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

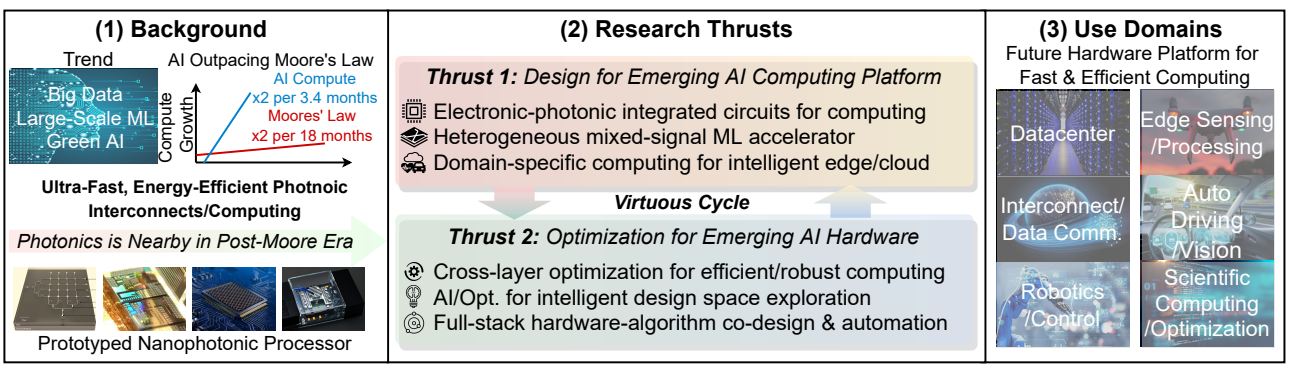

Figure 1: Overview of background, blueprint (two thrusts), and important use domains of my research.

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**.

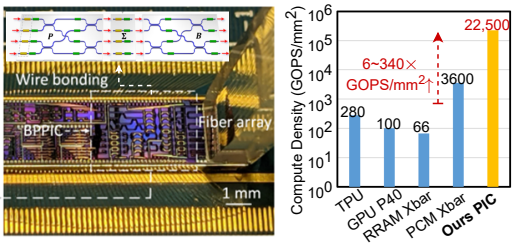

Figure 2: Our butterfly electronic-photonic co-packaged neural chip.

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.

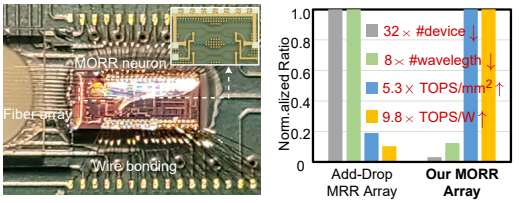

Figure 3: Our MORR-based single-device photonic tensor unit tape-out.

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.