Research Statement

Yuyang Wang

November 15, 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 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 craft **transformative connectivity solutions**, maximally harnessing the potential of integrated silicon photonics. To achieve this, I have devised a three-pronged approach for my proposed research agenda in the Department of Electrical and Computer Engineering at the Duke University. The first area of focus aims to accommodate the diverse and ever-changing data demands stemming from various emerging applications, calling for connectivity that is not only high-bandwidth but also runtime-reconfigurable. The second area re-envisions the conventional computing architectures in light of the expected advancements in dense, on-chip 3D optical I/Os, thereby pioneering new computational paradigms and interconnect functionalities enhanced by the manifolded reach of on-board optical connectivity. The third area reinforces the first two by tackling the design and optimization challenges that arise, thereby providing the necessary design tools and methodologies for the proposed connectivity solutions.

Research Foundation

My interdisciplinary research vision is positioned at the intersection of electronics/photonics, devices/systems, and design/applications. It necessitates a collaborative synthesis across various sub-disciplines, which I conceptualize along two orthogonal dimensions, as depicted in Fig. 1. The comprehensive skill set acquired through my doctoral and post-doctoral experiences has allowed me to integrate design methodologies and optimization techniques in both vertical and horizontal facets, thereby laying a solid groundwork for the research agenda that I propose. The expertise and insights gained from these endeavors will seamlessly inform and underpin the trio of research thrusts that I will elaborate as following.

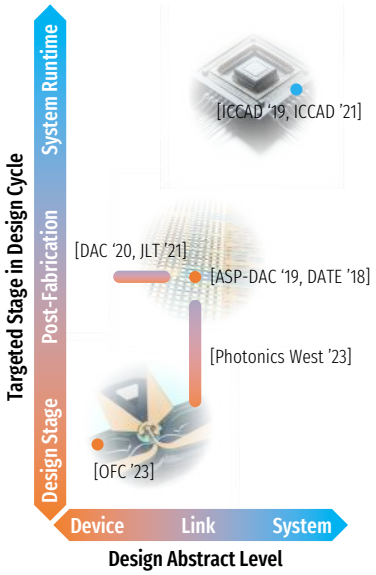

Figure 1: Cross-level design enablement and optimization techniques serving as research foundation.

Thrust 1: Traffic-Adaptable Optical Interconnects

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 been preferred, 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 not only larger volumes but also greater heterogeneity. This shift necessitates the next-generation of optical interconnects to provide traffic-adaptable runtime reconfiguration capabilities, in addition to high bandwidth capacities, to accommodate these evolving demands.

This research direction will leverage the pioneering work on integrated silicon photonics chip I/O from my postdoctoral studies[1], 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[2].

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.

[1] 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

[2] 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

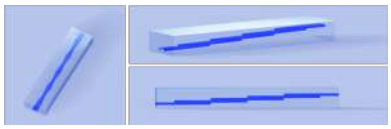

Figure 2: Conceptual rendering of compact 3D waveguide routing enabled by multiple layers of low-loss adiabatic tapers.

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

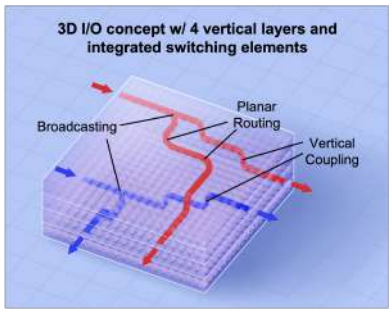

Figure 3: Conceptual rendering of a 3D optical I/O featuring switching and broadcasting capabilities for flexible die-to-die connectivity.

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.

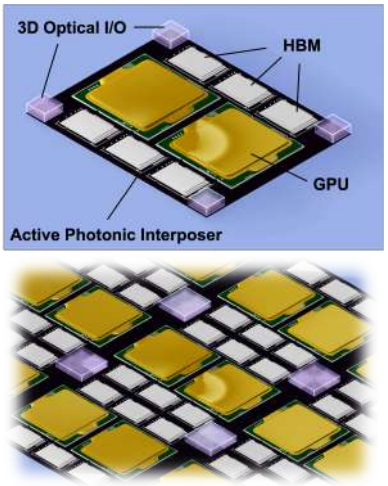

Figure 4: Conceptual rendering of densely interconnected computing units featuring fiber-less optical connectivity.

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[3,4,5], I am poised to leverage my expertise to forge a comprehensive design automation framework for the next generation of integrated silicon photonics interconnect systems.

Research Collaborations and Initiatives

My experience has been deeply rooted in multidisciplinary collaboration, a skill I mastered during my postdoctoral training at Columbia University, where I spearheaded research initiatives within our group, guided by my supervisor, and fueled by DARPA and SRC funding. These initiatives necessitated seamless teamwork with colleagues from academia, industry, and governmental bodies. Additionally, I have a proven track record of supporting my advisor in fundraising endeavors, notably contributing to the writing and visual elements of a successful $35M SRC JUMP 2.0 grant application. My role also extended to compiling regular reports and presenting at quarterly reviews to meet the obligations of our funded projects.

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.

[3] 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

[4] 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

[5] 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