Online Interview Notes

Yuyang Wang

January 15, 2024

Contents

1 Questions

1.1 Introduce your research at high-level; elaborate on the importance of your research.

- My research have been motivated by the communication bottleneck in today's distributed computing systems, which is becoming increasingly more significant because of the booming development of data-intensive AI/machine learning applications.
- The communication bottleneck is most exemplified at the chip boundary, where currently there is usually orders of magnitude discrepancy in terms of communication bandwidth between on-chip and off-chip links.
- My research therefore focuses on alleviating this communication bottleneck by integrating optical I/O deeply into the compute chip.
- The technology that I have extensively explored during my PhD and postdoc careers is dense wavelength-division multiplexing, or DWDM, based on silicon photonics micro-resonator devices.
- In particular, my postdoc research has focused on developing an optical link architecture that is scalable to hundreds of parallel wavelength channels and achieve a multi-Tbps aggregated data rate with only a moderate data rate per channel, which is the key to keeping the energy consumption low.
- Based on this link architecture, I also led the design of an optical transceiver chip, where transceiver stands for transmitter and receiver, that was designed for 3D integration with an electronic driver and co-packaging with the compute socket.
- This was a collaborative effort with another academia research group that designed the electronic driver circuitry, and an industry partner who leads in advanced packaging.
- Expanding from the link architecture, my research also encompasses the techniques that are essential to enable the design of such interconnect systems, including the modeling, simulation, and optimization across various abstract levels.
- For example, at the system level, in addition to providing a large total bandwidth, I also looked at the dynamics of the traffic in some published data center network traces, and proposed runtime power control strategies for the optical links for them to stay energy efficient in such a dynamic setting.
- I also investigated design automation methodologies at lower levels, including the modeling and simulation of photonic devices and circuits, and the characterization and management of fabrication process variations, with the goal of optimizing the system performance. I could go into more details with any of the works that I've mentioned if you're interested.
- In short, my past research was aimed at building a framework for the design of high-bandwidth and energy-efficient optical interconnects. This includes the hardware design, as well as the toolkits to tackle the associated design and optimization problems along the way.
- Meanwhile, I have also developed concrete skills to collaborate with people who complements my skills and expertise, which are needed for cross-disciplinary research at the system level.

1.2 What is your future research plan? What will you pursue in your research agenda?

- I plan to continue the investigation of optical interconnect technologies for applications at the system level, more deeply associated with the evolving data characteristics resulted from emerging computing applications, such AI/machine learning, edge/ubiquitous computing, etc.
- The key motivation is the growth of the traffic in future computing systems in terms of both volume AND heterogeneity, that calls for more dimensions and finer granularity of network reconfiguration.
- In addition to my PhD work, I have worked with a student that I mentored to add bandwidth reconfigurability to our transceiver architecture and study its implications to an optically connected compute cluster running distributed deep learning workloads.
- I imagine my future work in this direction would involve:
  - The profiling and characterization of traffic patterns from a more diverse range of computing applications
  - Leveraging the massive wavelength parallelism, to investigate additional reconfiguration knobs, such as the allocation of wavelengths (or physical channels) to various logic channels; and network functionalities such as multi-casting or broadcasting that can be achieved by putting the same data on a subset of the wavelengths and send them to different destinations.
- Another research direction that I plan to pursue in parallel is also at the system level, but addresses the communication bottleneck at the on-chip/off-chip interface from a different perspective.
- It aims at further increasing the I/O density, which is currently limited by the large pitch requirement of optical fiber arrays. Specifically, there have been some recent effort routing optical signals across multiple layers of waveguides. While I look forward to collaborating with device design experts to improve the loss of the vertical coupling and make the coupling elements as compact as possible, I also intend to consider it an enabling technology, and look into its system-level applications, such as having a multi-layered optical I/O that immediately manifolds the I/O bandwidth density.
- Furthermore, with the potential application of some emerging optical packaging technologies, such as photonic wire bonding, we could imaging having the compute chips, equipped with the 3D optical I/Os, sitting closely next to each other and connects through photonic wire bonds without the need for going into bulky optical fibers.
- Another option is to combine this with die-to-wafer bonding, and directly have multiple compute dies optically connected at wafer-level
- It is also a potential pathway to circumventing the silicon interposers currently used connect multiple compute chips on the same board, which is also reportedly reaching its bandwidth limit.
- And I see potential applications of such connectivity technology in emerging computing system architectures such as resource disaggregation.
- There are also some other research directions that I have been thinking of.
  - With the growth of edge and ubiquitous computing, there's an motivation of combining optical I/O with wireless communication, especially the receiving end of the massive antenna arrays. The limit on electrical pin numbers and the signal processing overhead prevent the number of antenna elements of a single 2-D array from scaling beyond 1024 or so. And there are initial looks into enabling the further scaling of antenna array systems through an optically connected backplane.

1.3 What are your most important accomplishments/work that you are most proud of?

- The work that I have been doing since my postdoc appointment—A silicon photonics optical link architecture, leveraging dense wavelength-division multiplexing (DWDM), to allow for massive wavelength parallelism and scalability, and achieve ultra-high bandwidth and energy efficiency for chip-to-chip communication.
- Significance: design with 3D integration and co-packageability in mind. Proof-of-concept for bringing optical I/O into the compute socket.
- Importance to my research agenda: at the center of my research; integrated much of the skills and expertise acquired during my PhD.