JIAQI GU

2501 Speedway, Austin, TX 78712 ◇ The University of Texas at Austin
jqgu@utexas.edu ◇ (512) 264-5470 ◇ jqgu.net
Ph.D. Candidate ◇ Department of Electrical & Computer Engineering

RESEARCH INTERESTS

Emerging Hardware for High-Performance, Efficient Computing

- Efficient AI hardware design
- Electronic-photonic mixed-signal computing platform design

Efficient Algorithm, Co-Design & Automation

- Hardware-software co-design & automation (for photonics, post-CMOS electronics, quantum)
- Efficient ML model/algorithm
- AI/ML for hardware design & design automation

EDUCATION

The University of Texas at Austin, TX, USA *Aug. 2018 – Present*
Ph.D. Candidate, Department of Electrical and Computer Engineering
Advisor: David Z. Pan
Co-advisor: Ray T. Chen
(GPA 4.0/4.0)

Fudan University, Shanghai, China *Sep. 2014 – Jun. 2018*
B.E., Department of Microelectronic Science and Engineering
(GPA: 3.91/4.0)
(Rank top 2/71)

AWARDS AND HONORS

Winner at Robert S. Hilbert Memorial Optical Design Competition	Synopsys	2022
Donald O. Pederson Best Paper Award	IEEE TCAD	2021
Cockrell School Graduate Student Fellowship	UT Austin	2021
First Place at ACM Student Research Competition Grand Finals	ACM	2021
Best Poster Award at NSF Workshop on Machine Learning Hardware	NSF Workshop	2020
First Place at ACM/SIGDA Student Research Competition	ACM/SIGDA	2020
7th Place at IWLS Contest on Machine Learning+Logic Synthesis	IWLS	2020
DAC Young Fellow	DAC	2020,2021
Best Paper Finalist (1 out of 6)	DAC	2020
Best Paper Award	ASP-DAC	2020
4th Place, System Design Contest on Low Power Object Detection	DAC-SDC	2019
First Prize Scholarship	Fudan University	2017–2018
2nd & 3rd Prize, National Mathematical Contest in Modeling	Fudan University	2016–2017

PROFESSIONAL EXPERIENCE

Nvidia Inc., CA, USA *May 2022 – Oct 2022*
Research Intern, ASIC & VLSI Research Team
- Hardware-efficient Transformer compression for natural language processing

Meta Platforms Inc., CA, USA *May 2021 – Dec 2021*
Research Intern, Meta reality labs, FAST AI team
- Efficient multi-scale Vision Transformer design for high-performance computer vision

SELECTED RESEARCH PROJECTS

Emerging Hardware for Efficient Computing

Open-source library for photonic AI computing: https://github.com/JeremieMelo/pytorch-onn [J11]
Contribute to library for quantum machine learning: https://github.com/mit-han-lab/torchquantum
Electronic-photonic NN accelerator [J13], [J10], [J7], [C16], [C9], [C5], [C2]
Photonic in-memory computing [J12], [C24]

Co-Design & Optimization for Emerging Hardware

Reliability and efficiency-driven model-circuit co-optimization flow [J13], [J10], [C34], [C27], [C5], [C1]
Machine learning-enabled hardware simulation & performance prediction [C37], [C33]
Automated circuit/architecture design [C34], [C26], [C25]
Efficient on-chip/on-device training for self-learnable AI hardware [C28], [C23], [C18], [C11], [C10]

PROFESSIONAL SERVICE

Working Group Member

- NSF AI Institute TILOS Ethics and Early Career Development, 2022.

Local Arrangement Co-Chair

- IEEE CAS Seasonal School: AI/ML for IC Design and EDA, 2022.

Program Committee Member

- Association for the Advancement of Artificial Intelligence (AAAI), 2023

Reviewer

- IEEE Transaction on Computer-Aided Design of Integrated Circuits and Systems (TCAD)
- ACM/IEEE Design Automation Conference (DAC)
- IEEE/ACM International Conference on Computer-Aided Design (ICCAD)
- IEEE Computer Society Annual Symposium on VLSI (ISVLSI)
- ACM Great Lakes Symposium on VLSI (GLSVLSI)
- Conference on Neural Information Processing Systems (NeurIPS)
- IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)
- European Conference on Computer Vision (ECCV)
- Association for the Advancement of Artificial Intelligence (AAAI)
- IEEE Journal of Selected Topics in Quantum Electronics (JSTQE)
- Applied Physics Letters (APL)
- IEEE Photonics Technology Letters (PTL)

TEACHING

Graduate Teaching Assistant	EE382M: VLSI Physical Design Automation	Spring 2022

SKILLS

Programming Languages

Python (PyTorch/TensorFlow), C/C++, CUDA, Matlab, Verilog

EDA Tools

Cadence Virtuoso, Synopsys Design Compiler, Xilinx Vivado Design Suite, Synopsys Optodesigner

PUBLICATIONS

Journal Papers

[J13] **Jiaqi Gu**, Chenghao Feng, Hanqing Zhu, Zheng Zhao, Zhoufeng Ying, Mingjie Liu, Ray T. Chen, and David Z. Pan, "SqueezeLight: A Multi-Operand Ring-Based Optical Neural Network with Cross-Layer Scalability," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, Jul. 2022.

[J12] Hanqing Zhu, **Jiaqi Gu**, Chenghao Feng, Mingjie Liu, Zixuan Jiang, Ray T. Chen, and David Z. Pan, "ELight: Towards Efficient and Aging-Resilient Photonic In-Memory Neurocomputing," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, Jun. 2022.

[J11] **Jiaqi Gu**, Chenghao Feng, Hanqing Zhu, Ray T. Chen, and David Z. Pan, "Light in AI: Toward Efficient Neurocomputing with Optical Neural Networks - A Tutorial," *IEEE Transactions on Circuits and Systems–II: Express Briefs (TCAS-II)*, Apr. 2022.

[J10] Chenghao Feng*, **Jiaqi Gu***, Hanqing Zhu, Zhoufeng Ying, Zheng Zhao, David Z. Pan, and Ray T. Chen, "A compact butterfly-style silicon photonic-electronic neural chip for hardware-efficient deep learning," *ACS Photonics*, 2022.

[J9] Chenghao Feng, Zhoufeng Ying, Zheng Zhao, **Jiaqi Gu**, David Z. Pan, and Ray T. Chen, "Towards high-speed and energy-efficient computing: A WDM-based scalable on-chip silicon integrated optical comparator," *Laser & Photonics Reviews*, Jun. 2021.

[J8] Zhoufeng Ying, Chenghao Feng, Zheng Zhao, **Jiaqi Gu**, Richard Soref, David Z. Pan, and Ray T. Chen, "Sequential logic and pipelining in chip-based electronic-photonic digital computing," *IEEE Photonics Journal*, Oct. 2020.

[J7] **Jiaqi Gu**, Zheng Zhao, Chenghao Feng, Zhoufeng Ying, Mingjie Liu, Ray T. Chen, and David Z. Pan, "Towards Hardware-Efficient Optical Neural Networks: Beyond FFT Architecture via Joint Learnability," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, 2020.

[J6] Chenghao Feng, Zhoufeng Ying, Zheng Zhao, **Jiaqi Gu**, Ray T. Chen, and David Z. Pan, "Wavelength-division-multiplexing (WDM)-based integrated electronic–photonic switching network (EPSN) for high-speed data processing and transportation," *Nanophotonics*, Aug. 2020.

[J5] Yibo Lin, Zixuan Jiang, **Jiaqi Gu**, Wuxi Li, Shounak Dhar, Haoxing Ren, Brucek Khailany, and David Z. Pan, "DREAMPlace: Deep Learning Toolkit-Enabled GPU Acceleration for Modern VLSI Placement," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, Jun. 2020. (**Best Paper Award**)

[J4] Zhoufeng Ying, Chenghao Feng, Zheng Zhao, Shounak Dhar, Hamed Dalir, **Jiaqi Gu**, Yue Cheng, Richard Soref, David Z. Pan, and Ray T. Chen, "Electronic-photonic Arithmetic Logic Unit for High-speed Computing," *Nature Communications*, Apr. 2020.

[J3] Yibo Lin, Wuxi Li, **Jiaqi Gu**, Mark Ren, Brucek Khailany, and David Z. Pan, "ABCDPlace: Accelerated Batch-based Concurrent Detailed Placement on Multi-threaded CPUs and GPUs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, Feb. 2020.

[J2] Ruoyao Wang, Zhenghan Fang, **Jiaqi Gu**, Yi Guo, Shicong Zhou, Yuanyuan Wang, Cai Chang, and Jinhua Yu, "High-resolution Image Reconstruction for Portable Ultrasound Imaging Devices," *EURASIP Journal on Advances in Signal Processing*, Dec. 2019.

[J1] **Jiaqi Gu**, Zeju Li, Yuanyuan Wang, Haowei Yang, Zhongwei Qiao, and Jinhua Yu, "Deep Generative Adversarial Networks for Thin-section Infant MR Image Reconstruction," *IEEE Access*, May 2019.

Refereed Conference Papers

[C37] **Jiaqi Gu**, Zhengqi Gao, Chenghao Feng, Hanqing Zhu, Ray T. Chen, Duane S. Boning, and David Z. Pan, "NeurOLight: A Physics-Agnostic Neural Operator Enabling Parametric Photonic Device Simulation," *Conference on Neural Information Processing Systems (NeurIPS)*, Dec. 2022. (**Spotlight**) (Acceptance Rate: 25.6%)

[C36] **Jiaqi Gu**, Ben Keller, Jean Kossaifi, Anima Anandkumar, Brucek Khailany, and David Z. Pan, "HEAT: Hardware-Efficient Automatic Tensor Decomposition for Transformer Compression," *Conference on Neural Information Processing Systems (NeurIPS), ML for System Workshop (MLSys)*, Dec. 2022. (**Spotlight**)

[C35] Wei Shi, Hanrui Wang, **Jiaqi Gu**, Mingjie Liu, David Pan, Song Han, and Nan Sun, "RobustAnalog: Fast Variation-Aware Analog Circuit Design Via Multi-task RL," *ACM/IEEE Workshop on Machine Learning for CAD (MLCAD)*, Aug. 2022.

[C34] Hanqing Zhu, Keren Zhu, **Jiaqi Gu**, Harrison Jin, Ray T.Chen, Jean Anne Incorvia, and David Z. Pan, "Fuse and Mix: MACAM-Enabled Analog Activation for Energy-Efficient Neural Acceleration," *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, Jul. 2022. (Acceptance Rate: 23.5%)

[C33] Hanrui Wang, Pengyu Liu, Jinglei Cheng, Zhiding Liang, **Jiaqi Gu**, Zirui Li, Yongshan Ding, Weiwen Jiang, Yiyu Shi, Xuehai Qian, David Z. Pan, Frederic T. Chong, and Song Han, "QuEst: Graph Transformer for Quantum Circuit Reliability Estimation," *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, Jul. 2022. (Acceptance Rate: 23.5%)

[C32] Chenghao Feng, **Jiaqi Gu**, Hanqing Zhu, Zhoufeng Ying, Zheng Zhao, David Z. Pan, and Ray T. Chen, "Optoelectronically Interconnected Hardware-Efficient Deep Learning using Silicon Photonic Chips," *Conference on Lasers and Electro-Optics*, Mar. 2022.

[C31] Chenghao Feng, **Jiaqi Gu**, Hanqing Zhu, David Z. Pan, and Ray T. Chen, "Design and Experimental Demonstration of A Hardware-Efficient Integrated Optical Neural Network," *Conference on Lasers and Electro-Optics*, Mar. 2022.

[C30] **Jiaqi Gu**, Hyoukjun Kwon, Dilin Wang, Wei Ye, Meng Li, Yu-Hsin Chen, Liangzhen Lai, Vikas Chandra, and David Z. Pan, "Multi-Scale High-Resolution Vision Transformer for Semantic Segmentation," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. (Acceptance Rate: 25.3%)

[C29] **Jiaqi Gu**, Hanqing Zhu, Chenghao Feng, Zixuan Jiang, Mingjie Liu, Shuhan Zhang, Ray T. Chen, and David Z. Pan, "ADEPT: Automatic Differentiable DEsign of Photonic Tensor Cores," *ACM/IEEE Design Automation Conference (DAC)*, Jul. 2022. (**Best-in-Track Paper**) (Acceptance Rate: 23%)

[C28] Hanrui Wang, Zirui Li, **Jiaqi Gu**, Yongshan Ding, David Z. Pan, and Song Han, "QOC: Quantum On-Chip Training with Parameter Shift and Gradient Pruning," *ACM/IEEE Design Automation Conference (DAC)*, Jul. 2022. (Acceptance Rate: 23%)

[C27] Hanrui Wang, **Jiaqi Gu**, Yongshan Ding, Zirui Li, Frederic T. Chong, David Z. Pan, and Song Han, "QuantumNAT: Quantum Noise-Aware Training with Noise Injection, Quantization and Normalization," *ACM/IEEE Design Automation Conference (DAC)*, Jul. 2022. (Acceptance Rate: 23%)

[C26] Zizheng Guo, Mingjie Liu, **Jiaqi Gu**, Shuhan Zhang, David Z. Pan, and Yibo Lin, "A Timing Engine Inspired Graph Neural Network Model for Pre-Routing Slack Prediction," *ACM/IEEE Design Automation Conference (DAC)*, Jul. 2022. (**Best-in-Track Paper**) (Acceptance Rate: 23%)

[C25] Hanrui Wang, Yongshan Ding, **Jiaqi Gu**, Yujun Lin, David Z. Pan, Frederic T. Chong, and Song Han, "QuantumNAS: Noise-Adaptive Search for Robust Quantum Circuits," *IEEE International Symposium on High Performance Computer Architecture (HPCA)*, Feb. 2022. (Acceptance Rate: 29%)

[C24] Hanqing Zhu, **Jiaqi Gu**, Chenghao Feng, Mingjie Liu, Zixuan Jiang, Ray T. Chen, and David Z. Pan, "ELight: Enabling Efficient Photonic In-Memory Neurocomputing with Life Enhancement," *IEEE/ACM Asia and South Pacific Design Automation Conference (ASPDAC)*, Jan. 2022.

[C23] **Jiaqi Gu**, Hanqing Zhu, Chenghao Feng, Zixuan Jiang, Ray T. Chen, and David Z. Pan, "L2ight: Enabling On-Chip Learning for Optical Neural Networks via Efficient in-situ Subspace Optimization," *Conference on Neural Information Processing Systems (NeurIPS)*, Dec. 2021. (Acceptance Rate: 22.7%)

[C22] **Jiaqi Gu**, Hanqing Zhu, Chenghao Feng, Mingjie Liu, Zixuan Jiang, Ray T. Chen, and David Z. Pan, "Towards Memory-Efficient Neural Networks via Multi-Level in situ Generation," *International Conference on Computer Vision (ICCV)*, Oct. 2021. (Acceptance Rate: 25.9%)

[C21] Zixuan Jiang, **Jiaqi Gu**, Mingjie Liu, Keren Zhu, and David Z. Pan, "Optimizer Fusion: Efficient Training with Better Locality and Parallelism," *International Conference on Learning Representations (ICLR) Workshop, Hardware Aware Efficient Training (HAET)*, May 2021. (Acceptance Rate: 28.7%)

[C20] Chenghao Feng, **Jiaqi Gu**, Hanqing Zhu, David Z. Pan, and Ray T. Chen, "Experimental Demonstration of a WDM-based Integrated Optical Decoder for Compact Optical Computing," *Conference on Lasers and Electro-Optics*, May 2021.

[C19] Jason Midkiff, Ali Rostamian, Kyoung Min Yoo, Aref Asghari, Chao Wang, Chenghao Feng, Zhoufeng Ying, **Jiaqi Gu**, Haixia Mei, Ching-Wen Chang, James Fang, Alan Huang, Jong-Dug Shin, Xiaochuan Xu, Michael Bukshtab, David Z. Pan, and Ray T. Chen, "Integrated Photonics for Computing, Interconnects and Sensing," *Conference on Lasers and Electro-Optics*, May 2021. (**Invited Paper**)

[C18] **Jiaqi Gu**, Chenghao Feng, Zheng Zhao, Zhoufeng Ying, Ray T. Chen, and David Z. Pan, "Efficient On-Chip Learning for Optical Neural Networks Through Power-Aware Sparse Zeroth-Order Optimization," *Association for the Advancement of Artificial Intelligence (AAAI)*, Feb. 2021. (Acceptance Rate: 21%)

[C17] Shubham Rai, Walter Lau Neto, Yukio Miyasaka, Xinpei Zhang, Mingfei Yu, Qingyang Yi, Masahiro Fujita, Guilherme B. Manske, Matheus F. Pontes, Leomar S. da Rosa Junior, Marilton S. de Aguiar, Paulo F. Butzen, Po-Chun Chien, Yu-Shan Huang, Hoa-Ren Wang, Jie-Hong R. Jiang, **Jiaqi Gu**, Zheng Zhao, Zixuan Jiang, David Z. Pan, *et al.*, "Logic Synthesis Meets Machine Learning: Trading Exactness for Generalization," *IEEE/ACM Proceedings Design, Automation and Test in Europe (DATE)*, Feb. 2021. (Acceptance Rate: 24%)

[C16] **Jiaqi Gu**, Chenghao Feng, Zheng Zhao, Zhoufeng Ying, Mingjie Liu, Ray T. Chen, and David Z. Pan, "SqueezeLight: Towards Scalable Optical Neural Networks with Multi-Operand Ring Resonators," *IEEE/ACM Proceedings Design, Automation and Test in Europe (DATE)*, Feb. 2021. (Acceptance Rate: 24%)

[C15] **Jiaqi Gu**, Zheng Zhao, Chenghao Feng, Zhoufeng Ying, Ray T. Chen, and David Z. Pan, "O2NN: Optical Neural Networks with Differential Detection-Enabled Optical Operands," *IEEE/ACM Proceedings Design, Automation and Test in Europe (DATE)*, Feb. 2021. (Acceptance Rate: 24%)

[C14] Chenghao Feng, **Jiaqi Gu**, Zhoufeng Ying, Zheng Zhao, Ray T. Chen, and David Z. Pan, "Scalable fast-Fourier-transform-based (FFT-based) integrated optical neural network for compact and energy-efficient deep learning," *SPIE Photonics West*, Mar. 2021.

[C13] Chenghao Feng, Zhoufeng Ying, Zheng Zhao, **Jiaqi Gu**, Ray T. Chen, and David Z. Pan, "Wavelength-division-multiplexing (WDM)-based integrated electronic–photonic switching network (EPSN) for high-speed data processing and transportation," *SPIE Photonics West*, Mar. 2021.

[C12] **Jiaqi Gu**, Zixuan Jiang, and David Z. Pan, "DREAMPlace 3.0: Multi-Electrostatics Based Robustness VLSI Placement with Region Constraints," *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, Nov. 2020. (Acceptance Rate: 27%)

[C11] Zixuan Jiang, Keren Zhu, Mingjie Liu, **Jiaqi Gu**, and David Z. Pan, "An Efficient Training Framework for Reversible Neural Architectures," *European Conference on Computer Vision (ECCV)*, Aug. 2020. (Acceptance Rate: 26%)

[C10] **Jiaqi Gu**, Zheng Zhao, Chenghao Feng, Wuxi Li, Ray T. Chen, and David Z. Pan, "FLOPS: Efficient On-Chip Learning for Optical Neural Networks Through Stochastic Zeroth-Order Optimization," *ACM/IEEE Design Automation Conference (DAC)*, Jul. 2020. (**Best Paper Finalists**) (Acceptance Rate: 23.2%)

[C9] Mario Miscuglio, Zibo Hu, Shurui Li, **Jiaqi Gu**, Aydin Babakhani, Puneet Gupta, Chee-Wei Wong, David Pan, Seth Bank, Hamed Dalir, and Volker J. Sorger, "Massive parallelism Fourier-optic convolutional processor," *Signal Processing in Photonic Communications (SPPCom)*, Jul. 2020.

[C8] Mario Miscuglio, Zibo Hu, Shurui Li, **Jiaqi Gu**, Aydin Babakhani, Puneet Gupta, Chee-Wei Wong, David Z. Pan, Seth Bank, Hamed Dalir, and Volker J. Sorger, "Million-channel parallelism Fourier-optic convolutional filter and neural network processor," *Conference on Lasers and Electro-Optics*, May 2020.

[C7] Chenghao Feng, Zhoufeng Ying, Zheng Zhao, **Jiaqi Gu**, Ray T. Chen, and David Z. Pan, "Integrated WDM-based Optical Comparator for High-speed Computing," *Conference on Lasers and Electro-Optics*, May 2020.

[C6] Chenghao Feng, Zheng Zhao, Zhoufeng Ying, **Jiaqi Gu**, David Z. Pan, and Ray T. Chen, "Compact design of On-chip Elman Optical Recurrent Neural Network," *Conference on Lasers and Electro-Optics*, May 2020.

[C5] **Jiaqi Gu**, Zheng Zhao, Chenghao Feng, Hanqing Zhu, Ray T. Chen, and David Z. Pan, "ROQ: A Noise-Aware Quantization Scheme Towards Robust Optical Neural Networks with Low-bit Controls," *IEEE/ACM Proceedings Design, Automation and Test in Europe (DATE)*, Mar. 2020. (Acceptance Rate: 24.5%)

[C4] Mingjie Liu, Keren Zhu, **Jiaqi Gu**, Linxiao Shen, Xiyuan Tang, Nan Sun, and David Z. Pan, "Towards Decrypting the Art of Analog Layout: Placement Quality Prediction via Transfer Learning," *IEEE/ACM Proceedings Design, Automation and Test in Europe (DATE)*, Mar. 2020. (Acceptance Rate: 24.5%)

[C3] Chenghao Feng, Zhoufeng Ying, Zheng Zhao, **Jiaqi Gu**, Ray T. Chen, and David Z. Pan, "Wavelength-division-multiplexing-based electronic-photonic network for high-speed computing," *SPIE, Smart Photonic and Optoelectronic Integrated Circuits XXII*, Feb. 2020.

[C2] **Jiaqi Gu**, Zheng Zhao, Chenghao Feng, Mingjie Liu, Ray T. Chen, and David Z. Pan, "Towards Area-Efficient Optical Neural Networks: An FFT-based Architecture," *IEEE/ACM Asia and South Pacific Design Automation Conference (ASPDAC)*, Jan. 2020. (**Best Paper Award**)

[C1] Zheng Zhao, **Jiaqi Gu**, Zhoufeng Ying, Chenghao Feng, Ray T. Chen, and David Z. Pan, "Design Technology for Scalable and Robust Photonic Integrated Circuits," *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2019. (**Invited Paper**)

REFERENCES

David Z. Pan, Ph.D., ACM, IEEE and SPIE Fellow (Advisor)
Silicon Labs Endowed Chair Professor
Department of Electrical and Computer Engineering
The University of Texas at Austin
+1 (512) 471-1436
EER 4.880, 2501 Speedway, Austin, TX, 78712
dpan@ece.utexas.edu

Ray T. Chen, Ph.D., IEEE, OSA and SPIE Fellow (Co-Advisor)
Keys and Joan Curry/Cullen Trust Endowed Chair Professor
Department of Electrical and Computer Engineering
The University of Texas at Austin
+1 (512) 471-7035
EER 3.808, 2501 Speedway, Austin, TX, 78712
chen@ece.utexas.edu

Duane S. Boning, Ph.D., IEEE Fellow
Clarence J. LeBel Professor & MTL Associate Director & Engineering Faculty Co-Director of the MIT Leaders for Global Operations (LGO) program
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
+1 (617) 253-0931
60 Vassar Street, Room 39-415A, Cambridge, MA 02139
boning@mtl.mit.edu

Song Han, Ph.D.
Associate Professor
Department of Electrical and Computer Engineering
Massachusetts Institute of Technology
+1 (707) 797-7288
50 Vassar Street, Room 38-344, Cambridge, MA, 02139
songhan@mit.edu

Brucek Khailany, Ph.D.
Senior Director
Nvidia Research
2788 San Tomas Expy, Santa Clara, CA, 95051
bkhailany@nvidia.com

Jiaqi Gu
The University of Texas at Austin
2501 Speedway, Austin, TX, 78712, USA
+1 (512)-264-5470
jqgu@utexas.edu
jqgu.net

December 6, 2022

Dear Faculty Search Committee,

I am applying for the tenure-track Assistant Professor position in the **Department of Electrical Engineering and Computer Science at The University of Tennessee - Knoxville**. I am currently a Ph.D. Candidate in the Department of Electrical and Computer Engineering at The University of Texas at Austin, supervised by Professor David Z. Pan and co-advised by Professor Ray T. Chen, and I will complete my dissertation by May 2022.

My research interests are **emerging post-Moore AI hardware, energy-efficient computing, and hardware-software co-design**. With my expertise in AI hardware design, electronic/photonic design automation, AI/ML algorithm, and optimization, my Ph.D. work focuses on building **next-generation AI accelerators with electronics and integrated neuromorphic photonics** toward revolutionary speed, energy efficiency, and adaptability. My approach is to build emerging AI hardware with customized circuit designs, perform AI-enhanced reliability- and efficiency-driven algorithm-hardware co-optimization, and experimentally demonstrate high-speed, low-power AI/ML inference/training on our electronic-photonic neural chip tapeout. My research has led to 16 first authored and more than 30 co-authored publications in premier conferences and journals, e.g., CAD (DAC, ICCAD, IEEE TCAD), AI/ML (NeurIPS, ICCV, CVPR, AAAI), and photonics (Nature Comm., Laser Photonics Reviews, ACS Photonics), and received Best Paper Awards at ACM/IEEE ASP-DAC 2020, IEEE TCAD 2021, Best Poster Award at NSF Machine Learning Hardware Workshop 2020, Best Paper Finalist at IEEE/ACM DAC 2020, First Place at ACM Student Research Competition Grand Finals 2021, and Winner at Synopsys Optical Design Competition 2022. I seek deep interdisciplinary collaborations in my research, especially with experts in photonics, device/circuits, systems, and AI/ML. My research direction will complement and extend UTK's impact in the area of **efficient post-Moore's law computing and electronic-photonic heterogeneous AI accelerator design**.

I look forward to an opportunity to continue my research efforts at a leading research institute via high-quality interdisciplinary collaborative activities. I also greatly enjoy teaching, served as a teaching assistant in VLSI physical design automation, and have experience in mentoring students doing research. In the next few years, I hope to develop introductory and advanced courses in computer/electrical engineering. I will commit to fostering a diverse and inclusive community by teaching and mentoring students with diverse backgrounds and interests to facilitate their career excellence. I have enclosed all requested application materials. Thank you for your consideration, and I look forward to hearing from you!

Yours sincerely,

Jiaqi Gu

Teaching Statement

Jiaqi Gu (jqgu@utexas.edu)

Teaching and advising are two of the most important activities of academia. I am enthusiastic about being a faculty member because it holds out the promise of engaging students in these activities, which is also essential to my academic career.

1 Teaching Experience

I have served as a Teaching Assistant (TA) for a graduate-level course, ECE 382M VLSI Physical Design Automation, at the University of Texas at Austin. This course covers advanced VLSI physical design automation topics and features a mid-term exam, bi-weekly problem sets, coding assignments, and a final research-oriented project. I updated course slides with new content, designed new questions for the mid-term exam, graded all assignments/exams, and held TA office hours to help students review class contents and answer questions regarding past assignments.

I helped develop the course by creating a new teaching mode, coding exercises, for this course, which requires students to design data structures and implement core electronics design automation (EDA) kernels with resource and runtime constraints, e.g., partitioning for floorplan and pathfinding for routing. I developed the open-source codebase from scratch that is tailored to the average coding capability of students and designed automatic testing and grading system. I held tutorial sessions to give a general introduction to the benchmark, codebase, and detailed requirements. This provides students with a great opportunity, complimentary to the in-class high-level introduction, for a deep understanding of the physical design algorithms and, most importantly, hands-on programming exercises.

For the research-oriented final project, many students find it challenging to develop novel ideas or implement new algorithms because they had limited research and coding experience at the beginning of the class, which I believe are critical skill sets for both academic and industry careers. I scheduled 1:1 meetings with each group to recommend reading resources of the literature of their selected topic and coding examples. I also guided brainstorming and encouraged students to express their thoughts, construct innovative ideas, and finally generate practical plans with a concrete goal and a clear division of responsibilities according to their capabilities. I also helped organize the final project presentation and gave feedback to the students.

While at the end of the semester, they gained EDA knowledge, practiced coding skills, and produced decent results in the final projects. I received an average student rating of 4.9/5 and enjoyed mentoring and interacting with students.

2 Teaching Philosophy and Approach

My teaching philosophy holds three cores.

First, lectures work best when they show strong intuition and motivation. Strong intuition and motivation mean the lecture itself tells a consistent story from the origin of the field and conveys a close connection with practical applications in broader disciplines. A good motivation for a certain topic, especially for highly-abstract theory/principles, can largely raise the students' curiosity, let them know why they should spend time learning this, and help them understand the content from multiple angles given the diverse background of students. Literature surveys, group conversations, and open-ended discussions are great approaches suitable for both undergraduate-level and graduate-level students.

Second, I believe the integration of theory and practice makes good teaching, especially for topics at the cross-section of hardware and software. It is critical to help students build the capability of problem formulation. The lectures can never cover all prior solutions to all problems, especially in the rapidly-advancing fields, e.g., efficient AI and emerging AI hardware. Instead, I plan to give students a systematic frame that pinpoints key areas and challenges and help them lay solid foundations by teaching foundational principles and classical approaches with milestone progress. Besides theoretical formulation, it will be more effective to learn through exercises to gain hands-on experience by solving real-world problems. Through programming, physical implementation, experiments, and comprehensive study, students can thoroughly understand the principles, working mechanisms, and their practical value. Above all, the key is to teach and inspire students to find critical problems and innovative solutions.

Effective approaches include laying a solid foundation by teaching students principles with detailed examples and encouraging collective creativity with paper reading, sharing, and group class projects. For lower-level undergraduate classes, labs and exercises are great modes to link theoretical knowledge to practical applications. For graduate-level classes, research-oriented projects, labs, programming assignments, and in-class contest can stimulate their potential and creativity that benefit their future career.

Third, I believe in teaching students in accordance with their aptitude and background. Due to individual interest diversity and education resource imbalance in different countries and regions around the globe, students' interest, expertise, and career plans can vary significantly. It is important to construct a teaching methodology that is customized to students' backgrounds to accommodate the pace of different learners. For example, I will make the course self-contained to allow beginners to quickly ramp up with pre-requisite basic knowledge. I will also provide optional problem sets, reading resources, and bonus questions in exams for both advanced and less-experienced students. Another effective approach is to assign students with complementary aptitude to the same group for homework and projects and encourage them to have more communication and learn from each other.

3 Teaching Interests and Plans

My research, teaching, and work experience cover circuits, architecture, and algorithms. I am qualified to and would readily commit the effort to teach both undergraduate-level and graduate-level courses, including digital logic design, machine learning, data structure and algorithms, VLSI, parallel computing, computer architecture, emerging AI hardware, domain-specific accelerator, VLSI CAD/physical design, and applied AI/ML for co-design. I will develop courses to convert advanced cutting-edge topics that my research focuses on, e.g., emerging computing hardware, photonic computing, cross-layer co-design, efficient AI/ML, to introductory lectures for students at all levels to learn foundational knowledge together with real-world use cases.

4 Advising and Mentoring

I am excited to advise undergraduate and graduate students and help them grow into innovative, thoughtful, and productive researchers. I have helped mentor one completed Master thesis and mentored 3 graduates on their research. These mentoring and learning experiences are mutual. I enjoyed helping students develop their research skills and taste. I also found it rewarding when I learned new things from them and deepened my understanding in a more systematic and organized way through conversation and discussion. For example, I have helped Yunfei Sun, an ECE Master student, to develop his Master thesis, "Design and optimization of multi-operand ring resonator based efficient optical recurrent neural network." He extended one of our photonic computing engine designs through this thesis into an optical recurrent neural network. He learned how to utilize integrated photonic devices to construct electronic-photonic computing engines, extend my open-source codebase for model training, and perform optical simulation for functional validation. The following quote is from the Acknowledgments section of Yunfei's Master thesis: "I would also like to appreciate Jiaqi Gu... for their help and useful suggestions when I face difficulties in the research. I benefit a lot from the discussions..."

I currently mentor a junior Ph.D. student Hanqing Zhu in our group on his research. I have worked with him individually with weekly brainstorming to guide him into the emerging AI hardware field, help him pick topics aligned with his interests, create research blueprints for his dissertation, and provide professional training on programming, paper writing, and presentation skills. He worked closely with me on photonic AI computing and hardware-software co-design. I am proud to see his rapid growth, great productivity, independent investigation ability after his first two years, and, most importantly, continuous momentum. I also mentored a female junior graduate student Zhili Xiong and a junior Ph.D. student Souradip Poddar to help them start off the research journey. Through 1:1 meetings, I encouraged them to explore foundational and practical topics in the field of VLSI design automation and shared my vision and advice with them. I was pleased to see they all found their focused areas in advanced VLSI placement for FPGA/ASIC and quickly ramped up to investigate and innovate.

The path to great research goes alongside great students. As a professor, I will recruit students excited about our research directions, share with them my knowledge, skills, and vision, focus my energy on their professional growth, and work with them to advance our shared research agenda.

Research Statement

Jiaqi Gu (jqgu@utexas.edu)

In the post-Moore era, conventional computing solutions of digital electronics have become a limiting factor in certain domains, most notably intelligent information processing. The proliferation of big data and artificial intelligence (AI) has motivated the investigation of *next-generation AI computing hardware* to support low-power, low-latency machine intelligence. AI computing platforms based on *emerging hardware* and *heterogeneous integration* can make transformative impacts in future datacenters, automotive, military applications, smart sensing, and intelligent edge, enabling foundational breakthroughs in real-time perception, control, decision-making, and learning. My research aims to **synergistically design emerging AI hardware and algorithms towards revolutionary speed, efficiency, and adaptability**.

Uniqueness: My work stands out from other research in efficient AI or co-design on three points: 1) my research focuses on *AI hardware with heterogeneous more-than-Moore technologies*, especially integrated photonics and non-CMOS devices; 2) my work *integrates theoretical innovation with experimental demonstration*. Our co-designed platforms are prototyped at leading semiconductor vendors and evaluated on practical AI tasks; 3) my work *explores the synergy among electronics, photonics, AI, and optimization* toward a virtuous cycle of hardware and software co-design for future heterogeneous computing platforms.

Summary of outcomes: My research is supported in part by *AFOSR* and *ONR*, and I led and participated in our collaborations with various academic institutions, e.g., *MIT*, *UChicago*, *Yale*, *GWU*, and *UCLA*. I have interned at *Meta* and *Nvidia* to materialize our proposed efficient machine learning (ML) methods for computer vision and language processing on virtual reality and future datacenter accelerators. My research deliverables lead to **16 first-authored publications** [13–28] and **24 co-authored publications** [1–3, 7–12, 29, 30, 32–34, 37–46] in premier CAD/ML/Arch/SPIE/Nature journals and conferences (Nature Communications, Laser & Photonics Review, ACS Photonics, TCAD, DAC, ICCAD, DATE, NeurIPS, CVPR, ICCV, ECCV, AAAI, HPCA, etc.). My research accomplishments on emerging AI hardware design and cross-layer co-optimization have been recognized by academia and industry, and received the **Best Paper Award** at ASP-DAC 2020, **Best Paper Finalists** at DAC 2020, **Best Poster Award** at NSF Machine Learning Hardware Workshop 2020, **First Place** in ACM Student Research Competition Grand Finals 2021, **Best Paper Award** at IEEE TCAD 2022, **Winner** in Synopsys Robert S. Hilbert Memorial Optical Design Competition 2022 and other Best Paper Nominations. I released an open-source photonic AI library TorchONN that implements optical neural networks with cross-layer co-optimization supports and received 164 stars on Github.

1 Past Research: Electronic-Photonic AI Platform & HW/SW Co-Design

My major research goals are summarized as two thrusts in Fig. 1. It is centered in **Efficient Computing** and expands to two synergistic branches: **Thrust 1: Design for Emerging AI Computing Platform** and **Thrust 2: Optimization for Emerging AI Hardware**. The virtuous cycle between algorithms and emerging hardware will push the limit of AI hardware performance and efficiency.

1.1 Thrust 1: Design for Emerging AI Computing Platform

Mixed-signal computing on heterogeneous hardware platforms is a disruptive technology that can bring orders-of-magnitude performance and efficiency improvement to important use domains. Integrated photonics has provided a complementary opportunity to extend electronic computing solutions, especially in the field of intelligent information processing, scientific computing, and combinatorial optimization, due to its sub-nanosecond latency and sub-fJ/MAC energy efficiency. However, the packaging density and reliability of photonic integrated

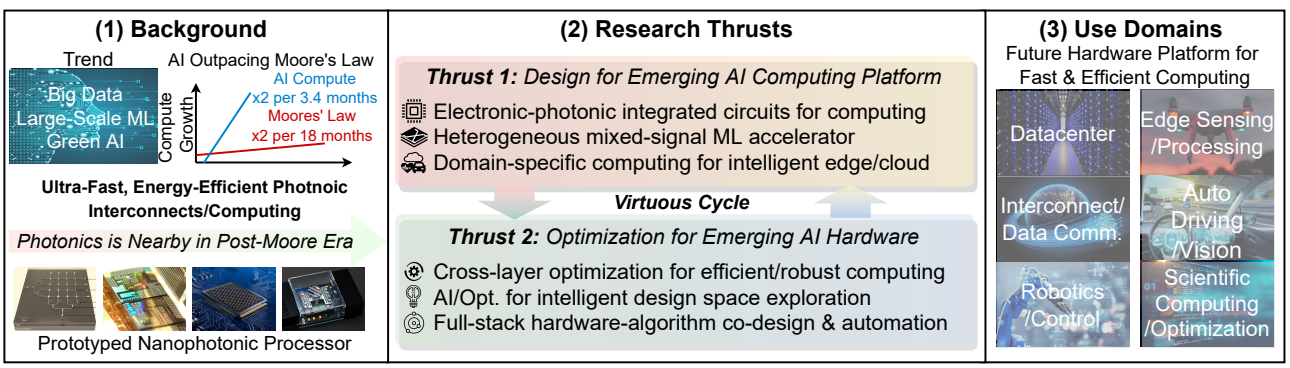


Figure 1: Overview of background, blueprint (two thrusts), and important use domains of my research.

circuits (PICs) often raise concerns due to the large spatial footprint of optical devices, limited computing precision, and noise robustness issues. My past research focuses on scalable, reliable, and adaptive electronic-photonic AI accelerator hardware with novel designs of computing units and circuit-model co-optimization techniques.

Hardware-Efficient Electronic-Photonic Neural Accelerator [1, 5, 22, 24]. Integrated photonic processors [35] have been demonstrated to accelerate general matrix multiplication, targeting a photonic substitution of GPUs/TPUs. However, the large spatial footprint of photonic circuits is the bottleneck for further scaling. Besides the continuous miniaturization from device shrinking, we propose to push the limit of scalability via **circuit compression**. To avoid using quadratically many optical devices, we design a compact photonic neural engine with a butterfly-style circuit topology that significantly **cuts down the optical device usage and realizes similar functionality**. Our team **taped out a programmable electronic-photonic co-packaged neural chip** at Advanced Micro Foundry. Our chip implements ResNet-20 and reliably achieves >85% accuracy on the CIFAR-10 image recognition dataset requiring only 3-bit voltage control precision. A single 4×4 photonic tensor core can achieve 225 TOPS/mm^2 compute density and 9.5 TOPS/W energy efficiency, which is orders-of-magnitude more powerful than modern GPUs and 2-3× more compact than the SoTA photonic tensor core, shown in Fig. 2. This compact photonic neural chip design and silicon prototype received **Best Paper Award** at ACM/IEEE ASP-DAC 2020 and **won the Synopsys Optical Design Competition 2022**.

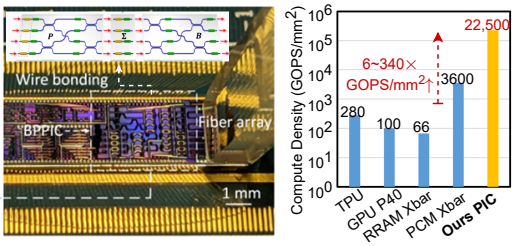


Figure 2: Our butterfly electronic-photonic co-packaged neural chip.

Ultra-Compact Electronic-Photonic Tensor Unit with Built-in Nonlinearity [4, 14, 16]. The compute density of conventional ML accelerators is typically upper-bounded by 1 multiply-accumulate operation (MAC) per device. Moreover, the system performance is often limited by the separate nonlinear activation circuitry. To break through this long-lasting performance bottleneck, we propose to **fuse tensor operations and nonlinearity in a single device**. For the first time, we squeeze an 8×8 matrix multiplication into a single 10×10 μm^2 multi-operand microring resonator (MORR) based on a novel usage of the underlying physics [14, 16]. Besides, the transmission of the device naturally supports **built-in reconfigurable nonlinearity**. Compared to previous photonic tensor cores based on standard microring (MRR) arrays [31, 36], we can realize comparable ML task performance with **quadratically fewer devices**, 8× fewer wavelengths, 5.3× higher compute density, 9.8× higher energy efficiency, and a 63.5% reduction in the simulated system energy consumption. Our team **taped out this MORR-based photonic neuron using AIM Photonics foundry**, shown in Fig. 3. This new design methodology implies an exciting research direction of *neuromorphic computing using nonlinear physical systems*, which shows great potential to push the compute density and efficiency to the extreme.

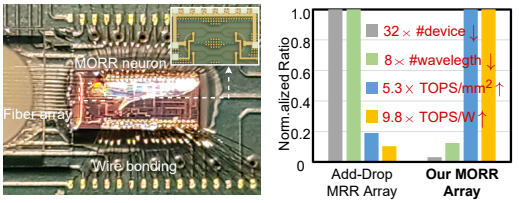


Figure 3: Our MORR-based single-device photonic tensor unit tape-out.

Heterogeneous Electronic-Photonic Mixed-Signal Accelerator Design [46]. Analog-to-digital (A/D) conversion and nonlinear activations gradually become the system performance bottleneck for emerging analog AI accelerators. I believe an electro-photonic hybrid system is the key to resolving this costly cross-domain signal conversion and nonlinear activation bottleneck. We propose a **heterogeneous electronic-photonic accelerator** [46] that adopts photonic engines for linear operations and electrical analog content-addressable memory (ACAM) to achieve **simultaneous A/D conversion and nonlinear activation in the analog domain**. Interestingly, we notice the analogy between ADCs and range-based lookup tables that essentially map the analog voltage signals to the corresponding digital levels with built-in nonlinearity. Therefore, we explore a much more efficient approach by fusing the functionality of ADCs and activation units into the discrete transmission of ACAM. Collaborated with device experts, we adopt magnetic tunnel junction (MTJ) devices to construct non-volatile ACAM cells, which can be orders-of-magnitude faster (ps-level) and more efficient (fJ-level) than traditional ADCs. To overcome the limited resolution and non-ideal variation of ACAM units, we extend the implication of *heterogeneity* by using **mixed dataflows of traditional ADCs and ACAM units** in the same architecture to balance the speed, energy efficiency, and computation fidelity. We automate the design process of this heterogeneous architecture, and our optimized system **saves 60% energy consumption** with marginal accuracy degradation compared to conventional ADC-based computing solutions.

1.2 Thrust 2: Optimization for Emerging AI Hardware

As the scale and heterogeneity of modern AI hardware platforms keeps growing, the design complexity and optimization difficulty have exponentially increased. We are encountering significant challenges in the **reliability, efficiency, and adaptability**. Focusing on those critical issues, my second research thrust aims

to pinpoint, formulate, and solve those three bottlenecks with **cross-layer circuit-architecture-algorithm co-optimization**. Specifically, I focus on (1) reliability boost via model-circuit co-optimization, (2) efficiency boost via intelligent design space exploration, (3) adaptability boost via on-chip self-learning.

Reliability Boost via Model-Circuit Co-Optimization. Functionality correctness is the first priority for emerging computing platforms. Non-ideal effects and limited computation precision often diminish the reliability and fidelity of analog computing platforms, most notably accuracy degradation or even malfunction in AI workloads. To resolve this reliability issue, the gap needs to be closed between theoretical simulation and physical deployment. I first analyze and understand the behavior of the analog computing engine and apply various customized solutions to the photonic AI hardware with co-optimization.

- **Device Quantization** [6, 23, 25, 43, 46]: We analyze the unique phase sensitivity of the photonic mesh and are *the first* that proposes to use customized training method to handle the limited precision in the photonic circuit inputs/outputs and device control signals. Our algorithm allows the gradients to propagate through the photonic circuits to the device configuration such that the training procedure is fully aware of the device quantization error. Our optimized photonic neural engine can tolerate such resolution limits and recover the inference accuracy comparable to its full-precision version.
- **Variation-Adaptive Training** [6, 14, 23, 25, 28]: We are the *first* to apply variation-adaptive training to boost the noise tolerance of the various photonic AI hardware designs. We inject the dynamic noise, static manufacturing variation, and thermal crosstalk in the optimization stage to mimic the physical hardware behavior. Besides, we explicitly apply protective regularization terms in the optimization objective based on analytical modeling to surpass the noise and crosstalk impacts. Our noise-aware training techniques help the system resume from low-fidelity or malfunction to achieve comparable accuracy to digital computers.
- **Circuit Sparsification** [6, 14, 24]: We are *the first* to sparsify the photonic circuit via optimization-based device pruning. Different from NN weight pruning, we consider the unique weight-to-device mapping and physical layout and structurally remove a proportion of devices while maintaining a similar functionality. We can reduce the circuit depth and noise sources to improve the robustness of the photonic neural system. When we map neural networks onto the photonic engine, our sparsified circuit shows superior noise resilience compared to the original unpruned counterpart.
- **Aging-Aware Optimization** [44, 45]: In-memory computing is a promising paradigm to resolve the data movement bottleneck. However, endurance, aging issues, and reprogramming cost are critical concerns for in-memory computing platforms based on non-volatile devices. We propose an optimization framework that encourages weight sharing and reorders the hardware mapping to minimize the redundant rewrites on non-volatile photonic phase-change material (PCM) memory cells. We can boost the endurance and energy efficiency of the photonic in-memory computing AI engine by 10×.

Efficiency Boost via Fast & Automated Design Space Exploration. Both the *hardware efficiency* and *design efficiency* matter in the lifecycle of AI computing platforms. Intelligent search-space exploration opens new opportunities to enable a faster design closure for more efficient hardware designs. My research focuses on two critical steps towards this goal: (1) fast simulation for performance evaluation and (2) automated search space exploration for efficient AI hardware designs.

- **ML-Enabled Ultra-fast Optical Simulation** [17]: Efficient design space exploration requires fast and accurate performance evaluation. As a key step in the evaluation, optical simulation is an important kernel that will be frequently queried. The time-consuming finite-difference Maxwell equation solving in the simulation becomes the bottleneck of scaling up photonic IC design. We propose to apply *AI for Physics* to the photonic IC design flow by using **GPU-accelerated ML methods to solve partial differential equations (PDEs)**. We introduce a data-driven framework `NeurOLight` [17] that learns the light propagation principles from simulation examples and ultimately can solve a family of parametric Maxwell equations. Our **differentiable** framework can perform **real-time** (million-second runtime, 120 FPS) parallel PDE solving, over 200× faster than multi-CPU numerical solvers, allow gradient-based inverse design, and is able to **generalize** to a wide range of simulation instances. This opens a wide research opportunity, including foundational AI for physics, circuit-level simulation acceleration, and differentiable simulator-in-the-loop optimization, to achieve orders-of-magnitude productivity boost in photonic IC design.
- **Automated Photonic Integrated Circuit Design** [27]: For decades, photonic IC mainly depends on hand-crafted designs. Such labor-intensive design flows are not scalable to handle the increasing scale and design complexity in heterogeneous integrated systems, leave a large design space unexplored, and lack adaptability to different design targets and constraints. We target an automated flow that directly generates the circuit design given the user specification, and efficiently explores the exponential design space of photonic circuits to push further the performance Pareto frontier. We propose a framework `ADEPT` [27] for **auto-circuit design**, which is **the first circuit-level design automation algorithm for photonic AI hardware**. Our

framework can easily adapt to device specifications from foundry PDKs and honor various chip design constraints, e.g., footprint, power, and latency. The searched circuits show 2-30× smaller area and much higher noise resilience than prior manual designs. Our team is working on the tape-out of the first auto-designed photonic neural chip using the AMF foundry for experimental demonstration. This work was **nominated from the track for the Best Paper Candidate** at DAC 2022.

- **Hardware-Efficient Model-Architecture Co-Optimization** [19, 20, 28]: Storage and data movement usually dominate the hardware cost of emerging AI accelerators. My solution is using model compression algorithms, e.g., quantization, tensor decomposition, to **trade redundant expressivity and high-cost data movement for low-cost computation**. We customize the architecture to support a special dataflow, such that the compressed operations can be **fused in the local processing units [19, 28] and computed on-the-fly**, even directly in the analog domain [28]. Our methods can significantly improve the efficiency of CNNs and attention-based Transformer models on advanced vision and NLP workloads, achieving over **100**× compression and over **5**× energy-delay product reduction. The proposed co-optimization methods have been adopted in the on-device vision inference in Meta reality lab and future datacenter language model accelerator in Nvidia research.

Adaptability Boost via On-Device Self-Learning. Besides inference acceleration, future AI systems, especially the intelligent edge, require on-device self-learnability. A self-learnable computing system can (1) address the robustness issues *in situ* and closes the gap between simulation and deployment of analog AI accelerators; (2) help with data privacy; (3) allow online learning and real-time adaptation on the edge with minimum communication cost; and (4) significantly reduce the training energy consumption. My research aims to address an extremely challenging task: **efficient in-situ training on an electronic-photonic AI accelerator**. We propose a series of on-chip training protocols [13, 21, 26] to enable self-learnable photonic AI chips with unprecedented training efficiency. Our hybrid framework integrates zeroth-order and first-order methods to overcome the limited controllability and observability of photonic integrated circuits for *in-situ* optimization. We also introduce mixed-training techniques with multi-level sparsity to reduce the training cost by approximating the gradients via matrix sampling and updating a tiny subnet of devices. We achieve **1,000**× improvement in training scalability to handle million-parameter photonic NNs and **30**× efficiency boost compared to prior protocols, enabling efficient self-calibration, online/lifelong learning, and edge training applications. Our work was selected as the **Best Paper Finalists** at DAC 2020 and won the **Best Poster Awards** at NSF Workshop on ML Hardware 2021.

2 Future Plans: Full-Stack Co-Design for Heterogeneous Computing Platform

The increasing requirement for computational speed and efficiency in emerging applications calls for continuous advancement in hardware design and optimization methodology. The emerging hardware design and software stack form a virtuous cycle with strong connection and mutual reinforcement. I will endeavor to push forward next-generation efficient computing through **intelligent co-design & automation** and **emerging hardware platform design**. I will leverage my strong background in hardware-algorithm co-design, design automation, AI/ML algorithms, and optimization to explore the efficiency-accuracy-robustness tradeoff in emerging computing platforms. As an interdisciplinary researcher, I have experience in **collaboration with the semiconductor industry and academic researchers in computer science, computer engineering, and circuit/device/material**. With joint efforts and domain knowledge, we will keep pushing the limit and lead the research frontier of next-generation computing.

2.1 Intelligent Co-Design and Design Automation for Emerging Hardware Platforms

In the post-Moore heterogeneous computing platforms, design complexities become extremely high. More intelligent hardware-software co-design technologies are needed more than ever to optimize performance and efficiency. As my **near-term and mid-term** plans, I intend to place great emphasis on the following aspects,

- **End-to-end software-to-hardware design automation infrastructure**: To standardize and streamline the design and development of electronic-photonic heterogeneous platforms, I will use my expertise in co-design and CAD and collaborate with other researchers to define and construct an end-to-end design-to-hardware infrastructure, including model-to-circuit mapping, hardware implementation with electronic-photonic design automation (EPDA), simulation, and performance evaluation. I will also develop intelligent compiling flows with algorithm-hardware co-optimization to map software applications to the hardware platform with high efficiency and robustness.

- **AI/ML for intelligent co-design technologies**: I plan to systematically explore AI/ML methods in the co-design flow towards unprecedented productivity and beyond-manual design quality, including automatic

circuit/architecture search given user spec., AI-accelerated simulation and performance evaluation, intelligent design space exploration, AI-guided hardware-in-the-loop optimization, and on-device learning protocols.

- **Full-stack support for emerging application deployment**: An exciting research avenue is to deploy our developed high-performance, low-power hardware platforms to support wide application domains, e.g., perception, control, and decision-making on autonomous driving, Internet of Things, smart cameras/UAV/VR, scientific computing, and combinatorial optimization. This application-oriented research requires domain-specific customization and will have a high impact in real-world use domains, which in turn helps accelerate the evolution and adoption of emerging computing hardware.

2.2 Electronic-Photonic Heterogeneous Computing Platform

Besides software stack design, my **mid-term and long-term** plans focus on emerging hardware platform development. Heterogeneous platforms with emerging technologies can represent a paradigm shift in future computing systems. I will research domain-specific computing hardware platforms with heterogeneous technologies. To be specific, I plan to work on:

- **Mixed-signal accelerator with heterogeneous integration**: The future computing platform will be 2.5D/3D mixed-signal ICs with heterogeneous technologies, e.g., CMOS, post-CMOS electronics, and integrated photonics. I will continue investigating such hardware platforms by leveraging advanced devices, manufacturing, and packaging. I will collaborate with hardware experts in academia and industry towards the system-level demonstration of 2.5D/3D *co-packaged heterogeneous platforms where laser, photo-detection, interconnects, computing engines, storage units, and electrical control logic are fully integrated.*
- **Near-data computing for intelligent edge**: I believe the future of intelligent edge will merge computation with data acquisition, storage, and networks with minimum cross-domain signal conversion. (1) Near-sensor computing with front-end processing and perception for intelligent sensing, e.g., integrate analog computing engines with optical/electronic sensors; (2) In-memory computing with intelligent storage units that support efficient in-place information processing, e.g., RRAM, MRAM, PCM; (3) In-network computing for intelligent interconnects and distributed processing, e.g., emerging computing engines inside interconnects and cross-node communication dataflows. I will collaborate with researchers in device, sensor, and network to explore the above directions to *resolve the bottleneck in cross-domain signal conversion (analog $\leftrightarrow$ digital, electrical $\leftrightarrow$ optical), data movement, and communication.*
- **Neuromorphic computing using physics**: Leveraging physics to compute is a promising trend to break through the compute density and efficiency limitation of the current hardware. I will extend my current work of *single-device tensor units* and demonstrate more use cases of utilizing the nonlinear response of physical systems for efficient neuromorphic computing. I will collaborate with researchers in device and physics to investigate such novel neuromorphic architectures with ultra-high compute density and efficiency.

The above topics present a flavor of the research I am inspired to work on both in the short term and long term. They all share the following theme at their core: software-hardware co-design solutions to build next-generation efficient computing platforms. I enjoy finding fundamental and critical problems, innovating for highly practical solutions, and contributing to the advancement of this field.

3 Potential Collaboration and Research Funding Plan

I believe my target research area is interdisciplinary by nature, integrating low-level devices, circuits, and high-level architecture and algorithm designs. Therefore, I seek close collaboration with researchers with backgrounds in integrated photonics, material/device, circuits, computer engineering, applied physics, or AI/ML. Besides, I will also connect with industry developers and foundries for chip tape-out, downstream applications, and integration of software stack and hardware platforms for highly impactful research. I have assisted my advisors' research funding in writing funding proposals, gaining experience in forming teams, composing white paper/proposal write-ups, and scheduling plans and budgets. I plan to apply for various government funding agencies, e.g., AFOSR, ONR, DARPA, NSF, and direct industrial funding from semiconductor companies, in the upcoming academic career.

4 Closing Remarks

Emerging hardware technologies, e.g., integrated photonics and non-CMOS devices, shed light on the future computing platforms in the post-Moore era. From algorithm and architecture to device, we are facing new opportunities and challenges across the full stack. My passion for hardware-software co-design, AI, and optimization drives me to investigate under-explored areas and make breakthrough contributions to the fields of emerging computing hardware and efficient AI.

References

[1] C. Feng, J. Gu, Z. Ying, Z. Zhao, R. T. Chen, and D. Z. Pan, "Scalable fast-Fourier-transform-based (FFT-based) integrated optical neural network for compact and energy-efficient deep learning," in *SPIE Photonics West*, Mar. 2021.

[2] C. Feng, J. Gu, H. Zhu, D. Z. Pan, and R. T. Chen, "Experimental Demonstration of a WDM-based Integrated Optical Decoder for Compact Optical Computing," in *Conference on Lasers and Electro-Optics*, May 2021.

[3] ——, "Design and Experimental Demonstration of A Hardware-Efficient Integrated Optical Neural Network," in *Conference on Lasers and Electro-Optics*, Mar. 2022. [Online]. Available: https://doi.org/10.1117/12.2610255

[4] C. Feng, J. Gu, H. Zhu, R. Tang, D. Z. Pan, and R. T. Chen, "Optically-interconnected, hardware-efficient, electronic-photonic neural network using compact multi-operand photonic devices," in *Optical Interconnects XXIII*, Jan. 2023.

[5] C. Feng*, J. Gu*, H. Zhu, Z. Ying, Z. Zhao, D. Z. Pan, and R. T. Chen, "A compact butterfly-style silicon photonic-electronic neural chip for hardware-efficient deep learning," *ACS Photonics*, 2021. [Online]. Available: https://arxiv.org/abs/2111.06705

[6] C. Feng, J. Gu, H. Zhu, Z. Ying, Z. Zhao, D. Z. Pan, and R. T. Chen, "A compact butterfly-style silicon photonic-electronic neural chip for hardware-efficient deep learning," *arXiv preprint arXiv:2111.06705*, 2021.

[7] ——, "Optoelectronically Interconnected Hardware-Efficient Deep Learning using Silicon Photonic Chips," in *Conference on Lasers and Electro-Optics*, Mar. 2022. [Online]. Available: https://doi.org/10.1117/12.2616217

[8] C. Feng, Z. Ying, Z. Zhao, J. Gu, R. T. Chen, and D. Z. Pan, "Integrated WDM-based Optical Comparator for High-speed Computing," in *Conference on Lasers and Electro-Optics*, May 2020.

[9] ——, "Wavelength-division-multiplexing-based electronic-photonic network for high-speed computing," in *SPIE, Smart Photonic and Optoelectronic Integrated Circuits XXII*, Feb. 2020.

[10] ——, "Wavelength-division-multiplexing (WDM)-based integrated electronic–photonic switching network (EPSN) for high-speed data processing and transportation," *Nanophotonics*, Aug. 2020.

[11] C. Feng, Z. Ying, Z. Zhao, J. Gu, D. Z. Pan, and R. T. Chen, "Towards high-speed and energy-efficient computing: A WDM-based scalable on-chip silicon integrated optical comparator," *Laser & Photonics Reviews*, Jun. 2021.

[12] C. Feng, Z. Zhao, Z. Ying, J. Gu, D. Z. Pan, and R. T. Chen, "Compact design of On-chip Elman Optical Recurrent Neural Network," in *Conference on Lasers and Electro-Optics*, May 2020.

[13] J. Gu, C. Feng, Z. Zhao, Z. Ying, R. T. Chen, and D. Z. Pan, "Efficient On-Chip Learning for Optical Neural Networks Through Power-Aware Sparse Zeroth-Order Optimization," in *Association for the Advancement of Artificial Intelligence (AAAI)*, Feb. 2021. [Online]. Available: https://arxiv.org/abs/2012.11148

[14] J. Gu, C. Feng, Z. Zhao, Z. Ying, M. Liu, R. T. Chen, and D. Z. Pan, "SqueezeLight: Towards Scalable Optical Neural Networks with Multi-Operand Ring Resonators," in *IEEE/ACM Proceedings Design, Automation and Test in Europe (DATE)*, Feb. 2021.

[15] J. Gu, C. Feng, H. Zhu, R. T. Chen, and D. Z. Pan, "Light in AI: Toward Efficient Neurocomputing with Optical Neural Networks - A Tutorial," *IEEE Transactions on Circuits and Systems–II: Express Briefs (TCAS-II)*, Apr. 2022.

[16] J. Gu, C. Feng, H. Zhu, Z. Zhao, Z. Ying, M. Liu, R. T. Chen, and D. Z. Pan, "SqueezeLight: A Multi-Operand Ring-Based Optical Neural Network with Cross-Layer Scalability," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, Jul. 2022.

[17] J. Gu, Z. Gao, C. Feng, H. Zhu, R. T. Chen, D. Boning, and D. Z. Pan, "NeurOLight: A Physics-Agnostic Neural Operator Enabling Parametric Photonic Device Simulation," in *Conference on Neural Information Processing Systems (NeurIPS)*, Dec. 2022.

[18] J. Gu, Z. Jiang, and D. Z. Pan, "DREAMPlace 3.0: Multi-Electrostatics Based Robustness VLSI Placement with Region Constraints," in *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, Nov. 2020.

[19] J. Gu, B. Keller, J. Kossaifi, A. Anandkumar, B. Khailany, and D. Z. Pan, "HEAT: Hardware-Efficient Automatic Tensor Decomposition for Transformer Compression," in *Conference on Neural Information Processing Systems (NeurIPS), ML for System Workshop (MLSys)*, Dec. 2022. [Online]. Available: https://arxiv.org/abs/2209.10098

[20] J. Gu, H. Kwon, D. Wang, W. Ye, M. Li, Y.-H. Chen, L. Lai, V. Chandra, and D. Z. Pan, "Multi-Scale High-Resolution Vision Transformer for Semantic Segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. [Online]. Available: https://arxiv.org/abs/2111.01236

[21] J. Gu, Z. Zhao, C. Feng, W. Li, R. T. Chen, and D. Z. Pan, "FLOPS: Efficient On-Chip Learning for Optical Neural Networks Through Stochastic Zeroth-Order Optimization," in *ACM/IEEE Design Automation Conference (DAC)*, Jul. 2020.

[22] J. Gu, Z. Zhao, C. Feng, M. Liu, R. T. Chen, and D. Z. Pan, "Towards Area-Efficient Optical Neural Networks: An FFT-based Architecture," in *IEEE/ACM Asia and South Pacific Design Automation Conference (ASPDAC)*, Jan. 2020.

[23] J. Gu, Z. Zhao, C. Feng, Z. Ying, R. T. Chen, and D. Z. Pan, "O2NN: Optical Neural Networks with Differential Detection-Enabled Optical Operands," in *IEEE/ACM Proceedings Design, Automation and Test in Europe (DATE)*, Feb. 2021.

[24] J. Gu, Z. Zhao, C. Feng, Z. Ying, M. Liu, R. T. Chen, and D. Z. Pan, "Towards Hardware-Efficient Optical Neural Networks: Beyond FFT Architecture via Joint Learnability," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, 2020.

[25] J. Gu, Z. Zhao, C. Feng, H. Zhu, R. T. Chen, and D. Z. Pan, "ROQ: A Noise-Aware Quantization Scheme Towards Robust Optical Neural Networks with Low-bit Controls," in *IEEE/ACM Proceedings Design, Automation and Test in Europe (DATE)*, Mar. 2020.

[26] J. Gu, H. Zhu, C. Feng, Z. Jiang, R. T. Chen, and D. Z. Pan, "L2ight: Enabling On-Chip Learning for Optical Neural Networks via Efficient in-situ Subspace Optimization," in *Conference on Neural Information Processing Systems (NeurIPS)*, Dec. 2021. [Online]. Available: https://arxiv.org/abs/2110.14807

[27] J. Gu, H. Zhu, C. Feng, Z. Jiang, M. Liu, S. Zhang, R. T. Chen, and D. Z. Pan, "ADEPT: Automatic Differentiable DEsign of Photonic Tensor Cores," in *ACM/IEEE Design Automation Conference (DAC)*, Jul. 2022. [Online]. Available: https://arxiv.org/abs/2112.08703

[28] J. Gu, H. Zhu, C. Feng, M. Liu, Z. Jiang, R. T. Chen, and D. Z. Pan, "Towards Memory-Efficient Neural Networks via Multi-Level in situ Generation," in *International Conference on Computer Vision (ICCV)*, Oct. 2021. [Online]. Available: https://arxiv.org/abs/2108.11430

[29] Z. Jiang, J. Gu, M. Liu, K. Zhu, and D. Z. Pan, "Optimizer Fusion: Efficient Training with Better Locality and Parallelism," in *International Conference on Learning Representations (ICLR) Workshop, Hardware Aware Efficient Training (HAET)*, May 2021. [Online]. Available: https://arxiv.org/abs/2104.00237

[30] Z. Jiang, K. Zhu, M. Liu, J. Gu, and D. Z. Pan, "An Efficient Training Framework for Reversible Neural Architectures," in *European Conference on Computer Vision (ECCV)*, Aug. 2020.

[31] W. Liu, W. Liu, Y. Ye, Q. Lou, Y. Xie, and L. Jiang, "Holylight: A nanophotonic accelerator for deep learning in data centers," in *IEEE/ACM Proceedings Design, Automation and Test in Europe (DATE)*, 2019.

[32] J. Midkiff, A. Rostamian, K. M. Yoo, A. Asghari, C. Wang, C. Feng, Z. Ying, J. Gu, H. Mei, C.-W. Chang, J. Fang, A. Huang, J.-D. Shin, X. Xu, M. Bukshtab, D. Z. Pan, and R. T. Chen, "Integrated Photonics for Computing, Interconnects and Sensing," in *Conference on Lasers and Electro-Optics*, May 2021. [Online]. Available: https://www.youtube.com/watch?v=HqR3YVC2CUI

[33] M. Miscuglio, Z. Hu, S. Li, J. Gu, A. Babakhani, P. Gupta, C.-W. Wong, D. Pan, S. Bank, H. Dalir, and V. J. Sorger, "Massive parallelism Fourier-optic convolutional processor," in *Signal Processing in Photonic Communications (SPPCom)*, Jul. 2020.

[34] M. Miscuglio, Z. Hu, S. Li, J. Gu, A. Babakhani, P. Gupta, C.-W. Wong, D. Z. Pan, S. Bank, H. Dalir, and V. J. Sorger, "Million-channel parallelism Fourier-optic convolutional filter and neural network processor," in *Conference on Lasers and Electro-Optics*, May 2020.

[35] Y. Shen, N. C. Harris, S. Skirlo *et al.*, "Deep learning with coherent nanophotonic circuits," *Nature Photonics*, 2017.

[36] A. N. Tait, T. F. de Lima, E. Zhou *et al.*, "Neuromorphic photonic networks using silicon photonic weight banks," *Sci. Rep.*, 2017.

[37] H. Wang, Y. Ding, J. Gu, Y. Lin, D. Z. Pan, F. T. Chong, and S. Han, "QuantumNAS: Noise-Adaptive Search for Robust Quantum Circuits," in *IEEE International Symposium on High Performance Computer Architecture (HPCA)*, Feb. 2022. [Online]. Available: https://arxiv.org/abs/2107.10845

[38] H. Wang, J. Gu, Y. Ding, Z. Li, F. T. Chong, D. Z. Pan, and S. Han, "QuantumNAT: Quantum Noise-Aware Training with Noise Injection, Quantization and Normalization," in *ACM/IEEE Design Automation Conference (DAC)*, Jul. 2022. [Online]. Available: https://arxiv.org/abs/2110.11331

[39] H. Wang, Z. Li, J. Gu, Y. Ding, D. Z. Pan, and S. Han, "QOC: Quantum On-Chip Training with Parameter Shift and Gradient Pruning," in *ACM/IEEE Design Automation Conference (DAC)*, Jul. 2022. [Online]. Available: https://arxiv.org/abs/2202.13239

[40] H. Wang, P. Liu, J. Cheng, Z. Liang, J. Gu, Z. Li, Y. Ding, W. Jiang, Y. Shi, X. Qian, D. Z. Pan, F. T. Chong, and S. Han, "QuEst: Graph Transformer for Quantum Circuit Reliability Estimation," in *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, Jul. 2022.

[41] Z. Ying, C. Feng, Z. Zhao, S. Dhar, H. Dalir, J. Gu, Y. Cheng, R. Soref, D. Z. Pan, and R. T. Chen, "Electronic-photonic Arithmetic Logic Unit for High-speed Computing," *Nature Communications*, Apr. 2020.

[42] Z. Ying, C. Feng, Z. Zhao, J. Gu, R. Soref, D. Z. Pan, and R. T. Chen, "Sequential logic and pipelining in chip-based electronic-photonic digital computing," *IEEE Photonics Journal*, Oct. 2020.

[43] Z. Zhao, J. Gu, Z. Ying, C. Feng, R. T. Chen, and D. Z. Pan, "Design Technology for Scalable and Robust Photonic Integrated Circuits," in *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2019.

[44] H. Zhu, J. Gu, C. Feng, M. Liu, Z. Jiang, R. T. Chen, and D. Z. Pan, "ELight: Enabling Efficient Photonic In-Memory Neurocomputing with Life Enhancement," in *IEEE/ACM Asia and South Pacific Design Automation Conference (ASPDAC)*, Jan. 2022.

[45] ——, "ELight: Towards Efficient and Aging-Resilient Photonic In-Memory Neurocomputing," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, Jun. 2022.

[46] H. Zhu, K. Zhu, J. Gu, H. Jin, R. T.Chen, J. A. Incorvia, and D. Z. Pan, "Fuse and Mix: MACAM-Enabled Analog Activation for Energy-Efficient Neural Acceleration," in *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, Jul. 2022.

The University of Texas at Austin
Chandra Department of Electrical and Computer Engineering
Cockrell School of Engineering

Engineering Education & Research Center
2501 Speedway
Austin, TX 78712
T: 512.471.6771
ece.utexas.edu

December 30, 2022

Re: Recommendation for Jiaqi Gu's faculty application

To Whom It May Concern:

It is my pleasure to give Jiaqi Gu my most enthusiastic support for a tenure-track assistant professor position at your esteemed institution. Jiaqi expects to get his PhD in May 2023 under my supervision at the ECE Department, UT Austin. As Jiaqi's research is very interdisciplinary, he is co-supervised by my colleague Prof. Ray Chen, a leading nanophotonics and optical interconnect expert, to build ultra-fast and energy-efficient optical AI accelerators with hardware and software co-design, and novel design/automation for emerging technologies.

Let me begin by stating my top line: Jiaqi is ***the top*** fresh PhD student who has ever graduated from my research group in the last 20 years. As a reference, I have graduated 39 PhD students and 5 post-docs at UT Austin so far. They have taken tenure-track faculty positions at major research universities such as Georgia Tech, Texas A&M, Chinese Univ. of Hong Kong, Peking Univ., Univ. of Nebraska-Lincoln, and key industry positions such as Nvidia Research, IBM Research, Google X, Meta Reality Lab, Apple, Cadence, Synopsys, etc. They have collectively won 19 best paper awards at premier venues and 5 Outstanding PhD Dissertation Awards from ACM/SIGDA and European Design and Automation Association (EDAA) – both are probably the most among all research groups in the world in the broad area of electronic design automation (EDA). Thus, it is truly exceptional for Jiaqi to excel in such an extraordinary cohort. I am extremely thankful and proud to have Jiaqi as my student.

Getting Jiaqi was pure luck, like running into a gem! Actually, when Jiaqi came to UT in Fall 2018, he first worked with another faculty. However, after one semester, he found his research interests aligned with mine better, so he gave up his original multi-year offer from that faculty and volunteered to work with me self-funded in Spring'19. Now I look back, even at that time, Jiaqi knew exactly what he wanted. He was so driven from the beginning. Since I have a collaborative project with Prof. Ray Chen on optical neural networks, I suggested Jiaqi to investigate this emerging area. It was quite a steep learning curve for Jiaqi, since he did not have much background on optical devices, and he knew only a bit on efficient machine learning. But this didn't stop Jiaqi from starting his amazingly innovative journey with me. He ramped up quickly, did exhaustive literature search, and built a solid foundation. Then he proposed a very novel optical neural network (ONN) leveraging fast Fourier transforms (FFT). Compared to the seminal work from MIT (Nature Photonics 2017) which uses bulky MZI devices, Jiaqi's FFT-ONN architecture directly makes use of primitive nanophotonic devices such as phase shifter, coupler, attenuator, combiner, etc., making his design much more compact. However, there is no free lunch, as his architecture can only perform circulant matrix multiplications (instead of general matrix multiplication). Yet, in the mainstream machine learning community, there's recent research showing that circulant matrices can have very good expressiveness too, but all these need good software and hardware co-design. Jiaqi then further developed novel software training and

pruning techniques to reduce FFT-ONN sizes. His experimental results demonstrate that the FFT-ONN architecture can achieve 10x improvement in terms of throughput and energy efficiency, and 2∼4x area improvement, compared to the MIT ONN work. Note that the reported optical inference accelerator from MIT (Nature Photonics 2017) was already orders of magnitude more energy efficient than the conventional electronic AI accelerators. Jiaqi's work, published at the IEEE/ACM Asian and South-Pacific Design Automation Conference (ASP-DAC, one of the top four EDA conferences) in January 2020, won the **Best Paper Award**. In its TCAD journal version, Jiaqi further extended FFT-ONN to general frequency-domain transforms and showed additional significant improvement. Not only that, Jiaqi and another student, Chenghao Feng, further put the FFT-ONN idea to the **real chip tape-out**. Since photonic IC design and testing are not yet well supported by commercial foundries, they have spent tremendous efforts in designing, packaging, testing, software co-design, and shown excellent silicon-proven results. The photonic IC tapeout work with new software training was published at *ACS Photonics*, a premier journal in photonics. It also won the 2022 **Robert S. Hilbert Memorial Optical Design Competition,** sponsored by Synopsys, the largest EDA company.

As mentioned earlier, ASP-DAC 2020 is just Jiaqi's first first-authored paper with me. It turned out that in 2020, Jiaqi published four first-authored papers, in all top-four EDA conferences, namely *Asian and South Pacific Design Automation Conference* (ASP-DAC), *Design and Test in Europe* (DATE), *Design Automation Conference* (DAC), and *International Conference on Computer Aided Design* (ICCAD). These are the most prestigious conferences in the broad area of electronic design automation (EDA), with as low as 20%-ish acceptance rate. It is not easy to publish in any of these conferences, but Jiaqi had a "home run" for all of them in 2020. More impressively, his ASP-DAC 2020 paper won **Best Paper Award** (one of two winners, out of around 300 submissions), and his DAC 2020 paper was a **Best Paper Finalist** (one of six finalists, out of nearly 1,000 paper submissions). In addition, Jiaqi is a key contributor to other collaborative papers, including *Nature Communications* (2020), *Nanophotonics* (2020), etc.

After DAC 2020 (July), Jiaqi told me that he was interested in participating in the ACM/SIGDA Student Research Competition (SRC), to be held at ICCAD in November 2020. SRC (graduate category) is mostly attended by senior PhD students close to graduation. It attracts many top PhD students in the SIGDA related areas to participate. Since Jiaqi was just a beginning third-year PhD student, I did not have very high expectations of him winning. But he surprised me by winning the **Gold Medal (First Place) at the SIGDA SRC.** He then represented SIGDA to participate in the ACM-wide SRC Grand Finals, competing with all Gold Medalists from other ACM SIGs in 2021. He wowed me again, by winning the **ACM Student Research Competition Grand Finals First Place**! This is extremely competitive as ACM has many special interest groups (SIGs), and each SIG holds its SRC at a flagship conference and selects the winners. Then the Gold Medalists from all SIGs compete at the ACM-wide SRC Grand Finals. So, **this first place is the champion among champions.** The award is usually presented at the annual ACM Awards Banquet, where the Turing Award and other major ACM Awards are presented.

Jiaqi also presented his work at the *2020 NSF Workshop for Machine Learning Hardware Breakthroughs Towards Green AI and Ubiquitous On-Device Intelligence* and won the **Best Poster Award**. This is a high-profile NSF workshop, attended by leaders of the field. It had 30 research posters in very broad areas related with hardware and machine learning. The **Best Poster Award (1 out of 30)** was based on both the audience vote and the organizing committee's evaluation. Jiaqi deeply impressed the audience and the organizing committee.

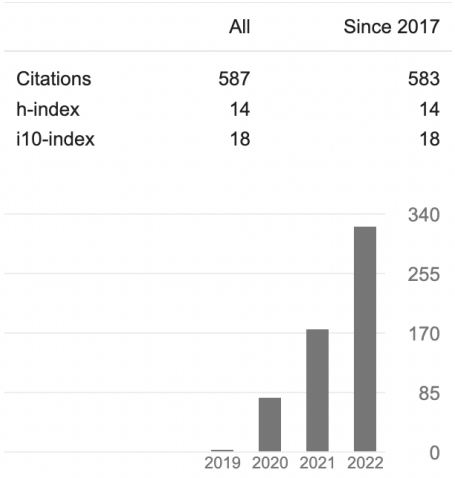
With this kind of record, Jiaqi could already get his PhD, but he is just passionate about expanding his research scope deeper and broader. And he is very clear about his long-term professional goal, i.e., to be a faculty member leading his own research group at a major research university. If you look at Jiaqi's publication record as of today, you will be amazed. He has published **16 first-authored and one co-first-authored papers in top conferences and journals in the last 3 years**:

- 2020: ASP-DAC, DATE, DAC, ICCAD, TCAD
- 2021: DATE x2, AAAI, ICCV, NeurIPS
- 2022: DAC, CVPR, NeurIPS (**Spotlight**), NeurIPS MLSys (**Spotlight**), ACS Photonics (co-first author), TCAS-II, TCAD

If we count all his publications, Jiaqi has total 13 journal articles and 37 conference papers at UT so far. I can attest that Jiaqi has contributed significantly to those non-first-authored papers too. This is unthinkable for just a bit over four years at UT, minus one semester not working with me, while juggling with his course work (btw, he indeed maintains a 4.0 GPA at UT!), and other responsibilities (such as TA). I don't know how he can manage to handle all these. Jiaqi is just extremely efficient and focused.

For me to even go into details of all Jiaqi's first authored papers would take tremendous space. I recommend you read his research statement. The key message that I would like to have here is that Jiaqi is super creative and innovative. He has a clear vision on the big picture and understands the limitations, e.g., due to the emerging devices. His goal is to build a holistic framework for emerging AI computing platform, while carrying out cross-layer optimizations and software/hardware co-design, and design automation around those issues. His innovative work has crossed device, circuit, architecture, algorithms, and application domains. He emphasizes both novel theoretical innovation and experimental demonstrations, including chip fabrications and running on practical AI tasks. Jiaqi has also open-sourced his photonic AI software library TorchONN at Github, and it has already received 184 stars. I think it is fair to say that Jiaqi is a clear leading researcher in bridging photonics and AI via cross-layer device-circuit-architecture-algorithm co-design. Besides his main PhD research of photonics-AI, Jiaqi has contributed significantly to other important research topics, e.g., ML for EDA (he is the first author of DREAMPlace3.0 - DREAMPlace is the leading academic placer which won **Best Paper Awards** from DAC'19 and TCAD'21), and quantum ML, among others.

Jiaqi's Google citation is already 587, with h-index 14. This is extraordinary for a fresh PhD (to graduate in May 2023) in our area. His citation growth is very steep, as shown below.



While Jiaqi is a very independent researcher, he has a very nice personality to work with. I can say that Jiaqi is one of the *most collaborative* students I have ever had, as you can see from his diverse publication record. In my lab, other students like to discuss with him, whether they are more senior or junior than him. Jiaqi has also helped me mentor several junior students in my group, doing a fantastic job. Jiaqi has also initiated and built very strong collaborations with researchers from MIT, Univ. of Chicago, Yale, etc. Jiaqi has taken two industry internships, at Meta Reality Lab and Nvidia Research – both have worked amazingly well and led to joint publications and technology transfer to industry.

In terms of professional services, Jiaqi is currently serving as an Ethics and Early Career Development Working Group member in the NSF National AI Institute TILOS. He is co-organizing an Industry Panel in Feb. 2023. Jiaqi has also served as a local arrangement co-chair for the IEEE CAS Society Seasonal School of "AI/ML for IC Design and EDA". He did an excellent job in helping with logistics and moderating some invited sessions. Jiaqi was also invited to serve as a Program Committee member for AAAI'23 and KDD'23, two top AI/CS conferences. This is very unusual for a typical PhD student.

In terms of teaching, Jiaqi was a Graduate Teaching Assistant of my graduate level course "VLSI Physical Design Automation" in Spring 2022. He did a fantastic job and helped revamp the programming assignments. Students really liked him and gave him a very high TA rating of 4.9/5.

In terms of funding and proposal experiences, Jiaqi has helped me write several proposals and deal with funding agencies / industry sponsors. Since Jiaqi is very open-minded and equipped with superb cross-layer training and implementation skills, he picks up new areas and makes solid contributions super quickly.

To summarize, Jiaqi is a super-star PhD graduate in emerging AI/ML systems and EDA, one of the very best that I have seen worldwide with such an amazing trajectory in my 25+ years as a researcher in the field. He is well-rounded in all aspects to be a very successful faculty member, including research idea generation, execution, communication, collaboration, personality, and so on. The sky is his limit! I give him my strongest possible recommendation. Please feel free to contact me if you have any questions.

Sincerely,

David Z. Pan, Fellow of ACM, IEEE and SPIE
Silicon Labs Endowed Chair Professor
Email: dpan@ece.utexas.edu
Phone: 512-471-1436

ELECTRICAL ENGINEERING AND COMPUTER SCIENCE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Duane S. Boning, Clarence J. LeBel Professor, EECS 617.253.0931
Room 39-415, 77 Massachusetts Avenue boning@mtl.mit.edu
Cambridge, MA 02139-4307

December 29, 2022

Faculty Search Committee

Dear Search Committee,

This letter is to enthusiastically support and recommend Jiaqi Gu for a position as Assistant Professor in Electrical Engineering and Computer Science. Jiaqi is of unusually high caliber in terms of intellectual ability, creativity, initiative, and research ability, with an outstanding record and outstanding potential for future research contributions.

I know Jiaqi's research and abilities from two perspectives. First, I have followed with great interest his research contributions in recent years, based on research area overlap with my own in two domains – machine learning for design and manufacturing automation, and silicon photonics. Second, I had the opportunity to collaborate with him and his co-workers at UT Austin on one highly sophisticated joint research effort sitting at the intersection of these two domains. These two perspectives inform my recommendation, not just on research results and contributions, but also on the skills and approaches that Jiaqi possesses and employs to enable his outstanding research.

I'll start with the contributions, and perspectives, gained through the research collaboration on NeurOLight. This work grew out of an interaction sparked by Jiaqi with one of my PhD students at MIT, Zhengqi Gao. We have a dual interest in machine learning and photonics, primarily with the view of developing and applying ML approaches for photonics design (while Jiaqi's interest is broader, involving also development of photonics for AI/ML acceleration). The NeurOLight work culminated in the recent paper at NeurIPS 2022 (J. Gu, Z. Gao, C. Feng, H. Zhu, R. T. Chen, D. S. Boning, and D. Z. Pan, "NeurOLight: A Physics-Agnostic Neural Operator Enabling Parametric Photonic Device Simulation"). This paper is an exceptionally sophisticated combination of both electromagnetics and machine learning, making multiple novel contributions and demonstrating impressive ability to *accelerate* and *generalize* EM-field calculations in photonic components. On the ML front, the paper extends the latest work in Fourier neural operators (in the domain of physics-inspired neural networks or PINNs) with a novel cross-shaped FNO+FFN block, to orthogonalize and interleave the vertical and horizontal EM-field modeling, with important reduction in number of parameters needed. Second, it synthesizes physics understanding with ML requirements by a novel data augmentation approach, leveraging mix-up superposition across multiple light sources to enable one-shot learning and prediction in multi-source problems using only single-source original data samples. Third, and especially important for generalization (i.e., not just learning to predict fields for a fixed geometry/material design) is a novel approach to encoding the physical domain with a novel joint PDE encoder. This includes an especially creative (and non-intuitive) approach based on a set of prior artificial wave patterns, to represent both short-range and long-range light interactions. Finally, the approach cleverly uses masked image ideas to learn how to "fill in" the field in the interior of the device, based on the input source fields serving as the "hint." These approaches combine to achieve ~50% lower prediction errors with ~40% fewer parameters than competing state of the art approaches. In summary, the work is impressive both in the impact for photonics simulation, and in the combination of multiple novel ML approaches.

This research collaboration has also given me the opportunity to see Jiaqi's research process in detail. First, he is eager to reach out and collaborate in an open and productive way, with both junior researchers (new PhD students in my group) and with me. This collaboration was very much at his initiative, and he coordinated and drove the multiple research discussions, brainstorming sessions, update presentations, paper drafting, and revisions. I know this interaction is only one of several cross-university research collaborations that Jiaqi has initiated (including separate ones with Prof. Song Han here at MIT); my sense is that Jiaqi is impressively efficient, effective, and creative in pursuing multiple research threads simultaneously.

One of the reasons that I was pleased to undertake the NeurOLight collaboration with Jiaqi, is that I was well aware of and impressed by his prior research contributions up to that point. One of the earliest contributions that really caught my attention, was his work as part of the team on DREAMPlace, which won the IEEE Tr. CAD best paper award in 2020. This work showed that clever application of ideas from machine learning (not just off-the-shelf application of ML models), particularly formulating design problems in such a way as to leverage gradient descent and GPU acceleration, can have an enormous impact in addressing large-scale VLSI design problems. This is indicative also of the broad and foundational experience – in both electronic CAD and deep machine learning – that Jiaqi has built on, as he has expanded into the photonics/ML intersection.

That intersection is now bi-directional: Jiaqi is both developing/employing ML for photonics design automation, and contributing impressive architectural ideas for using photonics to implement future AI hardware functionality. While my own work does not focus on AI hardware, I can appreciate the novelty and creativity of his work on optical neural networks using multi-operand rings, and other hardware-efficient optical neural network optimizations. Those works benefit from his deep understanding of photonics, computing architectures, *and* machine learning, e.g., in his outstanding work on automatic tensor decomposition for transformer compression (NeurIPS 2022 spotlight).

In reviewing Jiaqi's research statement, I'm impressed and excited by his research ambitions and goals. He is the rare talent that has the combined background in photonics, machine learning, and design automation, to make progress in this challenging intersection. He has the creativity, drive, organizational and collaborative skills, and research sense to deliver on his vision.

To summarize, I enthusiastically recommend and support Jiaqi Gu for an Assistant Professor position in Electrical Engineering and Computer Science. Jiaqi has already established an impressive record of highly creative and recognized excellence in the broad intersection of photonics (and other emerging technologies) and AI/ML. He has outstanding instincts for important problems and opportunities at this intersection, and possesses the unusual breadth of background to continue to make key contributions in these areas. He would be an outstanding addition to your faculty, and I very much look forward to continue being impressed by, and learning from, his research contributions.

Sincerely,

Duane S Boning

Duane S. Boning
Clarence J. Lebel Professor, Electrical Engineering and Computer Science, MIT

UNIVERSITY OF TEXAS AT AUSTIN
UT ECE
ELECTRICAL & COMPUTER ENGINEERING

November 23, 2022
Faculty Search Committee

Dear Committee Members:

It is my great pleasure to give my strong recommendation to Jiaqi Gu for a tenure-track faculty position at your esteemed institution. Jiaqi is about to receive his PhD degree under the co-supervision of Dr. David Pan and me from the Department of Electrical and Computer Engineering, The University of Texas at Austin (UT Austin) in 2023. Jiaqi is one of the very top students I have advised, and I strongly feel that he is a highly qualified and deserving faculty candidate.

Jiaqi received his BE degree in Microelectronics from Fudan University and joined UT Austin ECE in Fall 2018 to pursue his PhD. After he joined the AFOSR MURI team that I lead (https://muri2.engr.utexas.edu) in Spring 2019, he quickly demonstrated to be a genius researcher with ever-growing passions. Jiaqi mainly focused on algorithm and hardware design for integrated photonics computing, especially for artificial intelligence (AI) applications. Integrated photonics has shown extraordinary potential in optical computing mainly due to the bottleneck that Moore's law will eventually encounter. Analog photonic neural computing is a very promising technology for ultra-fast and efficient optical neural network (ONN) acceleration. Jiaqi has made seminal contributions to the design and optimization of next-generation photonic neuromorphic computing platforms in a cross-layer manner, from the device and circuit level up through architecture and algorithm level, from various perspectives of chip area, power, robustness, and trainability.

Jiaqi's PhD research covers several critical topics in the field of optical AI, addressing circuit area/power cost, noise robustness issues, and optical neural network training. I will give an introduction of some highlights of his research contributions:

- To reduce the circuit footprint and power consumption of optical neural network engines, Jiaqi proposed a novel butterfly-style ONN architecture. This new ONN design moves beyond the conventional designs for universal linear operations. Instead, it restricts the matrix expressivity of the photonic circuit to implement structured matrix multiplications. This work is also the first one in the field that investigates optical device pruning to further reduce the hardware cost with layout-aware structured sparsity. Jiaqi's design provides a new possibility to achieve more scalable optical deep learning with the same task accuracy but a smaller area cost, lower power consumption, and higher noise resilience compared to prior art. This work was first presented at ACM/IEEE ASP-DAC'20 (a premier conference in computer-aided design and design automation), and its journal version has been accepted by IEEE TCAD, the premier journal in electronic design automation. Later, our team taped out a silicon photonic neural chip at the AMF foundry to experimentally demonstrate the utility of the proposed butterfly-style ONNs, which was recently accepted to ACS Photonics (a premier journal in Photonics). The ONN architecture design and pruning methods have impacted the optical AI field, e.g., many later ONN review, architecture, co-design papers cited, followed, and compared with

this work. Jiaqi continues pushing the limit of photonic circuit compactness by using new devices (multi-operand optical ring resonators/MMIs/MZIs) and designing new architectures.

- Jiaqi also made breakthrough contributions to the ONN robustness issues against various physical variations and noises. As an analog computing platform, the computing fidelity and robustness of integrated photonic neural chips can be negatively influenced by non-ideal manufacturing and changing environments. Unlike conventional working flows that directly map software-trained ONN weights to the photonic hardware, Jiaqi investigated the noise sensitivity issues of ONN and proposed the first noise-aware device quantization framework ROQ, published at IEEE DATE 2020. He incorporates device non-ideality during training, so that the trained ONN can achieve high accuracy under various noises, even with 2- to 3-bit device programming resolutions, which proves the practicality of using photonics for reliable machine learning acceleration. It shall be noted that even with sophisticated noise modeling, the chip performance after physical deployment still has a gap from the software simulation results. Jiaqi later focuses on a more advanced training methodology, on-chip or hardware-in-the-loop training. He proposed an efficient zeroth-order on-chip training protocol FLOPS to train ONNs with in-situ optimization. Such a training paradigm naturally considers all physical non-ideality during on-chip computations, so the photonic neural chip can be self-learnable and noise-resilient in nature. This work was published at IEEE/ACM Design Automation Conference (DAC) in 2020, and its enhanced solutions were published at AAAI 2021 and NeurIPS 2021 (top conferences in AI/ML), respectively, which improved the training scalability by 25 times and 10,000 times compared to the prior art.
- Jiaqi also explores AI for optics, i.e., AI-enhanced design automation in photonic circuit design, to close the loop between photonics and AI. Jiaqi discovered that the manual design of photonic integrated circuits is not flexible and not necessarily the most efficient solution. Motivated by this, at DAC'22, he proposed the first automatic differentiable method to search for photonic circuit designs, achieving beyond-human performance in compactness and noise robustness. He also proposes a machine learning method to speed up frequency-domain optical simulation by 2 orders of magnitude for faster design space exploration and published at NeurIPS 2022.

Jiaqi's PhD research is exceptional. During his PhD at UT Austin, he has published more than 40 research papers in premier journals and conferences in design automation (DAC, ICCAD, DATE, ASP-DAC, TCAD), artificial intelligence (NeurIPS, CVPR, ICCV, AAAI), and photonics (Nat. Comm., LPR, ACS Photonics, Nanophotonics, CLEO, Photonics West), and so on. His research also led to impressive Best Paper Awards and Nominations at DAC, ASP-DAC, and TCAD, ACM Student Research Competition (SRC) Grand Finals, and Winner at Synopsys Optical Design Competition 2022.

I have supervised 39 postdocs and graduated 54 PhD students at UT Austin so far. Jiaqi Gu is definitely one of the very top students I have supervised, and I am proud to have him as my student. Jiaqi is now getting mature and full of visionary ideas, and he has a strong vision on promising research directions as well as hands-on skills to turn ideas into practice. He is a fantastic communicator and has good leadership who can initiate and organize group discussions, seek and lead cross-group collaborations, as well as mentoring junior PhD

students. His grand vision in future directions is to close the loop of optics and AI toward next-generation intelligent and efficient computing, and I am tremendously confident in his potential to attain prominent success.

Jiaqi Gu's interdisciplinary research on AI and integrated photonics is very unique and shall have growing impacts on emerging AI hardware and efficient computing. I am very confident you will find Jiaqi among the best candidates this year, and he will continue to be an invaluable asset to any institution that has him as a part. If you need any additional information, please feel free to contact me.

Sincerely,

Ray T. Chen
Keys and Joan Curry/Cullen Trust Endowed Chair Professor
Fellow Member of National Academy of Inventors
Fellow of IEEE, OSA and SPIE
Director of MURI Center for Power-Efficient Si Nanophotonics for Computing & Interconnects
Phone: (512)471-7035
chen@ece.utexas.edu
Nanophotonics and Optical Interconnects Research Lab
The University of Texas, Austin

Application Forms

Source of Applicants

Where did you learn of this opportunity?

Department Website