. Ye, C. Li, C. Wan, Y. Lin, “GPT4AIGChip: Towards Next-Generation AI Accelerator Design Automation via Large Language Models”, *ACM/IEEE International Conference On Computer Aided Design (ICCAD), 2023.*
9. Y. Zhao, **Y. Zhang**, Y. Fu, X. Ouyang, C. Wan, S. Wu, A. Banta, M. John, A. Post, M. Razavi, J. Cavallaro, B. Aazhang, Y. Lin, “e-G2C: A 0.14-to-8.31 uJ/Inference NN-based Processor with Continuous On-chip Adaptation for Anomaly Detection and ECG Conversion from EGM”, *IEEE Symposium on VLSI Technology and Circuits (VLSI), 2022.*
10. H. You, Y. Zhao, Z. Yu, C. Wang, Y. Fu, J. Yuan, S. Wu, S. Zhang, **Y. Zhang**, C. Li, V. Boominathan, A. Veeraraghavan, Z. Li, Y. Lin, “EyeCoD: Eye Tracking System Acceleration via FlatCam-Based Algorithm and Accelerator Co-Design”, *IEEE/ACM International Symposium on Computer Architecture (ISCA), 2022.*
11. H. You, T. Geng, **Y. Zhang**, A. Li, Y. Lin, “GCoD: Graph Convolutional Network Acceleration via Dedicated Algorithm and Accelerator Co-Design”, *IEEE International Symposium on High-Performance Computer Architecture* (HPCA), 2022.
12. **Y. Zhang**, H. You, Y. Fu, T. Geng, A. Li, Y. Lin, “G-CoS: GNN-Accelerator Co-Search Towards Both Better Accuracy and Efficiency”, *IEEE/ACM International Conference on Computer-Aided Design (ICCAD), 2021.*
13. **Y. Zhang**, Y. Fu, W. Jiang, C. Li, H. You, M. Li, V. Chandra, Y. Lin, DIAN: “Differentiable Accelerator-Network Co-Search Towards Maximal DNN Efficiency”, *ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED), 2021.*
14. **Y. Zhang**, A. Benta, Y. Fu, M. John, A. Post, M. Razavi, J. Cavallaro, B. Aazhang, Y. Lin, “RT-RCG: Neural Network and Accelerator Search Towards Effective and Real-time ECG Reconstruction from Intracardiac Electrograms”, *The ACM Journal on Emerging Technologies in Computing Systems (JETC), 2021.*
15. Y. Fu, **Y. Zhang**, H. You, Y. Lin, “Auto-NBA: Efficient and Effective Search Over The Joint Space of Networks, Bitwidths, and Accelerators”, *The International Conference on Machine Learning (ICML), 2021.*
16. Y. Fu, **Y. Zhang**, C Li, Z Yu, Y Lin, “A3C-S: Automated Agent Accelerator Co-Search towards Efficient Deep Reinforcement Learning”, *The 58th Design Automation Conference (DAC), 2021.*
17. Y. Fu, Z. Yu, **Y Zhang**, Y Jiang, C Li, Y Liang, M Jiang, Z Wang, Y Lin, “InstantNet: Automated Generation and Deployment of Instantaneously Switchable-Precision Networks”, *The 58th Design Automation Conference (DAC), 2021.*
18. T. Geng, C. Wu, **Y. Zhang**, C. Tang, C. Xie, H. You, M. Herbordt, Y. Lin, A. Li, “I-GCN: A Graph Convolutional Network Accelerator with Runtime Locality Enhancement through Islandization”, *IEEE/ACM International Symposium on Microarchitecture (MICRO), 2021.*
19. M. Li, Z. Yu, **Y. Zhang**, Y. Fu, Y. Lin, “O-HAS: Optical Hardware Accelerator Search for Boosting Both Acceleration Performance and Development Speed”, *IEEE/ACM International Conference on Computer-Aided Design (ICCAD), 2021.*
20. C. Li, Z. Yu, Y. Fu, **Y. Zhang**, Y. Zhao, H. You, Q. Yu, Y. Wang, Y. Lin, “HW-NAS-Bench: Hardware-Aware Neural Architecture Search Benchmark”, *The International Conference on Learning Representations (ICLR), 2021.*
21. H. You, X. Chen, **Y. Zhang**, C. Li, S. Li, Z. Liu, Z. Wang, Y. Lin, “ShiftAddNet: A Hardware-Inspired Deep Network”, *Conference on Neural Information Processing Systems (NeurIPS), 2020.*
22. Y. Zhao, C. Li, Y. Wang, P. Xu, **Y. Zhang**, Y. Lin, “DNN-Chip Predictor: A Multi-grained Graph-based Performance Simulator for DNN Accelerators”, *International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2020.*
23. P. Xu, Y. Zhao, C. Hao, X. Zhang, Z. Guan, **Y. Zhang**, Y. Wang, D. Chen, Y. Lin, “AutoDNNchip: An Automated DNN Chip Predictor and Builder for Both FPGAs and ASICs”, *ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA), 2020.*

Awards

- IEEE LAD Best Paper 2024
- 1st & 2nd Place University Demonstration at DAC 2022 & 2023
- Distinction in Research and Creative Work 2019