Research Statement

Yuyang Wang

The evolving field of distributed computing is seeing an exponential growth in traffic demands, fueled by a deluge of data-intensive workloads. This trend is prominently exemplified by the rapid expansion of data-intensive artificial intelligence (AI) applications. The recent advent of large language models is propelling the broad adoption of ever-larger deep learning 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, yet outpacing the evolution of the underlying communication infrastructure, rendering chip-to-chip data movement a formidable barrier impeding performance and energy efficiency. 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. My research agenda is situated at the system level, leveraging the latest breakthroughs in device designs and link architectures, while simultaneously informing their future advancements from a system application perspective. This cross-layer approach introduces unique design challenges, which I am equipped to tackle with my interdisciplinary research experience (Fig. 1), ensuring the readiness of essential design tools and methodologies for these advanced connectivity solutions.

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 research thrust, my objective is to significantly enhance traffic adaptability by co-designing reconfigurable link architectures along with dynamic reconfiguration strategies (Fig. 1-T1). Building upon the SiPh transceiver developed during my postdoctoral research [1, 2]—which stands out for its leading bandwidth capacity and energy efficiency among state-of-the-art solutions—I aim to incorporate greater reconfigurability into its design. My prior work, namely on runtime laser power scaling and link bandwidth reconfiguration [3–5], serves as a proof-of-concept for the effectiveness of traffic-adaptable tuning knobs in improving both the performance and the energy efficiency of optically connected computing systems. Moving forward, I anticipate the success of this research thrust to be contingent on the following critical tasks:

1. Profiling and characterizing the traffic patterns of a diverse range of distributed computing applications, expected to exhibit greater heterogeneity and temporal dynamics compared to the collective communications typically observed in current computing clusters, as referenced in [5].

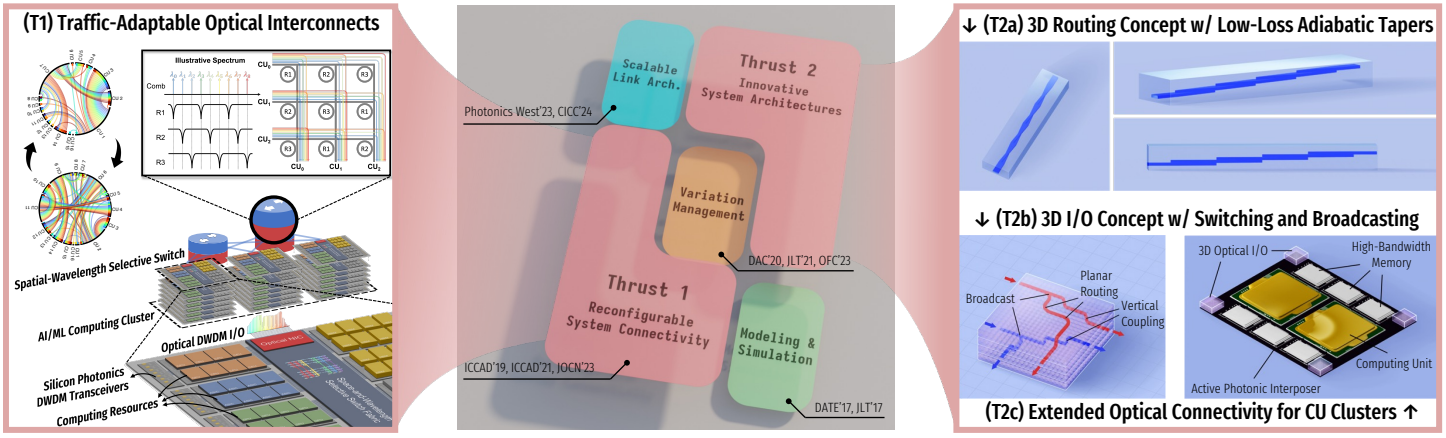

Figure 1: Overview of my research accomplishments and proposed research directions.

2. Introducing additional reconfigurable parameters beyond laser power and link bandwidth, such as wavelength allocation and switching/routing, and developing runtime reconfiguration strategies tailored to these characterized traffic patterns.
3. Conducting system-level simulations to assess the energy and performance impacts of the proposed reconfigurability, supported by credible performance models that accurately reflect real link designs.
4. Designing and integrating reconfiguration modules with state-of-the-art SiPh transceiver implementations, and validating the enhanced interconnect architecture with reconfiguration strategies on a hardware testbed driven by realistic/production network traces.

Throughout this endeavor, I also anticipate deriving valuable insights from a system application perspective. These insights will be instrumental in informing the design of SiPh devices and circuits, focusing on essential aspects such as tuning range and reconfiguration speed, to meet key performance metrics at the system level. This collaborative synergy across multiple design hierarchies is essential to maintain cutting-edge system connectivity in an ever-changing data landscape.

2 Thrust 2: Innovative System Architectures

Complementary to the first research thrust aimed at advancing chip-to-chip connectivity, the second thrust strives to address the notable gap between on-chip and off-chip communication bandwidths. This gap is particularly pronounced in accelerator systems comprising clusters of computing units (CUs) that frequently access data from both on-chip memory banks and off-chip memory pools. Expanding the number of on-chip high-bandwidth memory (HBM) stacks is increasingly impractical as the bandwidth capacity of electronic interposers approaches its limits. Conventional approaches using optical fibers to connect CU clusters and memory pools are also constrained by the size and pitch of fiber arrays. Nonetheless, the emerging concept of 3D optical I/Os, benefiting from dense waveguide routing across multiple layers, presents a promising avenue to scale up CU clusters with optical connectivity that stays on-board with extended reach (Fig. 1-T2a–c). My contribution to assisting the formulation of this concept, which was successfully showcased at the 2023 DARPA ERI Summit, has inspired me to further explore this cutting-edge area. The key challenges I plan to address in this research thrust include:

1. Formulating the 3D routing problem with objectives such as maximized density and minimized loss, and developing efficient routing algorithms that draw from traditional EDA expertise and the latest in machine learning techniques.
2. Informing the design of 3D routing elements with performance and area constraints, and optimizing their physical design employing recent advances in areas such as photonic inverse design and topology optimization.
3. Conducting system-level design space explorations for computing architectures with transformed memory connectivity to delineate optimal system configurations, such as the ideal size of CU clusters that benefit from the expanded reach of on-board optical connectivity, and the optimal balance between on-chip and off-chip memory capacities.

In addition to eliminating the bandwidth taper at chip boundaries and allowing for continued upscaling of CU clusters, this research thrust also promises to expand the role of optical interconnects beyond traditional data communication. For instance, certain computational tasks, such as matrix multiplication, can be offloaded to the optical domain, for which existing explorations have been limited by the vast difference in physical dimensions of electronic and photonic implementations. This limitation can be significantly alleviated by the manifolded density of optical components enabled by 3D routing. This thrust, therefore, not only addresses current technological limitations but also fosters the development of new computing paradigms, where optical interconnects assume a more dynamic and integral role in future computing 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, guided by my supervisor's mentorship 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. Given the interdisciplinary essence of my research agenda, I am enthusiastic about the opportunity to collaborate with the diverse faculty in the School of Computing and Augmented Intelligence and contribute my experience and enthusiasm to your esteemed institution. I keenly anticipate the chance to work with a community that resonates my commitment to innovation and making a meaningful impact on the future of technology.

References

[1] Y. Wang *et al.*, "Scalable architecture for sub-pJ/b multi-Tbps comb-driven DWDM silicon photonic transceiver," in *SPIE Photonics West*, Mar. 2023.

[2] Y. Wang *et al.*, "Silicon photonics chip I/O for ultra high-bandwidth and energy-efficient die-to-die connectivity," in *IEEE Custom Integrated Circuits Conference (CICC)*, 2024, under review.

[3] Y. Wang *et al.*, "Task Mapping-Assisted Laser Power Scaling for Optical Network-on-Chips," in *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, Nov. 2019.

[4] Y. Wang *et al.*, "Traffic-Adaptive Power Reconfiguration for Energy-Efficient and Energy-Proportional Optical Interconnects," in *IEEE/ACM International Conference On Computer Aided Design (ICCAD)*, Nov. 2021.

[5] Z. Wu, L. Y. Dai, Y. Wang *et al.*, "Flexible silicon photonic architecture for accelerating distributed deep learning," *J. Opt. Commun. Netw.*, 2023, to appear.