Research Statement

Yuyang Wang

November 29, 2023

In the dynamic landscape of distributed computing, the exponential growth in traffic demands within data centers and high-performance computing systems has been distinctive, fueled by a deluge of data-intensive workloads. This trend is prominently exemplified by the rapid expansion of machine learning, big data analytics, and most notably, deep learning (DL)–driven artificial intelligence (AI) applications. The recent advent of large language models, which has revolutionized natural language processing and creative content generation, is propelling the broad adoption of ever-larger DL models and datasets, marking a significant milestone toward the era of data ubiquity. The continued scaling of these applications has pushed the limits of computational hardware, notably via increased parallelism and specialization. Yet, this rapid progress has outpaced the evolution of the underlying communication infrastructure, rendering chip-to-chip data movement a formidable barrier impeding performance and energy efficiency. This communication bottleneck has become the grand challenge to the quest of upscaling the computing systems toward exascale.

My research endeavors to find **transformative connectivity solutions**, maximally harnessing the potential of integrated silicon photonics (SiPh). In this pursuit, I have devised a dual-thrust research agenda for my independent career. The first thrust focuses on **reconfigurable system connectivity**. It aims to develop optical interconnects that not only provide unprecedented bandwidth but also adapt in real-time to the ever-evolving demands of emerging applications. The second thrust looks into **innovative system architectures**. It targets redefining chip-to-chip communication with groundbreaking optical I/O technologies, thereby pioneering new computational paradigms and interconnect functionalities. The synergy of these thrusts introduces unique design challenges, which I am equipped to tackle with my interdisciplinary research experience, ensuring the readiness of essential design tools and methodologies for these advanced connectivity solutions.

My research agenda is situated at the system level, squarely fitting into the interdisciplinary nexus of the broad area of computer engineering that bridges silicon photonics, networked systems, computer architecture, design automation, and AI/machine learning. I look forward to the possibility of collaborating with the esteemed colleagues in the Department of Electrical and Computer Engineering at the Duke University to confront the grand challenge of data movement in future computing infrastructures across the full system stack.

1 Research Accomplishments and Skills

My doctoral and postdoctoral work have established a solid foundation for my anticipated research and equipped me with the skills necessary to address the upcoming research challenges. I was among the first to integrate accurate **compact models and simulation methodologies** of silicon photonic devices into widely used electronic design automation (EDA) platforms like Cadence Virtuoso [1, 2]. This integration is crucial for the accurate co-simulation of electronic and photonic components, enabling the efficient development of complex photonic integrated circuits (PICs) with reliable performance estimation. My expertise in **process variation characterization, mitigation, and tolerance** ensures the robustness and energy efficiency of fabricated designs [3–5]. This is particularly important for advanced technologies that often rely on emerging fabrication processes and require post-fabrication tuning. My work in these areas has been recognized in leading design automation conferences like DAC and ICCAD, respected photonics venues like OFC and JLT, and a forthcoming book chapter with Springer, effectively connecting the electronics and photonics research communities.

These design enabling techniques have been practically utilized in creating two generations of SiPh transceiver chips, featuring a **scalable link architecture** that facilitates unprecedented channel parallelism and delivers a chip I/O bandwidth of over 16 Tbps with energy consumption below 1 pJ/b [6, 7]. Fabricated in partnership with AIM Photonics through two full-wafer runs, each chip, measuring ~70 mm², densely integrates over 2,000 microresonators. The chip layout process was fully scripted and automated, showcasing not only significant technological advancements but also remarkable design efficiency. The highlighted link architecture was instrumental in securing a $35M SRC JUMP 2.0 grant with 23 principal investigators, a program to which I have contributed through proposal writing and ongoing research efforts. This work has also resulted in invited papers and presentations at both photonics and electronics design conferences (Photonics West and CICC), and an invited journal submission to *Nature Communications Physics*.

These accomplishments have advanced my research into exploring **traffic adaptability** for optical interconnects in distributed computing systems, grounding them in credible performance models and hardware validation. Notably, I have delved into runtime adjustments of parameters such as laser power and link bandwidth, aiming at accelerating distributed machine learning applications with reduced energy consumption [8–10]. These investigations underscore the significance of integrating architectural innovations and optimization strategies at the system level, a process which—without meticulous execution—could inadvertently counteract the advancements achieved at both device and link levels. This realization is a key driver behind my future research directions.

2 Research Vision and Agenda

My research agenda is set to continue at the system level, leveraging the latest breakthroughs at the device and link levels, while simultaneously informing their future advancements from a system application perspective. In light of the evolving data landscape, I plan to focus on two synergistic research thrusts in pursuit of groundbreaking connectivity solutions.

2.1 Thrust 1: Reconfigurable System Connectivity

With the advent of augmented reality (AR), virtual reality (VR), and Metaverse applications, distributed machine learning frameworks are seeing an increase in data privacy concerns that were previously confined to sectors with sensitive information, such as banking and healthcare. These sectors typically handle smaller volumes of data with more flexible latency requirements. In response, decentralized learning frameworks like federated learning have received growing popularity, as they allow the exchange of model parameters over raw data. Yet, certain applications still prioritize data parallelism to meet stringent requirements on model accuracy. Consequently, the data landscape in distributed computing is evolving toward both larger volumes and greater heterogeneity. This evolution, coinciding with the expansion of large models like GPTs, necessitates the next generation of optical interconnects to further excel in traffic adaptability, in addition to bandwidth and energy efficiency.

In this thrust, I will work on to greatly enrich the traffic adaptability through a co-design of reconfigurable link architectures and runtime reconfiguration strategies. I will build on the SiPh transceiver that I have developed in my postdoctoral studies, which provides a good starting point that is high-bandwidth and energy-efficient.

2.2 Thrust 2: Innovative System Architectures

3 Research Collaborations and Initiatives

My research experience has been deeply rooted in multidisciplinary collaboration, a skill I mastered during my postdoctoral training at the Columbia University. There, I led research initiatives within our group, benefiting from the mentorship of my supervisor and backed by funding from agencies like DARPA, SRC, and ARPA-E. These initiatives required seamless teamwork with colleagues from academia, industry, and governmental bodies. In addition, I have a proven track record in assisting both my doctoral advisor and postdoctoral supervisor with fundraising activities. My responsibilities also encompassed preparing and compiling reports and materials, as well as participating in presentations at quarterly reviews to fulfill the requirements of our funded projects.

4 Thrust 1: Traffic-Adaptable Optical Interconnects

This research direction will leverage the pioneering work on integrated silicon photonics chip I/O from my postdoctoral studies [6], which achieved ultra high-bandwidth and low-energy transmission through a scalable dense wavelength division multiplexing (DWDM) link architecture. Additionally, it will build upon preliminary investigations into bandwidth reconfiguration within distributed deep learning environments, where previously, traffic patterns exhibited less temporal variations [10].

The successful advancement of this research thrust will involve the following critical tasks:

1. Develop and incorporate a runtime reconfiguration module within the DWDM link architecture, facilitating dynamic bandwidth allocation that adjusts to varying traffic patterns and specific application needs. A preliminary off-chip prototype that divides wavelength channels between two ports has shown promise and is undergoing publication review, with an on-chip version already sent to AIM Photonics for fabrication and slated for testing in April 2024. Future iterations will focus on expanding port numbers, fine-tuning splitting ratios, accelerating reconfiguration times, and enabling channel reassignment.
2. Profile the traffic patterns of a selection of key distributed computing applications to shape the development of runtime reconfiguration strategies.
3. Conduct system-level simulations to assess the energy and performance impacts of the proposed runtime reconfiguration strategies.
4. Implement and test the envisioned interconnect architecture and reconfiguration strategies on a hardware testbed, using real production network traces to validate the approach.

By addressing the critical need for adaptive, high-performance optical interconnects that can keep pace with the ever-increasing demands of modern computing applications, this research thrust has the potential to reshape the landscape of future distributed computing infrastructures with tangible improvements in efficiency and versatility.

5 Thrust 2: 3D Optoelectronic Architectures

State-of-the-art accelerator systems, composed of clusters of computing units (CUs), are confronting a "memory wall" caused by the significant disparity between the bandwidth for intra- and inter-cluster communications. The option to expand the number of on-chip high-bandwidth memory (HBM) stacks is becoming less viable as the bandwidth capacity of electronic interposers nears saturation. Traditional approaches using optical fibers for interconnecting memory pools are impractical for densely arranged CU clusters due to the size and pitch limitations of fiber arrays. Nonetheless, the emerging concept of 3D optical I/Os, which facilitate dense waveguide routing in multiple layers, could unlock new possibilities for scaling up CU clusters with optical connectivity achieved directly through waveguides. Having contributed to the preparation for the concept's showcase at the 2023 DARPA ERI Summit, I am inspired to pursue this avant-garde research thrust, which has the potential to pioneer a novel computing architecture through a deeply integrated electronic-photonic synergy. I look forward to exploring several key research topics, in collaboration with field experts, including:

1. The device-level design and optimization of multi-layered 3D optical I/O modules for high-density, low-loss, and compact optical interfaces, incorporating innovative coupling mechanisms.
2. The architectural investigation of optically interconnected accelerator systems featuring dense, fiber-less connectivity to delineate the optimal configuration for system-level designs.
3. The pursuit of novel functionalities for on-chip silicon photonics, leveraging enhanced density and routing capabilities to perform computational tasks, extending beyond traditional data communication roles.

The fruition of this research direction promises not just to scale accelerator systems in alignment with the computational demands of the future, but also to broaden the scope for on-chip optical interconnects to assume a more dynamic and integral role in computing architectures.

6 Thrust 3: Design Automation for Future Integrated Photonics

Successfully navigating the design and optimization intricacies presented by the proposed research thrusts is crucial for actualizing the advanced connectivity solutions I envision. The anticipated challenges include:

1. Developing efficient yet accurate modeling and simulation methodologies for the envisioned connectivity solutions at device, circuit, and system levels, democratizing the design process and enabling rapid prototyping and design optimization.
2. Characterizing and mitigating process variations, along with creating designs that are robust against fabrication inconsistencies, especially crucial for the experimental processes involved in the second research thrust.
3. Creating novel design enablement technologies that exploit machine learning and artificial intelligence to expand the design capabilities for future integrated photonics.

With the cross-disciplinary design ecosystem that I have cultivated during my academic journey [5] [3] [4], I am poised to leverage my expertise to forge a comprehensive design automation framework for the next generation of integrated silicon photonics interconnect systems.

I am eager to bring my expertise and enthusiasm to your esteemed university, where I look forward to collaborating with a community that shares my dedication to innovation and to making a meaningful impact on the future of technology.

References

[1] R. Wu, Y. Wang, Z. Zhang, C. Zhang, C. L. Schow, J. E. Bowers, and K.-T. Cheng, "Compact modeling and circuit-level simulation of silicon nanophotonic interconnects," in *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017*. Lausanne, Switzerland: IEEE, Mar. 2017, pp. 602–605.

[2] Z. Zhang, R. Wu, Y. Wang, C. Zhang, E. J. Stanton, C. L. Schow, K.-T. Cheng, and J. E. Bowers, "Compact Modeling for Silicon Photonic Heterogeneously Integrated Circuits," *Journal of Lightwave Technology*, vol. 35, no. 14, pp. 2973–2980, Jul. 2017.

[3] Y. Wang, J. Hulme, P. Sun, M. Jain, M. A. Seyedi, M. Fiorentino, R. G. Beausoleil, and K.-T. Cheng, "Characterization and Applications of Spatial Variation Models for Silicon Microring-Based Optical Transceivers," in *2020 57th ACM/IEEE Design Automation Conference (DAC)*. IEEE, Jul. 2020, pp. 1–6.

[4] Y. Wang, P. Sun, J. Hulme, M. A. Seyedi, M. Fiorentino, R. G. Beausoleil, and K.-T. Cheng, "Energy Efficiency and Yield Optimization for Optical Interconnects via Transceiver Grouping," *Journal of Lightwave Technology*, vol. 39, no. 6, pp. 1567–1578, Mar. 2021.

[5] Y. Wang, S. Wang, A. Novick, A. James, R. Parsons, A. Rizzo, and K. Bergman, "Dispersion-Engineered and Fabrication-Robust SOI Waveguides for Ultra-Broadband DWDM," in *Optical Fiber Communication Conference (OFC) 2023*. Optica Publishing Group, 2023, p. Th3A.4.

[6] Y. Wang, A. Novick, R. Parsons, S. Wang, K. Jang, A. James, M. Hattink, V. Gopal, A. Rizzo, C.-P. Chiu, K. Hosseini, T. T. Hoang, and K. Bergman, "Scalable architecture for sub-pJ/b multi-Tbps comb-driven DWDM silicon photonic transceiver," in *Next-Generation Optical Communication: Components, Sub-Systems, and Systems XII*, G. Li, K. Nakajima, and A. K. Srivastava, Eds. SPIE, Mar. 2023, p. 55.

[7] Y. Wang, S. Wang, R. Parsons, A. Novick, V. Gopal, K. Jang, A. Rizzo, C.-P. Chiu, K. Hosseini, T. T. Hoang, S. Shumarayev, and K. Bergman, "Silicon photonics chip I/O for ultra high-bandwidth and energy-efficient die-to-die connectivity," in *2024 IEEE Custom Integrated Circuits Conference (CICC)*. IEEE, 2024, under review.

[8] Y. Wang and K.-T. Cheng, "Task Mapping-Assisted Laser Power Scaling for Optical Network-on-Chips," in *2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. Westminster, CO, USA: IEEE, Nov. 2019, pp. 1–6.

[9] Y. Wang and K.-T. Cheng, "Traffic-Adaptive Power Reconfiguration for Energy-Efficient and Energy-Proportional Optical Interconnects," in *2021 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*. Munich, Germany: IEEE, Nov. 2021, pp. 1–9.

[10] Z. Wu, L. Y. Dai, Y. Wang, S. Wang, and K. Bergman, "Flexible silicon photonic architecture for accelerating distributed deep learning," *Journal of Optical Communications and Networking*, 2023, to appear.