

Toward Large-Scale Photonics-Empowered AI Systems: From Physical Design Automation to System-Algorithm Co-Exploration

Ziang Yin^a, Hongjian Zhou^a, Nicholas Gangi^b, Meng Zhang^b, Jeff Zhang^a, Zhaoran Rena Huang^b, and Jiaqi Gu^a

^aSchool of Electrical, Computer and Energy Engineering, Arizona State University

^bSchool of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute

ABSTRACT

The continued scaling of artificial intelligence workloads is increasingly constrained by data movement, interconnect bandwidth, and energy efficiency in conventional electronic systems. Integrated photonics offers a promising pathway to address these challenges through high-bandwidth optical interconnects and energy-efficient photonic computing primitives. However, translating device-level photonic advances into large-scale, deployable AI systems remains difficult due to strong coupling between physical implementation, system architecture, and learning algorithms.

In this work, we identify three considerations that are essential for realizing practical photonic AI systems at scale: (1) **dynamic tensor operation support** for modern models rather than only weight-static kernels, especially for attention/Transformer-style workloads;¹ (2) **systematic management of conversion, control, and data-movement overheads**, where multiplexing and dataflow must amortize electronic costs instead of letting ADC/DAC and I/O dominate;² and (3) **robustness under hardware non-idealities** that become more severe as integration density grows.³ To study these coupled tradeoffs quantitatively, and to ensure they remain meaningful under real implementation constraints, we build a cross-layer toolchain that supports photonic AI design from early exploration to physical realization. **SimPhony**⁴ provides implementation-aware modeling and rapid cross-layer evaluation, translating physical costs into system-level metrics so architectural decisions are grounded in realistic assumptions. **ADEPT**⁵ and **ADEPT-Z**⁶ enable end-to-end circuit and topology exploration, connecting system objectives to feasible photonic fabrics under practical device and circuit constraints. Finally, **Apollo**⁷ and **LiDAR**^{8,9} provide scalable photonic physical design automation, turning candidate circuits into manufacturable layouts while accounting for routing, thermal, and crosstalk constraints. Together, these capabilities make our co-design loop both quantitative and physically grounded, bridging architectural intent and deployable photonic hardware.

Keywords: Photonics-empowered AI, electronic-photonic design automation (EPDA), system-algorithm co-exploration

Further author information: (Send correspondence to Jiaqi Gu)

Jiaqi Gu: E-mail: jiaqigu@asu.edu

1. INTRODUCTION

The rapid scaling of artificial intelligence (AI) workloads has exposed fundamental limitations in conventional electronic computing systems.^{10–13} While transistor scaling continues to deliver incremental improvements in logic density, system-level performance, and energy efficiency are increasingly constrained by data movement, memory bandwidth, and interconnect power consumption. These challenges become especially acute in large-scale AI systems, where communication and I/O frequently dominate both latency and energy budgets.

Photonics has emerged as a promising technology to relieve these bottlenecks, offering high bandwidth density, low propagation loss, and natural support for broadcast and wavelength-division multiplexing. In parallel with advances in optical interconnects and co-packaged/heterogeneous integration,^{14,15} recent demonstrations have shown photonic computing primitives that accelerate core AI operators (e.g., tensor/matrix computations) at impressive throughput and parallelism.^{1–3, 10, 16–18} Despite these device- and chip-level successes, a clear path toward *large-scale photonics-empowered AI systems* remains elusive.

A central challenge is that scaling beyond isolated accelerators requires system-level integration across devices, circuits, architectures, interconnect fabrics, and learning algorithms, under constraints that are qualitatively different from electronics. Photonic integrated circuit (PIC) / electronic photonic integrated circuit (EPIC) implementations must obey curvilinear geometries, limited routing resources, strict fabrication rules, and strong sensitivity to process variation and thermal effects, all of which directly impact loss, crosstalk, tuning power, and yield.^{19,20} Meanwhile, packaging and electronic-photonic interfacing introduce additional constraints and costs that can dominate deployment feasibility and module economics.^{14,15} As a result, manual photonic design does not scale to the complexity demanded by system-class AI, and architecture-only abstractions can be misleading unless they explicitly model physical and packaging realities.

In this work, we argue that realizing large-scale photonics-empowered AI systems requires two tightly coupled capabilities:

- **Photonic and electronic-photonic physical design automation (EPDA)** to enable scalable, manufacturable implementation of complex PICs/EPICs; and
- **System-algorithm co-exploration** that incorporates physical non-idealities, control/calibration limits, and packaging/interface costs into architectural design and learning optimization.

Drawing on our recent progress, we connect EPDA with cross-layer hardware/algorithm co-design and illustrate how these techniques together enable scalable photonic AI systems.

2. PHOTONICS-EMPOWERED AI SYSTEMS: A CROSS-LAYER VIEW

As argued in Sec. 1, scaling photonics beyond isolated accelerators requires co-optimization across devices, circuits, architectures, and learning algorithms under realistic physical constraints.

In this section, we ground these requirements through three photonic tensor-core (PTC) designs, Lightning-Transformer¹, TeMPO², and SCATTER³ that each target a different bottleneck regime in cloud/edge deployment and demonstrate how cross-layer co-design translates system constraints into implementable architectures.

2.1 Lightening-Transformer: Dynamic Tensor Operations for Transformer Inference

To support modern LLMs, particularly attention-based Transformer architectures, photonic computing cores must move beyond weight-static matrix units and enable *dynamic* tensor operations, while jointly optimizing signal conversion and data movement. Our prior architecture **Lightening-Transformer**¹ was the first photonic accelerator designed to efficiently execute high-throughput, dynamic optical matrix–matrix multiplications for self-attention. It replaces weight-static photonic matrix units with a *Dynamically-operated Photonic Tensor Core* (DPTC). At its heart is the **Dynamically-operated Dot-product (DDot)** engine, a coherent dot-product unit that enables picosecond-level operand switching and supports full-range (signed) matrix inputs without hardware duplication or multiple inference passes. Lightening-Transformer further integrates these computing cores with *photonic interconnects* for inter-core data broadcast. By exploiting WDM for spectral parallelism and optical broadcast for operand sharing, the cross-layer-optimized architecture achieves over a 12 \times latency reduction compared to prior photonic accelerators.

2.2 TeMPO: Amortizing Conversion Overheads for Edge-Efficient Photonic AI

While Lightening-Transformer targets cloud-scale throughput, edge AI faces a different constraint regime where area and energy budgets are highly restricted, and electronic interfaces can dominate total cost. To address this setting, we extend the dynamic tensor-core concept to **TeMPO**,² introducing an efficient, time-multiplexed dynamic photonic tensor core that improves utilization and amortizes overheads. At the device level, TeMPO employs customized, foundry-fabricated *slow-light Mach-Zehnder modulators* (SL-MZMs) that leverage enhanced light–matter interaction to achieve a footprint an order of magnitude smaller than standard PDK elements. At the circuit level, TeMPO tackles the long-standing ADC power bottleneck via *hierarchical partial product accumulation*. By aggregating photocurrents and using lightweight capacitive temporal integration in the analog domain, TeMPO reduces the required ADC sampling frequency by a factor of T (the integration time step, e.g., 60 cycles). This cross-layer co-design achieves 1.2 TOPS/mm² compute density and 22.3 TOPS/W energy efficiency, enabling real-time edge tasks such as voice keyword spotting and semantic segmentation.

2.3 SCATTER: Robust and Scalable Photonic Tensor Cores under Physical Non-Idealities

As photonic tensor cores scale in size and density, non-idealities and control constraints (e.g., loss, drift, thermal crosstalk, and calibration limits) become first-order design factors rather than second-order effects. Our recent architecture **SCATTER**³ exemplifies an extreme cross-layer co-design spanning device, circuit, layout, architecture, and algorithm, where a multi-step co-optimization pipeline jointly targets power/area minimization and robustness under realistic physical constraints.

- ❶ Starting from the bottom of the stack, SCATTER replaces communication-oriented foundry building blocks with *compute-tailored* low-power slow-light modulators, enabling substantial baseline reductions in footprint and energy.
- ❷ At the physical and circuit levels, SCATTER explores *circuit/weight-matrix co-sparsity* to enable crosstalk-aware layout that safely densifies the photonic tensor core without sacrificing robustness.
- ❸ At the architecture level, SCATTER introduces an on-chip *in-situ light redistribution* (rerouting) and power-gating mechanism that dynamically re-allocates optical power to active rows/columns, enabling high-efficiency structured sparse matrix multiplication while improving effective SNR by avoiding over-driving inactive channels.
- ❹ Finally,

SCATTER addresses dominant electronic overhead by upgrading conventional electrical DACs to a *hybrid electronic-optical segmented DAC*, combining high resolution with low power to preserve accuracy at reduced energy. Together, this cross-stack strategy turns performance, efficiency, and robustness into co-optimized objectives, yielding **511 \times area reduction and 12.4 \times power savings** while largely resolving thermal crosstalk, demonstrating a practical path toward robust, sparse, and scalable photonic AI acceleration.

3. ELECTRONIC-PHOTONIC DESIGN AUTOMATION: KEY ENABLER OF SCALABLE PHOTONIC AI SYSTEMS

Section 2 demonstrated that achieving high performance, energy efficiency, and robustness in photonic AI accelerators fundamentally requires *cross-layer co-design*, where device physics, circuits, architectures, and learning algorithms are optimized in concert rather than in isolation. However, while such cross-layer strategies are essential, they also expose a critical scalability challenge: as photonic AI systems grow in complexity, *manual or ad hoc co-design rapidly becomes untenable*. A photonic AI system’s performance is no longer a property of any single layer, but an emergent outcome of *tightly coupled interactions* across the full stack. For example, device- and circuit-level non-idealities, such as optical loss, thermal crosstalk, directly shape feasible architectures and inference/training strategies. Conversely, algorithmic choices, sparsity structures, and dataflow patterns feed back into physical layout, routing congestion, and electronic–photonic interface design.

This growing entanglement across abstraction layers motivates a fundamental shift: **cross-layer co-design must itself be elevated into an automated design paradigm**. To reliably translate architectural intent and algorithmic innovation into deployable photonic hardware, future photonic AI systems demand *full-stack, physics-aware, and closed-loop design automation*. In particular, electronic–photonic integrated circuits (EPICs) require design tools that can simultaneously reason about optical and electronic behaviors, propagate physical constraints upward into system-level models, and feed system-level specifications back down to circuits, layouts, and devices. Electronic–photonic design automation (EPDA) emerges as a key technological enabler to meet this challenge. In the following sections, we highlight three representative directions from our work that synergistically move toward full-stack EPDA: (i) **system-level modeling grounded in rigorous device/circuit/architecture/algorithm co-simulation**, (ii) **automated multi-objective exploration of photonic circuit topologies under realistic physical and architectural constraints**, and (iii) **physically realizable design closure through automated EPIC place-and-route (P&R)**.

3.1 SimPhony: Cross-Layer Modeling from Device Response to System Performance

Among recent progress in photonic AI system modeling,^{21–23} we emphasize our open-source cross-layer photonic AI system modeling tool, SimPhony,⁴ which serves as the core engine in the evaluation layer in the EPDA stack: it translates device/circuit responses into system-visible metrics so that architectural and algorithmic decisions are made under realistic physical constraints rather than ideal abstractions. Figure 1 shows the SimPhony framework, which integrates physical modeling, architectural analysis, and hardware-aware training to support automated system-algorithm co-exploration. Concretely, this means SimPhony provides a way to propagate device- and circuit-level effects, e.g., loss accumulation, drift, thermal behavior, and calibration/control costs, into the metrics that drive system decisions (throughput, latency, energy, and device non-idealities).

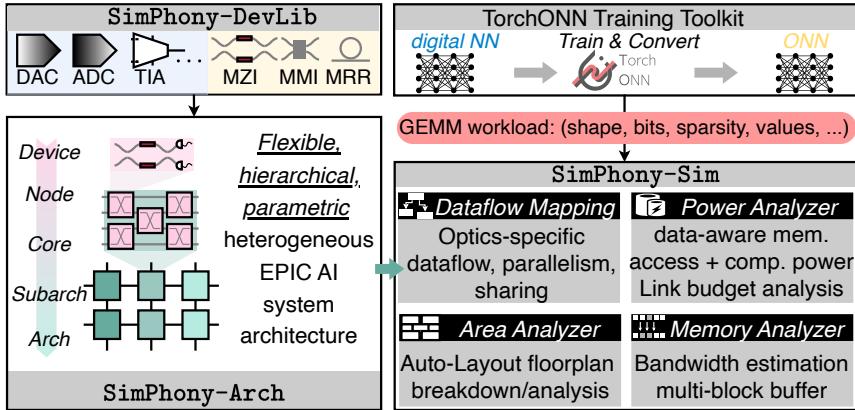


Figure 1: Overview of the SimPhony cross-layer modeling and co-exploration framework. Device- and circuit-level photonic models are integrated with architectural analysis, dataflow mapping, and power, area, and memory estimation. Hardware-aware training and conversion are supported through tight coupling with learning frameworks, enabling system–algorithm co-exploration under realistic physical constraints.

Within the overall workflow, SimPhony acts as the bridge from physical response → system response: candidate circuits/topologies are evaluated using device/circuit behavior as inputs, and the resulting system-level cost/performance projections guide which candidates should be further explored and physically implemented. This closes the loop between cross-layer intent and the constraints that will ultimately limit deployable hardware behavior.

3.2 ADEPT: Automated Multi-Objective PIC Topology Exploration

With system-level modeling that provides rapid hardware feedback, we leverage it to enable circuit topology exploration to generate high-performance photonic AI hardware. Existing work has developed automated architecture-level design space exploration²⁴ for efficient photonic AI accelerator designs. As a complementary direction, we have explored photonic tensor core circuit topology optimization using advanced optimization algorithms, ADEPT.^{5,6} Rather than hand-crafting circuit structures based on heuristics, this framework enables automated Pareto front search over a huge design space, producing Pareto-optimal candidates that can then be embedded in SimPhony as new computing core designs and physically realized by our layout synthesis tools introduced later.

A key challenge in photonic tensor core topology exploration is the *exponentially large, highly discrete* design space with multiple competing objectives and constraints (e.g., balancing area, power, latency, robustness, expressivity) that is beyond humans' design capability.

To enable efficient circuit topology design in this massive space *from weeks to hours*, we proposed both differentiable (ADEPT⁵) and multi-objective (ADEPT-Z⁶) optimization approaches. The differentiable version relaxes the discrete photonic component selection and construction problem as a soft probability learning problem, leveraging gradient-descent methods for rapid design space exploration while honoring chip footprint constraints. In this approach, we have successfully found non-intuitive designs that are simultaneously more compact, expressive, and robust compared to prior art, while keeping the whole process within 6 hours.

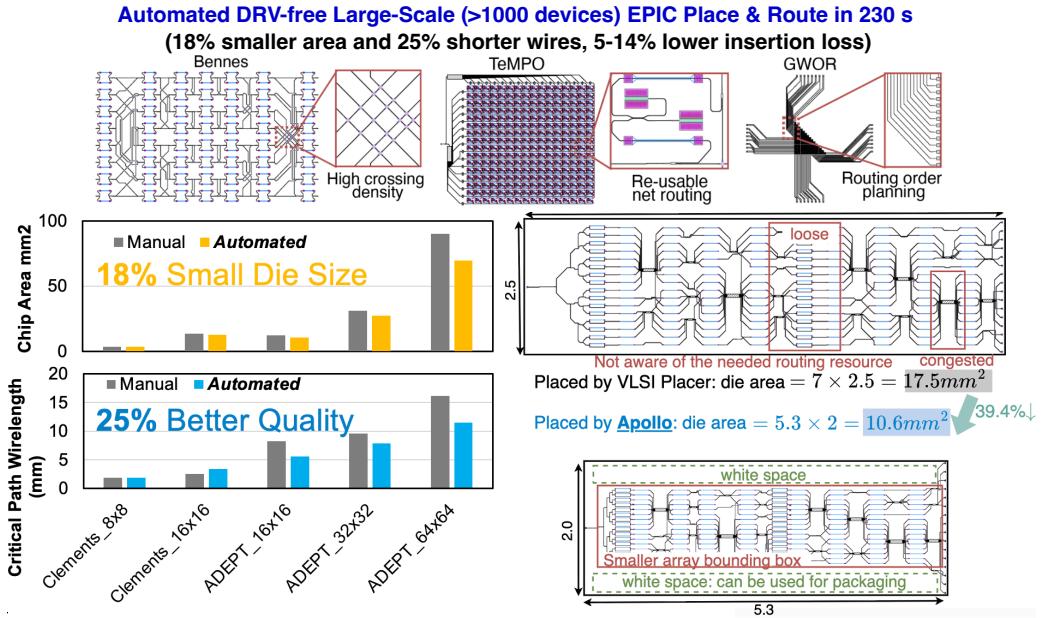


Figure 2: Our proposed automated PIC placement engine Apollo⁷ and router LiDAR-V2^{8,9} can generate compact, high-quality layout for large-scale PICs (over 1000 devices) within 230s.

Building on ADEPT, our latest framework **ADEPT-Z**⁶ extends to a *gradient-free* optimization formulation that is substantially more flexible than the differentiable relaxation used in ADEPT. This choice is important for photonic tensor core synthesis because the massive architecture design space is often *highly discrete*, including component types, port configurations, connectivity patterns, and placement decisions, which are difficult to faithfully encode in a differentiable parameterization. By operating directly in this discrete design space, ADEPT-Z can naturally support richer circuit grammars and constraints, while retaining efficient search. Equally critical, ADEPT-Z performs **multi-objective Pareto optimization** that *simultaneously* pushes key system metrics, e.g., **energy efficiency**, **compute density**, and **accuracy/expressivity**, instead of collapsing them into a single scalar objective. This multi-objective nature empowers ADEPT-Z with the capability of producing **tens of Pareto-frontier candidates in a single run within ~3 hours**, covering diverse area-power trade-offs that can be carried forward for downstream evaluation and selection.

More broadly, our ADEPT-series illustrates a central insight of EPDA: **advanced automation and optimization can outperform expert hand-design for complex photonic compute modules**. By systematically exploring an enormous combinatorial space, it can uncover *non-intuitive* circuit structures that exceed human heuristic design capability, while compressing an exploration process that traditionally takes experts **weeks** into a **few-hour** automated workflow. The resulting Pareto-optimal circuit candidates can then be embedded into system-level models (e.g., SimPhony) and passed to physical implementation tools, enabling a closed-loop path from topology discovery to deployable EPIC designs.

3.3 Apollo & LiDAR: Automated EPDA Flow for PICs and EPICs

After topology exploration (Sec. 3.2) identifies promising circuit netlists, the next bottleneck is design closure: translating these candidates into physically realizable layouts while accounting for curvilinear waveguide constraints, routing congestion, crossings, and layout-induced loss/crosstalk. Early demonstrations of PIC physical design largely relied on manual, ad hoc methodologies. While sufficient for proof-of-concept devices and small-scale circuits, such approaches do not scale to system-level PICs comprising thousands to millions of components. In practice, PIC development still follows a sequential, largely manual pipeline, device design, schematic capture, layout drafting, and verification.²⁵ Custom blocks (e.g., modulators, filters, and multiplexers) are typically handcrafted, assembled at the schematic level, and then translated into layout through manual or rule-based placement and routing. This fragmented workflow significantly slows iteration and becomes a fundamental bottleneck as photonic systems scale.

To overcome this barrier, electronic-photonic design automation (EPDA) is urgently needed for automated PIC layout synthesis. Recent research has explored automated placement and routing for photonic circuits considering various realistic constraints.^{7,8,26,27,27} An essential first step is to develop automation that can translate a designer-specified netlist and constraints into a high-quality physical layout by *automatically placing* photonic components while accounting for routability and layout-dependent effects. To this end, we propose **Apollo**,⁷ the first GPU-accelerated, routing-informed placement framework tailored for large-scale PICs. Rather than placing components independently of routing considerations, we explicitly model waveguide routing congestion and crossings during placement to preserve enough routing spacing for routability maximization. Figure 2 illustrates how routing-informed placement methodology substantially improves *layout regularity, area efficiency, and routability* in large-scale PICs within only **230s**. This approach is essential for large-scale PICs, where naive placement can render routing infeasible.

Following placement, our tool **LiDAR-V2**^{8,9} executes waveguide routing, distinguishing itself from existing methods by generating design-rule-violation (DRV)-free, GDSII layouts. Unlike traditional approaches that apply post-hoc smoothing, which often fail under congested constraints, LiDAR-V2 integrates a curvy-aware A* search that incorporates node orientation and bending radius directly into neighbor generation. And we propose an orientation-aware bitmap that enforces spacing rules and facilitates dynamic waveguide crossing insertion, eliminating the need for manual planning. As shown in Fig. 2, the result remains routable even under high crossing density and successfully generates valid layouts without DRVs while achieving a **5–14% insertion loss reduction** compared to the previous work.

In addition, metal routing is often overlooked in research prototypes, yet it can consume a substantial fraction of the overall layout closure time in practice. To this end, we explicitly incorporate metal routing²⁸ into our flow. Compared with conventional VLSI routing, PIC electrical routing must tightly coordinate with photonic structures and waveguides, account for packaging-driven pad breakout and long-distance interconnects, and obey more diverse keep-out and coupling constraints (e.g., to avoid optical loss, crosstalk, and thermal interference). By modeling these PIC-specific requirements, our flow delivers routing solutions that better match designers' needs while remaining scalable and design-rule compliant.

Recent automated frameworks from our group demonstrate that these stages can be unified into a **push-button EPDA flow**, translating high-level photonic circuit descriptions into fully routed layouts with minimal human intervention. Physical design automation is not merely a

productivity enhancement; it is a **key enabler of scalable photonic AI systems**. By automating layout synthesis and enforcing physical feasibility, EPDA fundamentally expands the architectural design space that can be explored within realistic time and resource budgets. As shown in Fig. 2, iterative placement–routing refinement achieves, on average, **an 18% reduction in die size and a 25% improvement in layout quality**. Overall, our EPDA toolflow allows system designers to evaluate larger and more complex photonic architectures, explore tighter integration densities, and interconnect strategies while maintaining predictable performance.

4. DISCUSSION

This work suggests that the main scaling bottleneck for photonics-empowered AI is not generating new circuit concepts, but closing the loop from cross-layer intent to implementable hardware under realistic loss, thermal/control, and layout constraints. The AI-assisted EPDA flow presented here helps reduce this gap by (i) evaluating system behavior from device/circuit responses, (ii) exploring topologies with implementation constraints in mind, and (iii) using scalable P&R to enforce physical feasibility.

Two practical directions could further strengthen this loop: (1) packaging/interface-aware modeling integrated earlier in evaluation so coupling and assembly constraints are reflected before topology and layout commitments, and (2) more systematic variability/test hooks (variation-aware metrics and calibration/test planning) so candidate designs are screened for robustness rather than relying on post-fabrication fixes.

REFERENCES

- [1] Zhu, H., Gu, J., Wang, H., Jiang, Z., Zhang, Z., Tang, R., Feng, C., Han, S., Chen, R. T., and Pan, D. Z., “Lightening-transformer: A dynamically-operated optically-interconnected photonic transformer accelerator,” in [*2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*]], 686–703, IEEE (2024).
- [2] Zhang, M., Yin, D., Gangi, N., Begović, A., Chen, A., Huang, Z. R., and Gu, J., “Tempo: Efficient time-multiplexed dynamic photonic tensor core for edge ai with compact slow-light electro-optic modulator,” *Journal of Applied Physics* **135**, 223105 (06 2024).
- [3] Yin, Z., Gangi, N., Zhang, M., Zhang, J., Huang, R., and Gu, J., [*SCATTER: Algorithm-Circuit Co-Sparse Photonic Accelerator with Thermal-Tolerant, Power-Efficient In-situ Light Redistribution*], Association for Computing Machinery, New York, NY, USA (2025).
- [4] Yin, Z., Zhang, M., Gangi, N., Huang, R., Zhang, J., and Gu, J., “Simphony: A device-circuit-architecture cross-layer modeling and simulation framework for heterogeneous electronic-photonic ai system,” in [*Proceedings of the 62nd Annual ACM/IEEE Design Automation Conference*, DAC ’25], IEEE Press (2025).
- [5] Gu, J., Zhu, H., Feng, C., Jiang, Z., Liu, M., Zhang, S., Chen, R. T., and Pan, D. Z., “Adept: Automatic differentiable design of photonic tensor cores,” in [*Proc. DAC*], (2022).
- [6] Jiang, Z., Ma, P., Zhang, M., Huang, R., and Gu, J., [*ADEPT-Z: Zero-Shot Automated Circuit Topology Search for Pareto-Optimal Photonic Tensor Cores*], 1077–1083, Association for Computing Machinery, New York, NY, USA (2025).
- [7] Zhou, H., Yang, H., Gangi, N., Huang, Z. R., Ren, H., and Gu, J., “Apollo: Automated routing-informed placement for large-scale photonic integrated circuits,” in [*2025 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*]], 1–9 (2025).

- [8] Zhou, H., Zhu, K., and Gu, J., “LiDAR: Automated Curvy Waveguide Detailed Routing for Large-Scale Photonic Integrated Circuits,” in [Proc. ISPD], 64–72 (2025).
- [9] Zhou, H., Yang, H., Ying, Z., Gangi, N., Ren, H., Matres, J., Gu, J., et al., “LiDAR 2.0: Hierarchical Curvy Waveguide Detailed Routing for Large-Scale Photonic Integrated Circuits,” *IEEE TCAD* (2025).
- [10] Shen, Y., Harris, N. C., Skirlo, S., et al., “Deep learning with coherent nanophotonic circuits,” *Nature Photonics* (2017).
- [11] Feng, C., Gu, J., Zhu, H., Ying, Z., Zhao, Z., Pan, D. Z., and Chen, R. T., “A compact butterfly-style silicon photonic-electronic neural chip for hardware-efficient deep learning,” *ACS Photonics* **9**(12), 3906–3916 (2022).
- [12] Gu, J., Zhao, Z., Feng, C., Li, W., Chen, R. T., and Pan, D. Z., “FLOPS: Efficient On-Chip Learning for Optical Neural Networks Through Stochastic Zeroth-Order Optimization,” in [Proc. DAC], (2020).
- [13] Gu, J., Zhao, Z., Feng, C., et al., “Towards area-efficient optical neural networks: an FFT-based architecture,” in [Proc. ASPDAC], (2020).
- [14] Tan, M., Xu, J., Liu, S., Feng, J., Zhang, H., Yao, C., Chen, S., Guo, H., Han, G., Wen, Z., Chen, B., He, Y., Zheng, X., Ming, D., Tu, Y., Fu, Q., Qi, N., Li, D., Geng, L., Wen, S., Yang, F., He, H., Liu, F., Xue, H., Wang, Y., Qiu, C., Mi, G., Li, Y., Chang, T., Lai, M., Zhang, L., Hao, Q., and Qin, M., “Co-packaged optics (cpo): status, challenges, and solutions,” *Frontiers of Optoelectronics* **16**, 1 (Mar 2023).
- [15] Ranno, L., Gupta, P., Gradkowski, K., Bernson, R., Weninger, D., Serna, S., Agarwal, A. M., Kimerling, L. C., Hu, J., and OBrien, P., “Integrated photonics packaging: Challenges and opportunities,” *ACS Photonics* **9**(11), 3467–3485 (2022).
- [16] Feldmann, J., Youngblood, N., Karpov, M., Gehring, H., Li, X., Stappers, M., Le Gallo, M., Fu, X., Lukashchuk, A., Raja, A. S., Liu, J., Wright, C. D., Sebastian, A., Kippenberg, T. J., Pernice, W. H. P., and Bhaskaran, H., “Parallel convolutional processing using an integrated photonic tensor core,” *Nature* **589**, 52–58 (Jan 2021).
- [17] Xu, S., Wang, J., Yi, S., and Zou, W., “High-order tensor flow processing using integrated photonic circuits,” *Nature Communications* **13**, 7970 (Dec 2022).
- [18] Ahmed, S. R., Baghdadi, R., Bernadskiy, M., Bowman, N., Braid, R., Carr, J., Chen, C., Ciccarella, P., Cole, M., Cooke, J., et al., “Universal photonic artificial intelligence acceleration,” *Nature* **640**(8058), 368–374 (2025).
- [19] Bogaerts, Wim and Chrostowski, Lukas, “Silicon photonics circuit design : methods, tools and challenges,” *LASER & PHOTONICS REVIEWS* **12**(4), 1700237:1–1700237:29 (2018).
- [20] Sharma, S. and Roy, S., “A survey on design and synthesis techniques for photonic integrated circuits,” *The Journal of Supercomputing* **77**, 4332–4374 (May 2021).
- [21] Liu, Y., Hu, B., Liu, Z., Chen, P., Du, L., Liu, J., Li, X., Zhang, W., and Xu, J., “Fiona: Photonic-electronic co-simulation framework and transferable prototyping for photonic accelerator,” in [Proc. ICCAD], (2023).
- [22] Yin, Z., Zhang, M., Gangi, N., Huang, R., Zhang, J., and Gu, J., “Simphony: A device-circuit-architecture cross-layer modeling and simulation framework for heterogeneous electronic-photonic ai system,” in [2025 62nd ACM/IEEE Design Automation Conference (DAC)], 1–7, IEEE (2025).
- [23] Andrulis, T., Chaudhry, G. I., Suriyakumar, V. M., Emer, J. S., and Sze, V., “Architecture-level modeling of photonic deep neural network accelerators,” (2024).

- [24] Li, M., Yu, Z., Zhang, Y., Fu, Y., and Lin, Y., “O-has: Optical hardware accelerator search for boosting both acceleration performance and development speed,” in [*Proc. ICCAD*], (2021).
- [25] Chrostowski, L. and Hochberg, M., [*Silicon photonics design: from devices to systems*], Cambridge University Press (2015).
- [26] Wu, Y., Guan, W., Tong, Y., and Ma, Y., “Automatic routing for photonic integrated circuits under delay matching constraints,” in [*Proc. DATE*], (2025).
- [27] Zheng, Z., Li, M., Tseng, T.-M., and Schlichtmann, U., “Topro: A topology projector and waveguide router for wavelength-routed optical networks-on-chip,” in [*2021 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*], 1–9, IEEE (2021).
- [28] Zhou, H., Yang, H., Gangi, N., Liu, B., Zhang, M., Ren, H., Wang, X., Huang, R., and Gu, J., “Photonics-aware planning-guided automated electrical routing for large-scale active photonic integrated circuits,” in [*Proc. ISPD*], (2026).