Silicon Photonic Multi-Chip Module Interconnects for Disaggregated Data Centers

Nathan C. Abrams
Department of Electrical Engineering
Columbia University
New York, NY, USA
nca2123@columbia.edu

Madeleine Glick
Department of Electrical Engineering
Columbia University
New York, NY, USA
msg144@columbia.edu

Keren Bergman
Department of Electrical Engineering
Columbia University
New York, NY, USA
bergman@ee.columbia.edu

***Abstract*—We present our development of 2.5D integrated multi chip module silicon photonic transceivers for disaggregated applications, such as big data and machine learning algorithms. Disaggregation of data center resources to improve application efficiency and performance can be achieved through photonic switching networks. Our four-channel transceiver provides the electro-optic interface between electrical data generation and photonic switching to enable disaggregation within data centers.**

Keywords—Silicon Photonics, Multi-Chip Module, Interposer, 2.5D Integration, Optical Interconnects, Disaggregation

I. INTRODUCTION

The combination of an exponential increase in global internet traffic with unsustainable increases in data center energy consumption [1] has accelerated the exploration of alternative data center architecture and interconnect technologies. Current resources in datacenters are largely organized according to legacy architectures with static configurations at the rack and server level. These static configurations of resources (compute, memory, storage) often result in the inefficient use of resources, with some being left idle while others are overtaxed. It has been shown that static networks are not the most efficient organization for data center traffic [2, 3]. The variable resource requirements will increase with increasing traffic devoted to machine learning algorithms where the requirements for the different stages of the machine learning algorithms of training and inference use significantly different mixes of compute and memory resources [4]. Disaggregation of the traditional server has been proposed as a solution to improve efficiency [5], in which similar resources are pooled, with the possibility of the resources being adaptively configured for optimized performance. Disaggregation also enables the economic advantage of independently upgrading resources (CPU, GPU, memory) which follow differing generation lifetimes rather than replacing entire servers. Disaggregation however, requires a high bandwidth, low latency interconnection fabric to carry the inter-resource traffic that in addition does not unduly disturb system performance [5]. The low latency interconnection fabric can be achieved with silicon photonic circuit switching fabrics, as silicon photonic switches support both high bandwidth and nanosecond reconfigurability [6]. Interfacing to the silicon photonic switches requires integrated transceivers to convert the electrical data to the optical domain of the silicon photonic switch and vice versa. Not only are optical transceivers needed to interface to optical switches, but electrical interconnects are being pushed to their limits where increasing the data rates also increases attenuation and requires additional equalization circuitry to address increased intersymbol interference. Silicon photonics is also an attractive solution for transceiver development, combining minimal signal attenuation, high modulation rates, energy efficiency, and parallelization through wavelength division multiplexing, all while building upon the mature CMOS ecosystem.

Integration of the photonics and electronics will enable higher bandwidth, lower power interconnects. The packaging and integration comes in various flavors with their own trade-offs as described in more detail below. Recent developments in electronic- photonic convergence are the interposer and the multi-chip module (MCM) [7, 8]. The interposer is a key enabling component for small footprint, low power, multi-terabit MCM's with optical interconnection of CPU, GPU, memory components.

II. INTEGRATION AND PACKAGING

For silicon photonics to be embedded in compute nodes and widely adopted into datacenters, careful consideration needs to be placed on how the photonics are integrated with both the driving electronics and computational electronics. Improper integration can nullify all the potential benefits of silicon photonics. If the integration introduces too large a parasitic capacitance or inductance, the electro-optic bandwidth could be impacted due to presence of parasitic poles. This may result in the need for equalization circuits which increase the energy consumption of the transceivers, and in extreme cases could ultimately limit the bandwidth of the transceiver. Impedance mismatch from improper integration can introduce signal

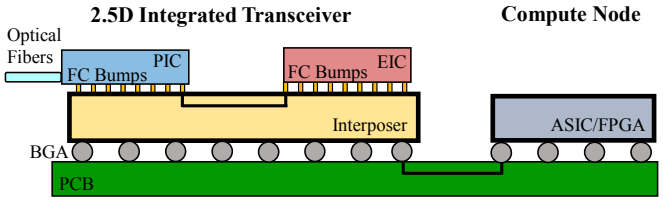

Fig. 1. The overview of how a 2.5D integrated MCM transceiver would interface to the compute node, using a thinned interposer with TSVs and BGA type connections to connect to a packaged substrate or PCB.

This work was supported in part by the U.S. Department of Energy Advanced Research Projects Agency—Energy under ENLITENED Grant DE-AR000843

978-3-903176-21-8 (c) 2020 IFIP

978-3-903176-21-8/20/$31.00 ©2020 IEEE

reflections, resulting in increased noise and decreased driving voltage.

Multiple approaches exist for integrating silicon photonic integrated circuits (PICs) with driving electronic integrated circuits (EICs). 2D integration is where the PIC and EIC are placed side by side and connected with wirebonds. While 2D integration is the simplest integration approach, the wirebond inductance can introduce significant parasitics, on the order of 1 nH/mm [9] and I/O bandwidth is limited as connections can only be made between one shared edge. 3D integration is where the EIC is flipped on top of the PIC, with connections being achieved with dense pitch copper pillars or solder bumps. Pitches have been demonstrated on the order of 50 μm [10], allowing for higher aggregate I/O bandwidth compared to 2D integration. Additionally, 3D integration is beneficial as the microbumps have minimal parasitics compared to 2D integration—typically on the order of 30 fF of parasitic capacitance [10]. A drawback of 3D integration is the I/O bandwidth from the compute EIC to the driving EIC. The I/O connections are typically achieved with wirebonds from the perimeter of the PIC, meaning the number of I/O connections are limited and will suffer from wirebond parasitic inductance. 2.5D integration is where the PIC and EIC are both flipped on top of an interposer using dense pitch copper pillars or solder bumps. The interposer serves as the connection interface between the PIC and EIC and serves to interface to the package substrate of the compute EIC, as seen in figure 1. Interposers can be constructed from a variety of materials—common choices are silicon, glass, and organic substrates. One benefit of the silicon interposer is that it allows for the fabrication of through silicon vias (TSVs) to provide connectivity between the front and back of the interposer if the interposer is thinned. A silicon interposer also supports the addition of a silicon nitride layer to allow for the fabrication of waveguides, allowing the interposer to provide both electrical redistribution and optical redistribution. 2.5D integration supports relatively high I/O connections between the PIC and driving EIC as well as from driving EIC to the compute EIC. Flip chipping allows for the full area of the PIC and driving EIC to be used for I/O connections and utilizing TSVs in the interposer allows for I/O connections to the compute EIC to be placed across the full area of the interposer. The parasitics for 2.5D integration are relatively low, but will be higher than 3D integration, as the connections between the PIC and EIC will require two bumps and an interposer trace compared to a single bump for 3D integration. An active interposer is an intersection between 2.5D integration and 3D integration. In an active interposer, active photonic components are fabricated within a thinned silicon interposer. The active interposer allows for similarly low parasitics as 3D integration, as the EIC can be flipped on top of the active interposer, but also allows for high I/O bandwidths between the MCM transceiver and the compute electronics, as the entire back of the active interposer can be for I/O connections. A final integration approach is monolithic integration, where the PIC and driving EIC are fabricated in the same process. The advantage of monolithic integration is reduction of the parasitics between the PIC and driving EIC to a minimum, as the photonics and electronics are fabricated in the same die. The disadvantages of monolithic integration are high development cost and reduced performance compared to separate photonic and electronic fabrication processes. For monolithic integration, the most common approach is to use a larger electronic node size to enable good photonic performance, which eliminates the ability to use cutting edge electrical nodes, such as 14 nm and below.

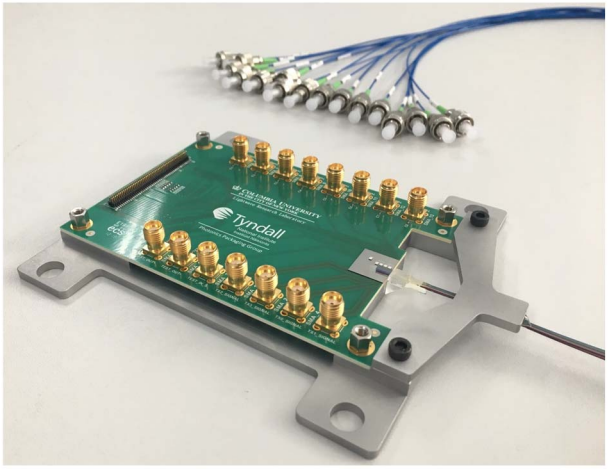

Fig. 2. The assembled MCM prototype. The PIC and EICs are flipped on top of the silicon interposer, which is placed on the edge of a PCB. A fiber array couples to the edge couplers on the PIC.

III. MCM Transceiver Design

A. Architecture

The PIC architecture is based on resonant microdisks coupled to bus waveguides. On the transmit side, four microdisk modulators are used, each fabricated with a different resonance so that the resonances are evenly spaced out over the free spectral range (FSR) to enable wavelength division multiplexing (WDM). The modulators have a reverse biased diode for depletion modulation and an integrated resistor for thermal tuning. On the receiver side, four microdisk demuxes are coupled to a bus waveguide. The drop port of the demuxes route to a high-speed photodiode. The demuxes have an integrated heater for thermal tuning. The PIC was fabricated on a multi project wafer (MPW) run through AIM Photonics using the AIM process design kit (PDK).

Each channel is targeted to operate at 10 Gbps. Increasing the total transceiver throughput will be achieved by increasing the number of cascaded modulators. While the number of microdisks that can be coupled to a single bus waveguide will be limited by the microdisk's FSR and insertion loss, scaling to higher channels can be achieved by using de-interleavers to direct bands of channels to separate bus waveguides before being interleaved back together. The benefit of scaling with relatively low channel rates is that it presents a path for high throughput transceivers with low energy consumption. The relatively low channel data rate reduces the need for SERDES and digital signal processing, which add significant energy consumption when required in the driving EICs. With large channel counts, it becomes architecturally feasible to dedicate one channel for clock forwarding, removing the need for clock recovery circuitry at the receiver and further reducing the EIC's energy consumption.

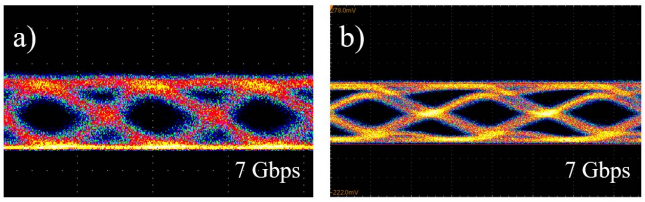

Fig. 3. The eye diagrams from an example channel of the transmitter (a) and receiver (b), both at 7 Gbps.

B. Integration

The MCM transceiver utilized 2.5D integration to provide the interconnection between the PIC and EIC. A custom silicon interposer was fabricated at SUNY CNSE with TSVs to provide connection between the front and back of the interposer. The EICs used for the transceiver were commercial Texas Instruments single channel TIAs and were designed for channel rates up to 11.3 Gbps. To interface to the four channel PIC, four separate bare die TIAs were flipped on to the interposer. Stud bumps were used to flip chip the PIC and the EICs to the interposer. Connections between the PIC and EIC were kept below the λ/4 length of the 10 GHz signals to reduce the impact of reflections due to impedance mismatch. Signals were routed to the back side of the interposer through TSVs, where ball grid array (BGA) type connections were used to provide connections to a PCB for further fanout for both DC and RF signals. The optical connection to the PIC was achieved via edge couplers on the PIC, which were coupled to a fiber array. The PIC overhangs off the interposer by 100 μm so that there is a visual sight of the edge couplers to aid in the alignment of the fiber array. The PCB and fiber were both connected to the same mechanical substrate for stability. The fully assembled MCM transceiver can be seen in figure 2.

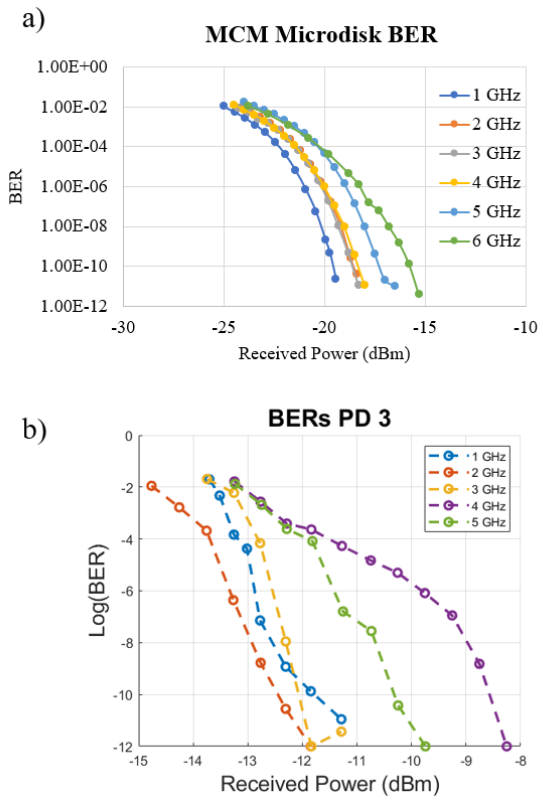

Fig. 4. The BER curves for both the transmitter (a) and the receiver (b), showing error free performance for 6 Gbps and 5 Gbps for the transmitter and the receiver, respectively.

C. Performance

To measure the performance of the MCM transceiver, the four channels on the transmitter and four channels on the receiver were independently tested for both bandwidth and bit error rate performance. The bandwidths for both the transmitter and the receiver show an electrical resonance at approximately 8 GHz—the cause of the resonance is currently being investigated. Example eye diagrams for both a transmitter channel and a receiver channel can be seen in figure 3. Error free (bit error rate of 1E-9) was achieved at 6 Gbps and 5 Gbps for the transmitter channels and receiver channels, respectively, as shown in figure 4.

IV. Conclusion

We provided an overview of our MCM transceiver's architecture, integration, and performance. The MCM transceiver provides the interface between silicon photonic switches and electronic resources (compute, memory, and storage), enabling the disaggregation of such resources via reconfiguring the silicon photonic switch.

References

[1] "Cisco Visual Networking Index: Forecast and Trends, 2017-2022," White Paper (2019).

[2] V. Shrivastav et al. "Shoal: A network architecture for disaggregated racks." 16th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 19). 2019.

[3] M. Ghobadi et al. "Projector: Agile reconfigurable data center interconnect." Proceedings of the 2016 ACM SIGCOMM Conference. 2016.

[4] W. J. Dally, C. T. Gray, J. Poulton, B. Khailany, J. Wilson, and L. Dennison, "Hardware-Enabled Artificial Intelligence," in 2018 IEEE Symposium on VLSI Circuits, 2018, pp. 3–6.

[5] P. X. Gao et al., "Network Requirements for Resource Disaggregation," in Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation, Berkeley, CA, USA, 2016, pp. 249–264.

[6] Cheng, Qixiang, et al. "Silicon Photonic Switch Topologies and Routing Strategies for Disaggregated Data Centers." *IEEE Journal of Selected Topics in Quantum Electronics* (2019).

[7] Y. Arakawa, T. Nakamura, Y. Urino and T. Fujita, "Silicon photonics for next generation system integration platform," in IEEE Communications Magazine, vol. 51, no. 3, pp. 72-77, March 2013.

[8] Q. Cheng, M. Bahadori, M. Glick, S. Rumley, and K. Bergman, "Recent Advances in Optical Technologies for Data Centers: A Review," *OSA Optica*, vol. 5, no. 11, pp. 1354-1370, 2018.

[9] I. Ndip, A. Öz, S. Guttowski, H. Reichl, K. Lang, H. Henke, "Modeling and Minimizing the Inductance of Bond Wire Interconnects," *IEEE Workshop on Signal and Power Integrity*, 2013.

[10] M. Rakowski *et al.*, "Hybrid 14 nm FinFET Silicon Photonics Technology for Low-Power Tb/s/mm^2 Optical I/O," *Symposium on VLSI Technology Digest of Technical Papers*, pp. 221-222, 2018.