Silicon Photonics Enabling the Disaggregated Data Center

Madeleine Glick, Sebastien Rumley, Keren Bergman
Department of Electrical Engineering, Columbia University, New York, NY, 10027, USA

Abstract: The interconnection network is a technology block to improved performance of the data center. Disaggregation enabled by silicon photonics is a route to low latency, cost and energy efficient solutions. © 2018 The Author(s)

OCIS codes: (060.4510) Optical communications, (060.4253) Networks, Circuit Switched

1. Disaggregation: The interconnection network is the bottleneck

The interconnection network has become a technology block to improved performance of the data center and high performance computing. Disaggregation of the traditional server has been proposed as a solution to improve efficiency and access to resources [1, 2]. The traditional data center is built around servers as building blocks. Each server composed of tightly coupled resources: CPU, memory, one or more network interfaces, specialized hardware such as GPUs and possibly some storage systems (hard disks or solid state disks). This design has been hitting several challenges. The various server elements follow different trends of cost and performance. As updated components become available, upgrading of the CPU or memory in the traditional design would require an entirely new server with new motherboard design at considerable expense and economic waste of adequately performing elements [3]. Traditional data centers also suffer from resource fragmentation in the cases where resources (CPU, memory, storage IO, network IO) are mismatched with workload requirements, for example for a compute intensive task that does not use the full memory capacity or a communication intensive task that does not fully use the CPU Data gathered from datacenters show that server memory is unused by as much as 50% or higher [2, 4]. These challenges become motivations for disaggregating the server, i.e. forming pools of similar resources that can be independently upgraded and adaptively configured for optimized performance of the data center or high performance computer.

The disaggregated data center requires an interconnection fabric. This fabric must carry the additional traffic engendered by the disaggregation and be high bandwidth and low latency in order to not only maintain, but also improve performance.

A related challenge is referred to as the “memory wall.” Due to differing performance improvements and trends (and fundamental physical constraints), there has been a growing gap between the CPU capabilities and the bandwidth available to access memory. For this reason, it is advantageous for the memory to be as close to the CPU as possible. However, as noted above, in many cases, there is a mismatch of resources which leads to inefficiencies. Per processor memory bandwidths range up to 200 GB/s. This is well beyond the capabilities of commercially available networks. Disaggregating memory in addition to hard disks and network IO will add considerably to the network traffic.

2. Metrics for the disaggregated interconnection network

A viable interconnection network for the disaggregated data center should be low cost, energy efficient and offer performance improvements to the traditional architecture. It should be noted that typical latency to memory, in a traditional server, where the memory is close to the CPU, is of the order of 10s of nanoseconds. Attention must be paid to performance degradation with any added latency. Several groups have developed metrics or guidelines to achieve these goals [1, 2]. The authors in [1] explore a limited set of applications and determine that, although perhaps not ideal, even current technologies with network bandwidth between 40-100 Gb/s and network latency of 3-5 microseconds would show performance advantages with disaggregated architectures. The main cause of the bandwidth and latency requirements is the application’s memory bandwidth demand. Reference [2] does a cost/performance analysis including cost of latency and bandwidth to determine when a data center disaggregated memory system would be cost competitive to a conventional direct attached memory system. The authors find that from a cost perspective, the current cost of an optically switched interconnect should be reduced by approximately a factor of 10 to be an economically viable solution.

3. Silicon photonics and disaggregation

Given the requirement for high bandwidth density at low cost and power consumption, it is not surprising that photonics, and especially silicon photonics, fabricated in high volume CMOS compatible foundries [5], is a prime candidate for the disaggregated interconnection network. At the link level, it is widely accepted that to achieve the required bandwidth density for the data center, the trend is towards on board silicon photonics with 2.5D integration on a multichip module (MCM) or with more advanced 3D integration using through silicon vias (TSV) for higher bandwidth and considerable energy savings compared to pluggable optics [6,7]. The disaggregated network also requires a switching fabric to adaptively provision the computing resources. Although the primary switching in remains packet switched, optical circuit switches are prime candidates for reconfiguration of resources in the disaggregated network. Various optical switching fabrics are being proposed for the data center [4, 8 - 10] with, most importantly, trade-offs in energy consumption and cost. With regards to energy consumption, special attention must be paid to the power penalty incurred by the laser source. The authors of [11], the Flexfly network, propose to use low to medium radix switches to rewire the interconnect, as required by the application, rather than overprovisioning transceivers or adaptively routing traffic to achieve a high bandwidth low energy interconnection network.

4. Conclusions

Performance improvements and energy consumption in the data center and high performance computing are currently both dominated by the cost of data movement. Integrated silicon photonic links at energies of less than 1pJ/bit and disaggregation of the server using novel switch architectures are the most promising route to achieving improved performance within the constraints of required energy efficiency and component cost reduction.

5. References

[1] Peter Xiang Gao et al. "Network Requirements for Resource Disaggregation." OSDI. Vol. 16. 2016.

[2] Bulent Abali et al. "Disaggregated and optically interconnected memory: when will it be cost effective?." arXiv preprint arXiv:1503.01416 (2015).

[3] https://www.intel.com/content/www/us/en/it-management/intel-it-best-practices/disaggregated-server-architecture-drives-data-center-efficiency-paper.html

[4] Georgios Zervas et al. "Optically Disaggregated Data Centers With Minimal Remote Memory Latency: Technologies, Architectures, and Resource Allocation." Journal of Optical Communications and Networking 10.2 (2018): A270-A285.

[5] David Thomson, et al. "Roadmap on silicon photonics." Journal of Optics 18.7 (2016): 073003.

[6] Ali Ghiasi, "Large data centers interconnect bottlenecks." Optics express 23.3 (2015): 2085-2090.

[7] http://onboardoptics.org/

[8] William M. Mellette, et al. "RotorNet: A scalable, low-complexity, optical datacenter network." Proceedings of the Conference of the ACM Special Interest Group on Data Communication. ACM, 2017.

[9] Monia Ghobadi, et al., ProjecToR: "Agile Reconfigurable Data Center Interconnect." Proceedings of the 2016 ACM SIGCOMM Conference. 2016, ACM: Florianopolis, Brazil. p. 216-229.

[10] Qixiang Cheng et al., Scalable photonic switching in high performance datacenters" to appear in Optics Express.

[11] Ke Wen, et al. "Flexfly: Enabling a reconfigurable dragonfly through silicon photonics." High Performance Computing, Networking, Storage and Analysis, SC16: International Conference for. IEEE, 2016.